Analyse factorielle multiple intra-tableaux. Application à l'analyse simultanée de plusieurs questions ouvertes.

Mónica Bécue-Bertaut[1] et Jérôme Pagès[2]

[1]Universitat Politècnica de Catalunya. c/ Pau Gargallo, 5 - 08028 Barcelona, Spain
[2]ENSAR/ INFSA. 65 rue de Saint- Brieuc, F-35042 Rennes cedex, France

Abstract

Textual data studies frequently use Correspondence Analysis (CA) applied to contingency tables. In the simultaneous study of several contingency tables having homologous rows, the usual practice consists in 1) CA on each table ; 2) CA on juxtaposed tables. We propose to complete this methodology with an analysis combining the principle of intra-sets CA (Escofier et Drouet 1983) and the ideas of Multiple Factor Analysis. This method offers a synthetic approach to several contingency tables. This article presents the principle of this method and also an application to the comparison of answers to several open-ended questions belonging to a same survey.

Keywords : Correspondence Analysis, Multiple Factor Analysis, Open-ended Questions

Résumé

L'étude de données textuelles recourt fréquemment à l'Analyse des Correspondances (AFC) pour analyser des tableaux de contingence. Pour l'étude simultanée de plusieurs tableaux de contingence aux lignes homologues, la pratique usuelle consiste à effectuer 1) l'AFC de chacun des tableaux ; 2) l'AFC des tableaux juxtaposés. Nous proposons de compléter cette méthodologie par un outil qui combine les principes de l'AFC intra-tableaux (Escofier et Drouet 1983) et les idées de l'Analyse Factorielle Multiple (AFM). Cette méthode offre une approche synthétique de plusieurs tableaux de contingence. Cet article présente les principes de cette méthode ainsi qu'une application à la comparaison des réponse à plusieurs questions ouvertes posées lors d'une même enquête.

Mots-clés : Analyse des Correspondances, Analyse Factorielle Multiple, Questions ouvertes.

1. Introduction

1.1 Introduction

L'étude de données textuelles recourt fréquemment à la construction de tableaux de contingence. Dans le cas de données d'enquête, un exemple courant de ce type de tableau croise d'une part des catégories de répondants (tranches d'âge, niveaux de diplômes, etc.) et d'autre part les mots utilisés dans les réponses à une question ouverte : le terme général x_{ij}, à l'intersection de la ligne *i* et de la colonne *j*, est le nombre de fois que les répondants appartenant à la catégorie *i* ont utilisé le mot *j*. Classiquement (Lebart, Salem, 1994), on étudie ce tableau en le soumettant à une Analyse Factorielle des correspondances (AFC), qui fournit principalement :

- une structure sur les mots (concrètement une représentation euclidienne), deux mots étant d'autant plus proches qu'ils sont utilisés par les mêmes catégories de personnes ;
- une structure sur les catégories, deux catégories étant d'autant plus proches qu'elles utilisent les mêmes mots.

1.2 Étude simultanée de plusieurs questions ouvertes

Un volet important du traitement des enquêtes concerne les relations entre les réponses à différentes questions. Lorsque ces questions sont ouvertes, en prolongeant la méthodologie précitée, on construit plusieurs tableaux de contingence (un par question) de type *catégories* × *mots*. Ceux-ci peuvent être analysés séparément (par AFC) mais cela est très lourd dès que le nombre de questions dépasse 2 et la synthèse de tels résultats est malaisée. Aussi, profitant de l'homologie des lignes, réalise t-on plutôt l'AFC de la juxtaposition de ces tableaux. Ce traitement fournit une structure sur les mots et les catégories qui apporte des éléments de réponse à des questions de type :

- Un même mot est-il utilisé par les mêmes catégories de personnes selon la question posée ?
- Quelles catégories de personnes utilisent les mêmes mots, au travers de l'ensemble des questions posées ?

