EXPLOITATION GRAPHIQUE DES PLANS FACTORIELS

par « Le Cercle Factoriel »[1]

Résumé

Les moyens informatiques graphiques actuels doivent faciliter la tâche d'interprétation des résultats d'Analyses des Données. Le « Cercle Factoriel » a travaillé à la définition d'un jeu de fonctionnalités qui devraient figurer dans les futurs logiciels d'exploration des tableaux de données.

1•INTRODUCTION

1•1 LES OBJECTIFS DU GROUPE DE TRAVAIL

Les moyens informatiques actuels permettent d'améliorer considérablement l'exploitation graphique des résultats d'analyse des données. Ceci est rendu possible grâce notamment à un affichage graphique de qualité et à l'interactivité. L'interactivité permet à l'utilisateur de modifier en temps réel les éléments du graphique par l'intermédiaire du clavier ou de la souris.

Le « Cercle Factoriel » — sous-groupe du Groupe ASU-Logiciels et Statistique — s'est fixé comme objectifs de réfléchir à l'exploitation graphique des plans factoriels en s'appuyant sur l'expérience de ses participants et en s'inspirant des éléments bibliographiques cités et de l'analyse de certains logiciels existants dont il a organisé les démonstrations.

On notera que la réflexion du groupe — les participants étant plus statisticiens qu'informaticiens — ne porte pas sur les moyens logiciels et matériels disponibles actuellement ou qui devraient être disponibles pour réaliser au mieux l'exploitation graphique des plans factoriels On peut cependant penser que les propositions faites ici sont réalisables concrètement avec les moyens actuels

1•2 PARTICULARITES DES GRAPHIQUES FACTORIELS

Le « Cercle Factoriel » a restreint son étude et ses propositions au cas des plans factoriels, sachant que les représentations à trois dimensions présentent des problèmes spécifiques certainement encore plus difficiles à traiter.

D'autre part on limite notre travail aux plans issus des méthodes classiques suivantes :

- l'analyse en composantes principales
- l'analyse des correspondances simples
- l'analyse des correspondances multiples
- optionnellement, l'analyse du triple

---

[1] **Membres rédacteurs**: D. Ambroise, J.M. Bernard, J.L. Blanchard, F. Goupil-Testu, D. Grangé, A. Morineau, F. Sermier, G. Thauront, N. Valette. Le groupe était animé par A. Morineau.

1•3 LES FONCTIONNALITES GRAPHIQUES ET STATISTIQUES

Les fonctionnalités graphiques sont les outils permettant de mettre en évidence l'information apportée par les plans factoriels. À ce titre, un logiciel d'exploitation graphique des plans factoriels devra par exemple permettre :

- la liaison dynamique de différents plans entre eux
- la liaison dynamique d'un plan avec le tableau des données
- la sauvegarde complète permettant de reprendre un graphique inachevé,
- l'habillage des éléments du plan factoriel, etc.

La distinction entre fonctions graphiques et statistiques n'est pas très nette. On peut cependant classer dans la zone statistique les fonctionnalités destinées à accroître l'information statistique extraite des plans factoriels ou, plus exactement, les procédures de type statistique conduisant à faciliter la lecture et l'interprétation des plans.

Dans la classe des fonctionnalités de type statistique, on trouvera par exemple, différentes procédures de sélection de points reposant sur des critères comme la contribution à l'inertie, le $\cos^2$ ou la valeur-test. Ces fonctionnalités pourront être déclinées suivant la nature de l'analyse, le rôle actif/illustratif des points et la nature même des points à représenter.

2•GENERALITES

2•1 CHOIX DU VOCABULAIRE

Il y a plusieurs façons de nommer les éléments intervenant dans les analyses de données. Nous adopterons arbitrairement un certain vocabulaire, le petit lexique ci-dessous permettant de faire les correspondances :

ligne du tableau ou **individu** (ligne est plus général)
unité statistique
observation

colonne du tableau ou **variable** (colonne est plus général)
question
paramètre
observation

variable **continue**
variable quantitative

variable **nominale**
variable qualitative
variable catégorielle
facteur contrôlé

variable en **classe**

modalité d'une variable
item
catégorie
classe
groupe
niveau du facteur

variable **fréquence** (ligne ou colonne)
variable d'effectifs
contingence
tableau statistique

individu **actif** ou variable **active**
principal

individu **illustratif** ou variable **illustrative**
supplémentaire
passif

2•2 LE PLAN FACTORIEL : UN PLAN TRES PARTICULIER

Un plan factoriel se différencie d'un graphique (x,y) usuel de plusieurs façons :

- Les axes sont « hiérarchisés »: un des axes est plus important que l'autre au sens où il indique une direction de plus grande dispersion (ou inertie ou allongement) que l'autre.
- Les points positionnés sur le graphique peuvent être de différents types (par exemple des points-individus et des points-variables). On parle alors de représentation simultanée.
- Les unités sur les axes dépendent dans certains cas de l'intention du graphique.
- La localisation exacte des points dans le plan factoriel ne permet pas d'apprécier exactement la distance réelle entre les points ni la distance au centre du graphique.
- Les zones du graphique ne sont pas équivalentes (par exemple le centre du graphique est souvent moins intéressant que la périphérie et les distances ne s'y lisent pas de la même façon).
- Certains points définissent plutôt des directions que des localisations. Dans ce cas on peut répéter pour les angles entre directions ce que l'on a dit pour les distances entre points.
- Les distances et les angles se lisent avec des règles de lecture qui diffèrent selon la nature de l'analyse et la nature des points.
- Le sens des axes est arbitraire (toutes les symétries sont possibles).

2•3 LES DIFFERENTS TYPES DE POINTS ET LEUR ROLE

Un tableau de données soumis à une analyse factorielle contient en général plusieurs types de « points » définis par les lignes et les colonnes, et les points peuvent jouer l'un ou l'autre deux rôles : actifs ou illustratifs.

2•3•1 Les points-individus

Ils correspondent en général aux lignes du tableau analysé par ACP ou ACM (et dans certaines applications de l'AFC). Il s'agit des « individus statistiques » c'est-à-dire des objets en général extraits d'une population et sur chacun desquels les observations ont été faites.

2•3•2 Les points-variables continues

Dans une analyse en composantes principales, un point-variable continue est défini par ses n coordonnées dans l'espace à n dimensions des individus.

On définit une distance entre deux variables continues actives. Suivant le cas cette distance s'exprime à partir des covariances (analyse non normée) ou à partir des corrélations (analyse normée). Le nuage des points-variables actives dans un plan factoriel permet de visualiser ce type de distances entre variables. Les points-variables continues illustratives sont positionnés dans le plan en fonction de leur relation aux axes factoriels.

Entre deux variables continues illustratives, ou entre une active et une illustrative, il n'y a jamais de calcul de distance.

2•3•3 Les points-anciens axes unitaires

Une analyse en composantes principales peut être considérée comme un changement de repère. Le nuage des points-individus est initialement construit dans le repère originel des variables, colonnes du tableau des observations. Un axe unitaire de ce repère donne dans ce nuage d'individus la direction dans laquelle la variable correspondante va croissant. Le transfert de ces anciens axes unitaires dans le nouveau repère des axes factoriels fournit des points associés aux variables et dont la position au sein des individus présente des interprétations intéressantes.

2•3•4 Les points-modalités

Ce sont les points représentant les modalités des variables nominales. Ils représentent donc des groupes d'individus. On verra que, dans tous les cas, les coordonnées de ces points sont les moyennes des coordonnées des individus du groupe (éventuellement à un coefficient près sur le plan factoriel, coefficient dépendant de l'axe).

2•3•5 Les points-fréquences

Ce sont les colonnes (et généralement aussi les lignes) d'un tableau de fréquences soumis à une AFC. Mais ce sont parfois des colonnes de fréquences dans un tableau « individus x variables ». Dans ce cas ce sont des variables illustratives pour une ACP ou une ACM.

2•3•6 Rôle actif ou illustratif

Un point est actif s'il participe au calcul de l'inertie du nuage de points. Il participe alors à la détermination des axes du plan factoriel. Sinon il est dit illustratif ou supplémentaire. Le choix du rôle joué par un point est essentiel dans la construction

de l'analyse et la connaissance de ce rôle est tout autant essentielle au moment de l'interprétation des résultats

2•4 NATURE DE L'ANALYSE

Dans ce paragraphe, nous rappelons brièvement la nature des principales analyses factorielles. Celle-ci dépend du type des éléments actifs. La discussion sur les éléments illustratifs lorsqu'ils sont d'un type différent des éléments actifs est faite au paragraphe concernant les représentations simultanées.

2•4•1 ACP : Analyse en Composantes Principales

2•4•1•1 ACP normée

Les variables actives sont des variables continues. La distance entre deux individus est la distance euclidienne usuelle calculée avec les valeurs centrées et réduites des variables. La distance entre deux variables actives ne dépend que de la corrélation linéaire entre les deux variables :

$$d^2(i,j) = 2\{(1-cor(i,j)\}$$

2•4•1•2 ACP non normée

L'analyse est non normée quand le calcul des distances entre individus se fait sans réduire chaque variable par son écart-type. Dans ce cas, la distance entre deux variables dépend des variances et des covariances :

$$d^2(i,j) = var(i) + var(j) - 2\ cov(i,j)$$

2•4•2 ACM : Analyse des Correspondances Multiples

Comme dans l'analyse en composantes principales, les lignes du tableau seront des points-individus. Les points-variables seront les modalités des variables nominales actives. Pour construire le nuage des individus, la distance entre deux points-individus actifs est la distance du c^2 entre les profils de ces deux points.

Pour construire le nuage des points-modalités actives, on utilise aussi la distance du c^2 entre les profils des deux modalités.

On démontre que, pour chaque axe et à un coefficient près (dépendant de l'axe), la coordonnée d'un point-modalité est égale à la moyenne des coordonnées des individus correspondants. De même, la coordonnée d'un point-individu sur un axe est égale (à un coefficient près) à la moyenne des coordonnées des modalités auxquelles il appartient.

Les points-individus et les points-modalités considérés comme lignes et colonnes d'un tableau restent dans des espaces distincts; les points-modalités considérés comme points moyens des individus sont alors dans le même nuage que les individus.

Ces dernières propriétés sont utilisables pour calculer d'une part les coordonnées d'un point-individu illustratif (point moyen des modalités actives auxquelles il appartient) et

d'autre part les coordonnées d'un point-modalité illustrative (point moyen des individus actifs qui la composent).

2•4•3 AFC : Analyse Factorielle des Correspondances

Les lignes actives et les colonnes actives du tableau sont des fréquences (des comptages). Les points-lignes sont les profils de distribution des fréquences en ligne et la distance entre deux points-lignes est la distance du c^2; le poids affecté à un point-ligne est proportionnel à la fréquence cumulée en ligne. Pour la construction des points-colonnes, les fréquences en colonne jouent exactement le même rôle que dans le cas des lignes : le tableau peut être transposé sans changer les résultats.

On démontre que la coordonnée sur un axe d'un point-ligne actif est, à un coefficient près ne dépendant que de l'axe, le barycentre des coordonnées (sur l'axe de même numéro) des points-colonnes actifs en utilisant comme poids les éléments du profil de la ligne. Cette règle est utilisée pour calculer la coordonnée sur un axe de toute ligne illustrative.

2•4•4 Analyse factorielle sur tableau de dissimilarités (ou analyse du triple)

Elle présente un seul type de points (qu'on appellera arbitrairement individus) donc pas de représentations simultanées Un individu dont la pondération est nulle équivaut, comme c'est toujours le cas, à un individu illustratif.

2•5 REPRESENTATIONS SIMULTANEES

Dans le cas le plus général, quelle que soit l'analyse effectuée, le tableau contient, en plus des variables actives, des variables et des individus de tous les types qui doivent pouvoir être positionnés en éléments illustratifs sur des plans factoriels.

La richesse des plans factoriels est liée en particulier à la possibilité de représentation simultanée de divers types de points sur un même graphique Les règles de positionnement dépendent des analyses ; elles conditionnent les règles de lecture des informations sur le plan factoriel.

3•FONCTIONNALITES GRAPHIQUES

3•1 INTRODUCTION

On peut distinguer deux usages des graphiques factoriels :

- les graphiques à des fins **exploratoires** des données et
- les graphiques de **production** pour publication.

En fait, en y regardant bien, on se rend compte que les fonctionnalités nécessaires à ces deux types de graphiques sont sensiblement les mêmes: dans les deux cas on cherche à mettre en évidence l'information contenue dans les données. Ainsi l'interactivité nécessaire dans la phase exploratoire l'est aussi dans la mise au point du graphique. On peut donc raisonnablement envisager un seul produit dont certaines

fonctionnalités seront plus vitales dans la phase exploratoire et d'autres dans la phase de production.

3•1•1 Les fonctionnalités de base

Que ce soit en exploration ou en production, on doit en permanence visualiser le graphique en cours. Ceci permet d'apprécier immédiatement les modifications effectuées. On doit également pouvoir sauvegarder, à tout moment, le graphique et les éléments qui ont permis sa constitution de façon à autoriser la reprise ultérieure du travail soit pour le terminer soit pour le modifier.

L'utilisateur doit pouvoir définir des feuilles de style « plans factoriels » qui seront utilisées pour tous les plans que l'on désire représenter.

Dans la phase exploratoire une grande interactivité avec les données de base est nécessaire. Il doit être possible de faire apparaître toutes les informations concernant un point en cliquant sur le point.

Dans la phase de production, le graphique doit être construit en tenant compte du type de sortie que l'on veut obtenir :

- un graphique pour publication papier en noir et blanc,
- un graphique pour publication papier en couleur,
- une sortie sur transparent,
- une sur diapositive,
- une sur écran d'ordinateur.

Il est souhaitable que l'utilisateur puisse imprimer un graphique « par morceaux » (impression multi-pages).

Un graphique achevé pour un type de sortie (transparents couleurs par exemple) doit pouvoir être repris pour être adapté automatiquement à un autre type de sortie (publication papier noir et blanc par exemple) sans avoir à être retravaillé manuellement.

En fonction du type de support souhaité (sur papier noir et blanc, sur papier couleur, transparents, diapositives), le logiciel devra proposer un graphique initial en faisant des choix, pour les libellés et leurs attributs des points, adapté au type de sortie demandé. Sur ce graphique initial, la plupart des problèmes de recouvrement des points auront été résolus automatiquement. L'utilisateur doit pouvoir alors retravailler le graphique selon son goût.

3•1•2 La notion de groupes

Les différents points d'un plan peuvent être regroupés pour faire des traitements communs au niveau de leur représentation, de leur sélection ou dé-sélection. Les groupements à envisager sont :

- les variables continues **actives** ou **illustratives**
- les modalités **actives** ou **illustratives**
- les individus **actifs** ou **illustratifs**
- les individus en fonction de l'appartenance à une classe (suite à une classification)

- les individus en fonction des modalités d'une variable.

Il y a autant de groupes d'individus qu'il y a de classes ou de modalités. Il peut être intéressant « d'éclater » le plan factoriel en autant de fenêtres graphiques, sur un même écran que de modalités ou classes. Dans le cas d'une classification des individus, ou d'une représentation suivant les modalités d'une variable, un autre groupe apparaît : celui des centres de gravité des classes d'individus.

On pourra, de plus, avoir besoin de définir des groupes points, sélectionnés suivant des critères de qualité de représentation sur le plan, qui devront être représentés de façon identique.

3•1•3 Représentation des variables

Les points-variables d'un graphique factoriel doivent être affectés d'un symbole et d'un libellé. Le symbole doit être placé aux coordonnées exactes du point. Le libellé doit être placé le plus près possible du symbole correspondant. Le logiciel doit calculer la position de ce libellé en tenant compte de son environnement et en évitant les recouvrements de libellés. Le libellé doit être amovible autour du symbole mais ne doit pas être autorisé à trop s'en éloigner. La possibilité de mettre une flèche qui pointe du libellé vers le symbole permet aussi de clarifier le graphique si ce libellé est trop éloigné.

Les attributs des libellés des variables doivent pouvoir être traités point par point ou pour tout un groupe. Le texte du libellé et ses attributs doivent pouvoir être repris à tout moment.

3•1•4 Représentation des individus

On peut vouloir représenter le nuage des individus de différentes façons :

- l'individu est représenté par son **identifiant**. On distinguera alors les individus actifs des illustratifs par des polices de caractères et/ou des couleurs différentes,
- l'individu est représenté par un **symbole** différent, selon qu'il est actif ou illustratif,
- selon la **valeur** d'une variable continue,
- selon les **modalités d'une variable** : les individus seront séparés en autant de groupes qu'il y a de modalités pour la variable choisie et représentés par des symboles différents et/ou des couleurs différentes. On distinguera également dans ce cas les individus actifs des illustratifs.
- selon les **résultats d'une classification** : les individus seront séparés en autant de groupes qu'il y a de classes. On doit distinguer clairement les individus actifs des illustratifs (qui sont en général affectés à la classe la plus proche au sens de la distance utilisée).

Dans ces deux derniers cas les individus pourront être représentés par le symbole de leur groupe d'appartenance. Mais il faut aussi avoir la possibilité de faire apparaître et représenter l'identifiant des individus isolément, en cliquant sur le point, si nécessaire. Cette fonctionnalité est particulièrement intéressante dans la phase exploratoire mais peut avoir son intérêt également dans la phase de production si l'on désire mettre un individu en exergue.

Dans le cas d'un grand nombre d'individus on préférera la représentation des individus par les centres de gravité des classes ou des modalités. Ces « centres de gravité » seront traités comme des points illustratifs et dotés de symboles et de libellés. Il peut être intéressant dans certains cas de mettre les symboles des centres de gravité dans une taille proportionnelle à l'importance de la classe. Il est aussi intéressant de relier, par des segments, chaque centre de gravité aux points de sa classe ce qui permet d'avoir une idée de la dispersion de la classe.

3•2 FONCTIONS D'HABILLAGE DES ELEMENTS

Elles concernent tout ce qui est attaché aux points représentés dans le plan comme les libellés, les symboles, le tracé des segments entre divers points.

3•2•1 Les libellés

Il faut avoir le choix entre des libellés courts, longs, ou la concaténation des deux et ceci aussi bien pour les variables continues que pour les modalités.

Si plusieurs points sont superposés on doit pouvoir afficher lisiblement tous les libellés (ou certains seulement) près du symbole associé.

3•2•1•1 Attributs des libellés

Ils doivent être contrôlés par :

- le choix des **polices de caractères,** fait en proposant une liste de polices de caractères disponibles auxquelles il sera possible d'associer des attributs (gras, italique) et de visualiser le résultat avant confirmation.
- le choix des **couleurs,** fait en proposant une palette de couleurs. Huit couleurs franches sont à ce niveau amplement suffisantes. Par exemple le jaune sort bien sur un écran ou une diapo fond noir mais est sans intérêt pour une sortie papier blanc ou des transparents usuels.
- le choix de la **taille** des polices de caractères. Il devra être proposé en % de l'écran afin de conserver les proportions du graphique au moment de la sortie sur imprimante ou autre support.

3•2•1•2 Le déplacement des libellés

Chaque libellé sera mobile pour éviter la superposition de libellés.

3•2•1•3 L'effacement des libellés

Les libellés seront effaçables (et réaffichéables) individuellement, ou globalement à l'intérieur ou à l'extérieur d'une zone délimitée par l'utilisateur, ou pour tout un groupe.

3•2•2 Les symboles

Pour les symboles, dont on doit avoir le choix de la forme, la couleur et la taille, seront utilisés pour la représentation graphique des différents points. L'utilisateur doit pouvoir aussi affecter un symbole à tout groupe défini de points.

3•2•2•1 Attributs des symboles

- Le choix des **symboles** devra être fait en proposant une liste de symboles disponibles dont on peut contrôler la taille et la couleur. L'utilisateur avant de se décider doit pouvoir visualiser le choix fait avant de le confirmer. Aux symboles usuels : carré, cercle, triangle, étoile, losange doivent être ajoutés des symboles particuliers associés à certaines polices spéciales.
- Comme pour les attributs des libellés : le choix des **couleurs** des symboles devra être fait en proposant une palette de couleurs.
- la **taille** des symboles devra être proposé en % de l'écran afin de conserver les proportions du graphique au moment de la sortie sur imprimante ou autre support.

3•2•2•2 L'effacement des symboles

Les symboles seront effaçables individuellement, ou globalement à l'intérieur ou à l'extérieur d'une zone délimitée par l'utilisateur, ou pour tout un groupe. Les points sont alors représentés par leurs libellés. Cette fonction est de peu d'intérêt en phase exploratoire mais peut être souhaitée en phase de production.

3•2•3 Les flèches

Des flèches pourront être utilisées :

- pour marquer les points qui sont ramenés sur le bord du cadre (leur longueur pourrait être proportionnelle à la distance du point au cadre)
- pour pointer la position exacte d'un point quand le libellé est trop éloigné (elle pourrait apparaître automatiquement dés que la distance atteint une certaine valeur)
- Il est aussi utile de pouvoir mettre des flèches pour représenter des anciens axes unitaires en ACP par exemple.

On pourra contrôler les valeurs et attributs des flèches.

3•2•4 Le tracé de lignes

Deux types de lignes sont nécessaires :

- **les segments de droites** (pour indiquer des trajectoires entre éléments de même nature, pour joindre des modalités d'une même variable, pour joindre le centre de gravité du graphique aux variables en ACP, pour joindre les centres de gravité des classes aux individus appartenant à cette même classe...).
- Les **courbes** (pour entourer un groupe de points).

3•2•4•1 Attributs des lignes

- Le **type de ligne** sera fait en proposant une liste de lignes disponibles : ligne pleine, pointillée, tirets longs...
- L'**épaisseur** de la ligne devra être proposée en % de l'écran afin de conserver les proportions du graphique au moment de la sortie sur imprimante ou autre support.

- La **couleur** de la ligne sera accessible via une palette de couleurs sur laquelle il suffira de cliquer

3•3 MANIPULATION DES ELEMENTS

Les éléments du plan factoriel seront traités individuellement ou par groupe. Les principales fonctions souhaitées sont énumérées ci-dessous :

3•3•1 La sélection individuelle

Cette fonction est nécessaire pour permettre la sélection ou (désélection) des divers points du plan. En particulier pour :

- identifier individuellement chaque point du plan, par un simple clic de la souris (sur le symbole ou sur le libellé quand ils existent).
- afficher des informations concernant un élément du plan, par exemple ses coordonnées sur les axes factoriels, sa contribution à l'inertie des axes, la qualité de sa représentation sur les axes.

