Chapitre 3

L'analyse des correspondances des tableaux lexicaux

Le tableau croisé, ou table de contingence, est l'une des formes de structuration des données les plus courantes dans l'analyse des variables qualitatives. En confrontant deux partitions d'une même population ou d'un même échantillon, le tableau croisé permet en effet de travailler sur des *variations par catégorie*, éléments indispensables en vue d'une première interprétation des résultats. Les analyses et descriptions de tableaux croisés sont d'ailleurs à la base du traitement statistique des données d'enquêtes.

Les méthodes d'analyse des données (ou encore : analyses descriptives multidimensionnelles) sont dévolues, pour l'essentiel, à la description de tableaux : tableaux de mesures, de classement, d'incidence, de contingence, ou de présence-absence. Elles ont profondément modifié les méthodes de traitement des données d'enquêtes en mettant à la disposition des utilisateurs des panoramas globaux par thème, des outils pouvant éprouver la cohérence de l'information, des procédures de sélection des tabulations les plus pertinentes.

Les typologies (classification des individus en prenant en compte simultanément plusieurs réponses ou plusieurs caractéristiques de base), les outils de visualisation (plans factoriels) permettent de jeter un regard plus *macroscopique* sur l'information de base.

Les tableaux lexicaux auxquels nous allons appliquer ces techniques sont des tables de contingence particulières ; l'individu statistique donnant lieu à des comptages pour chaque case du tableau sera l'*occurrence* d'une unité textuelle : forme, lemme, segment répété, etc.

Les lignes du tableau correspondront par exemple aux formes graphiques dont la fréquence dans le corpus est supérieure à un seuil donné, les colonnes aux parties du texte : locuteurs, catégories de locuteurs, auteurs, document, etc. Dans le cas général, la case (i,j) contiendra donc le nombre des occurrences de la forme i dans la partie de texte j.
Le présent chapitre rappelle brièvement les principes de base et les modalités pratiques d'utilisation de ces méthodes, avant de mettre en relief leur

contribution propre au renouvellement des techniques de dépouillement d'enquêtes.

3.1 Principes de base des méthodes d'analyse des données

Le principe commun à toutes les méthodes de statistique descriptive multidimensionnelle est schématiquement le suivant :

Chacune des dimensions d'un tableau rectangulaire de données numériques permet de définir des distances (ou des proximités) entre les éléments de l'autre dimension. Ainsi, l'ensemble des colonnes (qui peuvent être des variables, des attributs) permet de définir à l'aide de formules appropriées des distances entre lignes (qui peuvent être des individus, des observations). De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes.

On obtient ainsi des tableaux de distances, auxquelles sont associées des représentations géométriques complexes décrivant les similitudes existant entre les lignes et entre les colonnes des tableaux rectangulaires à analyser.

Le problème est alors de rendre assimilable et accessible à l'intuition ces représentations, au prix d'une perte d'information de base qui doit rester la plus petite possible. Brièvement, on peut dire qu'il existe deux familles de méthodes qui permettent d'effectuer ces réductions :

- *Les méthodes factorielles*, largement fondées sur l'algèbre linéaire, produisent des représentations graphiques sur lesquelles les proximités géométriques usuelles entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes. C'est à cette famille de méthodes qu'appartient l'analyse des correspondances, présentée dans ce chapitre.

- *Les méthodes de classification* qui opèrent des regroupements en classes (ou en familles de classes hiérarchisées) des lignes ou des colonnes, sont présentées au chapitre suivant.

Ces deux familles de méthodes peuvent d'ailleurs être utilisées avec profit sur les mêmes tableaux de données, et se compléter utilement. Il va de soi que les règles d'interprétation des représentations obtenues par le biais de ces techniques de réduction n'ont pas la simplicité de celles de la statistique descriptive élémentaire.
L'interprétation des histogrammes, les diagrammes en bâtons, les graphiques de séries chronologiques ne nécessitent qu'un apprentissage rudimentaire;

alors que dans le cas de l'analyse des correspondances, par exemple, il sera nécessaire de connaître des règles de lecture des résultats plus contraignantes que ne le laisse croire le caractère souvent suggestif des représentations obtenues. Il faudra au lecteur non seulement un apprentissage, mais aussi une véritable expérience clinique (on veut dire par là une expérience variée et difficilement formalisable : l'analyse de chaque tableau de données présentant quelques aspects particuliers).

Un paradoxe pédagogique

Pour bien assimiler le fonctionnement d'une méthode, il est nécessaire de raisonner sur un exemple de dimension réduite, parfaitement maîtrisable par l'utilisateur... Mais les techniques d'analyse exploratoire multidimensionnelle ne présentent de l'intérêt que parce qu'elles permettent de traiter de très grands ensembles de données. Une image résume cette situation pédagogique paradoxale : il est difficile de démontrer l'efficacité d'un filet de pêche dans un aquarium de salon!

Nous allons néanmoins montrer comment fonctionnent ces méthodes sur des tableaux de dimensions très réduites qui illustreront les principes, mais ne permettront pas d'apprécier pleinement l'efficacité des outils utilisés. Les exemples du chapitre 5 mettront sans doute mieux en évidence leurs possibilités réelles en matière de synthèse et d'exploration.

3.2 L'analyse des correspondances

L'analyse des correspondances est une technique de description des tables de contingence (ou encore : tableaux croisés) et de certains tableaux binaires (dits tableaux de "présence-absence"). Cette description se fait essentiellement sous forme de représentation graphique des associations entre lignes et entre colonnes.

3.2.1 Bref historique

Présentée et étudiée de façon systématique comme une technique souple d'analyse exploratoire de données multidimensionnelles par Benzécri (1973) dans un ouvrage qui reprend les travaux de son laboratoire au cours des dix années précédentes, l'analyse des correspondances s'est trouvée depuis de nombreux précurseurs, et a donné lieu à des travaux dispersés et indépendants les uns des autres. On peut citer sans être exhaustif les noms de Guttman (1941) et de Hayashi (1956) qui parlent tous deux de *méthodes de*

quantification, Nishisato (1980) qui parle de *dual scaling*, Gifi (1990) qui parle d'*homogeneity analysis* pour désigner l'analyse des correspon-dances multiples.[1]

3.2.2 L'analyse des correspondances exposée à partir d'un exemple numérique simple

Le tableau 3.1 que l'on va s'efforcer de décrire avec l'aide de l'analyse des correspondances est un tableau de fréquences croisant, en ligne, 14 formes utilisées dans leurs réponses à une question ouverte (la question ***Enfants*** introduite au paragraphe 2.2 du chapitre 2) par 2 000 personnes, et en colonne, 5 niveaux de diplômes[2] déclarés par chacune de ces personnes.

Tableau 3.1

Croisement (formes-Niveau de Diplôme). Effectifs bruts

	Sans Dipl.	CEP	BEPC	Bacc	Univ	Total
Argent	51	64	32	29	17	193
Avenir	53	90	78	75	22	318
Chômage	71	111	50	40	11	283
Conjoncture	1	7	5	5	4	22
Difficile	7	11	4	3	2	27
Economique	7	13	12	11	11	54
Egoisme	21	37	14	26	9	107
Emploi	12	35	19	6	7	79
Finances	10	7	7	3	1	28
Guerre	4	7	7	6	2	26
Logement	8	22	7	10	5	52
Peur	25	45	38	38	13	159
Santé	18	27	20	19	9	93
Travail	35	61	29	14	12	151
Total	323	537	322	285	125	1592

Rappelons que l'unité statistique retenue ici n'est pas l'individu, mais l'occurrence d'une forme graphique. Les colonnes constituent bien une partition de l'ensemble des personnes interrogées, mais pas les lignes : un répondant peut utiliser dans ses réponses plusieurs formes de la liste retenue

[1] Cf. aussi l'historique de Benzécri (1982), les synthèses de Tenenhaus et Young (1985) et de Escoufier (1985). On trouvera dans Greenacre (1984, 1993), Lebart et al. (1984), Benzécri (1992c) des exposés et des exemples en langue anglaise.

[2] Sans diplôme, CEP (Certificat d'études primaires), BEPC (Brevet d'études du premier cycle ou équivalent), Bac (Baccalauréat ou équivalent), Univ. (Université, grandes écoles ou équivalent).

ici. Les lignes constituent cependant une partition de l'ensemble des occurrences de formes.

On lit ce tableau de la manière suivante : la forme *Argent*, par exemple, a été utilisée 51 fois dans leurs réponses par les personnes appartenant à la catégorie "sans diplôme".

Les totaux de chaque ligne représentent les nombres d'occurrences de chaque forme alors que les totaux de chaque colonne représentent les nombres totaux de formes (de la liste) utilisées par les diverses catégories d'enquêtés.

Le tableau 3.2 représente les profils-lignes, exprimés ici en pourcentages : ils sont obtenus en divisant chaque élément du tableau par la somme de la ligne correspondante : le profil-ligne de la forme *Peur*, par exemple, est obtenu en divisant chaque terme de la ligne par 159.