Une telle AFC fournit des résultats riches et précieux et constitue, à nos yeux, la méthodologie de référence. Il faut toutefois remarquer que la structure qu'elle produit sur les catégories dépend :

- Des différences entre les profils des marges-en-ligne (sommes des valeurs d'une même ligne) des différents tableaux, dues au fait que certaines catégories s'expriment plus largement sur certaines questions que sur d'autres ;
- De l'importance relative des tableaux dans l'analyse, mesurable au travers des contributions des colonnes, elle-même due :
  - ♦ à des différences entre les nombres totaux de mots des tableaux (certaines questions «engendrent» des réponses plus longues que d'autres) : «toutes choses égales par ailleurs», un tableau influence d'autant plus l'analyse globale que son effectif total est important ;
  - ♦ à des différences d'intensité de structure entre les tableaux (certaines questions différencient plus les catégories que d'autres) : «toutes choses égales par ailleurs», un tableau influence d'autant plus l'analyse globale que sa structure est forte.

Toutes ces informations doivent bien sûr être notées et commentées mais on peut souhaiter qu'elles ne masquent pas les différences de profils de mots «intra-question» dans la mise en évidence d'une structure sur les catégories.

1.3 Analyse factorielle multiple intra-tableaux

Par ailleurs, l'AFC des tableaux juxtaposés ne fournit pas d'indications concernant deux points importants :

- La comparaison des structures sur les catégories induites par chacun des tableaux : e.g. quelles catégories sont à la fois proches du point de vue d'une question et éloignées du point de vue d'une autre ?
- La définition d'une structure sur les questions : quelles questions induisent la même structure sur les individus ? Lesquelles induisent des structures différentes ?

Pour aborder ces deux points, la méthodologie de référence est l'AFC séparée de chacun des tableaux mais, comme déjà dit, la comparaison de ces résultats est toujours longue et souvent inextricable.

Dans ce contexte, une méthodologie d'analyse simultanée d'un ensemble de tableaux de contingence ayant des lignes homologues a été progressivement mise au point (Bécue 1998, Bécue et Pagès 1999). Elle intègre les principes de l'analyse intra (Escofier et Drouet 1983), ce qui élimine l'effet des différences de marges-en-ligne, et l'Analyse Factorielle Multiple

(Escofier et Pagès 1998), ce qui équilibre l'influence des différents tableaux et fournit des représentations graphiques complémentaires. D'où le nom d'AFM intra-tableaux.

L'objet de cet exposé est de présenter, au travers d'un exemple, l'apport de cette méthodologie dans le traitement simultané des réponses à plusieurs questions ouvertes.

2. Exemple

L'exemple est extrait d'une enquête permanente sur les conditions de vie et les aspirations des français (Lebart, 1987). A chaque répétition de l'enquête, 2000 français de 18 ans et plus sont interrogés. Nous utilisons ici l'enquête effectuée en 1988 et trois des questions ouvertes.

Les deux premières questions ouvertes étaient formulées ainsi :

1. *Le nombre de divorces augmente actuellement en France, à votre avis pourquoi ?*
2. *Quelles sont les raisons qui peuvent faire hésiter un couple ou une femme au moment d'avoir un enfant ?*

La troisième question suivait une question fermée demandant aux individus dans quel sens pensaient-ils voir se modifier leurs conditions de vie dans les cinq prochaines années ; cinq niveaux de réponse étaient prévus depuis « s'amélioreront beaucoup » jusqu'à « empireront beaucoup ». Après cette question fermée, la question ouverte

3. *Pourquoi ?*

leur demandait d'expliciter leur choix.

Nous désignons ces trois questions par *Divorce*, *Enfants* et *Avenir*.

Les réponses sont regroupées selon les 9 modalités de la variable *Âge* × *Diplôme* (3 niveaux d'âge : *moins de 30 ans, de 30 à 50 ans, plus de 50 ans* ; 3 niveaux de diplôme : *études élémentaires, études secondaires, études supérieures*). Pour chacune des questions, seuls les mots de fréquence supérieure à 15 sont conservés : 159 mots pour la question *Divorce*, 126 mots pour la question *Enfants* et 154 mots pour la question *Avenir*. Les tableaux *catégories* × *mots* sont construits et juxtaposés, formant ainsi le tableau multiple à analyser (cf. Figure 1).