3•3•2 La Sélection d'un groupe

La sélection (ou dé-sélection) d'un groupe de points du plan se fera de plusieurs façons soit selon la nature du groupe, soit à partir d'un critère. On peut vouloir par exemple sélectionner:

- un groupe de points situé à l'intérieur (ou à l'extérieur) d'une zone délimitée par l'utilisateur, pour une identification globale par exemple.
- un groupe d'éléments actifs ou un groupe d'éléments illustratifs (variables ou individus).
- une classe d'individus associés à une modalité d'une variable nominale.
- une classe de points obtenue par une méthode de classification
- le groupe des centres de gravités des classes.
- un groupe de points selon un critère de plus forte contribution au plan.
- un groupe de points selon un critère de bonne représentation sur ce plan

3•3•3 Le zoom

La fonction zoom sera disponible avec indication visible des bornes. La zone à agrandir doit être délimitée par l'utilisateur et réaffichée en conservant toutes les informations qu'elle contenait antérieurement : identifiants d'éléments, segments, par exemple. Le compteur d'éléments figurant dans la légende doit être réaffiché en conséquence.

Le zoom doit pouvoir être utilisé « en cascade ». Le niveau de zoom sera alors indiqué à l'écran.

Les limites du zoom seront mémorisées à chaque étape pour permettre le retour arrière par les mêmes étapes. On doit pouvoir revenir à l'état initial du graphique à partir de n'importe quel niveau de zoom. Enfin, toute manipulation qui cache des points donnera lieu à l'ouverture d'une fenêtre situant la position de la zone agrandie à l'intérieur du graphique.

3•3•4 Zoom particulier

Un zoom particulier permet de se concentrer sur la partie la plus dense du graphique tout en conservant l'ensemble des points. (Les points extérieurs à la zone dense sont ramenés à la périphérie du graphique dans leur direction par rapport à l'origine).

Le zoom ordinaire doit être disponible à l'intérieur de ce zoom

3•3•5 Étirement (stretch)

Le rectangle contenant le graphique sera étirable dans le sens de la longueur ou de la largeur dans la limite de la place disponible sur l'écran.

3•3•6 Graduations

Le choix du type de graduation sur le graphique sera fait selon que l'on souhaite avoir des échelles identiques ou utiliser au mieux la surface de l'écran. Pour ne pas charger le graphique, il est préférable que les graduations figurent sur le cadre.

En analyse en composantes principales, on pourra faire apparaître sur les directions des anciennes variables des graduations correspondant aux unités des variables ou aux écarts-types.

3•4 FONCTIONS D'HABILLAGE DU PLAN

Les fonctions d'habillage concernent la présentation du graphique et plus particulièrement:

3•4•1 Les axes

Les axes figureront ou non sur le graphique, la structuration par quadrant n'ayant pas toujours une signification.

On doit pouvoir également faire apparaître ou non, à proximité des axes :

- leur numéro
- les valeurs propres associées
- les taux d'inertie de ces axes.

3•4•2 Le cadre

Le cadre entourant le graphique sera affichéable ou non et on aura le contrôle de l'épaisseur, et de la couleur de ce cadre. Les graduations pourront figurer ou non sur ce cadre.

3•4•3 Le quadrillage

Il peut être intéressant de faire figurer ou non un quadrillage sur le graphique. Il sera utile pour faciliter le repérage.

3•4•4 L'ajout de texte ou de titre

Il est intéressant :

- **ajouter** un titre ou un texte, introduit au clavier par l'intermédiaire d'un mini-éditeur, à un endroit quelconque du graphique indiqué par l'utilisateur, à l'intérieur ou à l'extérieur du cadre. Ce titre ou texte doit pouvoir être affiché horizontalement ou verticalement avec contrôle de la police de caractères, la couleur et la taille.
- **modifier** un texte déjà existant sur le graphique et contrôler la taille, la couleur, la police de caractères.
- **déplacer** ou **effacer** un texte.

3•4•5 La légende

Dés que l'on donne des significations sémantiques à des éléments graphiques, il faut les expliquer dans des cartouches de légendes déplaçables et éditables.

Par exemple :

- les individus sont représentés par des symboles, les variables par des libellés
- le type de symbole dépend de la modalité de la variable x
- la couleur du symbole dépend de la modalité de la variable y
- les variables illustratives sont en italique

La légende pourra contenir des statistiques comme par exemple les effectifs correspondants.

3•4•6 La définition d'une zone

Entourer une zone de points du graphique à l'aide de la souris avec possibilité d'effacement du tracé et contrôle de la couleur, de l'épaisseur et du type de trait.

3•4•7 Le coloriage de zone

Colorier dans la couleur de son choix ou choisir une trame, pour une zone fermée délimitée à l'aide de la souris.

3•5 STYLES, FEUILLES DE STYLES ET MODELES

3•5•1 Styles, feuilles de styles

À chaque élément du graphique est associé un ensemble d'attributs qui peuvent être purement graphiques comme il a été exposé plus haut, ou qui présentent un caractère statistique (cf. chapitre 4).

Les attributs graphiques peuvent concerner :

- l'ensemble du graphique : par exemple, le fait que le rapport d'aspect du graphique est égal à 1 ou que l'unité de mesure soit commune aux deux axes, tracé ou non du cercle unité, présence ou non d'une légende.
- les axes : existence ou non de marques de graduation (principales ou secondaires), style de ces marques, police, format d'affichage des nombres,

rappel ou non de l'inertie de l'axe (ou du pourcentage de l'inertie totale), position des axes (passant par l'origine ou renvoyés sur le cadre du graphique).
- les points : nature du symbole, couleur, présence d'une étiquette et attributs de texte correspondant, existence d'une ligne reliant les points et style de cette ligne.

Les attributs à caractère statistique semblent surtout être des caractéristiques globales au graphique. Ils peuvent être, par exemple :
- l'association d'une troisième (voire d'une quatrième...) variable aux points représentatifs : ainsi une variable de classification (3-ème variable), qui gouvernerait le style de marques des points et une variable de pondération (4-ème variable) qui déterminerait la taille du symbole (p. ex. : poids, cos^2 ou contribution à l'inertie...)
- un critère statistique déterminant l'affichage, l'affichage estompé ou le masquage des points représentés (on représenterait ainsi les points dont le cos^2 est supérieur à un certain seuil, ou les n points contribuant le plus à l'inertie, du plan,...).

La notion de style permet de regrouper l'ensemble des ces attributs : à chaque ensemble d'attributs il est possible d'attribuer un nom. Le nom stocke la définition de chacun des attributs de l'élément du graphique auquel s'applique le style. Il est possible de stocker le nom du style et sa définition, soit dans le fichier informatique contenant le graphique, soit dans une feuille de styles autonome, soit dans un fichier d'options spécifique du logiciel. Ce mécanisme de stockage doit permettre de récupérer les définitions d'attributs et de les appliquer à des éléments graphiques d'un autre graphique de la même analyse ou plus généralement à une autre analyse.

Nous présentons, en annexe, ce que pourrait être un dialogue de définition et d'application du style pour un point. Les dialogues sont très largement inspirés de Microsoft Excel version 5.

3•5•2 Modèles de graphiques

L'ensemble des choix adoptés pour la réalisation d'un graphique doit pouvoir être stocké de manière indépendante des données sur lesquelles il a été mis au point. Là aussi, on peut envisager de stocker ces informations dans des fichiers de modèles indépendants les uns des autres ou de les regrouper dans une bibliothèque.

L'utilisateur peut à son gré créer de nouveaux modèles, modifier des modèles existants, les supprimer et bien sûr, les appliquer à ses analyses, tout en gardant une faculté de modification particulière des différents attributs.

On peut envisager de fournir un premier jeu de modèles adaptés :
- aux différentes analyses factorielles (ACP, AFC, ACM)
- dans différents cas de figure
  - petit nombre de points, très grand nombre de points
  - modalités qualitatives ordonnées,...
- pour différents supports
  - papier, transparent, écran vidéo, diapositive,...
  - noir et blanc ou couleur, dans ses nombreuses variantes.

3•5•3 Exemples de réalisation

Dans des styles très différents, nous décrivons quelques réalisations dans différents logiciels statistiques ou non... Nous présenterons rapidement quelques fonctionnalités de SAS, JMP et Excel.

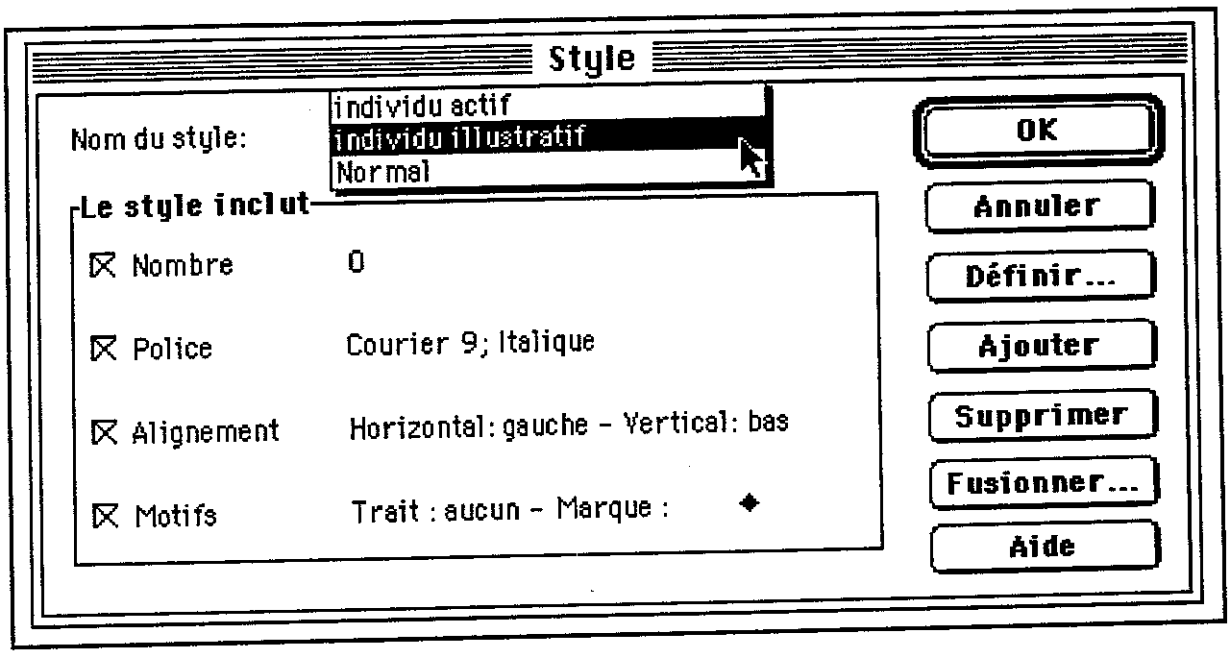
Dans le module SAS/GRAPH, de SAS Institute, il est possible de gérer une liste de définition d'axes, de motifs ou de symboles (instructions AXIS, PATTERN, SYMBOL,...). Ils peuvent être définis par programme ou de manière interactive dans des fenêtres spécialisées. Ces définitions sont ensuite utilisées dans les instructions qui génèrent les graphiques. Enfin, de très nombreuses options graphiques globales permettent de choisir les paramètres graphiques adaptés aux différents périphériques de sortie utilisés. Le premier travail de l'utilisateur de SAS/GRAPH consiste à déterminer un (ou plusieurs) choix de ces paramètres adapté à ses besoins. C'est un travail important, qui revient à gérer une description textuelle (programme SAS) d'un ensemble de styles et de modèles de graphiques.

De la même société, le logiciel JMP sur Apple Macintosh (et d'une manière similaire, le module SAS/Insight) propose une approche très différente (voir annexe descriptive de JMP). Les attributs graphiques et statistiques (point inclus ou exclus de l'analyse) des points sont des éléments associés à chaque ligne de la table de données et il est possible de les stocker, avec la table, comme autant de variables nouvelles. Le point fort de ce logiciel, sous l'angle qui nous intéresse ici, est la facilité avec laquelle il est possible d'éditer l'habillage d'un graphique et surtout la très grande facilité de stocker cet habillage et de le rappeler ultérieurement. En revanche, cette facilité ne s'applique qu'à l'intérieur d'un même tableau de données. Il n'est pas possible de définir des styles au sens où nous les avons définis, pas plus que de créer des documents types.

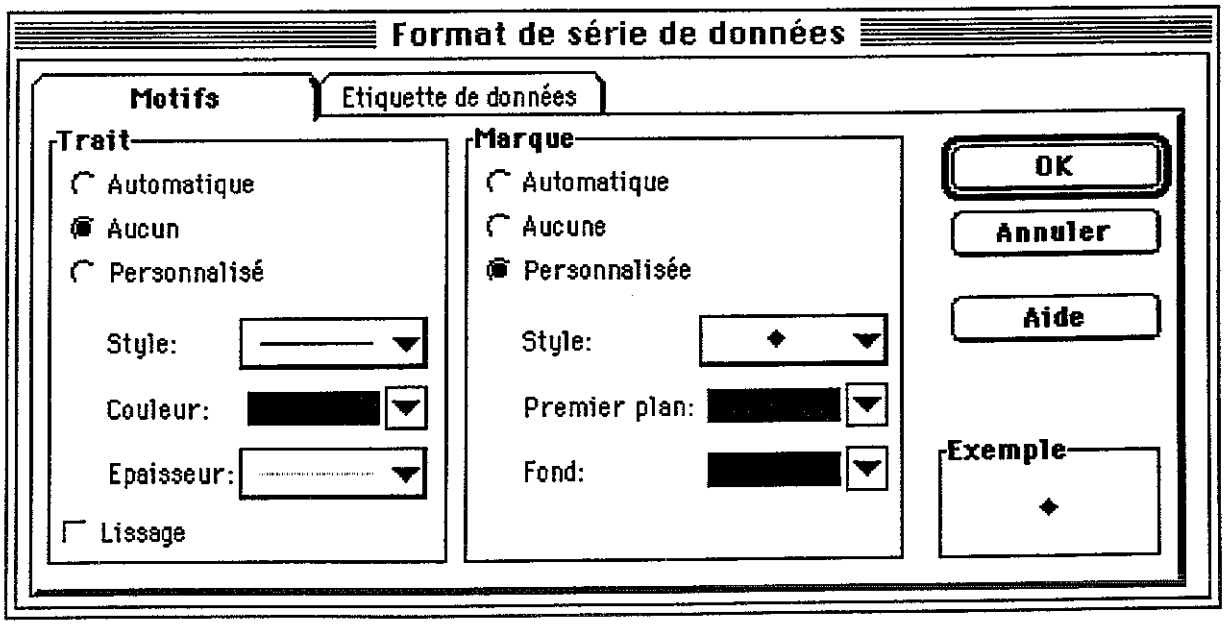
Dans le tableur Excel de Microsoft, il est possible de définir des styles qui sont attachés à chaque document. Il est possible de récupérer en bloc l'ensemble des styles défini dans un document et de les transporter dans un autre document (l'opération s'appelle Fusionner). Il est également possible de créer des documents modèles. Enfin, on peut définir des graphiques types, repérés par un nom et par une description sommaire. Ils sont gérés dans un fichier d'options (ou de préférences) unique du logiciel et il est extrêmement facile d'appliquer le graphique type ultérieurement à toute série de données.

3•5•4 Annexe

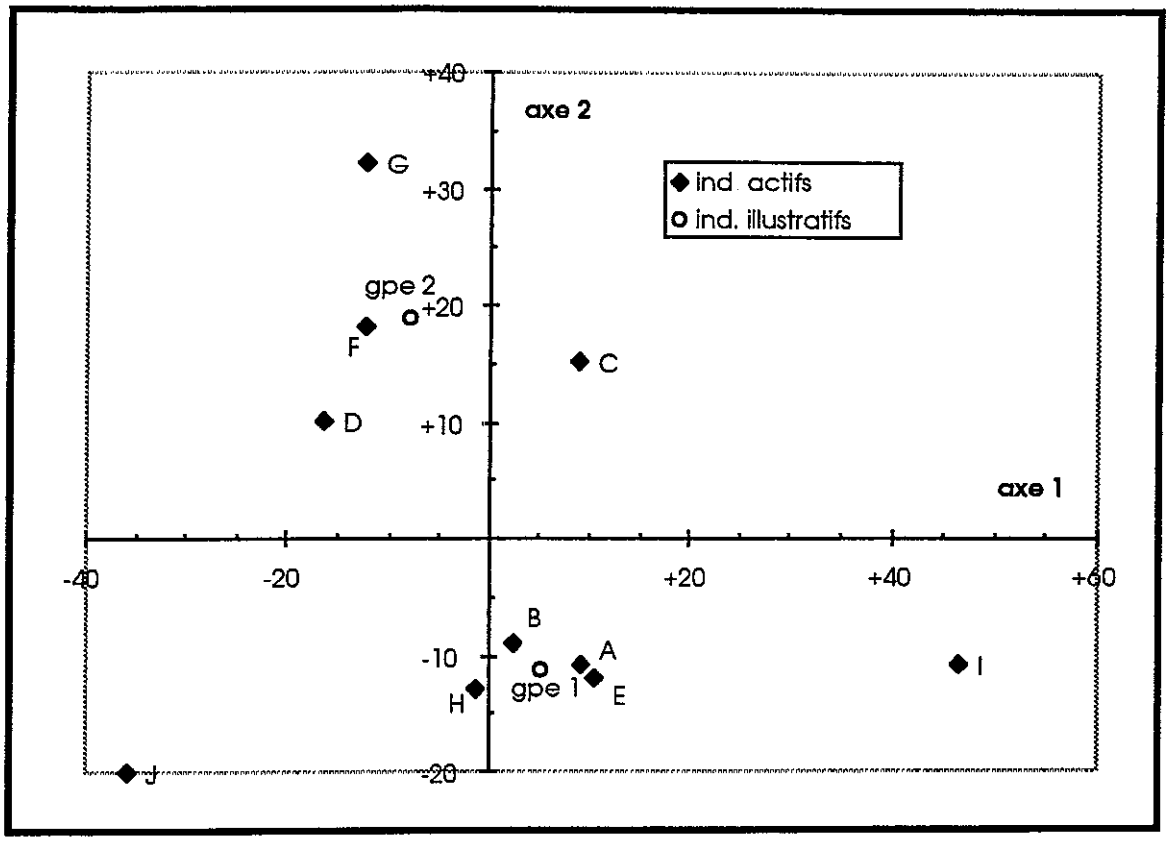
Exemple de dialogue de définition du style d'un point d'après Microsoft Excel 5



La définition du motif (trait et marque) débouche sur un dialogue de ce type :



Exemple de plan factoriel



Utilisation d'un document Excel avec deux styles de points : individus actifs et individus illustratifs

4•FONCTIONNALITES STATISTIQUES

4•1 DEUX POINTS DE VUE SELON L'UTILISATION

4•1•1 Exploration à l'écran

L'exploration des données à l'écran va permettre d'analyser les données de façon interactive en utilisant des fonctionnalités statistiques spécifiques

4•1•1•1 Exploration d'une classification

Exemple 1 : On veut analyser sur le plan factoriel le nuage des individus après une classification. Les différentes classes sont repérées par un symbole ou une couleur différente.

Exemple 2 : Sur ce même plan factoriel, on pourra vouloir représenter le centre de gravité des classes À partir d'un des centres de gravité choisi par l'utilisateur, on pourra, de façon interactive à l'aide d'un bouton curseur, relier successivement par des segments le centre de gravité aux points les plus proches. À chaque étape la fonction identification des points doit être disponible. Il est intéressant également de

repérer les points les plus éloignés d'un centre de gravité d'une classe, donc on doit pouvoir faire l'opération inverse, c'est-à-dire relier au centre de gravité (ou faire apparaître) d'abord les points les plus éloignés puis les points de plus en plus proches.

4•1•1•2 Sélection interactive de points sur un plan

Plusieurs critères statistiques peuvent être utilisés:

Par exemple, pour le critère du cos^2, on part d'un plan factoriel vide et on fait apparaître successivement à l'aide d'un bouton curseur le point le plus significatif selon le critère du cos^2, puis progressivement les points suivants. On doit pouvoir identifier à tout moment les points.

De même, on peut vouloir faire cette opération de façon descendante, c'est à dire avoir tous les points représentés sur le plan factoriel et faire disparaître successivement les moins significatifs.

Il serait intéressant d'avoir à chaque étape la valeur du critère et le nombre de points représentés sur le plan.

On pourrait procéder de même avec le critère de la contribution.

4•1•1•3 Sélection interactive de points illustratifs

On pourrait éliminer les points illustratifs dont la valeur-test (au sens défini dans SPAD) est inférieure à 2 écarts-types, 3 écarts-types ou 4 écarts-types de façon à faire apparaître les points illustratifs les plus significatifs ou à l'aide d'un curseur faire apparaître successivement les points illustratifs les plus significatifs en affichant la valeur-test.

4•1•1•4 Interactivité des calculs

On pourrait offrir la possibilité d'éliminer certains points pour effectuer le recalcul de l'analyse et le réaffichage du plan factoriel.

4•1•2 Production

Produire un graphique c'est représenter une image graphique de la phase d'exploration à un instant donné. On doit pouvoir fixer l'image avec ses caractéristiques : cos^2, nombre de points représentés, fonctionnalités statistiques associées, etc.

4•2 FONCTIONNALITES STATISTIQUES SELON LE TYPE DE POINTS

4•2•1 Généralités

On verra qu'il y a peu de fonctionnalités liées à la nature de l'analyse (ACP, ACM ou AFC) sinon les règles de représentations simultanées. Par contre de nombreuses fonctionnalités sont liées à la nature des points (individus, variables continues, modalités ou classes, fréquences).

Au paragraphe 4•3 on différenciera points actifs et illustratifs car les règles d'interprétation des proximités ne sont pas les mêmes (même si les calculs des positionnements sont identiques).

Chaque fois que l'on calcule une distance ou une inertie (ou une quantité dérivée comme une contribution, un cos^2, un plus proche voisin, etc.) pour habiller un graphique, il faut spécifier — ou faire spécifier — le support servant au calcul : un axe, le premier plan, l'espace jusqu'à l'axe de plus haut rang du plan représenté ou l'espace complet. Il y aura souvent une valeur par défaut naturelle (par exemple on calcule les k plus proches voisins dans l'espace complet — ou dans l'espace factoriel maximum — pour les identifier sur un plan factoriel quelconque).

On distingue le point (ou le symbole) représentant la localisation du point sur le plan et l'étiquette qui peut lui être adjointe. Pour l'un et pour l'autre, il faut résoudre les problèmes de points superposés ou empiétant et assurer une lisibilité optimale (en particulier pour les étiquettes).

S'il y a sélection de points selon un critère continu (ex: la contribution), on actionnera des « jauges » ou des curseurs. En général les points s'estomperont mais on pourra garder un « fantôme » de la position des points non sélectionnés (un pixel résiduel). Il en sera de même pour une sélection logique (une classe, un filtre logique plus général). En effet il y a souvent intérêt à garder une image de l'ensemble du nuage des points au sein duquel on veut mettre en valeur un sous-ensemble de points.

4•2•2 Pointeur d'information

En pointant sur un point on obtient toute l'information factuelle et statistique sur l'élément choisi.

Pour un individu, on obtient (à la demande) ses valeurs dans le tableau de données, son poids, ses coordonnées, contributions et cos^2 sur les axes factoriels, sa distance au centre, sa participation à l'inertie globale, ses classes d'appartenance dans les partitions disponibles, éventuellement les points les plus proches, etc.