Tableau 3.2

Les profils-lignes du tableau 3.1

	Sans Dipl.	CEP	BEPC	Bacc	Univ	Total
Argent	26.4	33.2	16.6	15.0	8.8	100.0
Avenir	16.7	28.3	24.5	23.6	6.9	100.0
Chômage	25.1	39.2	17.7	14.1	3.9	100.0
Conjoncture	4.5	31.8	22.7	22.7	18.2	100.0
Difficile	25.9	40.7	14.8	11.1	7.4	100.0
Economique	13.0	24.1	22.2	20.4	20.4	100.0
Egoisme	19.6	34.6	13.1	24.3	8.4	100.0
Emploi	15.2	44.3	24.1	7.6	8.9	100.0
Finances	35.7	25.0	25.0	10.7	3.6	100.0
Guerre	15.4	26.9	26.9	23.1	7.7	100.0
Logement	15.4	42.3	13.5	19.2	9.6	100.0
Peur	15.7	28.3	23.9	23.9	8.2	100.0
Santé	19.4	29.0	21.5	20.4	9.7	100.0
Travail	23.2	40.4	19.2	9.3	7.9	100.0
Total	20.3	33.7	20.2	17.9	7.9	100.0

La comparaison de deux profils-lignes va nous renseigner sur la façon dont les formes correspondantes s'associent aux catégories[1].
Cette comparaison est plus difficile à partir du seul tableau 3.1, car les fréquences des formes sont très variables. Ainsi, il n'apparaît pas immédiatement à la lecture du tableau 3.1 que *Conjoncture* est particulièrement citée par les personnes diplômées d'université, ce que le tableau 3.2 permet de lire facilement

[1] Pour rendre ce tableau plus lisible, les nombres obtenus ont été multipliés par 100.

Lorsque deux formes seront représentées par des points voisins en analyse des correspondances, cela signifiera que les profils-lignes correspondants sont voisins.

Le tableau 3.3 représente les profils-colonnes : ceux-ci sont obtenus de façon tout à fait analogue en divisant les éléments de chaque colonne par leur somme, et en multipliant par 100 le résultat obtenu.

Tableau 3.3

Profils-colonnes du tableau 3.1

	Sans Dipl.	CEP	BEPC	Bacc	Univ	Total
Argent	15.8	11.9	9.9	10.2	13.6	12.1
Avenir	16.4	16.8	24.2	26.3	17.6	20.0
Chômage	22.0	20.7	15.5	14.0	8.8	17.8
Conjoncture	.3	1.3	1.6	1.8	3.2	1.4
Difficile	2.2	2.0	1.2	1.1	1.6	1.7
Economique	2.2	2.4	3.7	3.9	8.8	3.4
Egoisme	6.5	6.9	4.3	9.1	7.2	6.7
Emploi	3.7	6.5	5.9	2.1	5.6	5.0
Finances	3.1	1.3	2.2	1.1	.8	1.8
Guerre	1.2	1.3	2.2	2.1	1.6	1.6
Logement	2.5	4.1	2.2	3.5	4.0	3.3
Peur	7.7	8.4	11.8	13.3	10.4	10.0
Santé	5.6	5.0	6.2	6.7	7.2	5.8
Travail	10.8	11.4	9.0	4.9	9.6	9.5
Total	100.0	100.0	100.0	100.0	100.0	100.0

La comparaison de deux profils-colonnes nous renseigne sur les proximités existant entre les différentes catégories de diplômes vis-à-vis du vocabulaire employé.

En analyse des correspondances, le voisinage de deux points représentant des catégories de diplômes traduira une similitude de profils-colonnes.
L'analyse des correspondances va donc décrire simultanément les similitudes de profils-lignes et de profils-colonnes, et fournir une représentation schématique des informations contenues dans les tableaux 3.2 et 3.3.

La figure 3.1 (plan factoriel engendré par les deux premiers axes de l'analyse des correspondances du tableau 3.1) donne une représentation visuelle des associations entre lignes et colonnes de ce tableau.
Pour une personne qui en maîtrise bien les règles de lecture, c'est un procédé rapide d'assimilation de l'information. On ne peut cependant espérer de la description d'une table dont les dimensions sont aussi modestes des grandes surprises ou des révélations. L'analyse des correspondances a ici une

fonction purement descriptive : la consultation des résultats est simplement plus aisée.

Montrons, à propos de cet exemple, la simplicité des principes et des règles d'interprétation de la méthode.

On notera f_{ij} le terme général de la table des fréquences dont les éléments sont préalablement divisés par l'effectif total k (k =1 592 pour le tableau 3.1). Cette table possède n lignes et p colonnes (ici, n =14 et p =5).

Selon les notations usuelles, $f_{i.}$ désigne la somme des éléments de la ligne i et $f_{.j}$ la somme des éléments de la colonne j de cette table.

Le profil de la ligne i est l'ensemble des p valeurs :

$$\frac{f_{ij}}{f_{i.}}, \quad j = 1,...p.$$

Le profil de la colonne j est l'ensemble des n valeurs :

$$\frac{f_{ij}}{f_{.j}}, \quad i = 1,...n.$$

Les éléments des profils, multipliés par 100, ne sont autres que les pourcentages-lignes et les pourcentages-colonnes qui figurent dans les tableaux 3.2 et 3.3.

Comment lire la figure 3.1 ?

Si deux points-lignes i et i' ont des profils identiques ou voisins, ils seront confondus ou proches sur chacun des axes factoriels. De façon tout à fait analogue, si deux points-colonnes j et j' ont des profils identiques ou voisins, ils seront confondus ou proches.

L'origine des axes correspond aux profils moyens (marges de la table de fréquences).

Le profil-ligne moyen a pour composantes : $f_{.j}$, $j = 1,..p.$
Le profil-colonne moyen a pour composantes : $f_{i.}$, $i = 1,..n.$

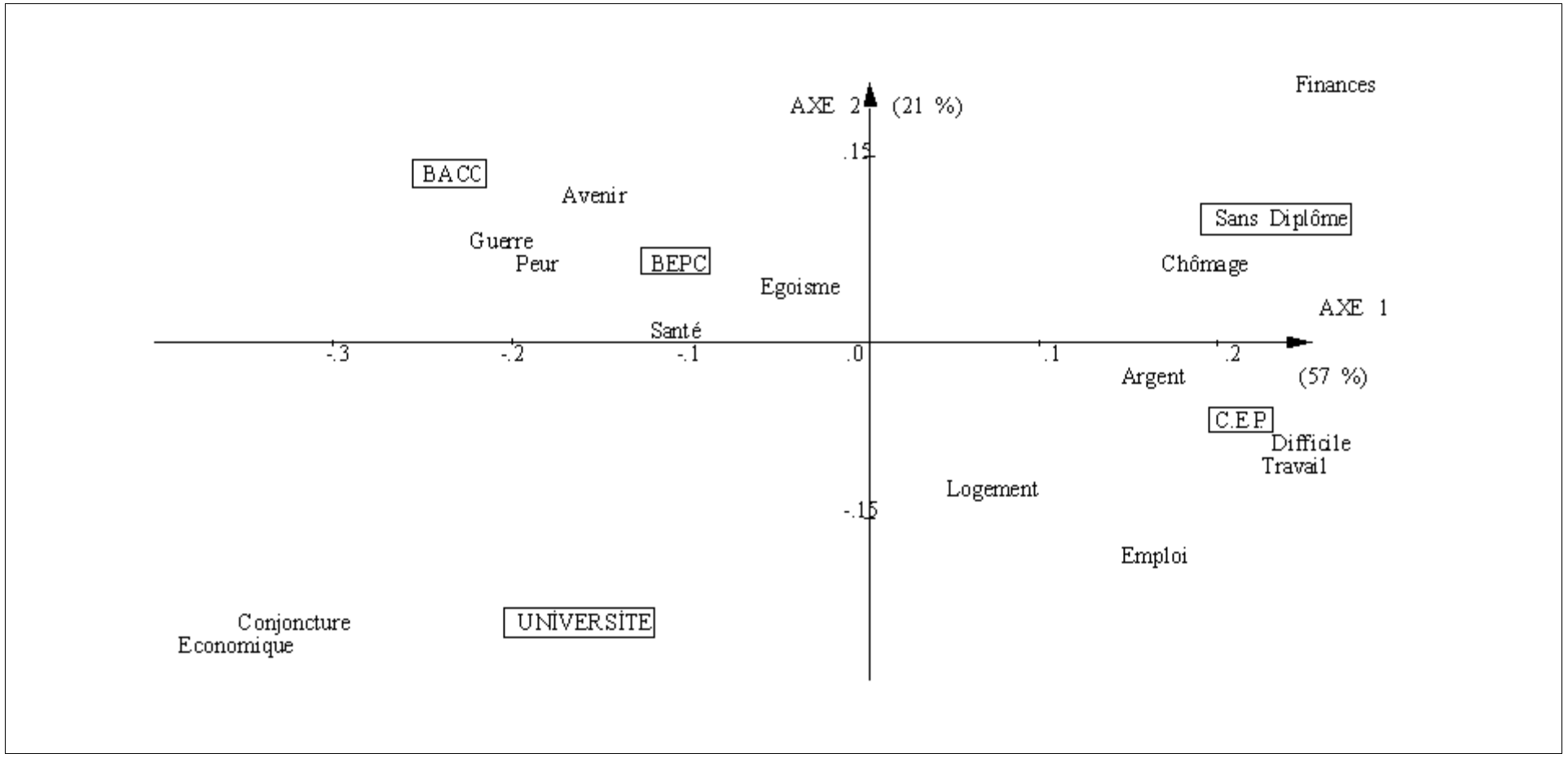

Figure 3.1

Proximités entre formes et entre diplômes (Réponses libres du corpus *Enfants*)
Analyse des correspondances du tableau 3.1

Ainsi un point-colonne comme *BEPC,* assez proche de l'origine, a un profil voisin de celui de la colonne *Total* (marge verticale) du tableau 3.3. De la même façon, un point-ligne comme *Santé* a un profil voisin de la ligne *Total* (marge horizontale) sur la dernière ligne du tableau 3.2.