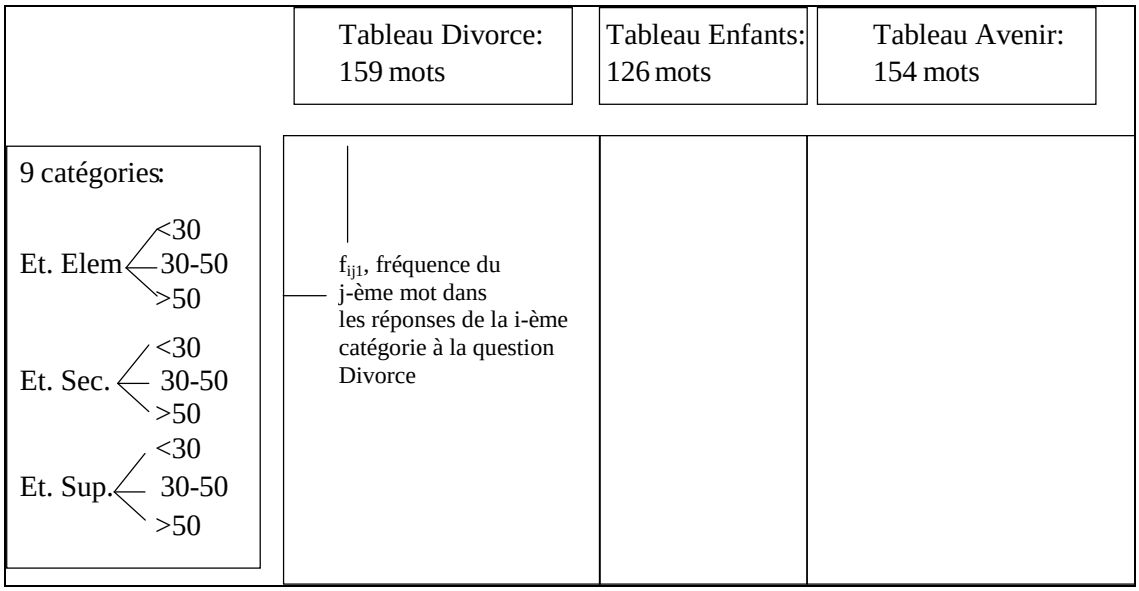


Figure 1. Juxtaposition des trois tableaux lexicaux agrégés correspondant aux trois questions

3. La méthodologie : AFM-intra du tableau juxtaposé

Notations. f_{ijt} : fréquence associée à la ligne i et la colonne j du tableau t ; un indice remplacé par un point indique la sommation sur cet indice.