Pour une variable continue (ou même un axe factoriel qui est une variable continue particulière), on obtient sa moyenne, son écart-type, son minimum, son maximum, le nombre de données manquantes, son histogramme, ainsi que ses coordonnées, contributions et cos^2 sur les axes factoriels.

Pour une variable « effectifs », on obtient sa moyenne, son écart-type, son minimum, son maximum, son histogramme, ainsi que ses coordonnées, contributions et cos^2 sur les axes factoriels.

Pour une variable nominale (ou une partition, qui est une variable nominale particulière), on obtient l'effectif (et le poids) de chaque modalité, les coordonnées, contributions et cos^2 sur les axes factoriels, ainsi que les valeurs-test sur les deux axes.

4•2•3 Cadre et échelles

Il y a deux types de représentation : la forme classique des plans (x,y) avec les axes en bordure et la forme « plan factoriel » avec les axes se croisant au centre de gravité ou

au point de coordonnées nulles. Cette seconde présentation est fournie par défaut. On peut passer de l'une à l'autre.

Noter le problème particulier du nuage des variables continues en Composantes Principales: il est souvent indispensable d'imposer l'origine (origine des vecteurs) dans le graphique — en particulier pour les ACP non normées. De plus il est important de « centrer le graphique » dans un cercle de corrélation dans le cas d'une analyse normée, même si le nuage des points est « non symétrique ».

Le choix des échelles et des unités sur les axes factoriels est un choix important. Un cercle de rayon unité pourra être déformé volontairement pour avoir la forme d'une ellipse pour rendre lisible un plan factoriel. Ce problème est traité ultérieurement.

Le sens des axes d'une AF étant arbitraire, on pourra inverser le graphique pour des raisons esthétiques, culturel (mettre les partis de gauche à gauche) ou de comparaison de plusieurs analyses.

4•2•4 Mise en valeur de groupes de points

Une opération statistique importante sur le plan factoriel est la sélection de points opérée pour améliorer ou guider la lecture et l'interprétation du graphique : reconnaître la spécificité des zones du plan, les caractéristiques des directions factorielles, estomper ou effacer les points qui encombrent et perturbent la lecture, etc.

Les principaux critères classiques de sélection sont la contribution à l'inertie (sur un axe, sur le plan, jusque sur le plan), le cos^2 (sur un axe, sur le plan, jusque sur le plan), la (vraie) distance à l'origine, le poids. Pour les points représentatifs d'un groupe d'individus (modalité ou classe), on peut ajouter la sélection en fonction de la valeur-test.

En général la sélection ne s'opère pas en oui/non ; on procède plutôt à un habillage qui met en valeur les points en fonction de la valeur du critère retenu. L'habillage peut porter sur l'écriture du libellé ou — de préférence — sur le symbole qui donne la localisation du point. Si les points individus sont représentés par des petits cercles, on jouera sur le diamètre des cercles. En agissant sur un curseur, on pourra par exemple estomper de plus en plus les points les moins intéressants au sens du critère choisi.

4•2•5 Connexions entre points

Relier deux points par une ligne introduit souvent un élément d'interprétation ou une aide de lecture du graphique. Cette opération est décrite dans plusieurs exemples de ce rapport.

On pourra par exemple relier chaque point d'une classe au centre de la classe (graphique en étoile) ce qui permet de juger de la dispersion de la classe dans le plan.

Pour des points-catégories d'une variable ordinale (les catégories de revenu par exemple) on pourra relier les points dans l'ordre naturel pour faciliter l'interprétation. S'il n'y a pas d'ordre naturel entre les catégories, on pourra tracer le chemin de longueur minimal (meilleure image des proximités réelles) ou au contraire le chemin de longueur maximal (dispersion des catégories dans le plan).

Pour avoir une représentation de la densité du nuage autour d'un point, on peut relier celui-ci à ses k plus proches voisins par un graphique en étoile. Une autre façon de procéder est de fixer le rayon d'une sphère et de marquer (ou mettre en surbrillance) les points intérieurs à la sphère centrée sur le point.

4•2•6 Cercles, ellipses, enveloppes

Autour d'un point représentant un groupe d'individus, on peut tracer des « ellipses de confiance » (dont les calculs sont parfois complexes). On peut aussi tracer l'enveloppe convexe des points projetés dans le plan.

Dans une analyse des correspondances simples, on peut tracer autour de chaque point un cercle dont le rayon mesure (d'une certaine façon) de combien le profil correspondant diffère du profil moyen, centre du graphique.

4•2•7 Cas de l'analyse des correspondances simples

La représentation simultanée des lignes et des colonnes de cette analyse est particulière : chaque point d'un nuage étant « quasi-barycentre » de l'ensemble des points de l'autre nuage. Cette représentation simultanée sera fournie par défaut.

Cependant on devra pouvoir obtenir à la demande le positionnement des lignes en vrais barycentres des colonnes et vice-versa.

4•3 FONCTIONNALITES LIEES AU TYPE « ACTIF/ILLUSTRATIF »

4•3•1 Généralités

La distinction actif/illustratif est essentielle et doit se traduire par des fonctionnalités graphiques particulières. Il faut rappeler en effet que, entre deux points actifs, il existe toujours une distance interprétable. Par contre entre deux points illustratifs — ou entre un point illustratif et un point actif — il n'y a pas de distance directement interprétable.

Cependant il est intéressant en général de représenter les deux types de points sur le même graphique ou sur des graphiques proches. Il est donc essentiel de bien distinguer les points actifs des points illustratifs sur un même graphique ou sur des graphiques proches. La distinction peut se faire par une typographie différente pour les libellés ou par des symboles différents (ou par la couleur). On pourra utiliser par exemple italique, gras, souligné, et/ou couleur pour habiller différemment les étiquettes des points.

4•3•2 Superposition des plans

Il est intéressant de disposer séparément du plan « actif » et des différents plans de points illustratifs. Il est intéressant aussi de disposer des représentations simultanées.

On peut imaginer que ces différents graphiques existent sur des supports transparents et sont superposables (donc dessinés sur les mêmes échelles). Ces opérations de superposition (avec adaptation automatique des échelles et repositionnement des étiquettes) devraient être réalisables facilement.

4•3•3 Cas de l'analyse des correspondances multiples

Dans une analyse des correspondances multiples, on peut considérer que chaque individu représente une certaine combinaison des modalités des variables actives. Lorsqu'il y a beaucoup de points individus et relativement peu de configurations de modalités actives réellement observées, il y aura donc beaucoup de points-individus confondus, chaque point multiple représentant une certaine combinaison de modalités observée plusieurs fois.

Il est intéressant alors de faire apparaître graphiquement l'abondance des points multiples — par exemple en faisant varier le rayon du cercle symbolisant la position de chaque point-multiple.

4•3•4 Cas de l'analyse en composantes principales

Considérons la représentation simultanée des individus et des variables en ACP. On sait que dans ce cas les variables sont en fait des individus artificiels représentant les extrémités des anciens axes unitaires (porteurs des variables actives). En projection sur un plan factoriel, on obtient des points à l'intérieur d'un cercle unitaire. En réalité on représente des directions, projections des vecteurs unitaires sur le plan. Mais il est clair qu'il n'y a aucune commune mesure entre les coordonnées des individus (réels) et les coordonnées de ces individus artificiels que sont les vecteurs unitaires.

Par conséquent, pour rendre lisible la figure, on sera amené en général à procéder à une dilatation de l'un des deux nuages, soit le nuage des vecteurs unitaires, soit le nuage des individus. Cette dilatation doit être facile à opérer.

De plus on prendra garde au fait que dans cette représentation simultanée très particulière des individus et des variables, il ne doit pas être possible de faire apparaître des variables illustratives (ce ne sont pas d'anciens axes unitaires porteurs du nuage des individus).

Sur toute direction représentative d'une variable continue dans l'espace des individus (direction de la variable active correspondance), on peut faire figurer des graduations dans l'unité de mesure de la variable. La moyenne de la variable se trouve au centre du graphique.

Dans le cas de l'ACP, le nuage des variables est une représentation graphique de la matrice des corrélations (ou des covariances si l'analyse n'est pas normée). Dans ce nuage la distance entre deux points s'exprime en fonction de la corrélation (ou des variances et covariance). On prendra donc soin de rendre non superposables tout nuage des individus avec tout nuage des variables.

4•3•5 Cas de l'analyse des correspondances simples

Un point étant un profil (un histogramme de répartition), on peut faire apparaître cet histogramme à la demande.

On doit pouvoir faire figurer sur tout plan factoriel d'une AFC des variables continues illustratives et des variables nominales illustratives.

4•4 REPRESENTATION SIMULTANEE PAR BIPLOT

4•4•1 Généralités

Le BIPLOT est une méthode de représentation graphique **simultanée** des lignes et des colonnes d'un tableau de données. Elle est souvent utilisée pour illustrer à l'aide d'un seul graphique les résultats d'une méthode factorielle : ACP, AFC, ...

La méthode du BIPLOT a été développée par le professeur K.R. Gabriel [voir réf.]. Elle consiste à faire une approximation de la matrice de données par un produit matriciel dont la dimension commune est 2 (afin de faire une représentation dans le plan).

La technique de décomposition est identique à celle de l'ACP: décomposition en valeurs singulières (valeurs propres et vecteurs propres).

L'ACP normée, hors représentation simultanée déjà discutée, fournit deux graphiques:

- le cercle de corrélations où ne sont représentées que les variables (espace des variables)
  (le cosinus de l'angle entre deux variables est égal à la corrélation entre ces variables)
- le tracé des individus sur le plan factoriel (espace des individus) (la distance entre les individus approxime la distance euclidienne du tableau de données initial)

le BIPLOT superpose individus et variables sur un même graphique :

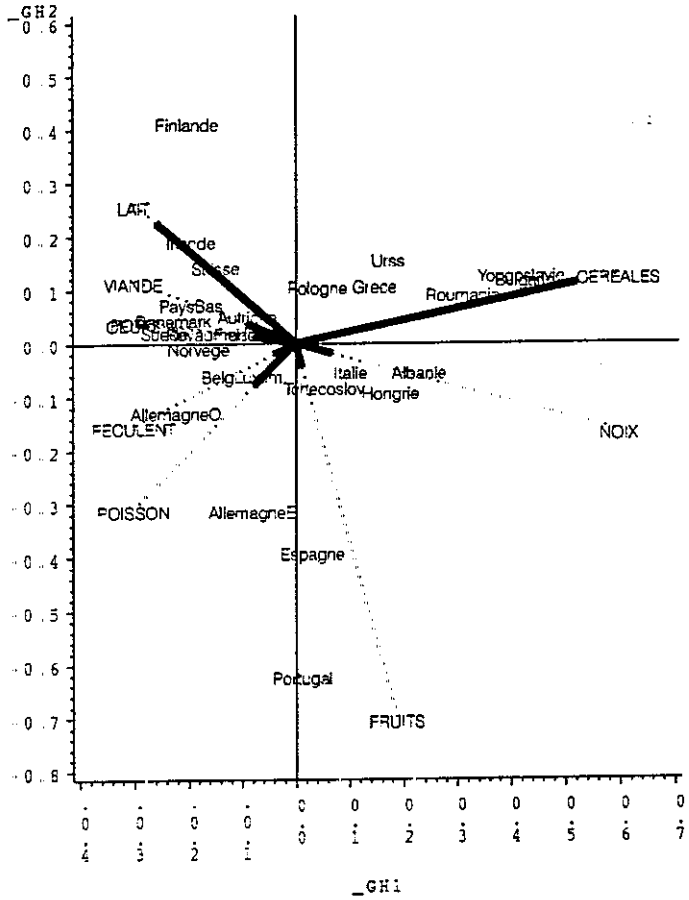
soit dans l'espace des variables
(le cosinus de l'angle entre deux variables est égal à la corrélation entre ces variables, la distance entre les individus approxime la distance de Mahalanobis, et non la distance euclidienne, du tableau de données initial)

- soit dans l'espace des individus
  (pas de propriétés sur les variables, la distance entre les individus approxime la distance euclidienne du tableau de données initial)
- soit dans un espace intermédiaire
  (aucune propriété sur les variables ni sur les individus, mais obtention d'un graphique "équilibré")

Sur les trois graphiques, la projection des individus sur une variable respecte la répartition des données initiales pour les variables.

Comme pour l'ACP, le biplot est généralement utilisé en centrant et réduisant les variables (calculs sur la matrice de corrélation), mais il est également possible de travailler sur des variables non réduites si les variables sont comparables entre elles (calculs sur la matrice de variance-covariance).

Exemple de Biplot (variables non réduites) dans l'espace des variables :



4•4•2 Description de la Méthode

Le principe initial est basé sur la reconstitution de la matrice initiale à l'aide du graphique, c'est à dire que la projection des individus sur les variables respecte la répartition des données initiales pour cette variable.

Toute matrice Y peut être décomposée en

$Y = AB'$

$(n,p) = (n,k) \times (k,p)$ avec k = rang de Y

Le Biplot consiste à approximer Y par :

$Y \approx AB'$

$(n,p) \approx (n,2) \times (2,p)$

A est la matrice des individus

B est la matrice des variables

Sur un graphique Biplot, on représentera

- les individus par des points

- les variables par des vecteurs

Pour trouver une décomposition, on utilise la décomposition en valeurs singulières :

$Y = ULV'$

avec

V: vecteurs propres de Y'Y

$$\Lambda = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & 0 \\ 0 & \sqrt{\lambda_2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_r} \end{bmatrix}$$

$U = YVL^{-1}$

r = rang de Y

En ne gardant que les 2 premières valeurs propres, on approxime Y par :

$Y \approx \hat{Y} = U_{[n\,2]}\Lambda_{[2.2]}V'_{[2.p]}$

On peut alors définir différentes approximations de Y en AB'

	A (individus)	B' (variables)
GH' (espace des variables)	U	LV'
Symétrique	$UL^{½}$	$L^{½}V'$
JK' (espace des individus)	UL	V'

GH' : la décomposition est effectuée dans l'espace des variables

JK' : la décomposition est effectuée dans l'espace des individus

En répartissant Λ sur A et B' de manière symétrique, la décomposition est effectuée dans un espace intermédiaire

$Y \gg A\,B'$

$y_{ij} \gg a_i'\,b_j$

Ce produit scalaire montre que la répartition des projections des points a_i sur les droites b_j approxime la répartition des données initiales pour la variable y_j, quel que soit la décomposition effectuée.

On peut calculer la qualité d'approximation par le biplot de :

la matrice centrée Y	$\frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{k=r} \lambda_k}$
la matrice de covariance S	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{k=1}^{k=r} \lambda_k^2}$
la matrice des distances de Mahalanobis pour les paires d'individus	$\frac{\lambda_1^0 + \lambda_2^0}{\sum_{k=1}^{k=r} \lambda_k^0} = \frac{2}{r}$

4•4•2•1 Les propriétés de la méthode GH' (espace des variables)

Y » A B' = (U) (LV')

Le carré de la norme du vecteur B_j représente n fois la variance de la colonne y_j

Le cosinus de l'angle entre les vecteurs b_j et $b_{j'}$ correspond à la corrélation entre les variables y_j et $y_{j'}$ (comme pour l'ACP, ceci n'est vrai que pour les variables actives, pour les variables illustratives, cette propriété n'est vraie que vis à vis des axes).

La distance euclidienne entre les points a_i et $a_{i'}$ est proportionnelle à la distance de Mahalanobis entre les individus y_i et $y_{i'}$ du tableau de départ.

La distance de Mahalanobis est une distance multidimensionnelle, qui tient compte des corrélations entre les variables ainsi que de leurs variances. Elle tend à transformer un nuage de points de forme allongée en un nuage de forme ronde.

$d^2_{Mahalanobis}(ii')=(y_i-y_{i'})' S^{-1}(y_i-y_{i'})$ où S est la matrice de variance-covariance

4•4•2•2 Les propriétés de la méthode JK' (espace des individus)

Y » A B' = (UL)(V')

La distance euclidienne entre les points a_i et $a_{i'}$ approxime la distance euclidienne entre les individus y_i et $y_{i'}$ du tableau de départ.

Il n'y a pas dans ce cas de propriétés spécifiques aux variables.

4•4•2•3 Les propriétés de la méthode Symétrique (espace intermédiaire)

Y » A B' = $(UL^{½})(L^{½}V')$

Cette méthode égalise l'effet des lignes et des colonnes :

Pour chaque axe, $\sum$ écarts^2 à l'axe sont égaux pour les individus et les variables.

On obtient alors un « beau » graphique (équilibré).

Hormis la propriété générale à toutes les décompositions (les projections des individus sur les variables approximent les répartitions initiales), il n'y a pas de propriétés spécifiques à l'interprétation des individus ni des variables

4•4•2•4 Lien avec l'ACP

Le biplot est basé sur les mêmes principes de décomposition que l'ACP. On retrouve donc les mêmes résultats à un coefficient multiplicatif près.

Le biplot fait une recherche des vecteurs et valeurs propres de Y'Y, tandis que l'ACP fait la recherche sur Y'DY, avec D matrice des poids (les termes de la diagonale valent 1/n)

	A (individus)	B' (variables)
GH'	U	LV'
Symétrique	$UL^{1/2}$	$L^{1/2}V'$
JK'	UL	V'
ACP	U^*L^*	$L^*V^{*'}$

avec : ULV' décomposition à partir de Y'Y,

$U^*L^*V^{*'}$ décomposition à partir de Y'DY

On obtient donc les formules de passage :

1. pour les individus :

$$GH_k(i) = \frac{Fact_k^*(i)}{\sqrt{n\lambda_k^*}}$$

$$JK_k(i) = Fact_k^*(i)$$

$$SY_k(i) = \frac{Fact_k^*(i)}{\sqrt{\sqrt{n\lambda_k^*}}}$$

2. pour les variables

$$GH_k(j) = \sqrt{n}Fact_k^*(j)$$

$$JK_k(j) = \frac{Fact_k^*(j)}{\sqrt{\lambda_k^*}}$$

$$SY_k(j) = \frac{\sqrt{\sqrt{n}}Fact_k^*(j)}{\sqrt{\sqrt{\lambda_k^*}}}$$

La principale différence entre l'ACP et cette représentation est due à la nature des opérations effectuées :

- l'ACP utilise des projections
- le BIPLOT utilise des approximations

Un tracé sur le plan 2-3 d'un BIPLOT n'a plus les propriétés d'un BIPLOT car on perd dans ce cas la reconstitution des données initiales. C'est pourquoi le BIPLOT est généralement limité au plan 1-2.

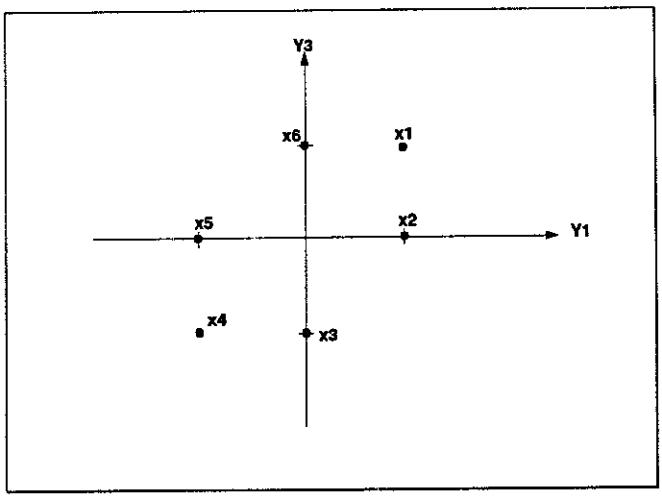
Par contre, il est possible de faire une représentation dans l'espace en prenant les 3 premiers axes factoriels. On garde dans ce cas toutes les propriétés du BIPLOT (appelé dans ce cas BIMODEL).

4•4•3 Un cas d'école (si l'on veut faire les calculs manuellement !)

On cherche à représenter :

$$Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 \\ -1 & -1 & -1 & -1 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Cette matrice représentant un nuage de points dans un espace de dimension 4 est formée de deux variables dédoublées. La représentation sur le plan Y1-Y3 permet d'avoir une idée de la forme du nuage de points.



Les valeurs caractéristiques de Y sont :

Moy(Y1) = Moy(Y2) = Moy(Y3) = Moy(Y4) = 0

Var(Y1) = Var(Y2) = Var(Y3) = Var(Y4) = 4/6

Corrélation(Y1,Y3) = 1/2

Rang(Y) = 2

Valeurs propres de Y'Y :
l1 = 12
l2 = 4
l3 = 0
l4 = 0

Vecteurs propres

$$\Delta_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix} \qquad \Delta_2 = \begin{bmatrix} -1/2 \\ -1/2 \\ 1/2 \\ 1/2 \end{bmatrix}$$

On peut alors décomposer Y (Y = ULV') :

$$V = [\Delta_1 \Delta_2] = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} = \begin{bmatrix} 2\sqrt{3} & 0 \\ 0 & 2 \end{bmatrix}$$

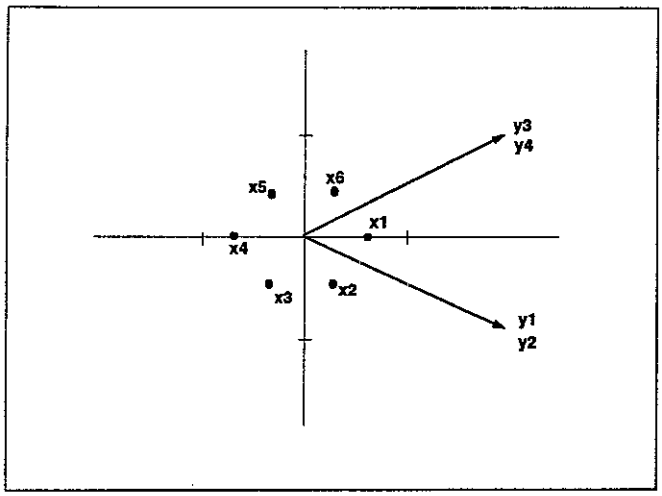
$$U = YV\Lambda^{-1} = \begin{bmatrix} 1/\sqrt{3} & 0 \\ 1/2\sqrt{3} & -1/2 \\ -1/2\sqrt{3} & -1/2 \\ -1/\sqrt{3} & 0 \\ -1/2\sqrt{3} & 1/2 \\ 1/2\sqrt{3} & 1/2 \end{bmatrix}$$

Il est maintenant possible de regarder les trois représentations possibles de Y : (les résultats complets se trouvent en annexe 5 : exemple 4)

4•4•3•1 Dans l'espace des variables

(Méthode GH' : $Y \approx \hat{Y} = AB' = (U_{[2]})(\Lambda_{[2]} V'_{[2]})$ ***)***

$$Y \approx \hat{Y} = \begin{bmatrix} 1/\sqrt{3} & 0 \\ 1/2\sqrt{3} & -1/2 \\ -1/2\sqrt{3} & -1/2 \\ -1/\sqrt{3} & 0 \\ -1/2\sqrt{3} & 1/2 \\ 1/2\sqrt{3} & 1/2 \end{bmatrix} \begin{bmatrix} \sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} \\ -1 & -1 & 1 & 1 \end{bmatrix}$$



On peut vérifier les propriétés de la méthode GH' qui sont en partie les mêmes que celles du cercle de corrélation de l'ACP :

Le cosinus de l'angle entre deux variables actives représente la corrélation

cos(y1,y3) = 1/2 » corr(Y1,Y3)

La longueur des variables est proportionnelle à la variance des variables d'origine.