Les points occupant des positions périphériques auront donc les profils les plus différents du profil moyen, et seront donc les plus typés. Tel est le cas pour des formes comme *Finances* et *Economique*, diamétralement opposées par la représentation, car correspondant respectivement aux formulations populaires et "instruites" de préoccupations sans doute voisines.

Que signifient les proximités ?

Il reste cependant à définir ce que l'on appelle "proche", c'est-à-dire à expliciter la façon dont sont calculées les distances dont des graphiques tels que la figure 3.1 fournissent des approximations planes.

Les profils qui sont des suites de *n* ou *p* nombres, selon qu'il s'agit de lignes ou de colonnes, permettent de définir des points dans des espaces à n ou p dimensions (plus précisément à *n-1* et *p-1* dimensions, car les sommes des composantes de chaque profil valent 1).

Les distances entre points seront donc définies dans ces espaces de dimensions élevées. La phase d'analyse proprement dite consistera à réduire ces dimensions de façon à permettre une représentation visuelle, tout en déformant le moins possible les distances.

La distance entre deux points-lignes i et i' est donnée par la formule:

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

De la même façon, la distance entre deux points-colonnes j et j' est donnée par :

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

Cette distance, appelée distance du chi-2, ressemble beaucoup à la distance euclidienne usuelle (somme des carrés des différences entre composantes des profils) à ceci près qu'une pondération intervient. Cette pondération est l'inverse de la fréquence correspondant à chaque terme : $(1 / f_{.j})$ pour chaque terme dans la somme qui définit $d^2(i, i')$, et $(1 / f_{i.})$ pour chaque terme dans la somme qui définit $d^2(j, j')$.

La distance du chi-2 possède une propriété remarquable, dite d'*équivalence distributionnelle.* Disons en bref que cette propriété assure une invariance

des distances entre lignes (resp. colonnes) lorsque l'on agrège deux colonnes (resp. lignes) ayant des profils identiques. Cette propriété d'invariance assure une certaine stabilité des résultats vis-à-vis des nomenclatures : ainsi, deux classes de diplômes ayant même profil lexical peuvent être considérées comme distinctes ou confondues sans que cela modifie les représentations obtenues.[1]

Pourquoi une représentation simultanée?

Nous savons maintenant interpréter la proximité entre deux points-lignes ou entre deux points-colonnes, ainsi que leurs positions respectives par rapport à l'origine des axes.

Mais la figure 3.1 nous montre des points-lignes et des points-colonnes simultanément, et donc des proximités qu'il est bien tentant d'interpréter : que le point-ligne *Chômage* soit au voisinage du point-colonne *Sans Diplôme* n'a peut-être rien de surprenant. Mais la proximité entre *Santé* et *BEPC* est moins évidente.

En fait, il n'est pas licite d'interpréter ces proximités croisées entre un point-ligne et un point-colonne, car les deux points ne sont pas dans le même espace au départ. En revanche, il est possible d'interpréter la position d'*un* point-ligne par rapport à l'*ensemble* des points-colonnes ou d'*un* point-colonne par rapport à l'*ensemble* des points-lignes. La principale justification de cette représentation simultanée est donnée par les *relations de transition*, liant les coordonnées d'un point dans un espace (celui des lignes par exemple) à celles de tous les points de l'autre espace (celui des colonnes pour notre exemple).

Si ϕ_i désigne la coordonnée d'un point-ligne i sur l'axe horizontal de la figure 3.1 (premier axe factoriel), et si ψ_j désigne celle d'un point-colonne j sur ce même axe, on a le système de relations remarquablement symétriques :

$$\phi_i = \beta \sum_{j=1}^{p} (f_{ij} / f_{i.}) \psi_j \qquad (1)$$

$$\psi_j = \beta \sum_{i=1}^{n} (f_{ij} / f_{.j}) \phi_i \qquad (2)$$

Le coefficient β est un coefficient positif, supérieur à 1.

Sans ce coefficient, on voit que les points-lignes seraient des barycentres des points-colonnes, et réciproquement. Une telle représentation doublement

[1] Pour une justification plus argumentée de ces pondérations, cf. par exemple Benzécri (1973), et Escofier (1978) qui présente plusieurs distances ayant cette propriété.

barycentrique n'est pas possible, car la prise de barycentre a un effet contractant. Pour que les relations (1) et (2) soient possibles simultanément, il faut donc un coefficient dilatateur β supérieur à 1.

Dans le cas des formules de transition correspondant au premier axe de l'analyse des correspondances, on trouve d'ailleurs le coefficient β le plus proche de 1 qui soit compatible avec ces deux formules.

On peut ainsi présenter directement l'analyse des correspondances comme la recherche des valeurs de Φ_i et de Ψ_j correspondant au plus petit coefficient dilatateur β.

Les formules (1) et (2) sont également valables pour les coordonnées sur l'axe vertical de la figure 3.1, avec bien sûr des valeurs différentes pour les Φ_i, Ψ_j, et pour le coefficient β.

Autre propriété des coordonnées des points-lignes (et des points-colonnes) sur les axes factoriels : elles sont centrées, c'est-à-dire vérifient les relations :

$$\sum_{i=1}^{n} f_i.\Phi_i = 0$$

$$\sum_{j=1}^{p} f_{.j}\Psi_j = 0$$

3.2.3 Validité de la représentation

Le tableau 3.4 donne les valeurs de deux séries de paramètres dont nous n'avons pas encore parlé : les *valeurs propres,* désignées par λ_α (comprises entre 0 et 1 en analyse des correspondances).

Tableau 3.4

Valeurs propres et pourcentages de variance (ou d'inertie)

	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.0354	57.04	57.04	**********
2	.0131	21.13	78.17	**********
3	.0073	11.76	89.94	*****
4	.0062	10.06	100.00	****

Ces valeurs propres valent ici $\lambda_1 = 0.035$ pour le premier axe et $\lambda_2 = 0.013$ pour le second ; les *pourcentages de variance*, appelés encore *pourcentages d'inertie* (rapport de chacune des valeurs propres à leur somme globale)

correspondant à ces valeurs propres valent ici 57% et 21% pour ces deux mêmes axes.

Un autre paramètre, la *trace t* (somme de toutes les valeurs propres) vaut ici 0.062 : il y a 4 valeurs propres non nulles. Alors que la trace représente l'inertie totale (ou variance totale) du nuage, les valeurs propres représentent les inerties (ou variances) correspondant à chaque axe.

Propriétés de la trace t

Lors de l'analyse des correspondances d'une table de contingence *(n,p)*, le produit de la trace *t* par l'effectif total *k* n'est autre que le classique chi-2 (χ^2 de Karl Pearson, avec *(n-1)(p-1)* degrés de liberté) utilisé pour tester l'indépendance des lignes et des colonnes de la table.

Ce paramètre est donc calculé par la formule:

$$\chi^2 = k\, t$$

de façon explicite : $$\chi^2 = k \sum_{i=1}^{n} \sum_{j=1}^{p} (f_{ij} - f_{i.} f_{.j})^2 / f_{i.} f_{.j}$$

On a, pour le tableau 3.1 :

$$k\, t = 1592 \times 0.062 = 98.7$$

pour *52* degrés de libertés [*52 = (14 - 1) x (5 - 1)*].

Bien entendu, l'hypothèse d'indépendance est rejetée : l'analyse des correspondances est précisément là pour nous aider à comprendre pourquoi cette hypothèse est rejetée.

On voit sur cet exemple la complémentarité entre statistique inférentielle et analyse exploratoire de donnée : dans le cas présent, celle-ci prend le relais de celle-là dès que la complexité du problème exclut de procéder à des tests d'hypothèses, ou à une modélisation qui s'avérerait inadaptée.