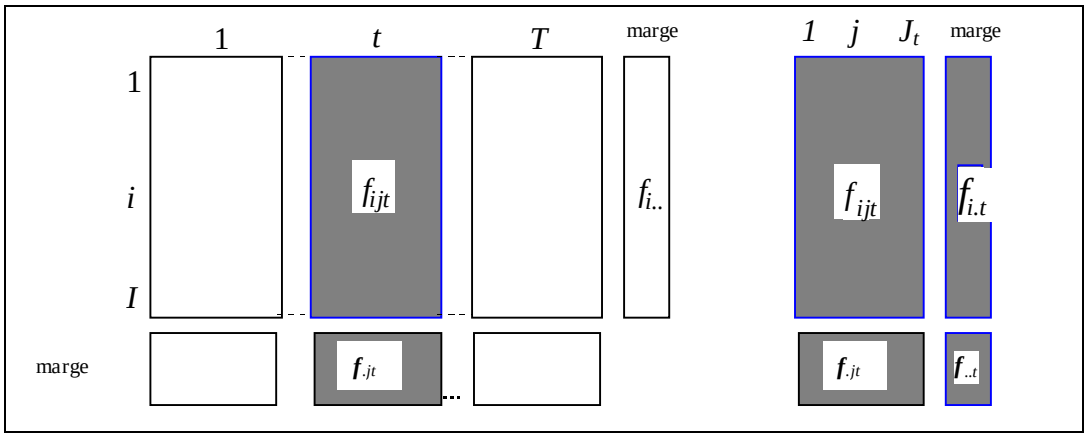


Figure 2. Le tableau de contingence multiple et ses marges : notations

Première étape : *analyses séparées*. L'AFC de chacun des tableaux permet d'obtenir une première vision des données et d'explorer l'existence de structures communes aux différents tableaux.

Deuxième étape : analyses pseudo-séparées. Elles consistent en l'AFC de chacun des tableaux, mais en imposant les marges-lignes $\{f_{i..}, i=1,\ldots,I\}$ et les marges-colonnes, $\{f_{.jt}, j=1,\ldots J\}$. La première valeur propre de chaque analyse, notée λ_1^t, est utilisée dans la 3e étape pour pondérer les colonnes afin d'équilibrer l'influence de chacun des tableaux dans l'analyse globale. Cette AFC du tableau t est équivalente à l'ACP du tableau de terme général :

$$\frac{f_{ijt} - \left(\frac{f_{i.t}}{f_{..t}}\right) \cdot f_{.jt}}{f_{i..} f_{.jt}} \qquad (1)$$

avec les poids ($f_{i..}$) pour les lignes et les poids ($f_{.jt}$) pour les colonnes. Les lignes ont ainsi le même poids pour toutes les analyses, i.e. le poids moyen calculé sur l'ensemble des tableaux.

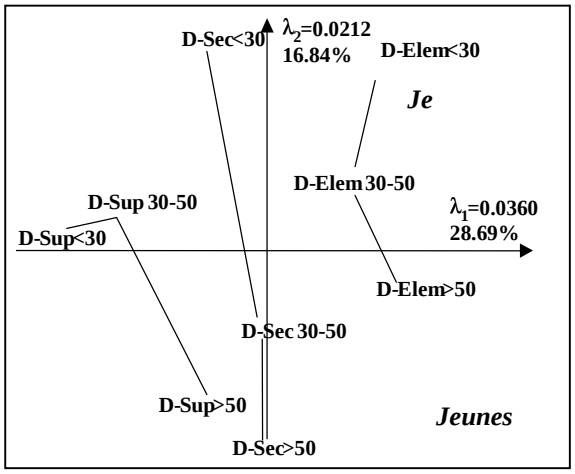
Troisième étape: analyse globale. Elle consiste en une Analyse Factorielle Multiple adaptée aux tableaux de contingence. On réalise une ACP non normée des tableaux juxtaposés, de terme général donné par (1), en donnant le poids $f_{i..}$ à la ligne i et le poids $f_{.jt}/\lambda_1^t$ à la colonne (j,t). Cette étape offre des résultats :

- analogues à ceux de l'AFC appliquée aux tableaux juxtaposés (principalement, une représentation globale des lignes-catégories et des colonnes-mots) ;
- spécifiques des tableaux multiples (principalement, la représentation superposée des structures des catégories induites par chacune des questions – structures partielles – et la représentation des facteurs dérivés des analyses pseudo-séparées).

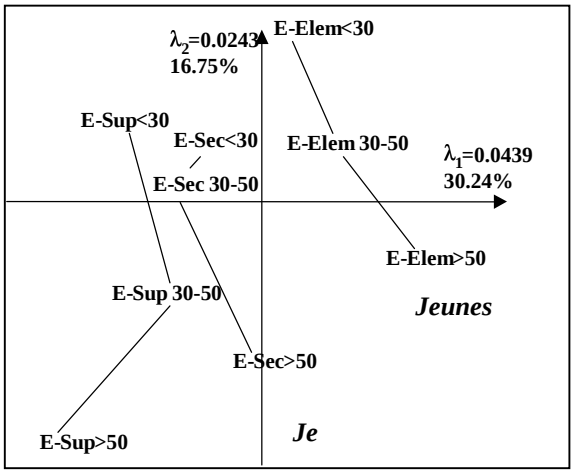
La lecture des résultats est facilitée par les nombreuses aides à l'interprétation de l'AFM.

