

Extraction d'informations stratégiques par Analyses Factorielle des Correspondances

Bernard DOUSSET
IRIT/ SIG, Université Paul Sabatier,
118 route de Narbonne, 31062 Toulouse cedex 04
dousset@irit.fr

1 Introduction

Elle s'applique aux données qualitatives : tableaux individus – variables qualitatives multimodales, matrices de présence – absence, matrices de contingences et de corrélations. Elle nécessite un pré traitement des données :

- Normalisation des lignes de la matrice (somme des poids unitaire),
- Ajout, au fond de la matrice, d'une matrice identité, afin de considérer les variables pures (colonnes) comme des individus complémentaires.

Chaque ligne correspond alors à un profil unitaire, il suffit donc de faire une analyse en composantes principales de ces profils. Comme les variables sont ici représentées sous forme d'individus complémentaires, une seule carte factorielle suffit pour rendre compte de façon globale des correspondances :

- Entre les individus,
- Entre les variables,
- Entre les individus et les variables.

Comme dans le cas des ACP, un observatoire graphique des caractéristiques de l'analyse nous permet de connaître la profondeur de l'exploration à entreprendre pour atteindre 80 à 90% de l'information. Aux histogrammes des modules des valeurs propres et du taux d'information contenu dans les espaces principaux, nous avons ajouté la répartition de densité des individus et des variables rapportée au plus dense, ainsi que la distribution des corrélations. D'un seul coup d'œil, il est alors possible de se faire une idée de la qualité de l'analyse en cours et de la présence ou non de faits remarquables [DOUS93]. Dans les histogrammes ci-dessous, nous pouvons détecter la présence de corrélations supérieures à 0,5 et donc de forts liens entre certains individus de la matrice analysée.

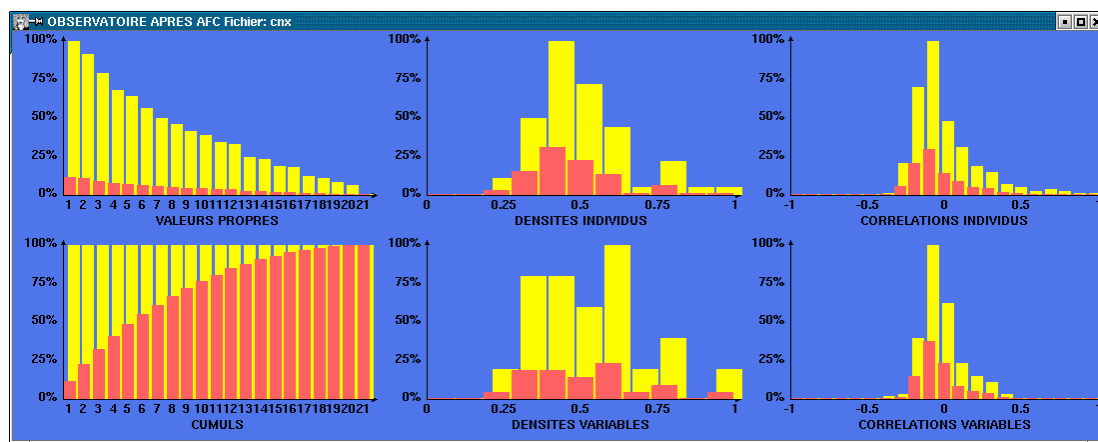


Figure 1. : Observatoire de la qualité d'une analyse multidimensionnelle.

2 Cartes produites

Les cartes factorielles 4D, ci-dessous, sont présentées sous un azimut permettant de distinguer un maximum d'associations journaux – pays.