Les valeurs propres

Pour un axe donné, il existe la relation suivante entre le coefficient β des relations de transition et la valeur propre:

$$\beta = 1/\sqrt{\lambda}$$

Une valeur propre proche de 1 assure donc une bonne représentation barycentrique le long de l'axe correspondant. Une telle situation permet d'interpréter les positions relatives des deux ensembles de points-lignes et de

points colonnes[1]. Mais cela n'est pas vraiment le cas pour notre exemple : si les points-diplômes étaient représentés comme de vrais barycentres des points-formes, ils seraient beaucoup moins excentrés... d'où le caractère peut-être trompeur, pour les non-initiés, de la représentation simultanée[2].

Les pourcentages de variance (ou d'inertie)

Ils mesurent l'importance relative de chaque valeur propre dans la trace. Il est assez exceptionnel que, comme c'est le cas ici, le premier plan factoriel "explique" 78% de la variance totale.

En général, ces pourcentages sont au contraire une mesure pessimiste de la part d'information représentée. De nombreux contre-exemples montrent que des pourcentages médiocres correspondent parfois à des représentations qui rendent compte de façon satisfaisante de la structure des données.

Autres aides à l'interprétation

Deux séries de paramètres permettent d'interpréter avec plus de sûreté les résultats, en complétant l'information donnée par les coordonnées des éléments sur les axes factoriels:

a) Les *contributions* (ou contributions absolues), qui décrivent la part prise par un élément (ligne ou colonne) dans la construction d'un axe factoriel.

b) Les *cosinus carrés* (ou contributions relatives) qui mesurent la qualité de la représentation de chaque élément par les axes.

---

[1] Sans pour autant permettre l'interprétation des proximités entre points lignes et points colonnes pris deux à deux.

[2] La racine carrée de la plus grande valeur propre peut également être interprétée comme le plus grand coefficient de corrélation existant entre les lignes et les colonnes de la table (corrélation canonique). Ce coefficient se calcule de la façon suivante : supposons qu'à chacune des 5 catégories de diplôme corresponde une valeur numérique (5 valeurs différentes possibles), et qu'il en soit de même pour les 14 formes utilisées (opération dite de codage ou de quantification). A chaque occurrence, on peut faire correspondre deux valeurs numériques (une pour les lignes, une autre pour les colonnes). On peut donc calculer un coefficient de corrélation entre ces deux ensembles de valeurs. Les 1592 occurrences peuvent être regroupées en 70 groupes (les 70 cases du tableau 3.1) correspondant à des couples de valeurs distinctes. La valeur maximale que peut atteindre ce coefficient de corrélation (qui est associée à un "codage optimal" des mots et des diplômes) est l. Cette recherche de corrélation maximale remonte à Hirschfeld (1935). On peut présenter l'analyse des correspondances à partir de cette approche.

Le tableau 3.5 présente les valeurs de ces différents paramètres pour l'exemple présenté plus haut.

Guide de lecture du tableau 3.5

La colonne P. REL (Poids relatifs) désigne les marges des lignes et des colonnes, qui figuraient déjà dans les tableaux 3.2 et 3.3 précédents.

La colonne DISTO (Distances à l'Origine) contient les carrés des distances à l'origine des axes, c'est à dire les distances de chaque profil au profil moyen (ou: marge). On peut vérifier par exemple sur le tableau 3.2 que les formes *finances* et *conjoncture* ont des profils très différents de la marge.

Les deux premières COORDONNEES sont celles des points, dont la figure 3.1 fournit une représentation approchée.

Les CONTRIBUTIONS (ou encore : contributions absolues), dont les sommes en colonne valent 100, traduisent l'importance des différents éléments dans la construction de chaque axe.

Les COSINUS CARRES (ou encore : contributions relatives), dont les sommes en ligne valent 1, montrent l'importance des différents axes dans l'explication de chaque élément.

3.2.4 Variables actives et illustratives

L'analyse des correspondances permet de trouver des sous-espaces de représentation des proximités entre profils. Mais elle permet aussi de positionner dans ce sous-espace des lignes ou des colonnes *supplémentaires* du tableau de données.

Une fois trouvés les axes factoriels Φ et Ψ et la valeur propre λ (et donc β), les formules de transition (1) et (2) ci-dessus peuvent être appliquées à des lignes ou des colonnes supplémentaires :

La formule (1) permet, à partir de $\Psi, \beta,$ de calculer, pour chaque axe factoriel, la coordonnée ϕ_{i+} d'une ligne supplémentaire $i+$, à partir du profil $(f_{i+j}/f_{i+.})$ (cf. formule 1').

Tableau 3.5

Principaux paramètres de l'analyse des correspondances du tableau 3.1
Les individus actifs sont les lignes du tableau 3.1 alors que les fréquences actives en sont les colonnes

			COORDONNEES				*CONTRIBUTIONS*				*COSINUS CARRES*			
COLONNES ACTIVES	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
Sans Diplôme	20.29	.06	.21	.08	-.07	.10	25.1	10.1	14.7	29.9	.68	.10	.08	.14
CEP	33.73	.03	.14	-.06	-.02	-.08	18.3	8.1	1.5	38.4	.64	.11	.01	.24
BEPC	20.23	.04	-.11	.03	.15	.06	6.8	1.3	59.9	11.9	.31	.02	.57	.10
Bacc	17.90	.10	-.27	.12	-.08	-.06	38.0	20.1	14.4	9.6	.76	.15	.06	.03
Univ	7.85	.17	-.23	-.32	.09	.09	11.9	60.5	9.5	10.3	.31	.59	.05	.05

			COORDONNEES				*CONTRIBUTIONS*				*COSINUS CARRES*			
Formes	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
Argent	12.12	.03	.12	-.02	-.10	.08	4.5	.4	16.9	13.9	.43	.01	.33	.23
Avenir	19.97	.04	-.18	.10	.05	-.01	17.6	14.6	7.6	.1	.72	.22	.06	.00
Chômage	17.78	.05	.21	.07	.00	-.04	22.6	6.8	.0	4.0	.87	.10	.00	.03
Conjoncture	1.38	.28	.40	-.33	.02	-.07	6.3	11.5	.0	1.0	.58	.40	.00	.02
Difficile	1.70	.07	.25	.07	-.06	.00	3.0	.6	.8	.0	.88	.06	.05	.00
Economique	3.39	.26	.35	-.32	-.08	.15	12.0	26.6	3.3	13.0	.48	.40	.03	.09
Egoisme	6.72	.05	-.06	.03	-.18	-.11	.7	.3	29.5	13.6	.07	.01	.66	.26
Emploi	4.96	.11	.14	-.22	.21	-.06	2.6	17.6	30.8	2.7	.16	.41	.40	.03
Finances	1.76	.20	.24	.21	.04	.32	2.8	5.7	.5	29.0	.28	.21	.01	.51
Guerre	1.63	.06	-.22	.07	.10	.02	2.2	.7	2.1	.2	.75	.09	.15	.01
Logement	3.27	.06	.01	-.13	-.09	-.19	.0	4.1	3.5	19.4	.00	.27	.13	.60
Peur	9.99	.05	-.20	.06	.03	-.01	11.7	2.6	1.5	.1	.90	.07	.02	.00
Santé	5.84	.02	-.11	.00	.02	.05	2.1	.0	.4	2.4	.80	.00	.03	.17
Travail	9.48	.06	.21	-.11	.05	.02	12.0	8.6	3.0	.6	.75	.20	.04	.01

$$\phi_{i+} = \beta \sum_{j=1}^{p} (f_{i+j} / f_{i+.}) \psi_j \qquad (1')$$

$$\psi_{j+} = \beta \sum_{i=1}^{n} (f_{ij+} / f_{.j+}) \phi_i \qquad (2')$$

La formule (2) permet de la même façon, à partir de Φ, β, de calculer la coordonnée ψ_{j+} d'une colonne supplémentaire $j+$ (cf. formule 2') dont le profil est $(f_{ij+} / f_{.j+})$.

On peut ainsi illustrer les plans factoriels par des informations supplémentaires n'ayant pas participé à la construction des plans, ce qui va avoir des conséquences très importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelés *éléments actifs* ou *variables actives*. Ils doivent former un ensemble homogène pour que les distances entre individus ou observations aient un sens, et donc que les proximités graphiques observées s'interprètent facilement.

Les éléments ou variables projetés a posteriori sur les plans factoriels sont les éléments *supplémentaires* ou *illustratifs*.

Ces éléments illustratifs (lignes ou colonnes) n'ont nul besoin de former un ensemble homogène dans la mesure où le calcul est effectué séparément pour chacun d'eux. Cette dichotomie entre variables actives et variables illustratives est fondamentale d'un point de vue méthodologique.

Exemple

Les quatre formes graphiques d'effectifs faibles dont on peut voir la ventilation sur le tableau 3.6 n'ont pas participé à l'analyse précédente.

Tableau 3.6

Quatre lignes supplémentaires (ou illustratives)

	SansDipl	CEP	BEPC	Bac	Univ	TOTAL
Confort	2	4	3	1	4	14
Mésentente	2	8	2	5	2	19
Monde	1	5	4	6	3	19
Vivre	3	3	1	3	4	14

On désire cependant savoir comment elles se situent par rapport aux autres formes déjà représentées sur le plan factoriel de la figure 3.1.