4. Résultats

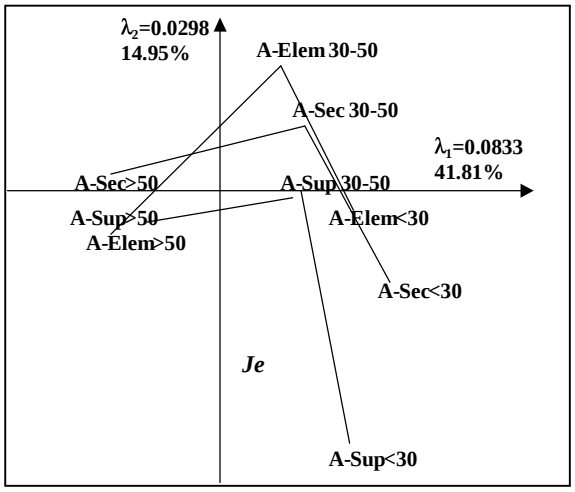
4.1 Analyses séparées des trois tableaux



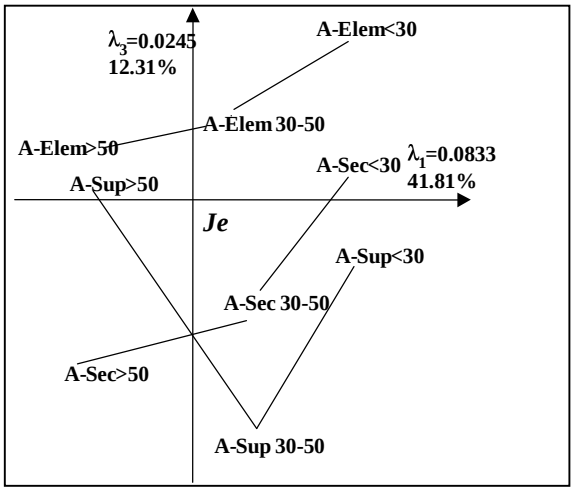
3.a Plan (1,2) de l'AFC de Divorce



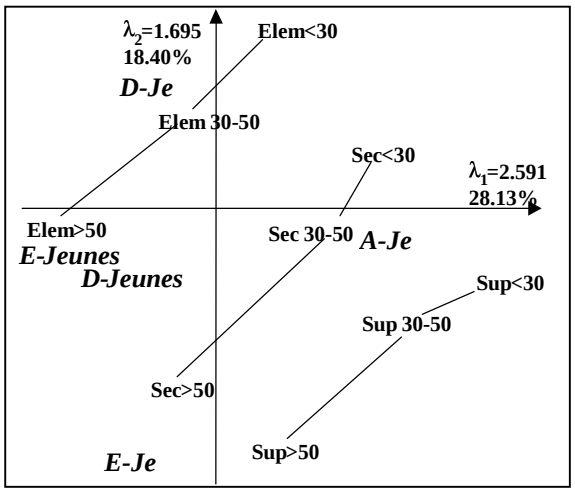
3.b Plan (1,2) de l'AFC de Enfants



3.c Plan (1,2) de l'AFC de Avenir



3.d Plan (1,3) de l'AFC de Avenir



3.e AFM-intra des 3 tableaux juxtaposés

LEGENDE:

D: Divorce	Elem: Etudes élémentaires
E: Enfants	Sec: Etudes secondaires
A: Avenir	Sup: Etudes supérieures

Sur les premiers plans issus des AFC de *Divorce* et *Enfants, le* premier axe est un axe de niveau d'études; le deuxième, un axe d'âge.