$||y3||^2 = 4$ » n Var(y3)

Comme pour tous les biplots, la projection orthogonale des individus sur les variables approxime la distribution de la variable d'origine.

Si on projette les individus sur y1, on retrouve la répartition d'origine à un coefficient multiplicatif a près:

proj(x1/y1) = proj(x2/y1) » $a\, x_{11} = a\, x_{21} = a$ (1)
proj(x3/y1) = proj(x6/y1) » $a\, x_{31} = a\, x_{61} = a$ (0)
proj(x4/y1) = proj(x4/y1) » $a\, x_{41} = a\, x_{51} = a$ (-1)

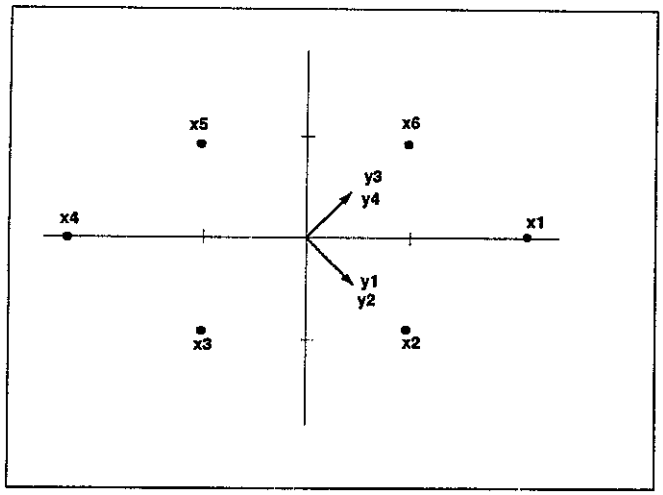
Si on projette les individus sur y3, on retrouve la répartition d'origine à un coefficient multiplicatif b près :

proj(x1/y3) = proj(x6/y3) » $a\, x_{13} = a\, x_{63} = b$ (1)
proj(x2/y3) = proj(x5/y3) » $a\, x_{23} = a\, x_{53} = b$ (0)
proj(x3/y3) = proj(x4/y3) » $a\, x_{33} = a\, x_{43} = b$ (-1)

4•4•3•2 Dans l'espace des individus

(Méthode JK' : $Y \approx \hat{Y} = AB' = (U_{[2]}\Lambda_{[2]})(V'_{[2]})$ *)*

$$Y \approx \hat{Y} = \begin{bmatrix} 2 & 0 \\ 1 & -1 \\ -1 & -1 \\ -2 & 0 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -1/2 & -1/2 & 1/2 & 1/2 \end{bmatrix}$$



On retrouve les distances euclidiennes entre individus de la matrice initiale

d(x1,x4) = 4 = d(X1,X4)

d(x1,x2) = $2^{1/2}$ = d(X1,X2)

Comme pour tous les biplots, la projection orthogonale des individus sur les variables approxime la distribution de la variable d'origine.

Si on projette les individus sur y1, on retrouve la répartition d'origine à un coefficient multiplicatif a près:

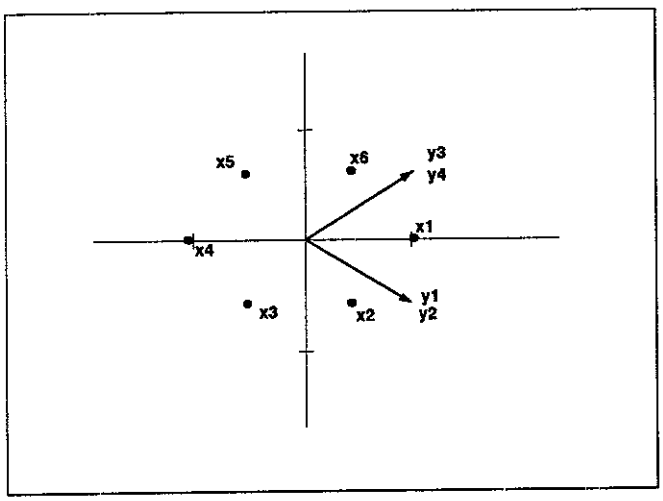
proj(x1/y1) = proj(x2/y1) » a x_{11} = a x_{21} = a (1)
proj(x3/y1) = proj(x6/y1) » a x_{31} = a x_{61} = a (0)
proj(x4/y1) = proj(x4/y1) » a x_{41} = a x_{51} = a (-1)

Si on projette les individus sur y3, on retrouve la répartition d'origine à un coefficient multiplicatif b près:

proj(x1/y3) = proj(x6/y3) » a x_{13} = a x_{63} = b (1)
proj(x2/y3) = proj(x5/y3) » a x_{23} = a x_{53} = b (0)
proj(x3/y3) = proj(x4/y3) » a x_{33} = a x_{43} = b (-1)

4•4•3•3 Méthode Symétrique : $Y \approx \hat{Y} = AB' = (U_{[2]}\Lambda_{[2]}^{1/2})(\Lambda_{[2]}^{1/2}V_{[2]}')$

$$Y \approx \hat{Y} = \begin{bmatrix} \sqrt{2\sqrt{3}}/\sqrt{3} & 0 \\ \sqrt{2\sqrt{3}}/2\sqrt{3} & -\sqrt{2}/2 \\ -\sqrt{2\sqrt{3}}/2\sqrt{3} & -\sqrt{2}/2 \\ -\sqrt{2\sqrt{3}}/\sqrt{3} & 0 \\ -\sqrt{2\sqrt{3}}/2\sqrt{3} & \sqrt{2}/2 \\ \sqrt{2\sqrt{3}}/2\sqrt{3} & \sqrt{2}/2 \end{bmatrix} \begin{bmatrix} \sqrt{2\sqrt{3}}/2 & \sqrt{2\sqrt{3}}/2 & \sqrt{2\sqrt{3}}/2 & \sqrt{2\sqrt{3}}/2 \\ -\sqrt{2}/2 & -\sqrt{2}/2 & \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$$



Les individus et les variables sont équitablement répartis autour de chaque axe.

Comme pour tous les biplots, la projection orthogonale des individus sur les variables reflète la distribution de la variable d'origine.

Si on projette les individus sur y1, on retrouve la répartition d'origine à un coefficient multiplicatif a près:

proj(x1/y1) = proj(x2/y1) » $a\,x_{11} = a\,x_{21} = a\,(1)$
proj(x3/y1) = proj(x6/y1) » $a\,x_{31} = a\,x_{61} = a\,(0)$
proj(x4/y1) = proj(x4/y1) » $a\,x_{41} = a\,x_{51} = a\,(-1)$

Si on projette les individus sur y3, on retrouve la répartition d'origine à un coefficient multiplicatif b près:

proj(x1/y3) = proj(x6/y3) » $a\,x_{13} = a\,x_{63} = b\,(1)$
proj(x2/y3) = proj(x5/y3) » $a\,x_{23} = a\,x_{53} = b\,(0)$
proj(x3/y3) = proj(x4/y3) » $a\,x_{33} = a\,x_{43} = b\,(-1)$

4•4•4 Un exemple d'interprétation : la consommation de protéines

Dans [Gabriel 81] et [Gabriel 86], K.R. Gabriel analyse la consommation de protéines de 25 pays européens (source : A. Weber 1973 Agrarpolitik im Spannungsfeld der internationalen Ernährungspolitik, Kiel, Institut für Agrarpolitik und Marktlehre)

Consommation de protéines en Europe (grammes par tête par jour)

	Viande rouge	Porc et Volaille	Oeufs	Lait	Poisson	Céréales	Féculents	Noix	Fruits et Légumes
Albanie	10.10	1.40	0.50	8.90	0.20	42.30	0.60	5.50	1.70
Autriche	8.90	14.00	4.30	19.90	2.10	28.00	3.60	1.30	4.30
Belg. Luxem	13.50	9.30	4.10	17.50	4.50	26.60	5.70	2.10	4.00
Bulgarie	7.80	6.00	1.60	8.30	1.20	56.70	1.10	3.70	4.20
Tchécoslov.	9.70	11.40	2.80	12.50	2.00	34.30	5.00	1.10	4.00
Danemark	10.60	10.80	3.70	25.00	9.90	21.90	4.80	0.70	2.40
Allemagne-E	8.40	11.60	3.70	11.10	5.40	24.60	6.50	0.80	3.60
Finlande	9.50	4.90	2.70	33.70	5.80	26.30	5.10	1.00	1.40
France	18.00	9.90	3.30	19.50	5.70	28.10	4.80	2.40	6.50
Grèce	10.20	3.00	2.80	17.60	5.90	41.70	2.20	7.80	6.50
Hongrie	5.30	12.40	2.90	9.70	0.30	40.10	4.00	5.40	4.20
Irlande	13.90	10.00	4.70	25.80	2.20	24.00	6.20	1.60	2.90
Italie	9.00	5.10	2.90	13.70	3.40	36.80	2.10	4.30	6.70
Pays-Bas	9.50	13.60	3.60	23.40	2.50	22.40	4.20	1.80	3.70
Norvège	9.40	4.70	2.70	23.30	9.70	23.00	4.60	1.60	2.70
Pologne	6.90	10.20	2.70	19.30	3.00	36.10	5.90	2.00	6.60
Portugal	6.20	3.70	1.10	4.90	14.20	27.00	5.90	4.70	7.90
Roumanie	6.20	6.30	1.50	11.10	1.00	49.60	3.10	5.30	2.80
Espagne	7.10	3.40	3.10	8.60	7.00	29.20	5.70	5.90	7.20
Suède	9.90	7.80	3.50	24.70	7.50	19.50	3.70	1.40	2.00
Suisse	13.10	10.10	3.10	23.80	2.30	25.60	2.80	2.40	4.90
Royaume-Uni	17.40	5.70	4.70	20.60	4.30	24.30	4.70	3.40	3.30
URSS	9.30	4.60	2.10	16.60	3.00	43.60	6.40	3.40	2.90
Allemagne-O	11.40	12.50	4.10	18.80	3.40	18.60	5.20	1.50	3.80
Yougoslavie	4.40	5.00	1.20	9.50	0.60	55.90	3.00	5.70	3.20

On utilise la méthode GH' (espace des variables) : $Y \approx \hat{Y} = AB' = (U_{[2]})(\Lambda_{[2]} V'_{[2]})$

Le résultat graphique de l'analyse de ces données est celui qui termine le paragraphe **4.4.1.**

- $\|b_i\|^2$ » $n\,\mathrm{Var}(y_i)$

La variance des consommations en protéines est élevée pour les céréales et le lait, et relativement faible pour les autres sources.

- $\cos(b_j,b_{j'}) \approx corr(y_j,y_{j'})$

La consommation de céréales est corrélée négativement avec la consommation de viande.

Les consommations d'origines animales (lait, viande, oeufs, porc) sont corrélées entre elles.

Il n'y a pas de lien entre la consommation de céréales ou de viande avec la consommation de légumes.

- $d^2(a_i,a_{i'}) = d^2_{Mahalanobis}(x_i,x_{i'})$

On distingue bien l'Europe de l'Est de l'Europe de l'Ouest (Est à droite, Nord-Ouest à gauche, Sud-Ouest en bas).

La consommation de protéines reflète bien l'emplacement géographique des pays.

- la projection des individus sur chaque variable approxime la distribution de celles-ci.

L'Europe de l'Est est regroupée autour de la consommation de céréales et en opposition à la consommation de viande animale. Les européens du Nord-Ouest ont une tendance inverse.

Les européens du Sud-Ouest semblent avoir une forte consommation de légumes, fruits, noix.

ANNEXES

Le « cercle factoriel » a étudié plusieurs logiciels qui proposent des graphiques factoriels. Nous ne présenterons ici que quelques logiciels qui nous ont paru offrir des fonctionnalités peu courantes et intéressantes. Mais aucun de ces logiciels ne dispose de l'ensemble des fonctionnalités décrites dans la première partie.

1. FONCTIONNALITES GRAPHIQUES DU LOGICIEL JMP (VERSION V3.0 POUR MACINTOSH)

1.1 REPRESENTATION DU TABLEAU DE DONNEES

Les données sont représentées par un tableau individus x variables, dans une grille de même type que celle d'un tableur. Cependant, étant donné la nature statistique des informations du tableau, les en-têtes de lignes et de colonnes jouent un rôle spécifique.

1.1.1 Colonne

Une colonne contient une variable, repérée par son nom. Un dialogue permet de définir la nature de la variable et de préciser les contraintes portant sur ses modalités. La variable peut contenir trois types de données :

- numérique, avec possibilité de fixer des bornes à son intervalle de variation,
- caractère, avec possibilité de définir la liste des modalités possibles,
- état de la ligne, dont nous parlerons plus bas.

Une variable peut être calculée en fonction des valeurs des autres variables. Dans ce cas, la définition est dynamique, c'est-à-dire qu'il y a recalcul immédiat des valeurs si les données dont elles dépendent sont modifiées. Il est également possible d'acquérir les données sur un instrument de mesure connecté à l'ordinateur.

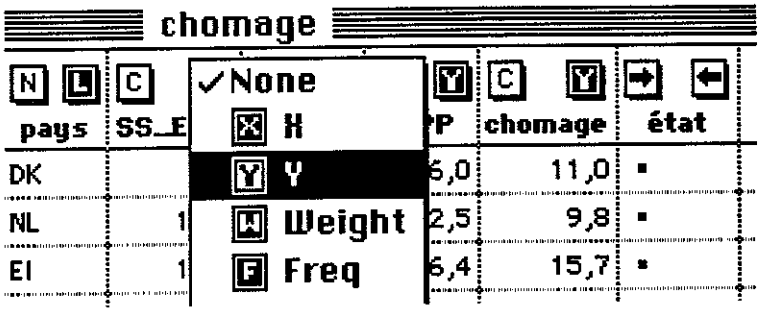
Dans chaque colonne au-dessus du nom, figurent deux boites contenant chacune un menu local. Celle de gauche permet de fixer le type de modélisation de la variable, c'est-à-dire comment la variable sera traitées dans les analyses :

- nominale, la variable est constituée d'une série de modalités, elle est considérée comme une variable discrète,
- ordinale, comme la précédente, mais les modalités (numériques ou non) ont un ordre,
- continue, la variable doit nécessairement être numérique.

La boite de droite permet de préciser le rôle de la variable dans les analyses. Á une variable il est possible d'affecter l'un des rôles suivants :

- aucun, pour mémoire,
- x, la variable joue le rôle d'une variable explicative,
- y, la variable sera « expliquée » par le modèle,

- poids, les individus peuvent être pondérés,
- effectif, comme ci-dessus, mais il est possible de faire des calculs de degré de liberté,
- libellé, le contenu de la variable sera utilisé comme identifiant, il ne peut y avoir qu'une seule variable identifiante.



Une fois que l'utilisateur a spécifié ces informations, il lui reste à choisir les analyses qu'il veut effectuer. C'est le logiciel qui se charge de déterminer, en fonction des choix opérés, la méthode à utiliser pour traiter les données. Ainsi, lorsque l'utilisateur décide d'analyser une variable Y en fonction d'une variable X, JMP utilise la méthode correspondant au type de modélisation des variables, selon le tableau ci-dessous :

Y \ X	**continu**	**nominal ou ordinal**
continu	régression et ajustement	analyse de la variance
nominal ou ordinal	régression logistique	table de contingence

1.1.2 Ligne

Une ligne décrit un individu. Á chaque individu est associé un ensemble d'attributs qui définissent l'état de la ligne. Ces attributs sont des attributs graphiques :

- couleur associée à l'individu et
- style de la marque représentant l'individu dans les représentations cartésiennes.

Il peut également être :

- sélectionné ou non,
- inclus ou exclus de l'analyse,
- caché ou non sur les représentations graphiques,
- affecté d'un libellé ou non (le libellé est toujours la valeur de la variable identifiante).

L'ensemble de ces informations constitue l'état de la ligne. Il est enregistré dans le fichier de données. Il est également possible de créer de nouvelles variables stockant

l'état de chaque ligne. Une telle variable sert à conserver une ou plusieurs configurations particulières des lignes. Il est ensuite possible à tout moment de transférer l'état courant des lignes dans une variable d'état et en sens inverse, d'une variable d'état vers la configuration courante. Les modifications des attributs graphiques sont immédiatement répercutées sur toutes les représentations des données.

L'utilisation d'une variable d'état est un moyen simple et élégant de stocker un ou plusieurs habillages graphiques et de les rendre actifs les uns après les autres.

Dans le tableau ci-dessous, figurent 11 pays, décrit par 2 variables nominales, dont une sert d'identifiant, 4 variables continues. La variable d'état décrit une situation dans laquelle, les États-Unis et le Royaume-Uni sont repérés par un carré, la France et le Japon par un losange et sont munis d'une étiquette, et l'Italie et la Belgique sont sélectionnées.

chomage

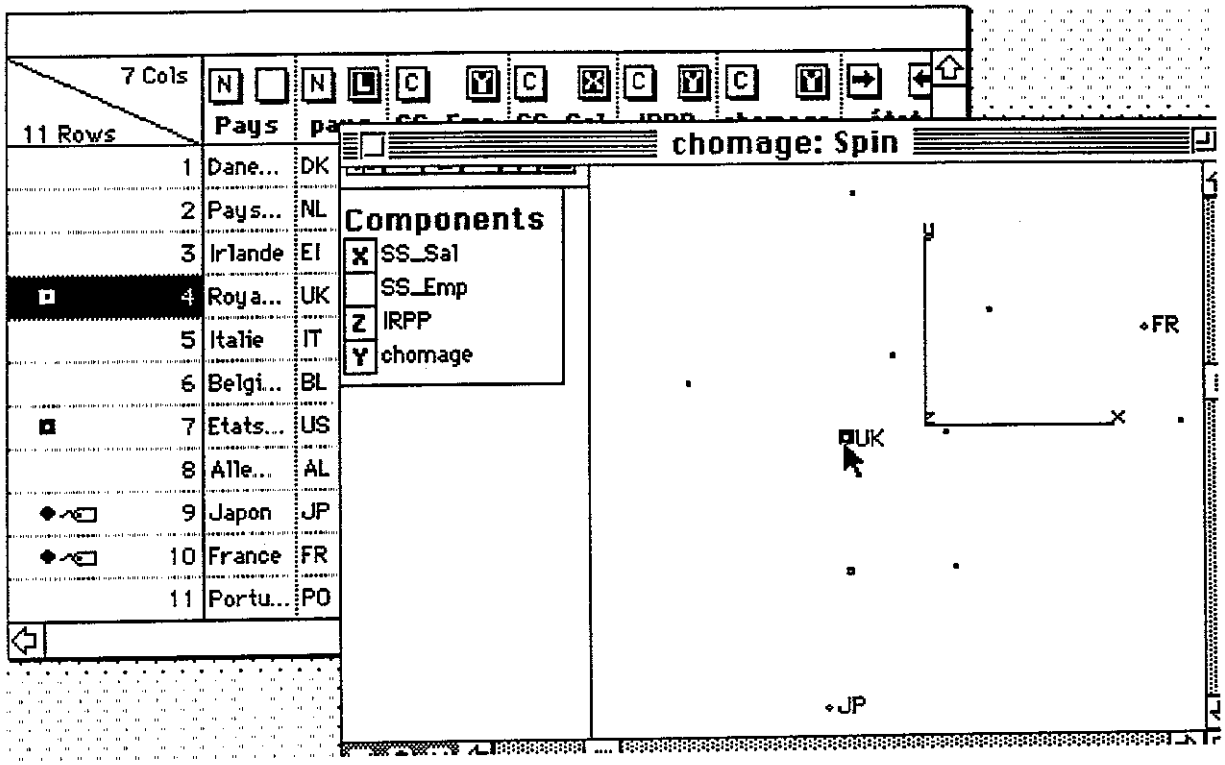
11 Rows \ 7 Cols	Pays	pays	SS_Emp	SS_Sal	IRPP	chomage	état
1	Danemark	DK	0,0	2,5	36,0	11,0	
2	Pays-Bas	NL	10,8	10,7	32,5	9,8	
3	Irlande	EI	12,2	7,8	16,4	15,7	
4	Royaume...	UK	10,4	7,6	15,5	9,6	
5	Italie	IT	50,1	9,0	14,2	11,7	
6	Belgique	BL	41,9	12,1	11,6	12,8	
7	Etats-Unis	US	7,7	7,7	11,3	6,3	
8	Allemagne	AL	18,2	18,2	8,7	10,0	
9	Japon	JP	7,6	7,0	2,4	2,9	
10	France	FR	43,8	17,1	1,0	12,3	
11	Portugal	PO	24,5	11,0	0,9	6,4	

Enfin, il est possible d'attribuer à chaque ligne une marque ou une couleur dépendant des modalités d'une variable nominale ou ordinale.