Leurs profils-lignes peuvent être positionnés dans le même espace à 5 dimensions, et peuvent donc subir la même projection sur le plan de la figure 3.1.

De façon analogue, le tableau 3.7 contient trois colonnes supplémentaires (des catégories d'âge), qui n'ont pas été incluses dans l'ensemble des colonnes actives en raison de l'hétérogénéité des thèmes : l'interprétation des proximités entre lignes, donc entre formes, aurait été plus délicate. Deux formes sont-elles proches à cause de leur répartition par diplômes ou par classes d'âge? Il n'est pas facile de trancher si les distances entre formes sont calculées à partir des deux variables simultanément[1].

Tableau 3.7

Trois colonnes supplémentaires (ou illustratives)

forme	Age-30	Age-50	Age+50
Argent	59	66	70
Avenir	115	117	86
Chômage	79	88	177
Conjoncture	9	8	5
Difficile	2	17	18
Economique	18	19	17
Egoisme	14	34	61
Emploi	21	30	28
Finances	8	12	8
Guerre	7	6	13
Logement	10	27	17
Peur	48	59	52
Santé	13	29	53
Travail	30	63	58
Total	433	575	663

La figure 3.2 ci-après nous montre que les trois premières formes supplémentaires appartiennent plutôt aux réponses des personnes dont le niveau de diplôme est élevé, alors que la quatrième, *mésentente*, est moins caractéristique.
On trouvera au tableau 3.8 un ensemble de paramètres plus techniques qui permettent d'apprécier la particularité de chacune de ces ventilations.
On ne s'étonnera pas de ne pas trouver de "contributions" dans ce tableau, puisque les éléments illustratifs ont par définition une contribution nulle : ils

[1] On peut être cependant amené à étudier la juxtaposition de plusieurs tableaux de contingence (cf. chapitre 5, paragraphe 5.4), pour obtenir une première visualisation d'un ensemble de caractéristiques (un ensemble de variables socio-démographiques par exemple, ou une batterie de questions fermées).

ne participent pas à la construction des axes, mais sont positionnés a posteriori.

Il est de bonne discipline de commencer par utiliser, en qualité de tableaux actifs, des tableaux homogènes, ne décrivant les proximités que d'un point de vue. On pourra, dans un second temps, illustrer la représentation obtenue par des informations supplémentaires.

Sur la figure 3.2, les trois classes d'âge s'ordonnent comme les diplômes le long de l'axe horizontal : aux âges croissants correspondent des niveaux de diplômes décroissants. Il s'agit d'un trait structurel de la population étudiée; les moins âgés sont les plus instruits, mais ceci complique l'interprétation en terme de causalité... Est-ce que l'effet du niveau d'instruction sur le contenu et la forme des réponses libres n'est pas un effet plus direct lié à l'âge ? Pour aller plus avant dans l'interprétation, il faudra croiser les variables Age et Diplôme, pour obtenir une partition des répondants en $3 \infty 5 = 15$ colonnes... mais on sort là de l'exemple d'école qui nous sert à illustrer le fonctionnement de la méthode.[1]

Tableau 3.8

Paramètres issus de l'analyse pour les éléments illustratifs

Colonnes illustratives

			COORDONNÉES			*COSINUS CARRES*		
Libellés des colonnes	P.REL	DISTO	*1*	*2*	*3*	*1*	*2*	*3*
moins de 30 ans	27.2	.08	-.11	.06	.10	.14	.04	.13
de 30 a 50 ans	36.1	.03	.02	-.05	.02	.01	.09	.01
plus de 50 ans	41.6	.11	.18	.05	-.10	.29	.02	.09

Lignes illustratives

			COORDONNÉES			*COSINUS CARRES*		
Formes	P.REL	DISTO	*1*	*2*	*3*	*1*	*2*	*3*
Confort	.8	.64	-.21	-.70	-.07	.07	.78	.01
Mésentente	1.1	.16	-.15	-.12	-.17	.13	.09	.18
Monde	1.1	.31	-.52	-.14	-.08	.88	.07	.02
Vivre	.8	.68	-.31	-.50	-.52	.14	.37	.40

[1] Une partition croisée Age-Diplôme en 9 postes sera utilisée dans l'exemple "grandeur réelle" du chapitre 5.

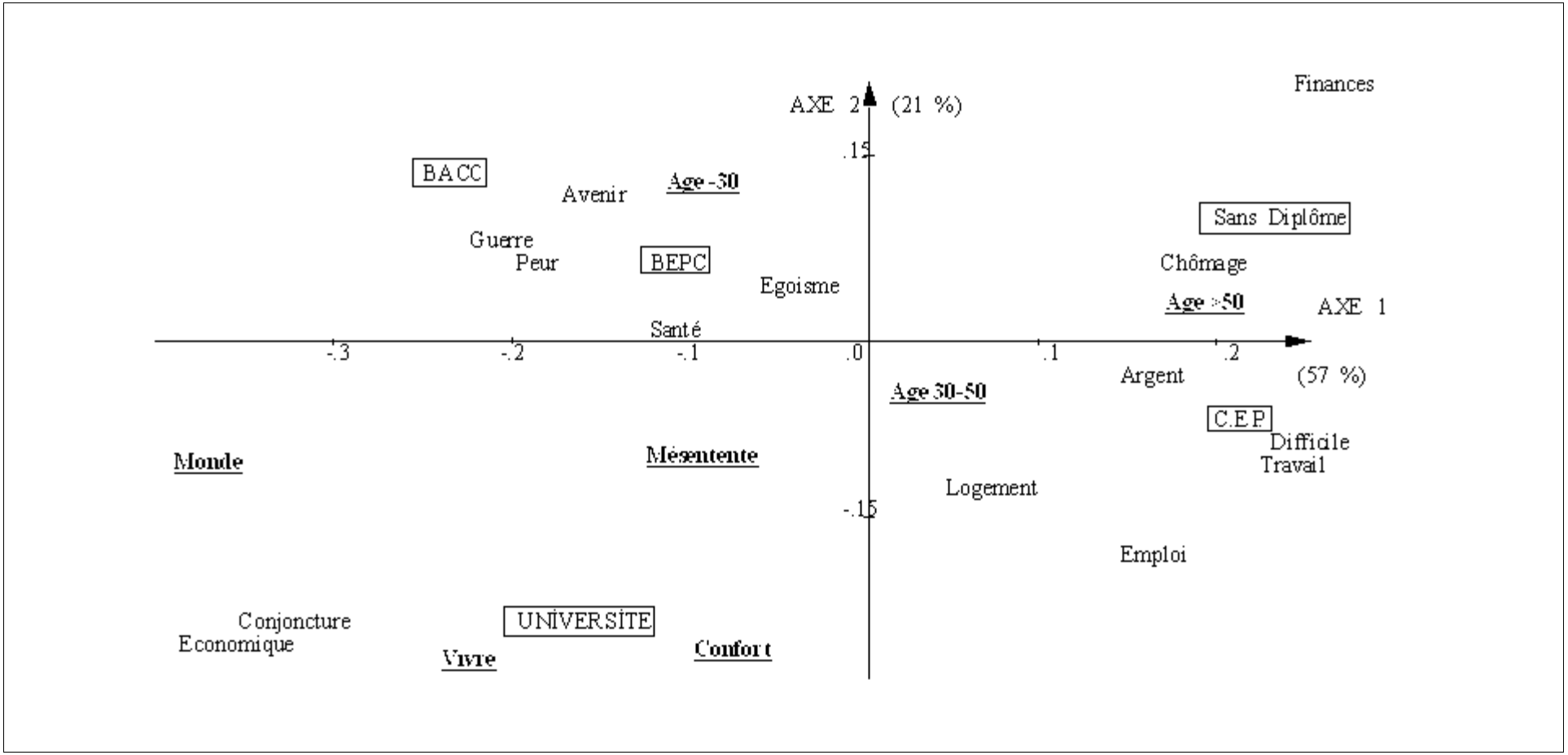

Figure 3.2

Associations entre formes et diplômes (suite)
Positionnement des éléments illustratifs (soulignés) dans le plan de la figure 3.1

3.3 Analyse des correspondances multiples

L'analyse des correspondances multiples permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple privilégié : les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers) ; les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions. On peut faire remonter les principes de cette méthode à Guttman (1941), mais aussi à Burt (1950) ou à Hayashi (1956). Il s'agit en fait d'une simple extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques.

Cette extension se fonde sur la propriété suivante. On pose à k individus deux questions ayant respectivement n et p modalités de réponse, et l'on suppose que pour chaque question, les différentes modalités s'excluent mutuellement (une seule réponse possible).

Il est alors équivalent de soumettre à l'analyse des correspondances la table de contingence (n,p) croisant les deux questions, ou d'analyser le tableau binaire **Z** *à k lignes et (n+p) colonnes qui décrit les réponses.*[1]

L'analyse de **Z** est plus coûteuse, mais plus intéressante, car elle se généralise immédiatement au cas de plus de deux questions.