On retrouve ces axes, inversés, sur le plan (1,3) de l'AFC de *Avenir* : le premier est un axe d'âge, le troisième est un axe de niveau d'études.

Figure 3 . Premier plan factoriel des trois analyses séparées

4.2 Analyses pseudo-séparées

Pour effectuer les analyses pseudo-séparées, on attribue à chaque catégorie son poids moyen calculé sur l'ensemble des questions (proportionnel à la longueur du sous-corpus formé par les réponses de la catégorie, toutes questions confondues). Dans cet exemple, les poids des catégories pour chacune des questions et les poids moyens sont notablement proches. On peut donc considérer que la déformation de la structure des lignes introduite par la modification des poids est négligeable. De fait, les valeurs propres obtenues sont très voisines des valeurs propres calculées dans les analyses séparées : $\lambda_1^1 = 0.0365$; $\lambda_1^2 = 0.0429$; $\lambda_1^3 = 0.0836$.

4.3 AFM intra-tableaux

Le troisième tableau (question *Avenir*) a une structure plus forte que les autres. En absence de pondération des colonnes, ce tableau aurait une influence prédominante sur la détermination des axes.

4.3.1 Les facteurs de l'analyse globale

L'AFM intra-tableaux procure deux valeurs propres dominantes : $\lambda_1 = 2.59$ et $\lambda_2 = 1.69$ (respectivement, 28.13% et 18.40% de l'inertie totale). Le tableau 1.a montre que chacun des trois groupes de mots-colonne, correspondant aux trois questions, fournit un apport important et équilibré à l'inertie du premier facteur. Les mots-colonne *Divorce* et *Enfants* contribuent nettement plus à l'inertie du deuxième facteur que les mots-colonne *Avenir*.

Les corrélations entre le premier facteur global et les projections des trois nuages-catégories, définis par chacune des questions sont élevées (tableau 1.b). En ce qui concerne le deuxième facteur, la corrélation est forte avec les projections des deux premiers nuages partiels (*Divorce* et *Enfants*), moindre mais néanmoins élevée avec celle du troisième (*Avenir*).

On en conclut que les deux premiers facteurs sont communs aux trois nuages-question. Le premier facteur de l'analyse globale et, dans une moindre mesure, le deuxième constituent des directions importantes d'inertie pour chacune des questions, particulièrement pour la question *Enfants*, cependant non confondues avec les principales directions de dispersion des trois nuages. L'apport moindre de la question *Avenir* provient de la plus faible dimensionalité de ce tableau qui ne présente qu'une valeur propre dominante.

	F1	F2
Inertie totale	2.59	1.70
Divorce	0.86	0.64
Enfants	0.92	0.61
Avenir	0.81	0.45

Tableau 1.a
Décomposition de l'inertie des deux premiers facteurs de l'AFM selon les trois questions

	F1	F2
Divorce	0.97	0.95
Enfants	0.98	0.97
Avenir	0.93	0.81

Tableau 1.b
Corrélations entre la projection du nuage global et celle de chacun des trois nuages partiels associés à chacune des questions

Tableau 1. Les facteurs de l'analyse globale, directions de dispersion des nuages partiels

La représentation des catégories issue de l'AFM est donnée figure 3-e. Prises en compte simultanément, les 3 questions confèrent aux catégories une structure très régulière, compromis

entre les représentations des AFC séparées. Dans le détail, on note que, globalement, les *30-50 ans* sont plus proches des *plus de 50 ans* que des *moins de 30 ans*.

Le calcul des corrélations entre les trois premiers facteurs normés des trois AFC pseudo-séparées et les deux premiers facteurs de l'AFM intra-tableaux permet d'étudier les relations entre les facteurs de ces quatre analyses. Les plans engendrés par les deux premiers facteurs de la question *Divorce*, par les deux premiers facteurs de la question *Enfants* et par les premier et troisième facteurs de la question *Avenir* sont tous trois très proches du plan engendré par les deux premiers facteurs de l'AFM. On peut aussi noter que le deuxième facteur de l'AFC de la question *Avenir* est très corrélé au troisième facteur de l'analyse globale.