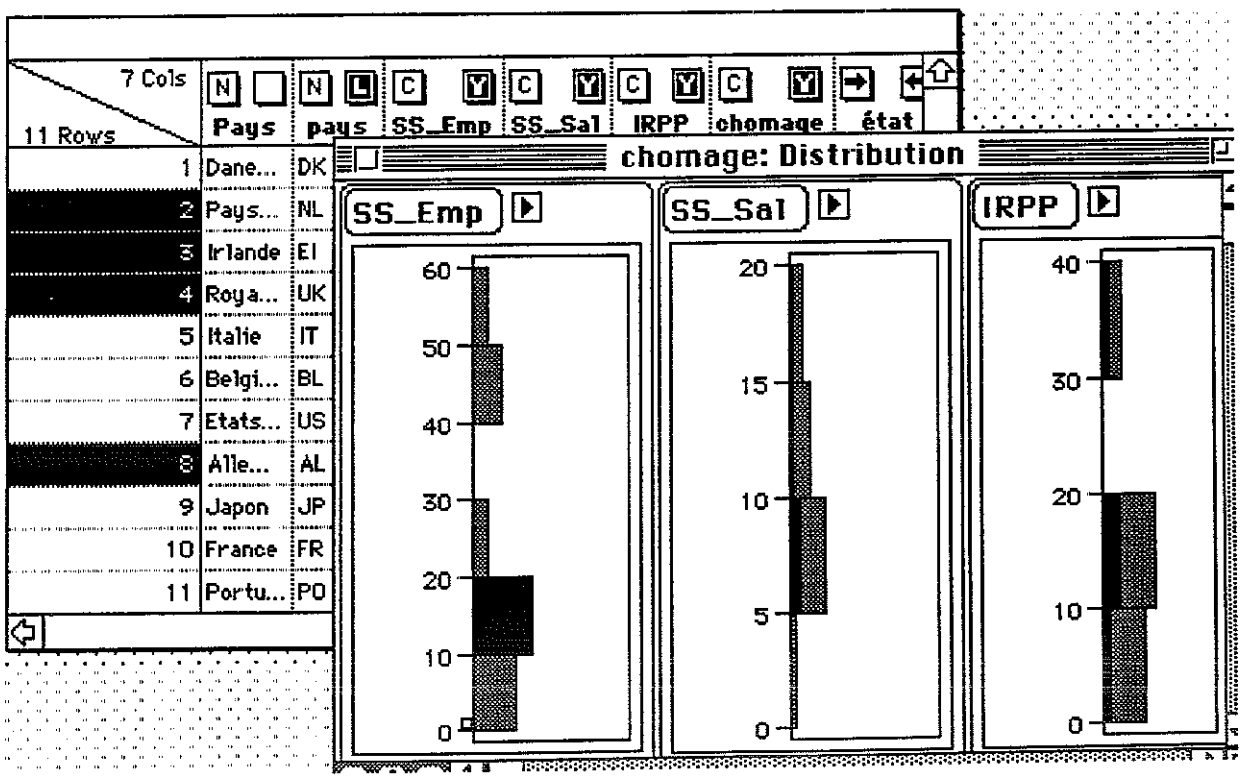
1.2 LIEN DYNAMIQUE ENTRE TOUTES LES VUES

Le tableau de données peut ensuite être visualisé par différentes méthodes. Chacune des représentations que l'on obtient est une « vue » sur le jeu de données qui reflète à tout moment l'état du tableau. Ainsi, un individu sélectionné l'est sur toutes les vues. Ceci fonctionne entre le tableau de données et un graphique cartésien, dans les deux sens, comme on peut le voir ci-dessous : en cliquant sur le point, on le sélectionne et

tant que le bouton reste enfoncé, la valeur de l'identifiant apparaît, bien que le point ne soit pas muni d'une étiquette.

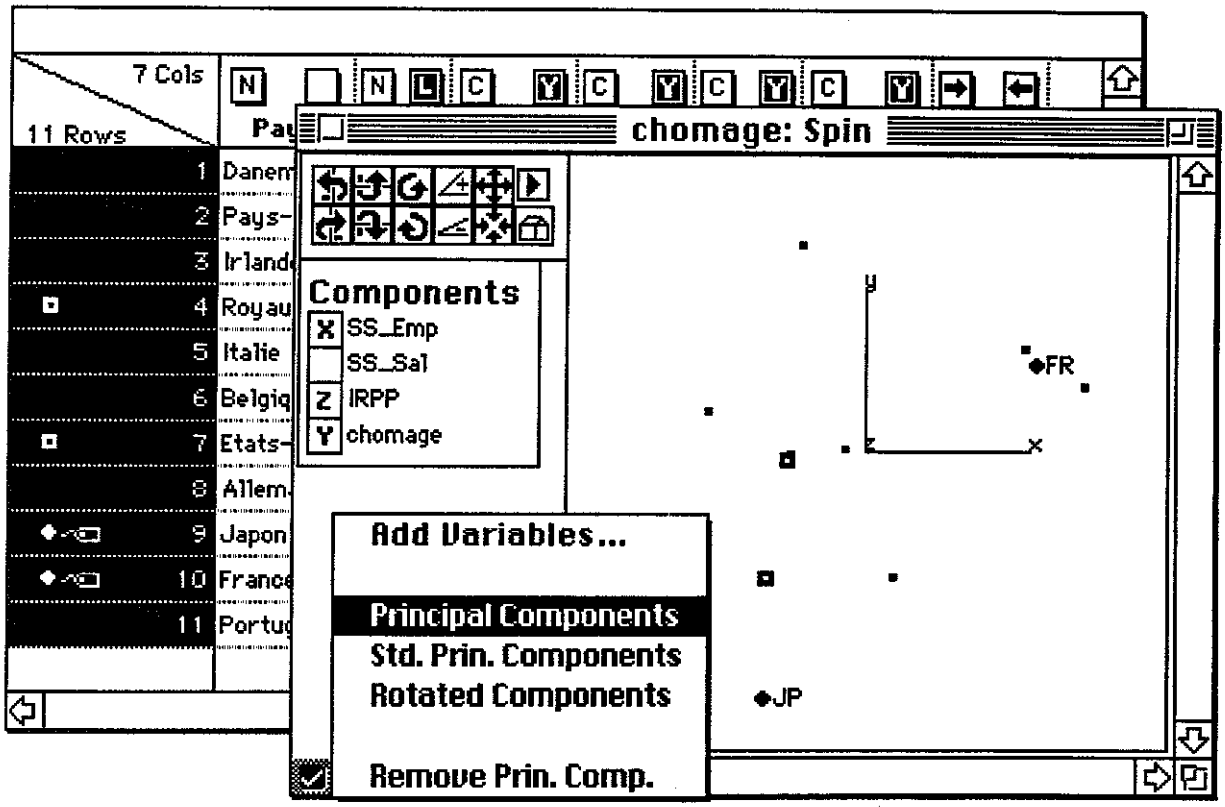


On peut également sélectionner une classe dans un histogramme. Les individus appartenant à la classe sont alors sélectionnés et les autres histogrammes reflètent la sélection en représentant la distribution conditionnelle de ces individus. Sur l'exemple suivant, on a cliqué sur la tranche 10–20 de la variable SS_Emp ; les 4 pays correspondant sont contrastés dans le tableau et leur répartition est représentée en surbrillance sur les histogrammes des autres variables.



1.3 INTERACTIVITE

Chaque fenêtre de résultats est munie d'un menu local qui propose d'autres traitements à réaliser. Ainsi, à partir d'un graphique dans l'espace, il est possible de demander directement à réaliser l'analyse en composantes principales.



1.4 SCRIPTABILITE

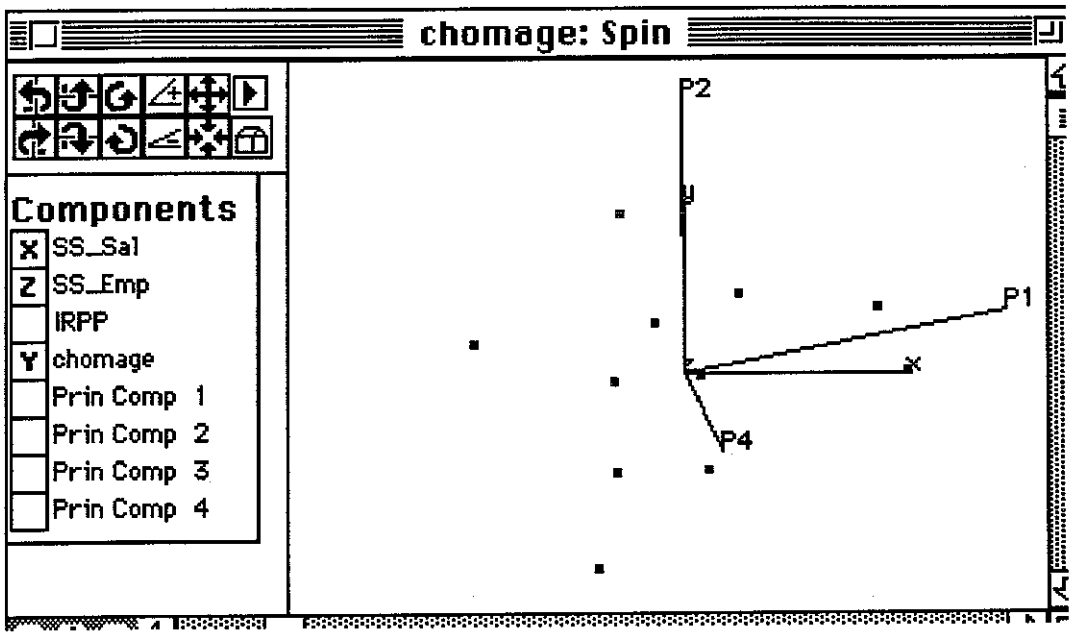
Bien qu'étant un logiciel interactif, JMP possède des capacités de programmation. En effet, il est piloté par des événements qu'il est possible de décrire à l'aide d'un véritable langage de programmation ; il s'agit en fait du langage de script dont le support est prévu en standard dans le système du Macintosh. Le langage permet donc d'enchaîner des traitements dans différents logiciels répondant au même standard ou d'échanger des données, texte, tableau ou graphique avec ces mêmes logiciels.

En principe, toute action effectuée par l'utilisateur peut être « enregistrée ». L'enregistrement génère un texte de programme qu'il est possible d'éditer et de modifier. Ce programme peut ensuite être « rejoué ».

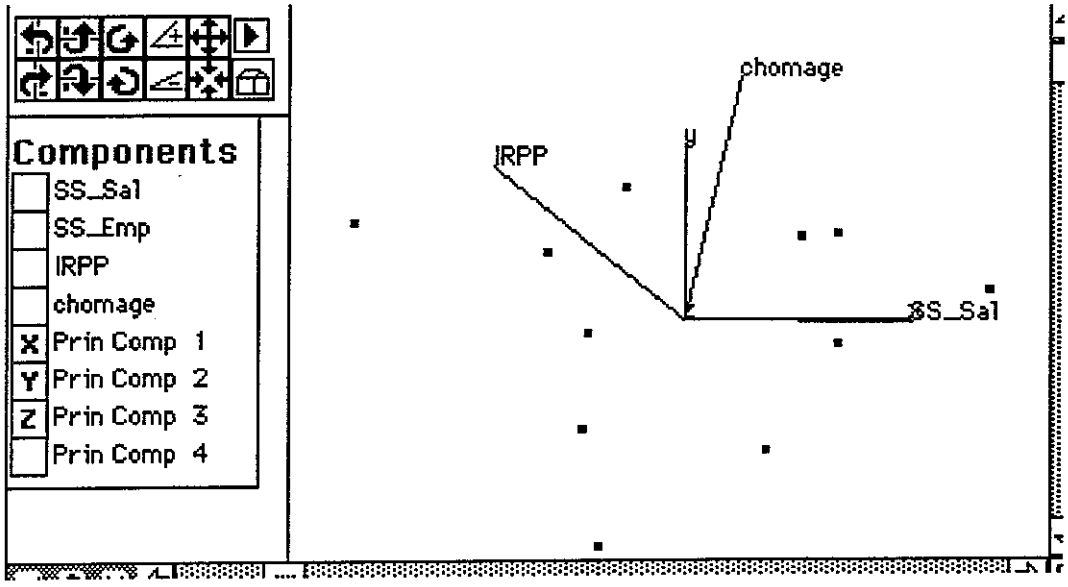
1.5 LES TRAITEMENTS FACTORIELS DE JMP

1.5.1 L'ACP

Elle est proposée comme une extension naturelle de la représentation d'un nuage de points tridimensionnel (spinning plot). Elle rajoute les projections des axes principaux dans le nuage de points sous forme de « biplot rays ». Il est possible de désactiver la représentation mixte des individus et des variables.



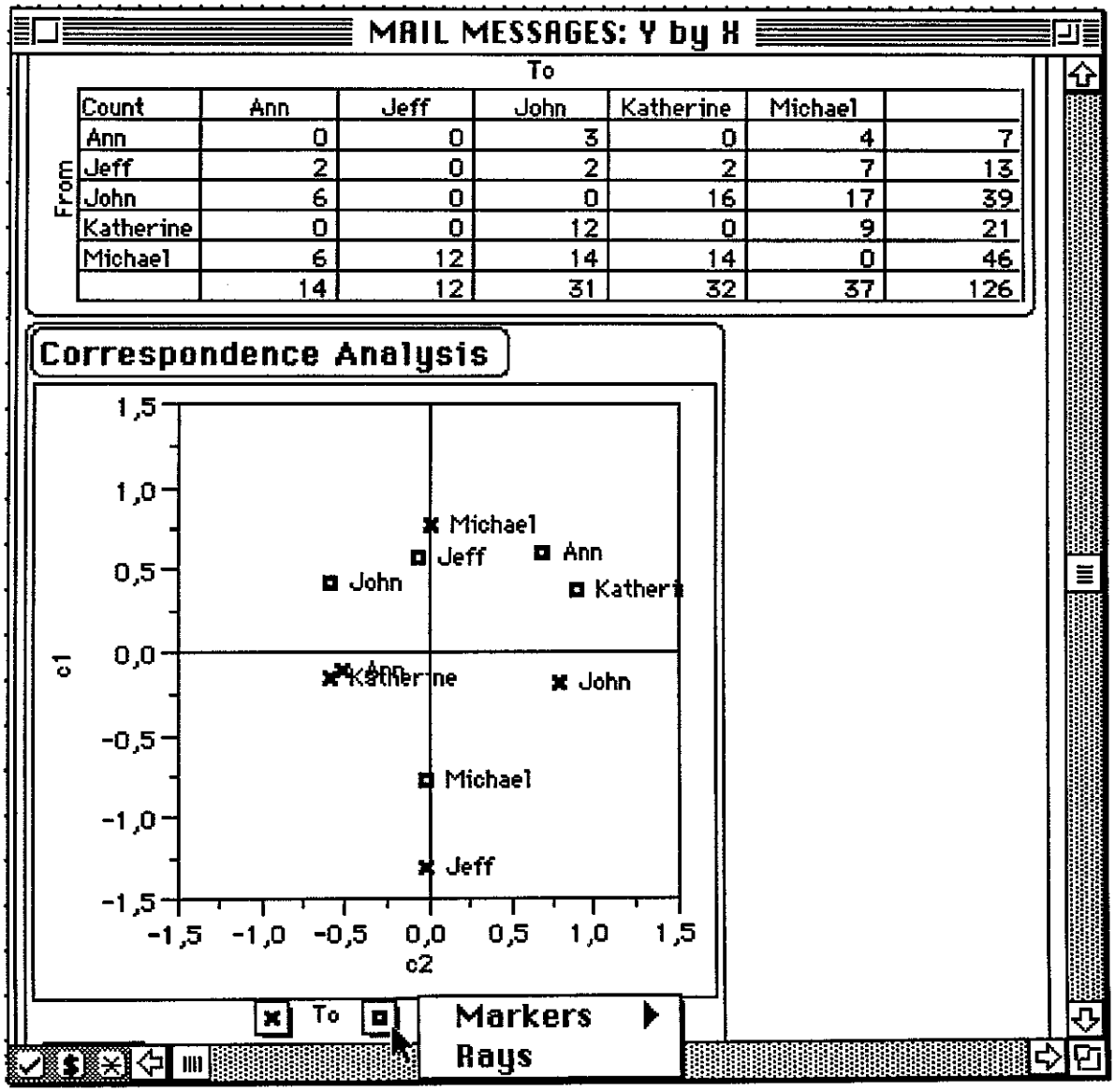
Pour obtenir la représentation du plan principal, il suffit de réaffecter les rôles des axes x, y, z dans les cases prévues à cet effet. Après permutation du rôle des axes, on retrouve le nuage de points projetés dans le premier plan factoriel, avec une représentation mixte de la projection des variables initiales.



Les composantes principales peuvent être ajoutées très simplement aux variables dans le tableau des données initiales. Elles sont alors stockées dans le tableau comme des variables calculées, ce qui permet de projeter aisément des points supplémentaires : l'addition d'une nouvelle ligne provoquera en effet le calcul des composantes du point.

1.5.2 L'AFC

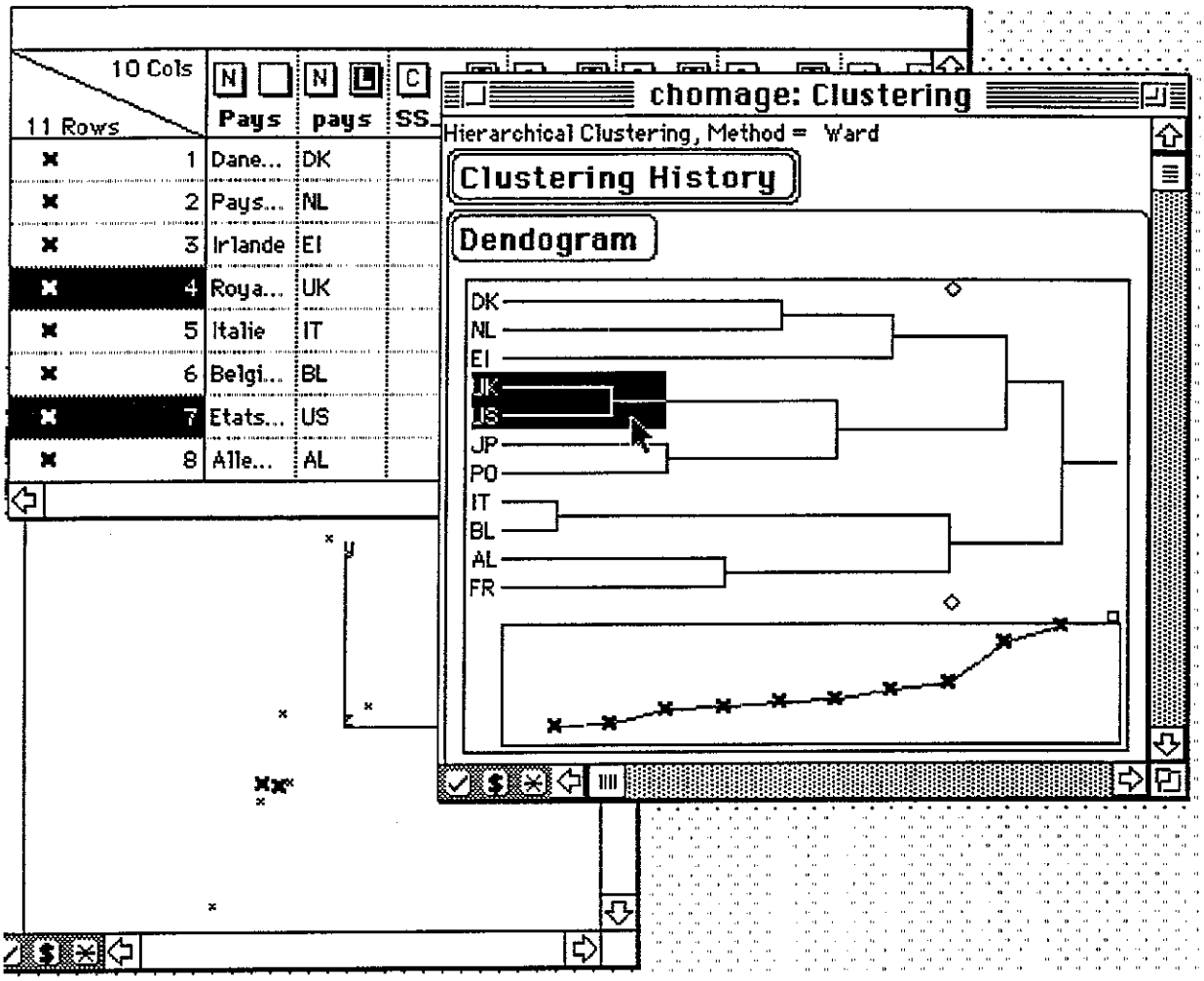
Elle s'obtient comme un traitement complémentaire de l'étude d'un tableau de contingence (ajustement de Y qualitatif par un X qualitatif). Dans la version actuelle de JMP, c'est une fonctionnalité nouvelle qui n'est pas encore aussi bien intégré au reste du logiciel. En particulier, le graphique obtenu n'est pas en liaison dynamique avec les autres vues de données et il est nettement plus difficile que dans l'ACP de récupérer les coordonnées factorielles dans un nouveau tableau de données.



Les deux jeux de modalités sont représentés simultanément par deux marques différentes qu'il est possible de modifier. Il est également possible de relier les modalités de l'une ou l'autre des variables ou des deux au centre du graphique.

1.5.3 La CAH et son exploitation

Elle est proposée comme méthode d'analyse à part entière. Elle fournit l'arbre d'agrégation des classes. Cet arbre est lié dynamiquement aux autres vues, tableau de données, représentation cartésienne ou plan factoriel. Ainsi la sélection d'un nœud de l'arbre sélectionne les points correspondants et permet de visualiser de manière interactive leur position sur les autres graphiques.



Les losanges encadrant l'arbre indiquent visuellement le niveau de segmentation dans la typologie, ils sont déplacés à l'aide du curseur. Ayant ainsi fixé un nombre de classes, il est alors facile de créer une nouvelle variable stockant la typologie obtenue dans le tableau de données. La nouvelle variable peut alors être utilisée pour affecter à chaque individu une marque définie par l'appartenance aux classes.

2. SPAD•GF

SPAD•GF est un logiciel graphique, entièrement interactif, qui permet l'exploitation des plans factoriels fournis par les logiciels **SPAD• N** et **SPAD•T.** Ces plans factoriels sont issus des méthodes classiques d'analyse des données : analyse en composantes principales, analyse des correspondances et analyse des correspondances multiples.

SPAD•GF possède un grand nombre des fonctions graphiques évoquées dans le chapitre 3.

Nous présentons ci-dessous diverses figures qui illustrent quelques unes des fonctions graphiques et statistiques de **SPAD•GF**.