Le tableau 3.9 représente ainsi le cas de trois questions ayant respectivement 4, 3 et 4 modalités de réponse.

Tableau 3.9

Tableaux R (Codage réduit), Z (disjonctif complet) .

R=

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

Z =

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

[1] Chaque ligne de **Z** comporte dans ce cas deux "1" dans les colonnes correspondant aux réponses choisies, et *n+p-2* "0".

Le *tableau disjonctif complet* **Z** comporte donc ici trois blocs. Les réponses des 12 individus (lignes) peuvent alors être codifiées dans un tableau réduit **R**, (12 lignes et 3 colonnes) qui contient simplement les numéros des modalités. Les procédures de calcul n'utiliseront en fait que ce tableau peu encombrant.

Représenté sur le tableau 3.10, le troisième tableau **B** est le produit du tableau disjonctif **Z** par son transposé :

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

Ce dernier tableau est symétrique (on l'appelle *tableau de Burt* ou *tableau de correspondance multiple*).

Il contient 9 blocs : les blocs diagonaux sont des matrices diagonales, dont les éléments diagonaux sont les effectifs de réponses correspondant à chaque modalité. Les trois blocs distincts parmi les blocs restant ne sont autres que les trois tables de contingence croisant les trois questions deux à deux.

Tableau 3.10

Tableaux B=Z'Z (de Burt)

B = Z'Z =

2	0	0	0	1	1	0	1	1	1	1
0	4	0	0	1	2	1	0	0	3	1
0	0	3	0	2	1	0	1	2	0	0
0	0	0	3	2	1	0	0	1	0	2
1	1	2	2	6	0	0	2	2	1	1
1	2	1	1	0	5	0	0	1	3	1
0	1	0	0	0	0	1	0	0	0	1
1	0	1	0	2	0	0	2	0	0	0
1	0	2	1	2	1	0	0	3	0	0
1	3	0	0	1	3	0	0	0	4	0
1	1	0	2	1	1	1	0	0	0	3

L'analyse des correspondances du tableau **Z** donne les mêmes axes factoriels normés que celle du tableau **B**, c'est-à-dire finalement des graphiques de proximités très similaires, aux échelles près ; mais les valeurs propres homologues sont différentes : à la valeur propre λ_α de l'analyse de **Z** correspond la valeur propre $(\lambda_\alpha)^2$ de l'analyse de **Z'Z**.

Notons que dans le cas de deux questions, le tableau **R** n'a que deux colonnes, le tableau **Z** est formé de deux blocs, $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ et **B** en comporte quatre.

Les deux analyses précédentes donnent également la même représentation que l'analyse des correspondances de la petite table de contingence **C**

croisant les deux questions[1], avec, cette fois, une valeur propre égale à : $(2\lambda - 1)^2$.

3.3.1 Structure de base d'un échantillon d'enquête.

Cet exemple illustre l'utilisation de l'analyse des correspondances multiples pour le traitement des données d'enquêtes. L'exemple précédent a attiré l'attention sur le fait que derrière un effet "niveau d'éducation" pouvait se cacher un effet d'âge... (dans notre échantillon, les moins instruits sont en moyenne les plus âgés). Étant donnée la structure de la population française, ces mêmes effets, on le sait, ne sont pas indépendants du genre (sexe), du niveau de vie... d'où l'idée de décrire le réseau d'interrelations entre toutes les caractéristiques de base des enquêtés, puis de positionner les autres thèmes de l'enquête en tant qu'éléments illustratifs.

Toujours pour faciliter la présentation des résultats, l'échelle réelle de l'application sera réduite : le tableau **Z** aura 144 lignes et 24 colonnes actives représentant les modalités de réponses à 5 questions (tableau 3.11).

Un tableau complémentaire $\mathbf{Z}^+$ comprendra en outre 25 questions illustratives notant la présence ou l'absence des formes graphiques extraites des réponses à une question ouverte posée aux 144 personnes en 1990. Cette question ouverte, inspirée par un travail antérieur de C. Baudelot (1988) a pour libellé : *Que signifie pour vous réussir sa vie ?*

(l'élément z^{+}_{ij} correspondant de $\mathbf{Z}^+$ vaut 1 si l'individu i a utilisé la forme j pour la réponse libre précitée, et 0 sinon. $\mathbf{Z}^+$ aura donc 50 colonnes).

Dans une exploitation en vraie grandeur, il n'est pas rare de positionner en éléments illustratifs plusieurs centaines de modalités. L'intérêt réel de la méthode réside dans ces possibilités de tri et de filtrage systématique à grande échelle.

Le paradoxe pédagogique évoqué au paragraphe 3.1 est encore valable... il y a conflit entre la taille et la pertinence de l'exemple.

Tableau 3.11

Structure de base.

[1] La matrice **C** est un bloc non-diagonal de **B**. Elle s'écrit $\mathbf{C} = \mathbf{Z}_1'\mathbf{Z}_2$. On notera que dans ce cas les *lignes et les colonnes* de la table de contingence **C** ne sont *que des colonnes* pour la matrice **Z**. La représentation simultanée des lignes et des colonnes n'est donc pas un simple artifice graphique: elle trouve une justification dans cette présentation de l'analyse des correspondances.

Liste des questions actives et illustratives

I - 5 questions actives (15 Modalités)

	nombre de modalités
1 . genre (Sexe)	2
2 . niveau de Diplôme	4
3 . statut matrimonial	4
4 . avez-vous eu des enfants?	2
5 . age en 3 classes	3

II - 25 variables illustratives

(Présence ou absence de la forme dans la réponse à la question ouverte)

	nombre de modalités		nombre de modalités
argent	2	mes	2
arriver	2	mari	2
atteindre	2	métier	2
bien	2	peau	2
bon	2	professionnelle	2
bonheur	2	réussir	2
boulot	2	réussite	2
équilibre	2	santé	2
familiale	2	sentir	2
famille	2	situation	2
fonder	2	travail	2
ma	2	travailler	2
me	2		

Lecture du tableau 3.12

Le tableau 3.12 ne représente que la moitié inférieure du tableau de Burt (lequel, rappelons-le, est symétrique) relatif aux 5 questions actives. On trouve dans ce tableau tous les tableaux de contingence (il y en a 10) croisant les 5 questions actives deux à deux.

Sur la diagonale se trouvent les questions croisées avec elles-mêmes, et donc les effectifs correspondant à chaque modalité : on peut ainsi lire qu'il y a 70 hommes et 74 femmes dans l'échantillon. Sur 78 célibataires, il y a 42 hommes et 36 femmes.

Lecture du tableau 3.13

Les règles de lecture du tableau 3.13 sont presque en tout point identiques à celles du tableau 3.5 précédent. Seuls les calculs de contributions cumulées pour les modalités de chaque question sont nouveaux. Leur interprétation est immédiate.

Tableau 3.12

Croisements deux à deux des cinq questions actives.

(Tableau de Burt B)

	masc	*femi*	*CEP**	*BEPC*	*Bacc*	*Univ*	*celi*	*mari*	*divo*	*veuf*	*enfa*	*enfn*	*A—30*	*A—50*	*A+50*
masc	70	0													
femi	0	74													
*CEP**	15	12	27	0	0	0									
BEPC	12	24	0	36	0	0									
Bacc	18	25	0	0	43	0									
Univ	25	13	0	0	0	38									
celi	42	36	10	19	24	25	78	0	0	0					
mari	27	28	12	13	18	12	0	55	0	0					
divo	1	6	2	3	1	1	0	0	7	0					
veuf	0	4	3	1	0	0	0	0	0	4					
enfa	23	29	15	17	15	5	2	42	5	3	52	0			
enfn	47	45	12	19	28	33	76	13	2	1	0	92			
A—30	39	35	8	16	23	27	65	8	1	0	3	71	74	0	0
A—50	26	23	12	10	17	10	9	35	5	0	34	15	0	49	0
A+50	5	16	7	10	3	1	4	12	1	4	15	6	0	0	21
	masc	*femi*	*CEP**	*BEPC*	*Bacc*	*Univ*	*celi*	*mari*	*divo*	*veuf*	*enfa*	*enfn*	*A—30*	*A—50*	*A+50*

Tableau 3.13

Paramètres issus de l'analyse des correspondances multiples pour les éléments actifs