La qualité de représentation des trois nuages de mots-colonne sur le plan principal de l'AFM est très proche de la qualité de représentation de ces mêmes nuages sur les plans principaux séparés : 43.4% au lieu de 45.4% pour *Divorce*, 44.9% au lieu de 46.6% pour *Enfants* et 53.6% au lieu de 57.6% pour *Avenir*. Dans ce dernier cas, il est intéressant de noter que la perte en qualité de représentation sur le premier axe est importante (34.6% au lieu de 42.7%) mais que cette perte est compensée par un gain sur le deuxième axe (19% au lieu de 14.86%).

4.3.2 Représentation superposée des nuages partiels

Afin de comparer les structures des catégories observées dans les différentes villes, on projette successivement, en tant que lignes supplémentaires, l'ensemble des lignes de chacun des tableaux $\mathbf{Y_t}$, complétées par des zéros. On obtient ainsi une représentation qui superpose la description globale des catégories et celles-ci induites par chaque $\mathbf{Y_t}$ (on parle alors de catégories partielles). La figure 4 reproduit un extrait de cette représentation : on peut ainsi observer les trajectoires des catégories *études élémentaires* et *études supérieures* telles qu'elles sont décrites à travers leurs réponses aux questions *Divorce* et *Avenir*.

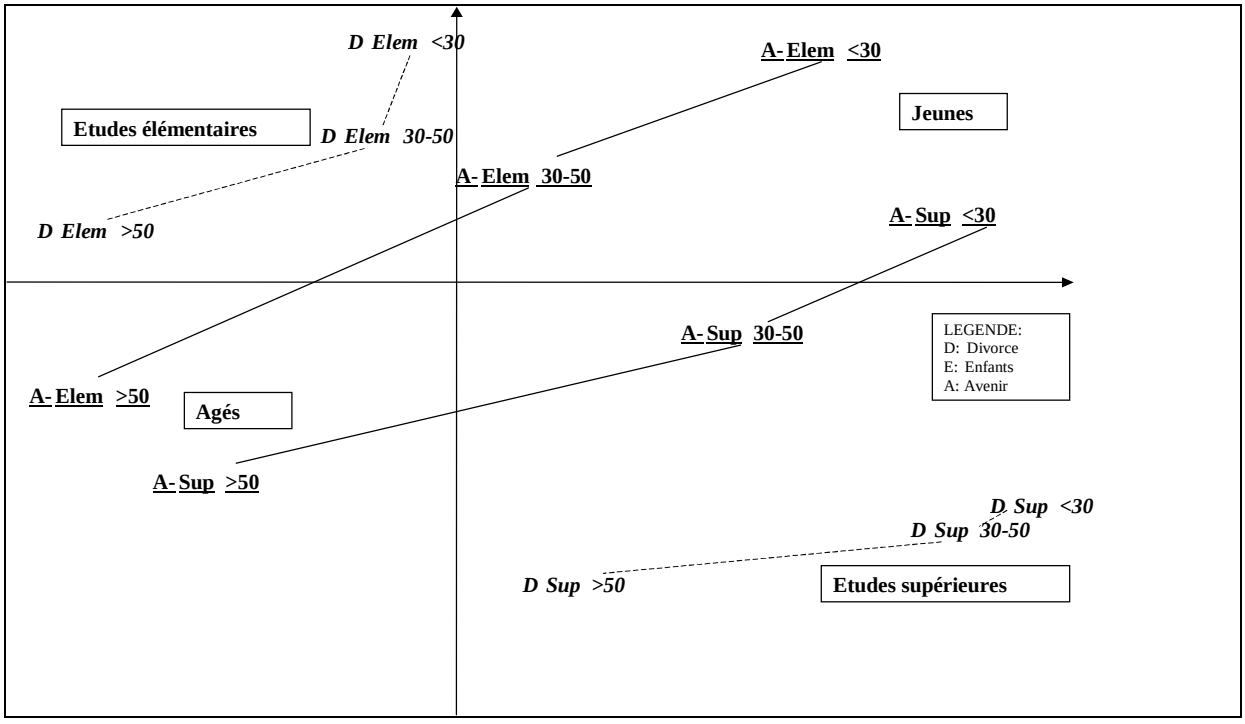


Figure 4. Extrait de la représentation superposée des nuages «partiels» : Trajectoires des catégories d'âge correspondant aux études élémentaires et supérieures pour les questions Avenir (A) et Divorce (D).