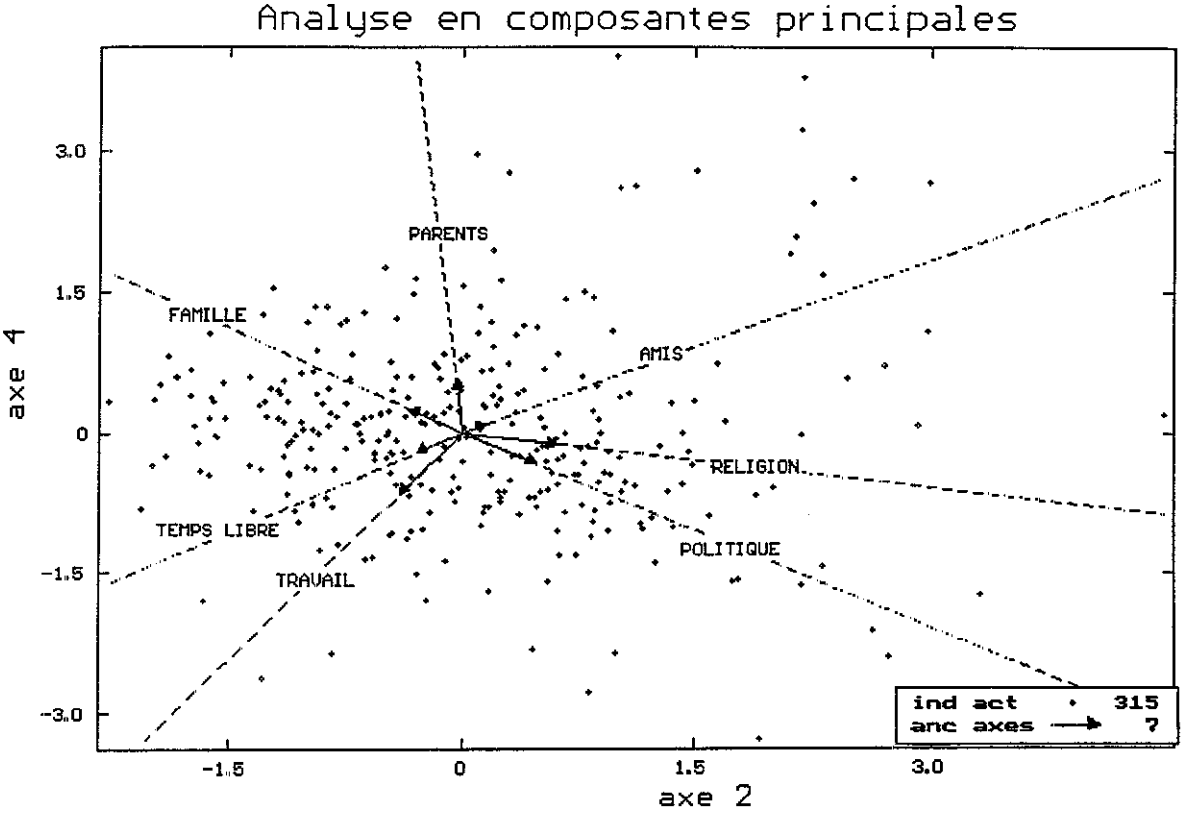


Figure 1 *Représentation simultanée dans le plan en Analyse en Composantes Principales.*

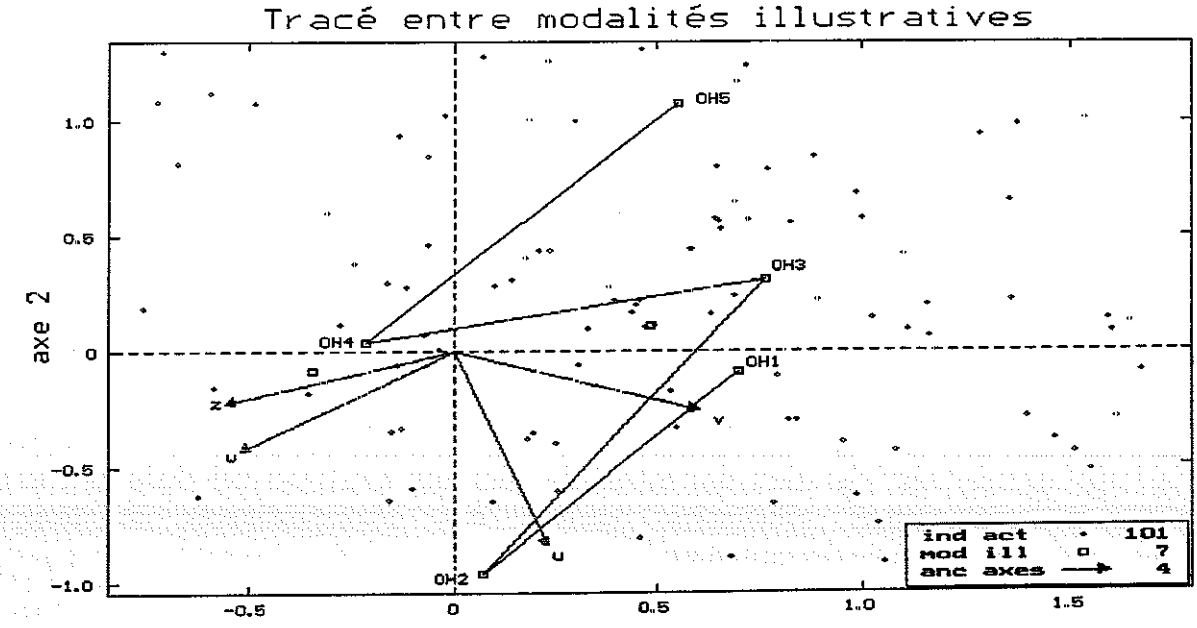


Figure 2 *Zoom sur le centre d'un graphique, les éléments présents auparavant sur le graphique (Identifiants, trajectoires) sont conservés.*

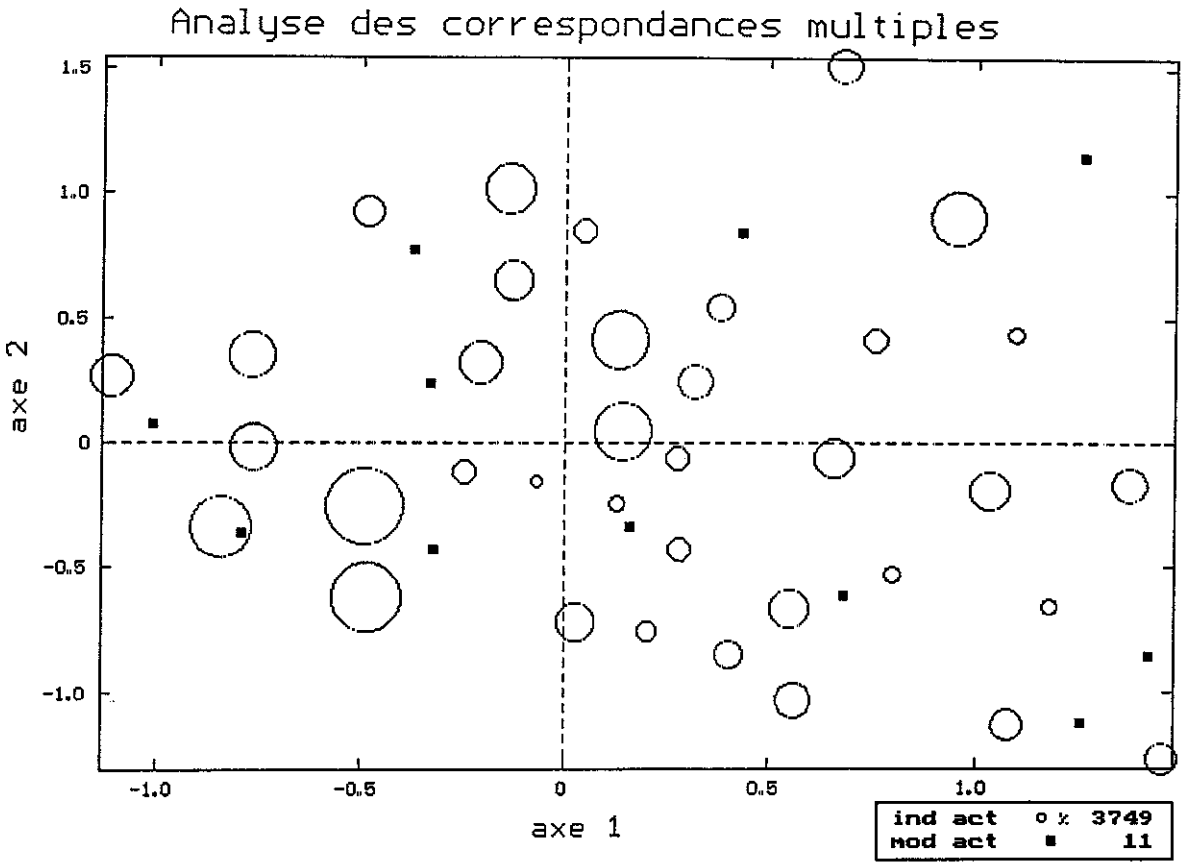


Figure 3 *Traitement des points superposés : on peut choisir de repérer des points superposés sur le plan en leur affectant un symbole circulaire dont la surface est proportionnelle au nombre de points superposés. C'est le cas dans l'analyse des correspondances multiples présentée ici.*

3. LOGICIEL EYELID-2 (VERSION 2.04 POUR DOS)

EyeLID est un logiciel statistique principalement tourné vers l'analyse descriptive et exploratoire et qui comporte un important module graphique. Il a été conçu et développé par J.-M. Bernard et H. Rouanet du Groupe Mathématiques et Psychologie (CNRS) en collaboration avec R. Baldy (BBT, Londres), et est diffusé par la société INDIA (voir Bernard & coll., 1993).

L'idée centrale d'EyeLID consiste à interroger un ensemble de données structurées, un « protocole », à l'aide du « langage d'interrogation des données » (LID). Développé initialement dans le contexte de l'analyse de la variance, sa conception générale en fait un outil d'exploration et d'analyse descriptive pour des types de données variés.

3.1 STRUCTURE DES DONNEES ANALYSEES

Un « protocole » que traite EyeLID est défini en termes d'« unités » statistiques, éventuellement pondérées, de variables numériques et de facteurs structurants (variables nominales). On utilisera ici le terme « facteurs structurants » à ne pas confondre avec les « facteurs » de l'analyse factorielle qu'on appellera plutôt « axes ». Cette structure générale convient aussi bien pour traiter :

- des données expérimentales; les facteurs structurants sont alors des facteurs manipulés par un expérimentateur, et les variables des variables numériques observées,
- des données issues d'une analyse multidimensionnelle (AFC, ACM, ACP, MDS, etc.) ; le protocole est alors un nuage de points, repéré par des axes factoriels (variables numériques), et les facteurs structurants pourront être n'importe quelle variable nominale (définie sur les points) jugée d'intérêt.

Toute variable qui peut s'avérer utile, lors de l'analyse, pour distinguer des groupes d'unités peut être introduite dans la structure des données en tant que facteur structurant : caractère actif ou illustratif, résultat d'une classification, indicateur de bonne représentation dans un axe ou dans un plan, etc. Ceci vaut pour tout type de nuage produit par une des méthodes multidimensionnelles, que les unités ou points du nuage représentent des individus, des modalités (ACM ou AFC) ou des variables numériques initiales (ACP). Le logiciel ne comporte pas de module propre de saisie/modification des données, ni d'analyse factorielle, mais des interfaces permettent de récupérer les données provenant d'autres logiciels (ADDAD, SAS, DS3, VAR3, PAC).

On illustrera seulement ici, sur un exemple d'ACM, quelques fonctionnalités d'EyeLID dans le cas des analyses multidimensionnelles, en insistant sur les aspects graphiques. Il s'agit d'une enquête sur « l'ouvrier français en 1970 » (Adam & coll., 1970) portant sur 1049 ouvriers (les individus) ; on a retenu ici les questions syndicales et politiques (dont 4 actives et 1 supplémentaire). Après ACM et récupération des données au format EyeLID (à l'aide d'un des utilitaires, ADD2LID ou QUEST), on obtient le nuage des individus. Ce nuage est composé de 1049 points de poids 1 (individus « I »), dont certains sont confondus, ou, de façon équivalente, de 319 points pondérés distincts (patrons de réponse « J ») ; il est structuré comme suit (on a retenu les 4 premiers axes) :

Facteurs structurants :

I	1049	Individus
J	319	Patrons de réponse différents observés sur les questions actives
VOTSYN	8	Vote syndical
ADHSYN	8	Adhésion syndicale
PRESA	8	Vote présidentiel 1969 1-er tour
PARTI	8	Sympathie pour un parti politique
PRESB	3	Vote présidentiel 1969 2-eme tour

GROUP 2 Permet de distinguer certains "patrons typiques" (group1) des autres.

Variables (axes factoriels) :

V1, V2, V3, V4

Les facteurs structurants et les variables sont désignés par des codes qui s'utilisent dans les demandes d'analyse LID. On a choisi ici des codes suffisamment explicites pour donner un caractère quasi-naturel aux demandes d'analyse.

3.2 FONCTIONNALITES STATISTIQUES : LE LANGAGE LID

3.2.1 Nuages dérivés

Le langage de d'interrogation de données (LID) permet de définir des nuages dérivés du nuage de base. Les dérivations possibles sont de deux types fondamentaux: « ensemblistes » et « résiduelles ».

3.2.1.1 Dérivations ensemblistes

Tout facteur, e.g. VOTSYN, définit implicitement plusieurs ensembles, dérivés de l'ensemble I :

- des sous-ensembles de I correspondant à chaque modalité de VOTSYN, désignés par : « I/votsyn1 », « I/votsyn2 », etc.;
- l'ensemble de ses modalités : « VOTSYN » équivalent à « votsyn1, votsyn2,....votsyn8 ».

Avec le langage LID, on peut désigner de façon compacte chacun de ces ensembles, ainsi que d'autres ensembles obtenus par les opérations ensemblistes usuelles (union, intersection, ensemble-produit, complément, etc.). Voici les principaux opérateurs ensemblistes du langage :

/	restriction
concat.	composition de modalités (et)
!	négation d'une modalité
_	regroupement de modalités (ou)
,	énumération de modalités
& ou *	ensemble-produit, croisement de deux ensembles
<>	emboîtement de deux ensembles
->	sélection des variables

Le nuage de base peut lui-même être désigné par une formule LID qui serait ici, par exemple:

I & J & VOTSYN & ADHSYN & PRESA & SYMP & PRESB -> V1,V2,V3,V4

Si on ne s'intéresse qu'aux seules coordonnées factorielles des individus, la formule « I -> V » suffirait (« -> V » désigne la sélection de toutes les variables de base) ; mais l'indication des facteurs structurants permettra d'« interroger » le nuage en terme de ces facteurs.

3.2.1.2 Dérivations résiduelles

Les dérivations de type résiduel permettent notamment de « gommer » du nuage de base les différences dues à un facteur structurant. Chacune se ramène à effectuer un ou plusieurs centrages sur un nuage :

- centrage d'un sous-nuage, e.g. « I(votsyn1) » ;
- dérivation « intra », i.e. centrage simultané de plusieurs sous-nuages, e.g. « I(VOTSYN) » ;
- dérivation d'interaction, i.e. double centrage.

3.2.1.3 Nuages résumés (moyens, etc.) et "mots-clés à droite"

Une formule telle que « VOTSYN » désigne le nuage des 8 points moyens associés au facteur structurant VOTSYN; elle peut aussi s'écrire de façon plus explicite « VOTSYN Mean I », le mot-clé « Mean » indiquant que les points dérivés doivent être calculés par moyennage équipondéré sur les individus (« I »). Mais on peut agréger ou résumer les individus par toute autre statistique: avec « VOTSYN Median I », on aura un nuage de points médians, avec « VOTSYN Variance I », on obtiendra les variances des 8 sous-nuages associés à VOTSYN, etc.

3.2.2 Représenter/résumer un nuage

Une formule du langage LID définit un nuage dérivé (éventuellement le nuage de base lui-même). Un mot-clé à gauche de cette formule indique le traitement à effectuer sur ce nuage. On peut, entre autres:

- représenter le nuage, sous forme de tableau (« Table », « Raw ») de graphique (« Graph »);
- calculer des statistiques globales sur le nuage (moyenne, médiane, inertie, variance, contribution à la variance globale, matrice de covariance, de corrélation, etc.).

Ainsi l'exploration avec EyeLID se fait typiquement en alternant des visualisations graphiques qui suggèrent une interprétation pour un axe donné, et des calculs statistiques qui permettent d'attester numériquement de la qualité du résumé ainsi obtenu (e.g., « le vote syndical rend compte de x% de la variance de l'axe 1 »).

3.2.3 Récursivité du langage LID

Une particularité d'EyeLID réside dans le fait que n'importe quel nuage dérivé est assimilable, du point de vue structurel, à un nuage de base ; tout nuage est donc susceptible d'être l'objet de traitements identiques, en particulier, graphiques : un nuage dérivé peut être considéré momentanément ou définitivement comme le nuage de base, être envoyé dans un fichier spécifique qui pourra être traité par EyeLID dans une session ultérieure.

3.3 REPRESENTATION GRAPHIQUE DES NUAGES

Le module graphique agit simultanément sur un ou plusieurs nuages qui, visuellement, apparaissent superposés à l'écran (en général 1 ou 2 nuages, mais il n'y a pas de limitations).

3.3.1 Objets et attributs graphiques

Le graphique est décrit en termes d'un certain nombre d'objets, caractérisés chacun par un certain nombre d'attributs :

- Unités (points) :
  - Marqueurs (pres./abs., type, taille, couleur)
  - Libellés (pres./abs., contenu, type, taille, couleur)
- Liens entre les points (pres./abs, type, couleur, flèches)
- Cadre, Axes, Commentaires, Graduations, Fenêtre de visualisation

À partir du graphique initial affiché à l'entrée du module graphique, un langage de « commandes graphiques » permet de spécifier, de définir, ou de changer tel ou tel attribut de tel objet. Pour les objets liés aux points du nuage, on peut définir des modifications spécifiquement pour une ou plusieurs modalités des facteurs structurants : ainsi « mkcol 3 » (marker color) affectera la couleur de tous les points, alors que « mkcol votsyn1,votsyn2 3 » ne le fera que pour les points indexés par « votsyn1 » ou « votsyn2 », et que « mkcol VOTSYN » fera varier la couleur des marqueurs selon le facteur structurant VOTSYN. Toutes les opérations ensemblistes du langage LID sont disponibles dans les commandes graphiques.

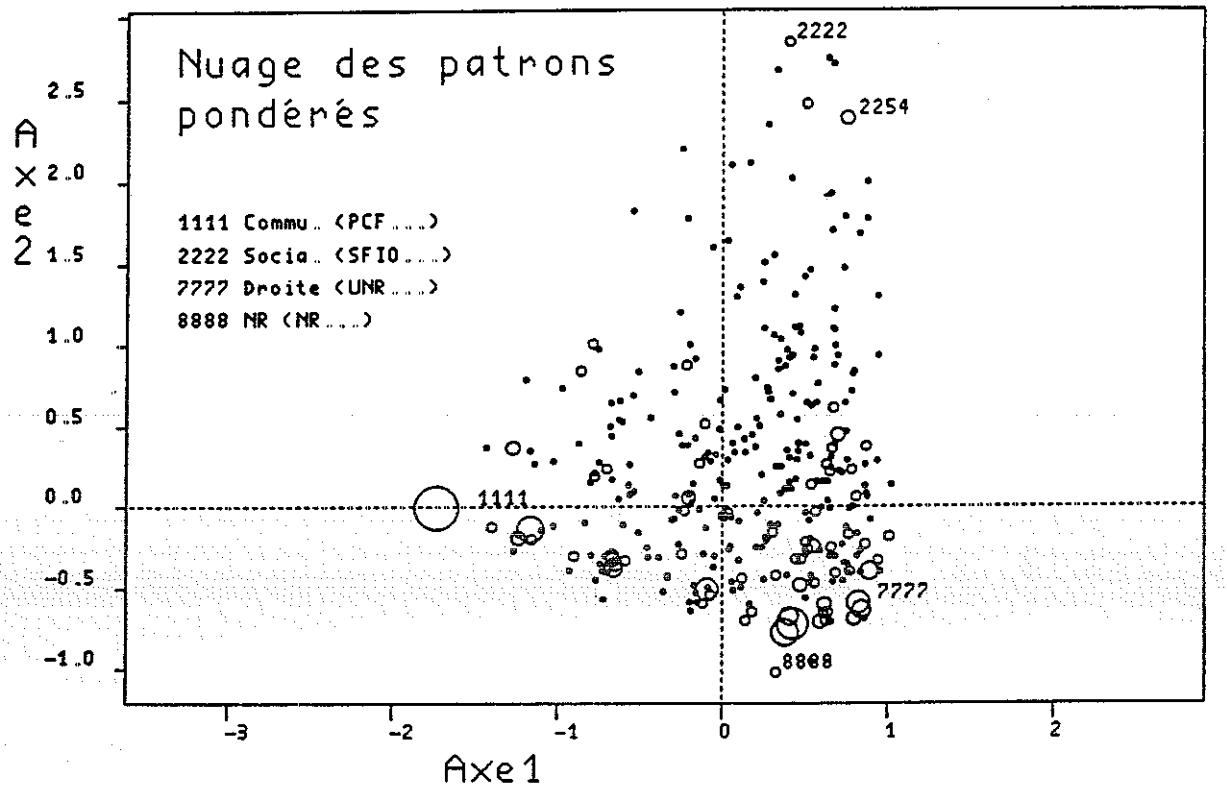


figure 1

Une succession de commandes graphiques peut être enregistrée et rejouée en bloc ultérieurement.

La fenêtre de visualisation consiste en une « caméra » dirigée vers le (ou les) nuage(s) qu'on peut déplacer, centrer, utiliser en « zoomant » ou en rétrécissant; ces manipulations peuvent porter séparément sur l'axe x, sur l'axe y ou sur les deux.

3.3.2 Exemples de graphiques

3.3.2.1 Voir le nuage de base

- La figure 1 représente le nuage de base ; il est obtenu par la demande « Graph J & GROUP -> V ». À partir du graphique initial, on a effectué les opérations suivantes :
- solidarisé la fenêtre (même échelle en x et y) ;
- représenté chaque point par un cercle de taille proportionnelle à son poids ;
- fait apparaître les libellés des patrons de réponse typiques (« group1 ») qui permettent de « baliser » le nuage :
  - 1111 vote et adhésion à la CGT, sympathie PCF, vote Duclos),
  - 2222 (vote et adhésion à la CFDT, sympathie SFIO, vote Deferre),
  - 7777 (vote non-syndiqués, pas d'adhésion syndicale, sympathie UNR, vote Pompidou),
  - 8888 (NR aux 4 questions).

Les libellés peuvent être formels (codes/numéros de modalités) comme ici ou repris d'un dictionnaire de libellés en clair.

On aurait obtenu une figure presque identique avec la demande « Graph I -> V », mais les individus présentant un même patron de réponse auraient été superposés, d'où un nuage non-pondéré, et l'absence du facteur structurant « GROUP » n'aurait pas permis de faire apparaître spécifiquement les libellés de « group1 ». Par contre, les deux protocoles ont bien entendu mêmes variances (valeurs propres de l'ACM):

		V1	V2	V3	V4
Variance J & GROUP	-> V	0.611	0.491	0.416	0.373
Variance I	-> V	0.611	0.491	0.416	0.373

En spécifiant « -> V » dans la demande d'analyse, EyeLID produit en fait un nuage à 4 dimensions « -> V1,V2,V3,V4 », et, par défaut, il utilise les deux premières, i.e. les axes V1 et V2. Une commande graphique (« varxy .... ») permet à tout instant de changer d'axes, en conservant l'ensemble des attributs des points du (ou des) nuage(s).

3.3.2.2 Regarder le nuage de base à la « lumière » des facteurs structurants:

EyeLID permet de nommer (et de stocker) tout nuage dérivé ; le nuage de base, structuré avec tous ses facteurs, est lui-même désigné par « $probase ». La demande « Graph $probase » conduira donc aussi au nuage de la figure 1. À l'aide des deux commandes « varxy... » et « unicol VOTSYN » (faire varier la couleur des points selon la modalité du facteur « VOTSYN »), on peut explorer efficacement le nuage dans n'importe quel plan, à la « lumière » du facteur structurant « VOTSYN ». Une bonne séparabilité visuelle des sous-nuages, ainsi distingués par leur couleur, correspond à une forte contribution à la variance de VOTSYN à un des deux axes ou au plan, ce qu'on peut attester numériquement par :

		V1	V2	V3	V4
Variance VOTSYN	-> V	0.362	0.294	0.129	0.242
Variance I(VOTSYN)	-> V	0.249	0.198	0.287	0.132

La première ligne donne la part de variance prise en compte par VOTSYN, la seconde la part de variance résiduelle ou « intra »: c'est la fameuse décomposition « inter-intra » de la variance. À chacune de ces demandes correspond un nuage qu'on pourrait aussi visualiser: « Graph VOTSYN -> V » (nuage des points moyens de VOTSYN) et « Graph I(VOTSYN) -> V » (nuage résiduel intra-VOTSYN).