MODALITES			*COORDONNEES*				*CONTRIBUTIONS*				*COSINUS CARRES*			
IDEN — LIBELLE	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
1 . Genre														
masc — homme	9.72	1.06	—.22	—.49	—.58	—.14	.9	7.3	13.2	.9	.04	.23	.32	.02
femi — femme	10.28	.95	.20	.46	.55	.14	.8	6.9	12.5	.9	.04	.23	.32	.02
			contribution cumulée =				1.7	14.3	25.7	1.8				
3 . Niveau de Diplôme														
CEP* — Sans diplôme ou CEP	3.75	4.33	.68	.31	—1.19	—.40	3.3	1.1	21.5	2.8	.11	.02	.32	.04
BEPC — BEPC	5.00	3.00	.29	.63	.79	—.40	.8	6.2	12.7	3.9	.03	.13	.21	.05
Bacc — Baccalauréat	5.97	2.35	—.07	—.37	.54	1.09	.1	2.6	7.2	33.2	.00	.06	.13	.50
Univ — Université	5.28	2.79	—.68	—.39	—.52	—.56	4.6	2.5	5.8	7.8	.16	.05	.10	.11
			contribution cumulée =				8.8	12.4	47.2	47.7				
7 . Etes vous actuellement														
celi — Célibataire	10.83	.85	—.80	.18	.02	.02	13.4	1.1	.0	.0	.76	.04	.00	.00
Mari — Marié(e)	7.64	1.62	.89	—.55	—.10	.32	11.6	7.3	.3	3.7	.49	.19	.01	.06
Divo — Divorcé(e)	.97	19.57	1.01	.14	1.82	—3.09	1.9	.1	13.2	43.9	.05	.00	.17	.49
veuf — Veuf(ve)	.56	35.00	1.65	3.82	—2.27	.53	2.9	25.5	11.6	.7	.08	.42	.15	.01
			contribution cumulée =				29.8	33.9	25.1	48.4				
8 . Avez—vous eu des enfants														
enfa — oui	7.22	1.77	1.19	—.21	.00	.01	19.5	1.0	.0	.0	.79	.03	.00	.00
enfn — non	12.78	.57	—.67	.12	.00	—.00	11.0	.6	.0	.0	.79	.03	.00	.00
			contribution cumulée =				30.6	1.6	.0	.0				
9 . Age en 3 classes														
A—30 — Moins de 30 ans	10.28	.95	—.84	.13	.04	.02	13.9	.6	.1	.0	.74	.02	.00	.00
A—50 — Entre 30 et 50 ans	6.81	1.94	.78	—.86	.10	—.16	8.1	15.8	.3	.8	.32	.38	.01	.01
A 50 — Plus de 50 ans	2.92	5.86	1.13	1.53	—.37	.29	7.1	21.5	1.6	1.2	.22	.40	.02	.01
			contribution cumulée =				29.1	37.8	2.0	2.0				

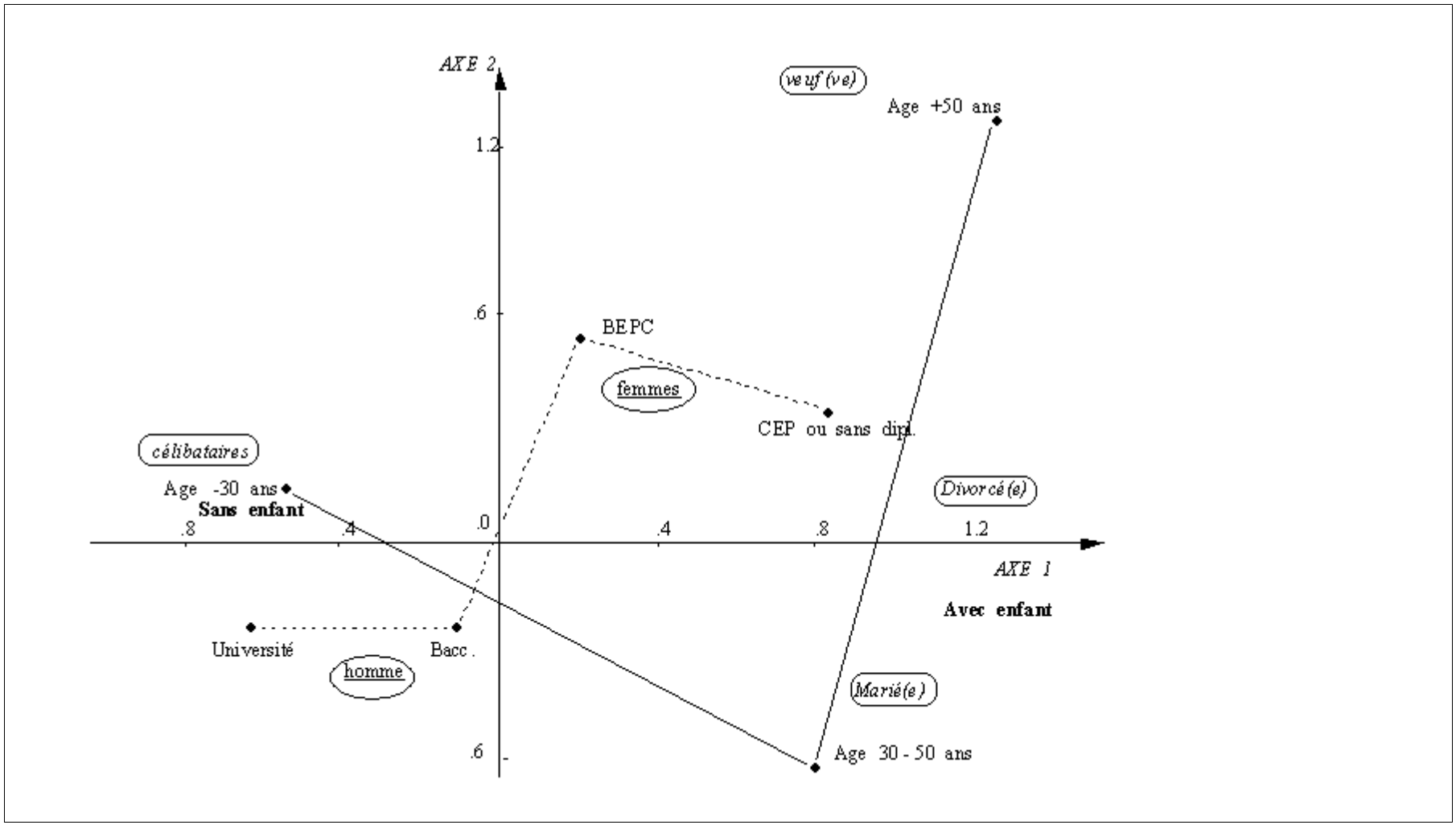

Figure 3.3

Structure de base de l'échantillon.

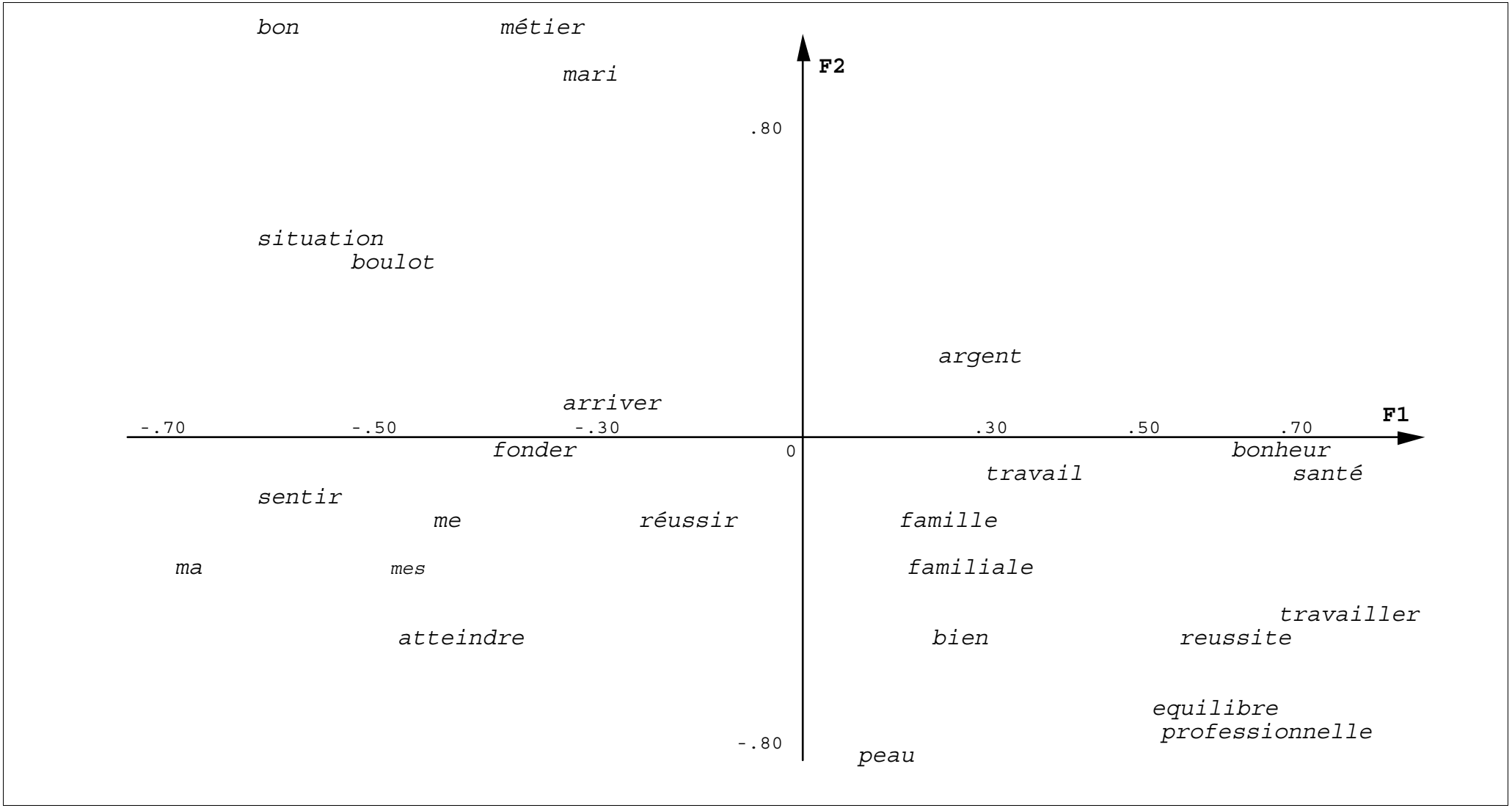

Figure 3.4

Positionnement des formes graphiques du tableau 3.15 dans le plan de la figure 3.3

Lecture de la figure 3.3

La structure du *nuage* des modalités actives est décrite par le plan factoriel de la figure 3.3, qui résume donc les 10 tableaux croisés.