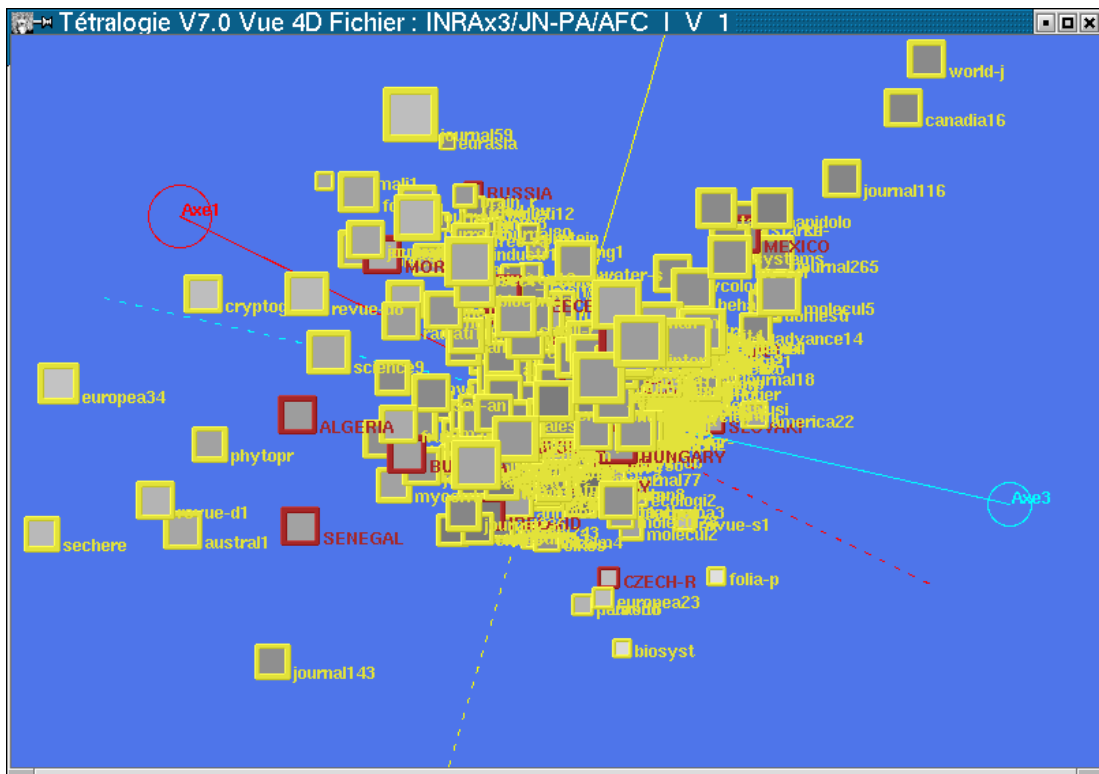


Figure 2. : Carte factorielle 4D d'une AFC réalisée sur une matrice Journaux - Pays.