3.3.2.3 Un ou plusieurs sous-nuage(s)

La figure 2 représente le sous-nuage des « votes Duclos » (modalité « presa1 ») dans lequel on a distingué les sympathisants PCF (« parti1 ») des autres. Elle s'obtient par la demande LID

« Graph J & parti1,!parti1 /presa1 -> V »,

dans laquelle « ! » signifie la négation, et « / » la restriction. À partir du graphique initial, on a effectué les opérations suivantes:

- choix d'une taille de marqueurs proportionnelle aux poids ;
- choix de types de marqueurs différents pour « parti1 » (PCF, losange plein) et pour « !parti1 » (autre que PCF, cercle vide);
- superposition du nuage de base, par la commande « recall $probase » qui, en recourant à des petits marqueurs (points) sert de fond de carte.

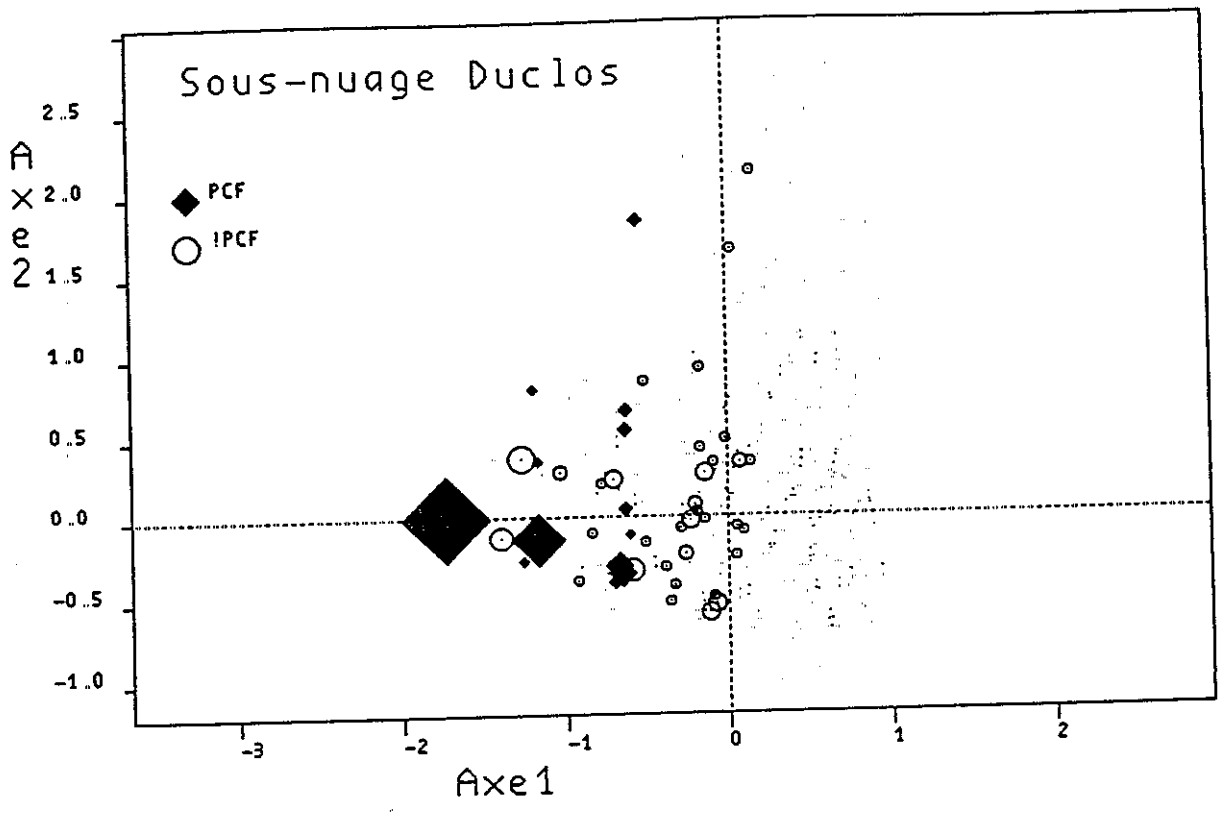
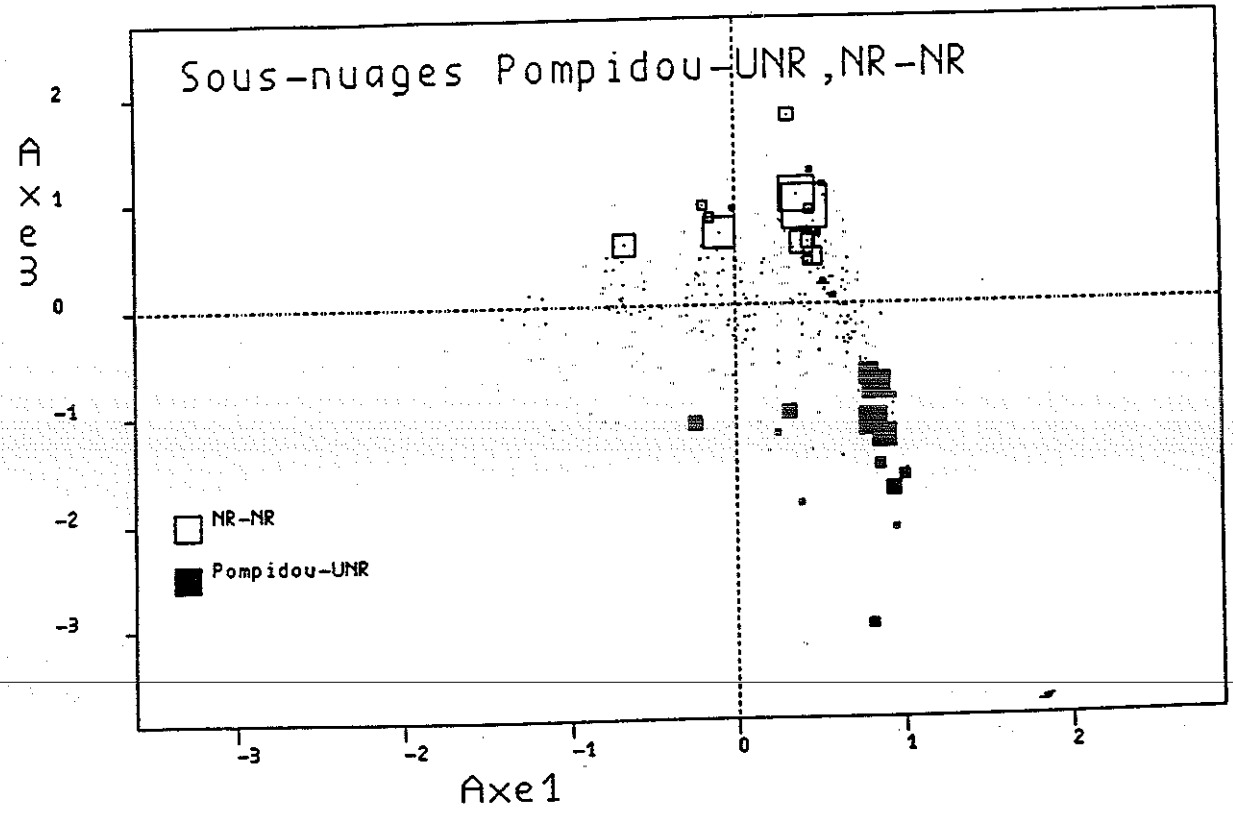


figure 2

La figure 3 représente aussi deux sous-nuages, mais chaque sous-nuage est ici défini par une combinaison de modalités de deux questions. La demande correspondante est

« Graph J & presa7parti7,presa8parti8 -> V »



Parmi les individus, elle distingue ceux qui sont, à la fois, votants pour Pompidou (« presa7 ») et sympathisants UNR (« parti7 »), de ceux qui n'ont répondu ni à la question présidentielle (« presa8 ») ni à la question de sympathie pour un parti (« parti8 »). Ici également on a mis le nuage de base en fond de carte (« recall $probase ») et changé d'axes (« varxy 1 3 »).

La figure 3, qui permet d'interpréter l'axe 3 comme un axe politique spécifique (qui oppose la « droite » aux « non-répondants »), a été choisie à partir des contributions des modalités composées de PRESA et PARTI au plan 1-3 qu'il est possible de visualiser en utilisant un protocole dérivé par mot-clé à droite « Cta » (contribution absolue) :

« Graph PRESA & PARTI Cta I -> V1,V3 »

Ce graphique (voir figure 4) comporte 50 points (sur les 8x8 possibles); chaque point a pour coordonnées la contribution absolue (« Cta ») du sous-nuage associé ; tous les points à faible contribution au plan 1-3 sont rassemblés près de l'origine et seuls « presa7parti7 » et « presa8parti8 » se distinguent sur l'axe 3.

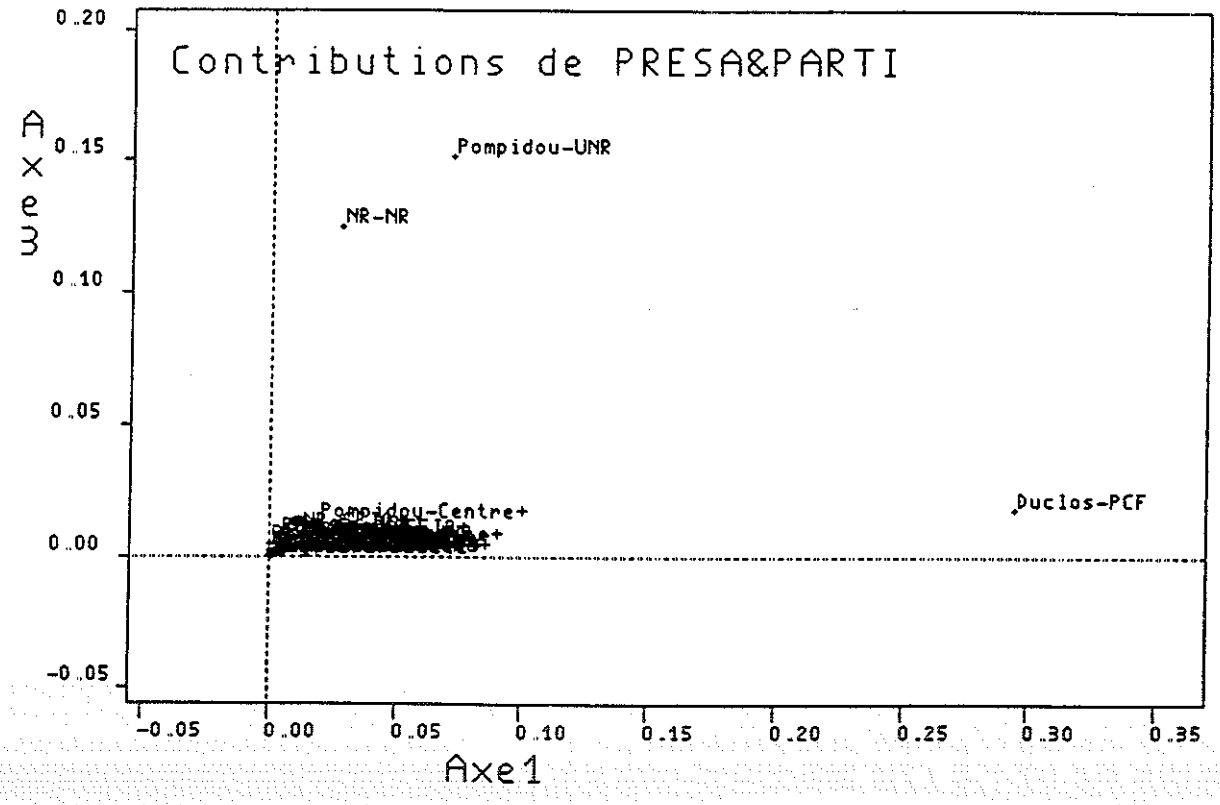


figure 4

3.3.2.4 Nuage de points-moyens associés à un facteur structurant

La figure 5 correspond à la superposition d'un nuage de 3 points moyens « adhsyn1, adhsyn2, adhsyn7 -> V » (CGT, CFDT, Non syndiqués) et du nuage de base « $probase ». Une seule commande graphique (« pjoin ») permet de joindre canoniquement les deux nuages : toute unité de base indexée par « adhsyn1 » est automatiquement liée au point moyen « adhsyn1 », d'où les étoiles de la figure 5 qui n'est, en fait, qu'une version graphique de l'idée décomposition inter-intra.

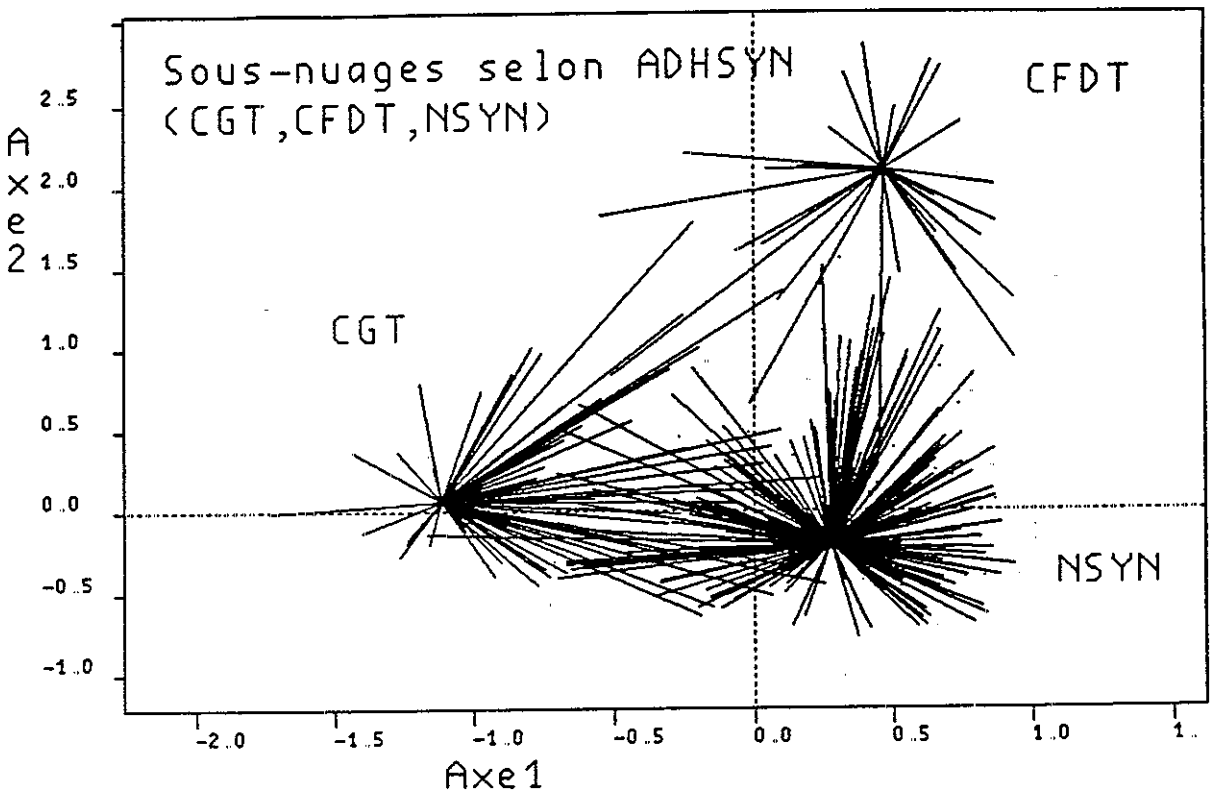


figure 5

Ce graphique suggère, par exemple, d'interpréter l'axe 1 comme celui de l'opposition entre adhsyn1 (CGT), et adhsyn2_adhsyn7 » (CFDT et Non-syndiqués regroupés). Le calcul de diverses contributions permet de quantifier numériquement la qualité de ce résumé, ainsi que la part de variance perdue (i) en regroupant « adhsyn2 » et « adhsyn7 » et (ii) en omettant les autres modalités de ADHSYN ("Cti" signifie "contribution interne" à la variance) :

Cti ADHSYN	-> V1	0.336
Cti adhsyn1,adhsyn2,adhsyn7	-> V1	0.324
Cti adhsyn1,adhsyn2_adhsyn7	-> V1	0.322

3.3.2.5 Nuage associé au croisement de deux facteurs structurants:

On peut également superposer deux nuages dérivés. Dans la figure 6, on a superposé le nuage « PRESA -> V » (8 points moyens pour le vote présidentiel, 1-er tour) au nuage « PRESA & PRESB -> V » (24 points moyens correspondant au croisement du vote 1-er tour et du vote second tour). En joignant canoniquement ces deux nuages (« pjoin »), on peut visualiser le « report des voix » du premier tour (« PRESA ») sur le second tour (« PRESB » = Pompidou, Poher, Abstention + NR).

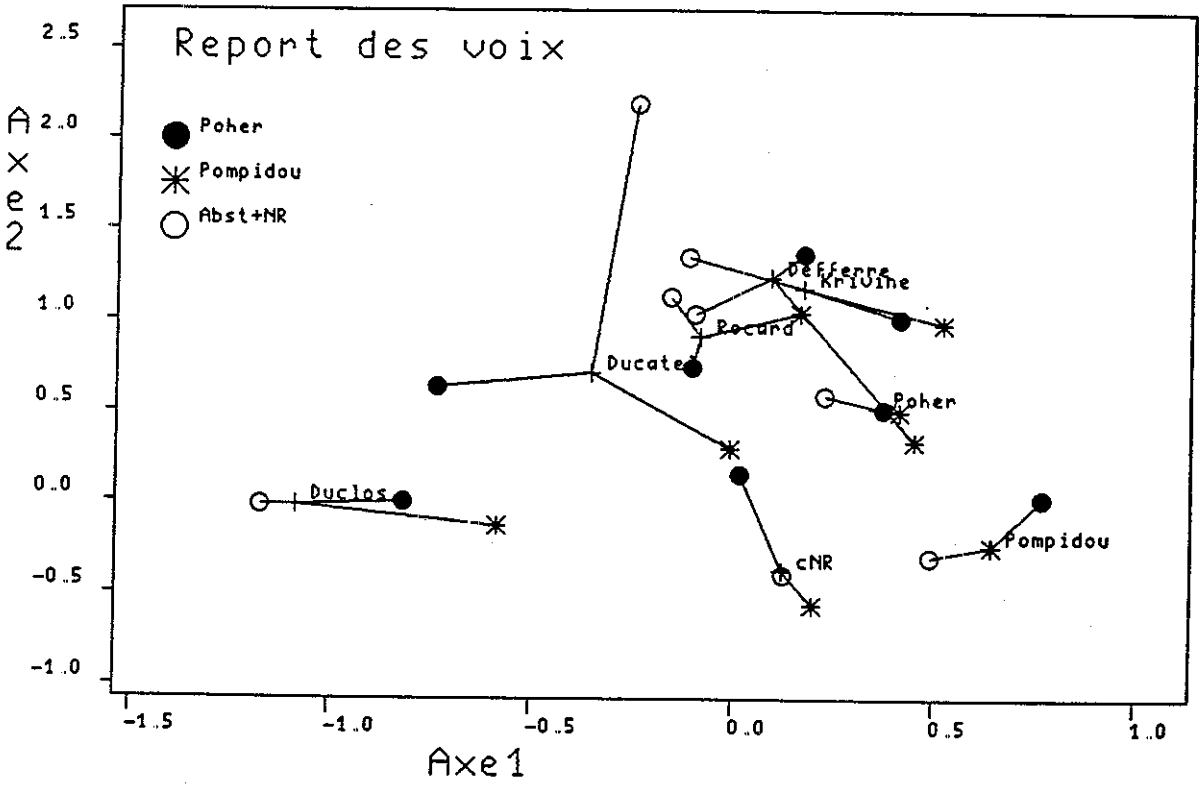


figure 6

3.3.2.6 Méthode EyeLINK

Le programme EyeLINK est uniquement graphique et fonctionne comme le module graphique d'EyeLID. Il dispose d'une fonctionnalité supplémentaire qui permet de visualiser, dans le plan choisi à un instant donné, les distances entre points calculées en prenant en compte toutes les dimensions du nuage. On peut, par exemple, voir apparaître successivement des liens (« link ») entre les points les plus proches, en ne mettant de libellés que pour les points ainsi reliés. La figure 7 représente les 14 premiers liens du nuage des modalités actives de l'enquête « Ouvrier »:

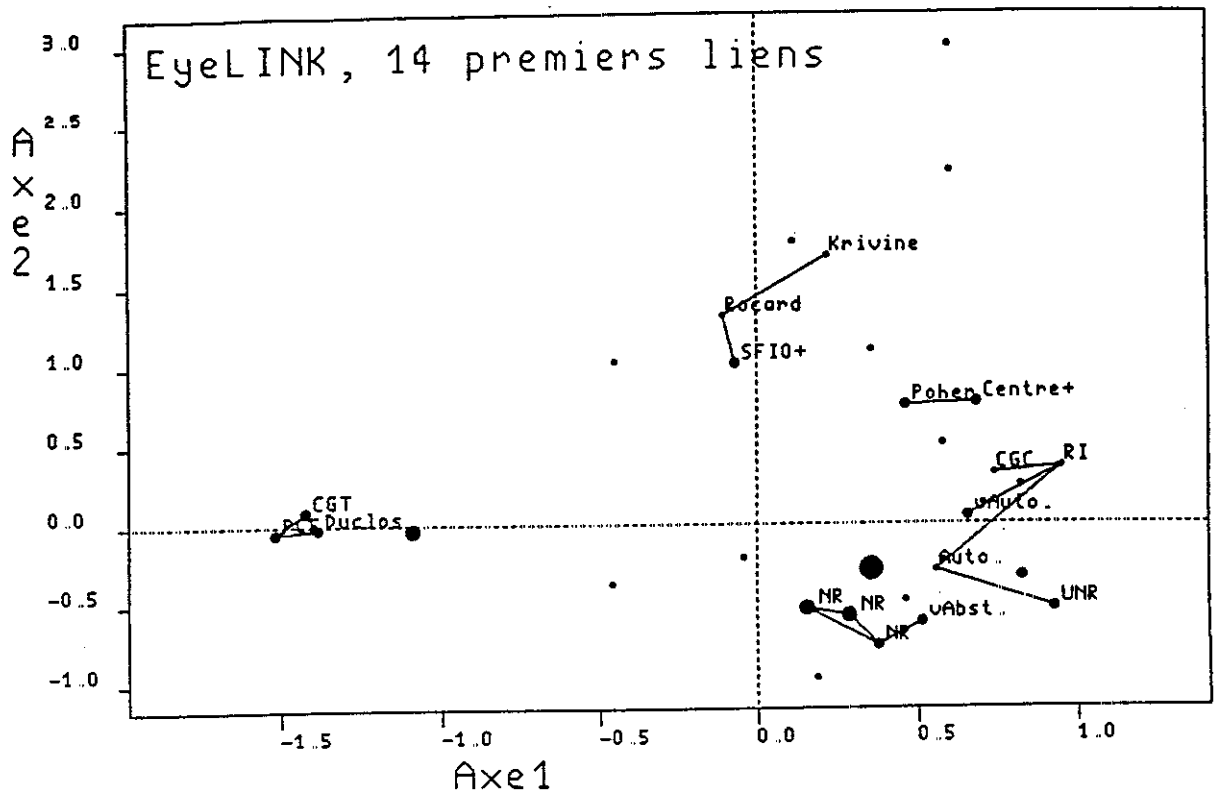


figure 7

3.4 CONCLUSION

EyeLID s'avère être un outil extrêmement efficace pour l'exploration rapide d'un nuage, bien que moins « fini » lorsqu'il s'agit de produire des graphiques publiables définitifs. La souplesse du langage LID permet aussi bien de dégager des traits généraux d'un nuage que de se focaliser sur des questions fines.