Malgré le caractère sommaire des variables utilisées, les liaisons entre niveau de diplôme, âge, genre et équipement sont plus complexes que l'on ne l'imagine a priori. Un effet de cycle de vie, et aussi de génération se cache derrière des variables à deux modalités comme la variable *genre* : la catégorie "féminin" est lestée par un surnombre de personnes âgées sans diplôme. Il faut donc en tenir compte chaque fois que l'on interprète un éventuel "effet" de cette variable sur telle opinion ou attitude.

La figure 3.3 n'est cependant qu'une approximation plane d'un ensemble de proximités qui ne peuvent pas toujours être représentées dans un plan... il y a donc certaines déformations.

Même si elle peut donner lieu à quelques commentaires, la figure 3.3 n'est pas un résultat en soi : elle n'est qu'une préparation de l'information destinée à accueillir les variables illustratives. Auparavant, disons quelques mots de la validité de la représentation.

3.3.2 Validité de la représentation

Dans le cas des correspondances multiples, on évitera d'utiliser les pourcentages de variance pour juger de la pertinence des axes : ces pourcentages n'ont pas le même sens que lorsqu'il s'agit d'une table de contingence car le codage binaire introduit un *bruit* qui réduit la part d'explication attachée à chaque valeur propre.

Tableau 3.14

Valeurs propres et pourcentages correspondants

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé	
1	.51	25.95	25.95	**********
2	.31	15.89	41.83	******
3	.24	12.27	54.11	*****
4	.21	10.59	64.69	****
5	.19	9.88	74.58	****
6	.16	8.03	82.61	***
7	.12	6.36	88.97	***
8	.11	5.61	94.58	**
9	.06	3.27	97.85	*
10	.04	2.15	100.00	*

Les deux premiers pourcentages (arrondis) sont respectivement de 26% et 16%, pour des valeurs propres de 0,51 et 0.31 (cf. tableau 3.14).

En pratique, la stabilité du plan factoriel s'éprouve par des techniques de simulation (perturbations aléatoires du tableau de données) ou de validation par échantillon-test (positionnement d'individus n'ayant pas participé à la construction des axes).

Ici, la taille réduite de l'échantillon (144 individus) ne permet pas de procéder à de fines inductions à partir de proximités observées.

3.3.3 Positionnement des variables illustratives.

La représentation simultanée des lignes et des colonnes liée à l' analyse des correspondances n'a pas été utilisée : seules les colonnes du tableau **Z** ont été portées sur la figure 3.3. Les 144 points-lignes n'ont pas été tracés pour des raisons d'encombrement graphique, mais surtout parce que les individus correspondants sont anonymes... seules leurs caractéristiques présentent de l'intérêt. Ce sont précisément les autres informations disponibles sur les personnes interrogées qui vont être "projetées" en éléments illustratifs.

Le fait que le tableau **Z** ne comporte que des "0" et des "1" va permettre une interprétation particulièrement simple de la disposition relative des colonnes illustratives.

La formule de transition (2) précédente nous montre en effet que la coordonnée Φ_j d'une colonne j (réponse illustrative) s'obtient en multipliant par β la simple *moyenne arithmétique* des coordonnées des individus qui ont choisi la réponse j. C'est ainsi que l'on positionne toute catégorie illustrative.

Lecture de la figure 3.4

On lit sur la figure 3.4 le positionnement des formes graphiques dont les libellés et les coordonnées figurent dans le tableau 3.15. Ces éléments de réponse sont donc projetés comme variables illustratives sur le plan factoriel de la figure 3.3 (les modalités des variables actives n'ont pas été reproduites sur la figure 3.4 afin d'en alléger la lecture).

Les liaisons existantes entre l'emploi des mots et les caractéristiques des personnes qui répondent sont visibles immédiatement dans un cadre *qui tient compte des interrelations* entre ces caractéristiques.
Les consultations classiques de tableaux croisés sont en effet hypothéquées par le fait "qu'une variable peut en cacher une autre"... elles sont de plus largement redondantes lorsque les caractéristiques successives sont liées

entre elles. Le système de projection de variables supplémentaires permet donc d'économiser du temps et d'éviter des erreurs d'interprétation.

La série de variables illustratives portée sur la figure 3.4 est constituée par 25 formes graphiques utilisées par les mêmes individus dans leurs réponses à la question ouverte : "*Que signifie pour vous réussir sa vie ?*". Ces 25 formes ont été sélectionnées en raison de leurs positions excentrées, et donc caractéristiques.

Lors d'une application en vraie grandeur, des centaines de formes seront projetées et triées de cette façon.

Tableau 3.15

Coordonnées factorielles et contributions pour les formes supplémentaires

Formes	MASSES	DISTO	COORDONNEES			COSINUS CARRES		
			F1	F2	F3	F1	F2	F3
argent	.009	7.68	.23	.17	.39	.007	.004	.020
arriver	.004	21.18	-.25	.09	.38	.003	.000	.007
atteindre	.003	43.28	-.50	-.36	-.25	.006	.003	.001
bien	.013	6.67	.32	-.34	.15	.016	.017	.004
bon	.004	30.55	-.70	.84	.77	.016	.023	.020
bonheur	.003	32.85	.78	.03	.47	.019	.000	.007
boulot	.005	25.45	-.57	.36	.23	.013	.005	.002
équilibre	.004	13.49	.52	-.46	.58	.020	.016	.025
familiale	.005	14.10	.10	-.24	.22	.001	.004	.004
famille	.009	8.22	.17	-.15	.03	.004	.003	.000
fonder	.002	61.00	-.40	.01	.53	.003	.000	.005
ma	.005	17.09	-.89	-.23	-.23	.047	.003	.003
me	.002	34.56	-.47	-.18	-.50	.006	.001	.007
mes	.002	39.23	-.54	-.30	.03	.007	.002	.000
mari	.003	30.40	-.30	.70	1.13	.003	.016	.042
métier	.005	11.97	-.31	.91	.24	.008	.070	.005
peau	.006	11.09	.06	-.57	.21	.000	.029	.004
professionnelle	.006	13.13	.58	-.55	.70	.026	.023	.038
réussir	.024	2.41	-.20	-.14	.06	.016	.009	.002
réussite	.004	14.54	.52	-.36	.57	.019	.009	.022
santé	.004	22.75	1.30	-.09	-.84	.075	.000	.031
sentir	.003	22.88	-.74	-.08	-.17	.024	.000	.001
situation	.007	16.72	-.74	.42	.36	.033	.010	.008
travail	.014	5.07	.37	-.06	.30	.027	.001	.018
travailler	.004	41.29	.75	-.29	-.73	.014	.002	.013

On trouve par exemple des possessifs (*ma, mes*) dans une zone où les personnes interrogées sont jeunes et instruites. Ils sont accompagnés de verbes *(atteindre, arriver, réussir*). Seul, le verbe *travailler* caractérise les personnes plus âgées, qui s'expriment beaucoup moins à la première personne.

Toutes ces proximités suggèrent en fait des conjectures à vérifier ensuite sur les données initiales, et sur un échantillon plus important.

En résumé...

On peut résumer la démarche suivie lors de cette application de l'analyse des correspondances multiples par les deux étapes :

a) établissement d'une typologie de la population selon un point de vue caractérisé par un ensemble homogène de variables actives (ici : les caractéristiques de base des personnes interrogées).

b) identification des sous-groupes de cette population à partir de toutes les autres informations pertinentes disponibles. En fait, c'est la méthode qui permettra de sélectionner les variables supplémentaires ayant des coordonnées "significatives" sur les axes factoriels, ce qui permet d'envisager des explorations systématiques, avec de nombreux croisements de variables.

Les analyses des correspondances simples et multiples permettent donc d'obtenir des images suggestives, mais fragiles des tables de contingences et des tableaux de codages binaires.

L'utilisation conjointe des méthodes de classification présentées au chapitre suivant permettra à l'utilisateur de procéder à des inductions plus sûres et à des lectures plus confortables.