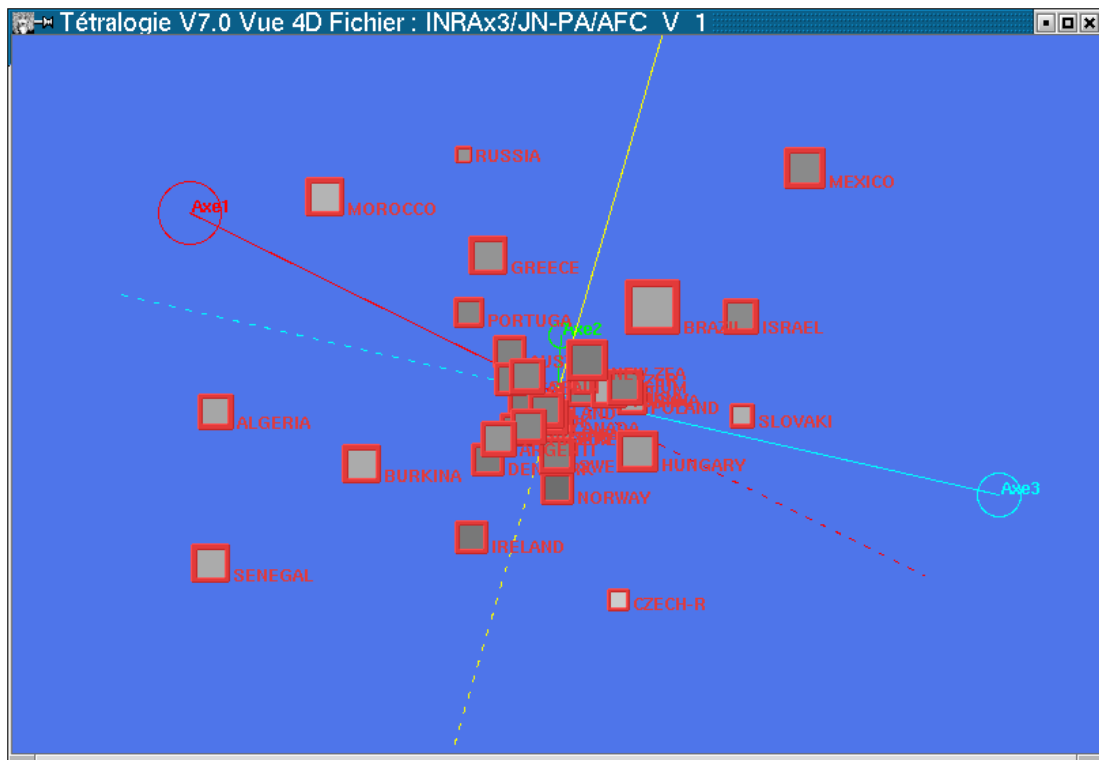


Figure 3. : Même carte factorielle 4D ne visualisant que les pays.

Extraction d'informations stratégiques par Analyses Factorielle des Correspondances

Les cartes factorielles 4D, ci-dessous, sont présentées sous un azimut permettant de distinguer un maximum d'associations journaux – pays.

La notion de proximité entre individus et variables a maintenant un sens. Comme ce sont les typologies qui sont analysées, les plus fortes sortent en premier sur le sous-espace principal. Une analyse descendante (par visualisation des axes secondaires ou glissement d'axes) est souvent nécessaire pour découvrir toutes les informations pertinentes. Une deuxième solution est d'éliminer progressivement dans le tableur les cas les plus typiques en vérifiant à chaque fois que la matrice produite est connexe. Une nouvelle AFC permet de réaliser un zoom avant sur les typologies les plus marquées issues du centre du nuage.

Ci-dessous, un autre exemple de carte factorielle obtenue après AFC sur une matrice de cooccurrence Thèmes – Auteurs pour un grand laboratoire de recherche en informatique. Nous pouvons remarquer des groupes d'auteurs corrélés par leur thématique de recherche, des thèmes interfaces (entre deux groupes d'auteurs), certains experts des interfaces, des thèmes interfaces avec ou sans expert, des variations continues de thématiques. Après vérification, les équipes thématiques détectées correspondaient parfaitement avec celles de la plaquette de présentation du laboratoire en question. Trois constatations sont alors possibles :

- Le laboratoire concerné a réalisé une présentation objective de sa structure,
- Il est possible de connaître parfaitement cette structure depuis l'extérieur,
- De plus, les qualités et les défauts de cette structure sont maintenant connus.