L'utilisation de langages de commandes (langage LID + commandes graphiques), et la récursivité en font un outil puissant, même si l'interactivité est moins immédiate qu'avec les menus déroulants, mais ceci est le prix d'un surcroît de liberté dans l'analyse. Un autre avantage du fait de recourir à des langages est de permettre d'automatiser une partie du traitement: EyeLID peut aussi fonctionner en « batch » ou en mode « démonstration » à partir de demandes LID et de commandes graphiques saisies dans un fichier texte.

4.CUMULUS II

4.1 RESUME

Cumulus permet la représentation graphique de plans factoriels issus d'analyses factorielles effectuées par SAS ou par SPADN. Les variables et les observations, actives et illustratives peuvent être représentées ainsi que les caractéristiques du plan (valeurs propres, taux d'inertie des axes...).

Le nuage des individus peut être éclaté en sous nuages suivant les modalités d'une variable qualitative, ou suivant les modalités d'une variable qualitative, ou suivant les classes issues d'une classification.

Cumulus permet :

- d'obtenir les labels des points sur 20 caractères,
- de séparer les labels se recouvrant,
- de choisir les symboles des points et leurs attributs,
- de travailler tous les éléments du graphique : titres, notes de bas de page, axes, taux d'inertie...,
- de contrôler les polices de caractères, leur taille, leur couleur.

Une première version de ***Cumulus*** a été développée en 1986 avec SAS version 5.18 et a été entièrement réécrite pour la version 6.06. Les facilités du langage SCL ont été utilisées pour améliorer l'interactivité, la convivialité, ainsi que le temps de réponse. Il permet de produire des résultats tels que présentés en figure 1.

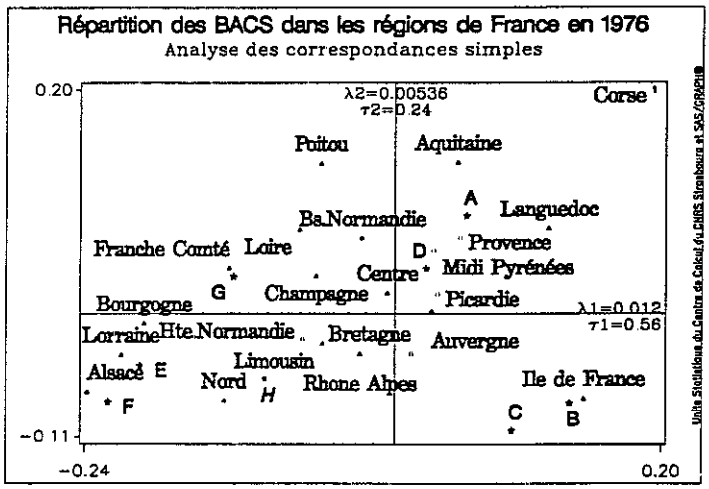


figure 1

4.2 PRESENTATION

Cumulus est une application graphique développée sous forme de menus. L'utilisateur peut agir à l'aide d'une souris ou d'un clavier standard.

Quand ***Cumulus*** a besoin d'informations complémentaires une fenêtre s'ouvre automatiquement et une question en clair demande à l'utilisateur d'entrer l'information nécessaire : par exemple, le nom du fichier contenant les résultats de l'analyse factorielle. Mais il y a très peu de fenêtres dans lesquelles il est nécessaire

d'entrer de l'information. La plupart du temps il suffit de cliquer sur-le-champ souhaité.

Sur les premiers écrans de ***Cumulus,*** il faut préciser le logiciel qui a produit l'analyse dont on veut représenter les résultats. Il faut alors entrer le nom des fichiers dans lesquels ont été sauvegardés les résultats, et préciser le numéro des axes factoriels du plan à représenter. Dans le cas de résultats venant de SAS il faut encore préciser si les résultats proviennent d'une analyse en composantes principales ou d'une analyse des correspondances.

Riche de toute cette information ***Cumulus*** affiche à l'écran un premier graphique factoriel du plan souhaité (figure 2). Un titre est fourni par défaut et donne le nom de l'analyse effectuée. Les valeurs propres et les taux d'inerties sont inscrits au bout des axes correspondants. Les points éloignés de plus de 2.3 écarts-types sont ramenés sur le bord du cadre, afin d'éviter que les points trop excentrés ne regroupent l'ensemble du nuage des points autour du centre de l'origine.

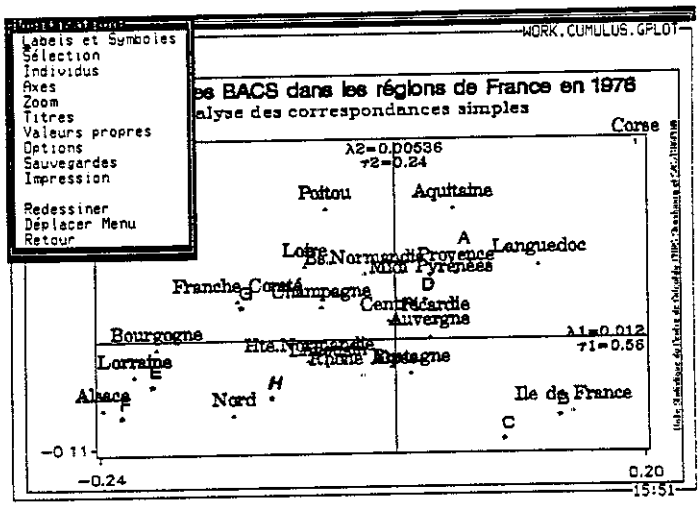


figure 2 : menu de modifications

Ces points sont accompagnés d'une flèche qui indique la direction de leur position réelle. Sur ce premier graphique, des polices de caractères ainsi que des couleurs différentes par défaut permettent de distinguer les variables des individus, aussi bien actifs qu'illustratifs. D'autre part, tous les noms de variables (longs de 20 caractères) sont affichés au-dessus d'un symbole qui représente la position exacte de la variable. Il peut donc y avoir sur le premier graphique recouvrement des noms de variables.

Une petite fenêtre « Modifications » est affichée dans un coin du graphique et présente toutes les options de modification du graphique. Cette fenêtre pouvant cacher certains points du graphique il est possible de la déplacer par l'option « Déplacer Menu ». L'utilisateur peut ainsi choisir les modifications souhaitées tout

en visualisant le graphique et pourra à tout moment, quand il le désire, demander la mise à jour du graphique pour voir l'effet des modifications demandées.

L'option "Label et symboles" permet de modifier la couleur, la police et la taille des symboles et des noms des points. Ces modifications peuvent être faites pour tout un groupe de points c'est à dire pour toutes les variables actives, ou illustratives, ou tous les individus actifs ou illustratifs ou centre de gravité des classes d'individus. Pour éviter le recouvrement des informations, les noms des points pourront être déplacés autour de leur symbole comme indiqué plus bas. Le mécanisme de déplacement des informations sera décrit plus loin.

Dans le cas d'un grand ensemble de données le recouvrement des points peut être insoluble en particulier près de l'origine. Il est alors possible de demander une sélection des points les plus représentatifs par l'option « sélection ». Ceci permettra d'améliorer la lisibilité du graphique tout en ne gardant que les points les plus significatifs. La sélection peut être effectuée suivant les critères de l'inertie ou de la corrélation du point au plan.

L'option « individus » permet des traitements spécifiques du nuage des points individus. Elle sera décrite plus loin.

L'option « Axes » permet de donner ou non un titre aux axes, mais également de modifier les bornes les axes. Tous les points sélectionnés sont représentés sur le plan.

L'option « Zoom » permet de ne visualiser qu'une certaine zone du plan et seuls les points de cette zone sont affichés et accessibles.

L'option « Valeurs propres » permet de déplacer les valeurs propres et les taux d'inertie d'un bout de l'axe à l'autre ou de modifier légèrement leur position. Mais les valeurs ne peuvent être changées.

Enfin, « Options » offre des options spécifiques aux analyses, par exemple : cercle de corrélation en analyse en composantes principales. Elle sera décrite plus loin.

Le graphique final peut être sauvegardé dans un catalogue SAS et/ou imprimé sur l'une des imprimantes reconnues par SAS.

Mais on peut aussi sauvegarder toutes les modifications déjà effectuées dans un tableau SAS, afin de pouvoir reprendre le graphique dans une session ultérieure.

4.3 MODIFICATION DES ATTRIBUTS

Le paragraphe suivant décrit comment le problème de recouvrement des points peut être résolu en déplaçant les noms des points (dits « labels ») autour de leur symbole.

Tout d'abord, dans une zone de haute densité de points, il peut être intéressant de faire un zoom sur cette région afin de voir clairement les points à déplacer. Une option est prévue à cet effet.

Une fenêtre « Zoom » s'ouvre alors (figure 3) dans laquelle l'utilisateur peu soit choisir un quadrant soit choisir une zone explicite du plan.

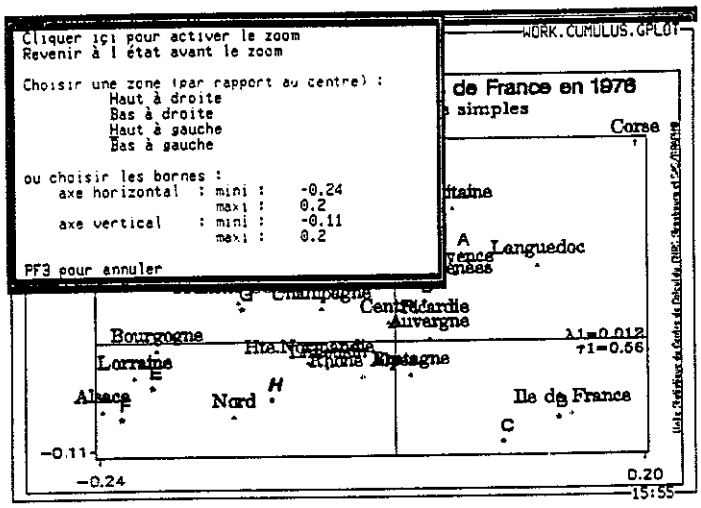


figure 3 : option Zoom

Seuls les points de cette zone seront représentés. En cliquant sur la zone « haut gauche » par exemple, on obtient le graphique (figure 4) qui distingue les points qui se recouvraient. On constate qu'il s'agit du quadrant haut-gauche de la figure 2

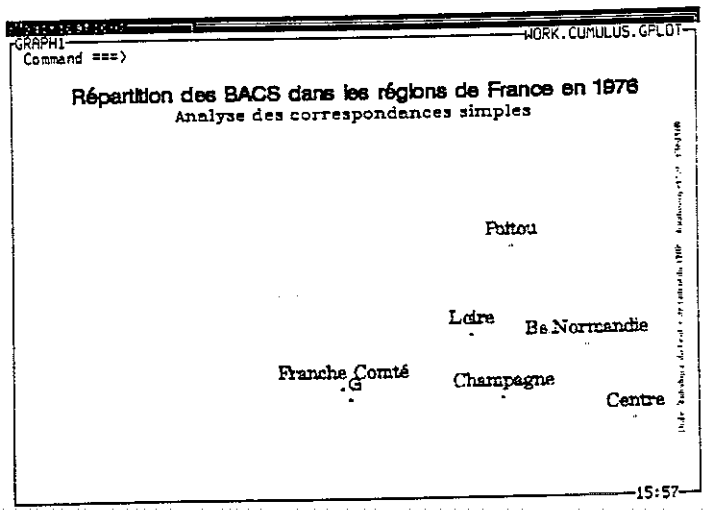


figure 4 : Zoom du quadrant haut gauche de la figure 1

À partir de là on pourra déplacer les points en activant l'option « label et symboles » du menu modifications (figure 2). Cette option ouvre une nouvelle fenêtre « label et symboles » (figure 5) à partie de laquelle on va pouvoir modifier les symboles ou les

libellés des points. Cette fenêtre est découpée en deux parties, une partie supérieure qui offre différentes options et une partie inférieure qui présente les attributs d'un point, attributs pouvant tous être modifiés. Le petit pavé en bas de la fenêtre décrit la position du label autour du point, et pour le déplacer il suffit de cliquer sur la position désirée. Quinze positions sont disponibles sachant que le symbole du point ne peut être déplacé et se trouve au centre.

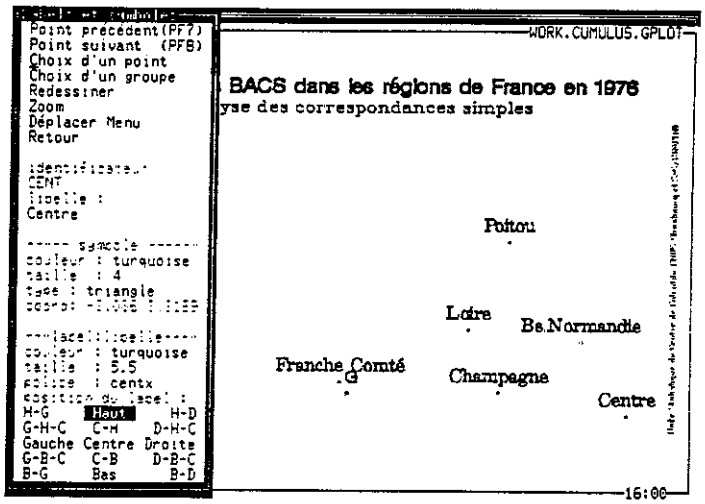


figure 5 : option "labels et symboles"

Dans la partie haute du menu, les trois premières options permettent de modifier les attributs de tout un groupe, lignes actives ou illustratives, variables actives ou illustratives, centres de gravité de classes s'il y a lieu.

L'option « choix d'un point » affiche la liste des points du graphique, filtrés en fonction de la sélection ou du zoom (figure 6). Il suffit alors de cliquer sur le point dont on désire modifier les attributs. L'option « Redessiner » permet à tout moment de voir l'effet des modifications demandées.

On arrive ainsi progressivement à résoudre tous les problèmes de recouvrement des labels jusqu'à obtenir le graphique final de la figure 2.

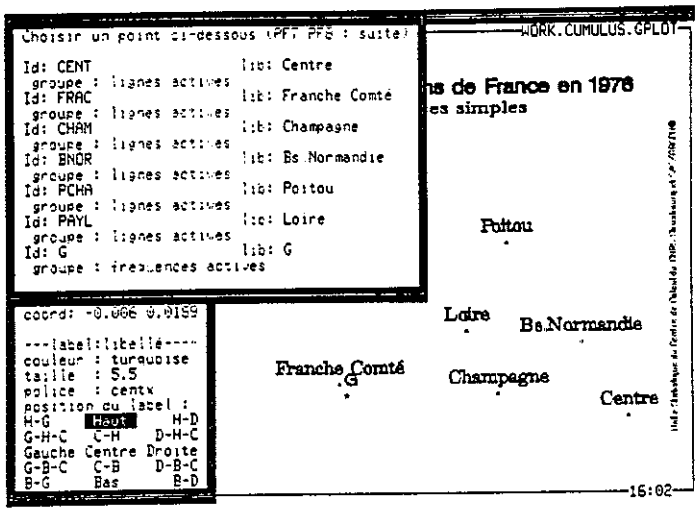


figure 6 : Choix du point à modifier

4.4 AUTRES OPTIONS

Cumulus dispose d'options spécifiques à chaque analyse. Pour une analyse en composante principale, par exemple, il est possible de tracer le cercle de corrélation (figure 7) ainsi que des traits joignant les variables à l'origine des axes. Dans le cas d'analyse des correspondances, les modalités des variables nominales peuvent être reliées par des segments.

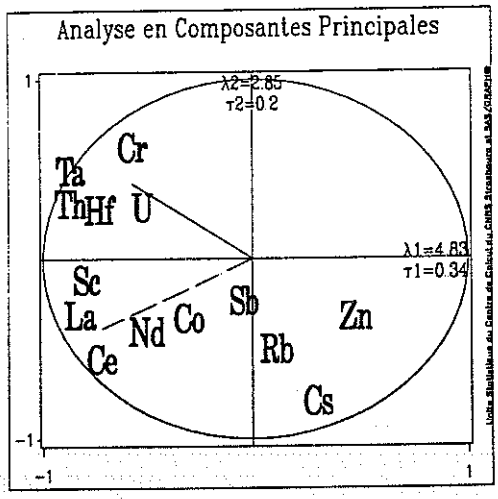


figure 7 : Cercle de corrélation

4.5 REPRESENTATION DES INDIVIDUS

Cumulus permet de représenter les centres de gravités des classes obtenues (ou des modalités) par des symboles différents et éventuellement dans une taille proportionnelle au poids de la classe (ou de la modalité). Le calcul de la taille du symbole de la classe tient compte du fait que les individus peuvent être eux même pondérés (figure 8). Cette figure représente le centre de gravité des classes obtenues après classification sur facteurs du tableau de données représenté figure 1. On peut également donner des noms à ces centres de gravité.

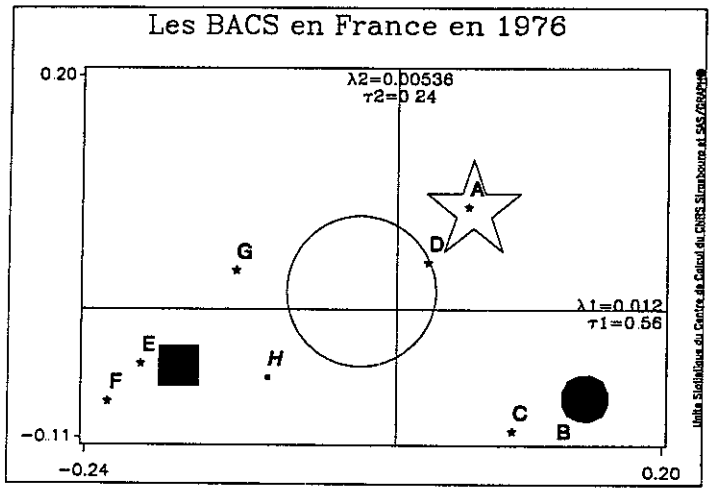


figure 8 : Nuage des individus représentés par les centres de gravité des classes

Mais les individus peuvent aussi être représentés par un symbole et par leur nom. De plus, le nuage des individus peut être éclaté en sous nuages de symboles différents suivant les modalités d'une variable nominale ou les résultats d'une classification (figure 9).

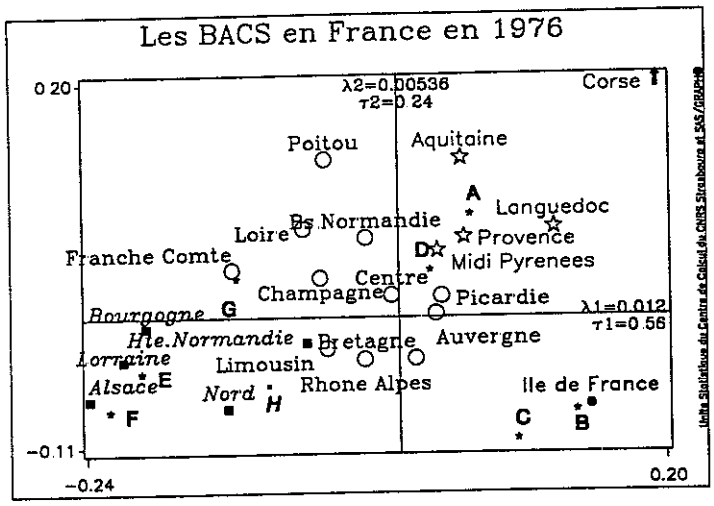


figure 9 : Nuage des individus représentés suivant leur classe d'appartenance

Cette figure correspond au même résultat de classification présenté figure 8, mais donne une autre représentation du nuage des individus. Chaque région de France a un symbole et une police de caractère qui correspond à sa classe d'appartenance.

4.6 CONCLUSION

Avec ***Cumulus*** il faut entre une demi-heure et une demi-journée suivant la complexité du problème pour obtenir le graphique d'un plan factoriel directement utilisable pour publication.

5.QUELQUES REFERENCES BIBLIOGRAPHIQUES

REFERENCES sur le graphique statistique

Becker R. A., Cleveland W. S., Wilks A. R. (1987) *Dynamic Graphics for Data Analysis*. Statistical Science, Vol 2 n° 4, pp. 355-395.

Cleveland W. S. (1993) *A Model for Studying Display Methods of Statistical Graphics*. Journal Comput. Graphical Stat., Vol 2 n° 4, pp. 323-343.

Goupil-Testu F. (1995) *Un Outil Graphique Interactif d'Aide à l'Interprétation de Résultats d'Analyse de Données*. Revue MODULAD, n° 15.

Grangé D., Ringenbach M. (1991) *Cumulus II ou le Nuage des Analyses Factorielles vu par SAS*. Club SAS, Cannes.

Weihs H. Schmidli (1990) *Online Multivariate Graphical Analysis: Routine Searching for Structure*. Statistical Science, Vol 5 n°2, pp. 175-226.

American Statistical Association (1992, 1993) *Proceedings of the Section on Statistical Graphics*.

Computing Science and Statistics (1992) *Graphics and Visualization*. Proceedings 24th Symposium on the Interfaces.

REFERENCES sur les méthodes factorielles

Saporta G., *Probabilités, analyse des données et statistique*. Editions Technip - Paris 1990

Tenenhaus M., *Méthodes statistiques en gestion*, Dunod Entreprise - Paris 1994

Lebart L., Morineau A., Piron M., *Statistique exploratoire multidimensionnelle, Dunod _ Paris 1995*

REFERENCES EYE-LID

Adam G., Bon F., Capdevielle J., Mouriaux R., (1970) - *L'ouvrier français en 1970*, Presses de la FNSP, Paris.

Bernard J.-M, Rouanet H. et Baldy R., (1993) - *« EyeLID-2, Manuel de référence et Guide de l'utilisateur »*, édité et diffusé par INDIA S.A., 22 rue de Douai, 75009 Paris.