Cette représentation permet de retrouver les grands traits des représentations issues des AFC séparées. Ainsi :

- Parmi les *plus de 50 ans*, les *diplomés élémentaires* et *supérieurs* sont peu différenciés par la question *Avenir* ; autrement dit, les *diplômés supérieurs de plus de 50 ans* emploient un vocabulaire moins intellectuel que le niveau d´études ne le laissait présager lorsqu'ils répondent à la question *Avenir*.
- Parmi les *diplomés supérieurs*, les *moins de 30 ans* et les *30-50 ans* sont peu différenciés par la question *Divorce* ; ces deux catégories ont un vocabulaire très marqué par le niveau d'études lorsqu'ils répondent à la question Divorce et, pour les 30-50 ans, plus jeune que ce qui correspond à leur âge.
- Plus généralement, cette représentation met en évidence un plus grand effet de l'âge sur les réponses à la question *Avenir* et un plus grand effet du diplôme pour la question *Divorce*.

4.3.3 Représentation superposée des mots et des catégories

La représentation des mots permet d'étudier les proximités entre les mots utilisés pour répondre à une même question ou à des questions différentes. On peut superpos er la représentation des mots et celle des catégories : il existe en effet entre ces deux représentations des règles de transition (Bécue et Pagès, 1999).

En particulier, il est intéressant d'étudier comment les mêmes mots sont ou non choisis par les même catégories selon la question posée. Ainsi, *Jeunes* est employé dans les questions *Divorce* et *Enfants* par les mêmes catégories (répondants de plus de 30 ans, de formation élémentaire ou secondaire), tandis que l'emploi de *Je* présente un usage différencié selon la question : il est plutôt employé par les catégories peu diplômées pour répondre à la question *Divorce* (relativement souvent pour indiquer une absence d'opinion avec *je ne sais pas*) ; il est principalement employé par les répondants de plus de 50 ans dans la question *Enfants* (d'une façon dominante pour exprimer *je ne sais pas*) ; enfin, il est très employé par toutes les catégories mais légèrement plus par les répondants d'âge inférieur 30 ans de niveau d'études secondaire ou supérieur, pour donner un ton plus personnel à la réponse à la question *Avenir* (*je ne m'attends pas à voir mon salaire augmenté, je vais peut-être quitter mon travail, parce que je suis ambitieux, je suis vieux, etc.*)

Références

Bécue, M. (1998). Three-way textual data analysis in*: Advances in Data Science and Classification,* Rizzi A., Vichi M., Bock H.-H., Springer, 457-464.

Bécue, M., Pagès J. (1999). Intra-Sets Multiple Factor Analysis. Application to textual data. *Proc. of the 9th International Symposium on Applied Stochastic Models and Data Analysis*, J. Jansen et al. (eds), Universidade de Lisboa Editor, 51-60.

Escofier, B., Drouet, D. (1983). Analyse des différences entre plusieurs tableaux de fréquence, *Les Cahiers de l'Analyse des Données*, VIII, 4, Dunod, Paris, 491-499.

Escofier, B., Pagès, J. (1988-1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation,* Dunod, Paris.

Lebart, L. (1987). Conditions de vie et aspirations des français, évolution et structure des opinions de 1978 à 1986, *Futuribles*, septembre 1987, 25-56. 0,4

Lebart, L., Salem, A. (1994) *Statistique Textuelle,* Dunod, Paris.