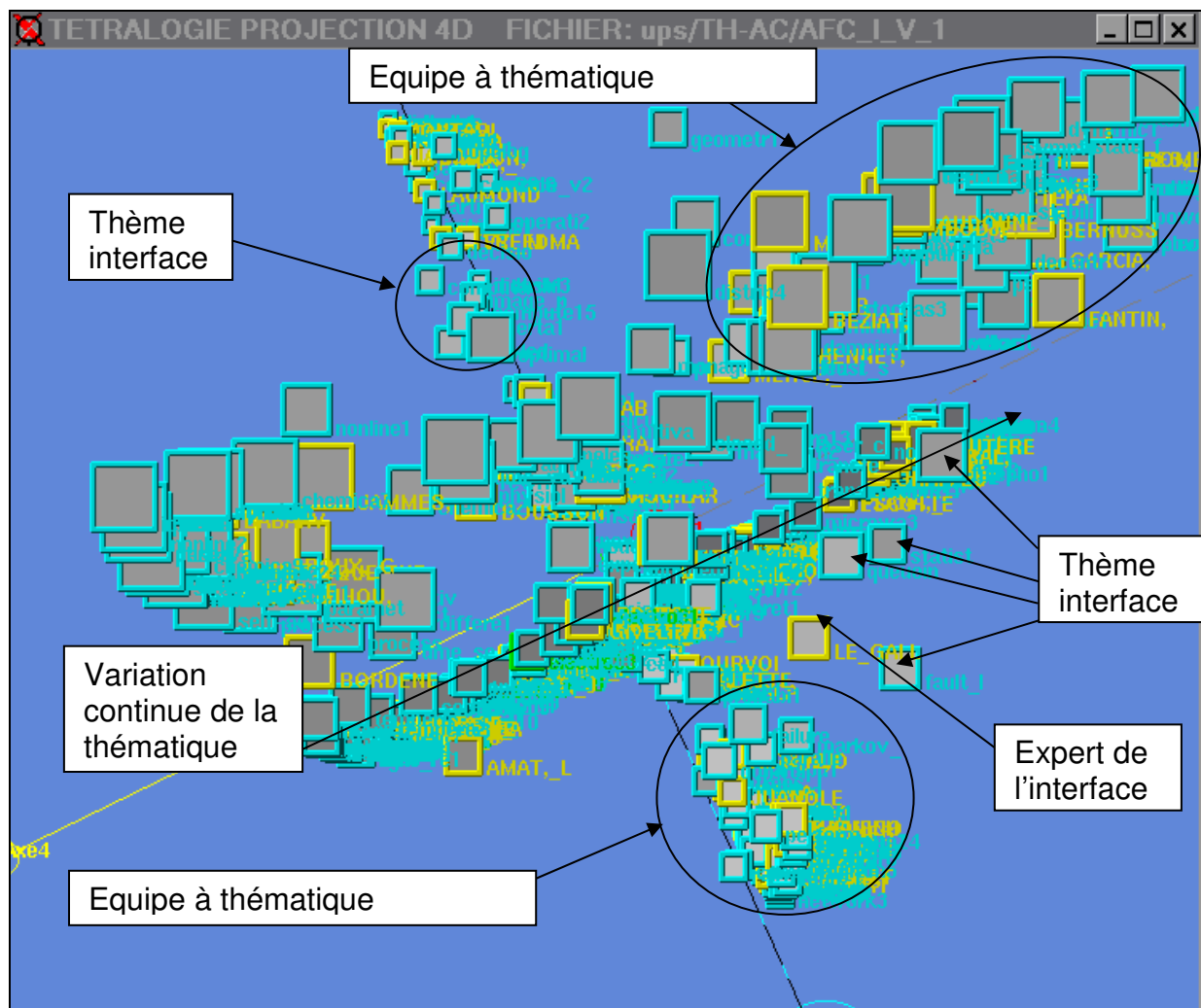


Figure 4. : Carte factorielle en 4D d'une AFC Thématique – Auteurs.

3 Mise en valeur d'une hypothèse

Lorsque nous nous intéressons à un élément particulier (période, acteur, thème) ou à un groupe, il est possible de le mettre en évidence en l'amenant par le biais des rotations, choix des axes et zooms sur un côté de la vue. Les éléments qui lui sont corrélés le suivent, il est donc facile d'illustrer une hypothèse image à l'appui. En 4D, il est même possible de cumuler les effets discriminants de 4 axes pour réaliser cette mise en valeur. Dans certains cas (trop d'items à afficher), on peut ne manipuler que la classe des items recherchés. Dans l'exemple illustré ci-dessous, nous avons choisi de manipuler le résultat d'une AFC réalisée sur une matrice de cooccurrence entre mots-clés et auteurs principaux.

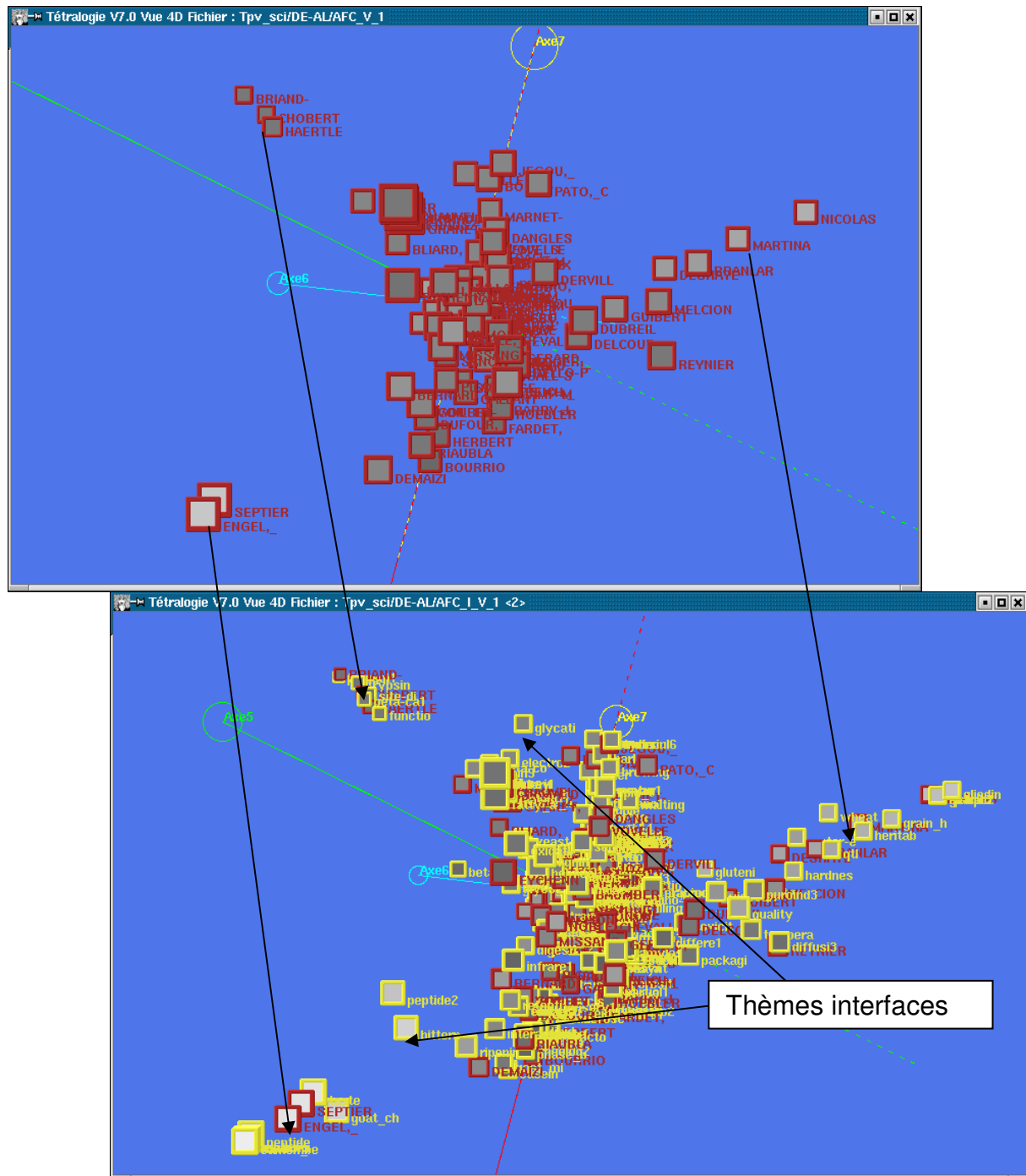


Figure 5. : Utilité d'exporter l'azimut de la carte simplifiée vers la carte générale.

- **En haut à gauche** nous cherchons à extraire une ou plusieurs équipes aux thèmes cohérents (carte des auteurs seuls : colonnes de la matrice),
- **En bas à droite**, après exportation de l'azimut convenablement choisi, nous visualisons la carte complète avec les mots-clés.

C'est ainsi qu'apparaissent des liens thématiques entre auteurs, donc des groupes thématiques. Il est possible de trouver des thèmes interfaces entre deux groupes (ensembles de mots-clés simultanément utilisés par les deux groupes) et parfois des spécialistes de ces interfaces. Enfin, le croisement de la sémantique avec des acteurs de même type met en évidence une structuration du domaine beaucoup plus synthétique qu'avec les documents. En effet, un acteur donné produit dans le temps des documents complémentaires et dévoile ainsi sa stratégie. On peut facilement imaginer l'utilité d'une carte qui cumule simultanément tous ces indices, surtout s'il est possible de la manipuler à sa guise.

4 Conclusion

L'analyse factorielle des correspondances s'applique essentiellement au croisement de variables qualitatives (acteurs ou éléments sémantiques), mais elle n'est réellement interprétable que si l'espace d'analyse est de dimension assez réduite (quelques dizaines de variables principales). Comme il s'agit d'une analyse de typologies, les correspondances les plus typiques apparaissent en premier. Pour visualiser les signaux forts (souvent moins typés) il faut avoir recours aux glissements d'axes afin de les découvrir au niveau des axes secondaires. Dans ces conditions, il s'agit certainement de la plus puissante méthode de détection et de synthèse des relations entre tous les types de variables rencontrées dans les informations textuelles.