UNE INTRODUCTION A L'ANALYSE FACTORIELLE DES CORRESPONDANCES AVEC *SPSS pour Windows*

Dominique Desbois

INRA-ESR Nancy et SCEES
4 avenue de Saint-Mandé, 75570 Paris Cedex 12.
Fax : +33 1 49 55 85 00 E-mail :desbois@jouy.inra.fr

Résumé. Cette note initie l'utilisateur débutant à la mise en oeuvre de l'Analyse Factorielle des Correspondances au moyen de la procédure ANACOR du logiciel *SPSS pour Windows*. Cette mise en oeuvre concerne l'analyse des tableaux de contingence à partir d'un exemple basé sur des données individuelles et d'un exemple basé sur des données agrégées. Le listage des résultats obtenus est commenté par la présentation du formulaire de l'analyse des correspondances associé à chacun des résultats obtenus.

MOTS-CLÉS : Analyse Factorielle des Correspondances, logiciel statistique, mise en oeuvre.

Abstract. This note introduces the beginner to the use of Correspondence Analysis by means of the ANACOR procedure from the *SPSS for Windows* software. This practical use concerns the analysis of contingency tables, stated from examples based upon either individual casewise data or aggregated data. The listing of results for each output is annotated with the main mathematical formulae of Correspondence Analysis.

KEY WORDS : Correspondence Analysis, Statistical Software, Introductory Use.

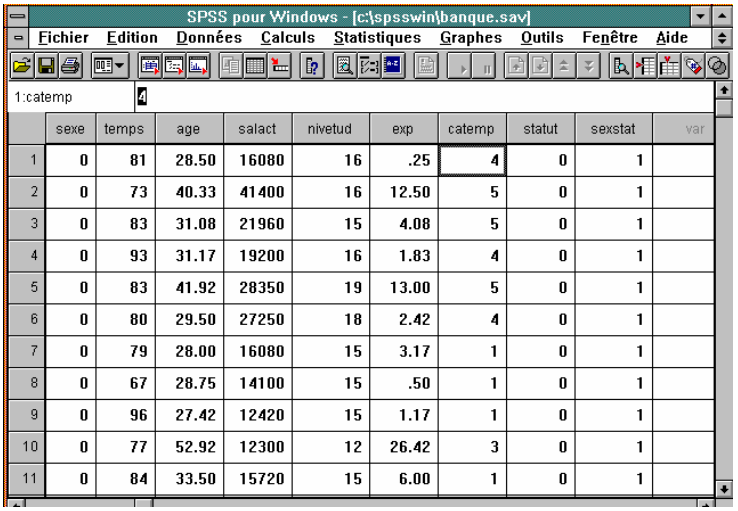
1. Introduction

Cette note a pour but d'aider les utilisateurs débutants de *SPSS pour Windows* dans la mise en oeuvre de l'**Analyse Factorielle des Correspondances**, méthode d'analyse multidimensionnelle des données statistiques. La procédure **ANACOR** d'Analyse Factorielle des Correspondances (**AFC**) de SPSS permet essentiellement d'analyser des tableaux de contingence. Un **tableau de contingence** est un tableaux à deux dimensions constitué par le croisement de deux variables qualitatives à catégories nominales (e.g. le sexe, statut matrimonial, ...) ou ordinales (e.g. le niveau d'études, la tranche de salaire, ...) dont les cases contiennent le comptage d'occurrences conjointes des caractères présents dans une population d'individus.

1. Un exemple d'analyse sur données individuelles

1.1 Les données

Les données sont constituées par un fichier du personnel fictif de salariés d'une société commerciale tout aussi imaginaire livré avec le logiciel pour servir de jeu d'essai (fichier SPSS `banque.sav`) . Ce fichier comporte des renseignements sur la fonction de ces salariés et leur statut. La fonction des salariés (variable `catemp`) est classée selon 7 catégories ou **modalités** : employé de bureau, employé stagiaire, agent de sécurité, rédacteur stagiaire, personnel vacataire, cadre stagiaire, personnel technique. On distingue également 4 statuts différents de salarié (variable `sexstat`) selon leur sexe (homme ou femme) et leur appartenance ethnique (majoritaire ou minoritaire).



	sexe	temps	age	salact	nivetud	exp	catemp	statut	sexstat	var
1	0	81	28.50	16080	16	.25	4	0	1	
2	0	73	40.33	41400	16	12.50	5	0	1	
3	0	83	31.08	21960	15	4.08	5	0	1	
4	0	93	31.17	19200	16	1.83	4	0	1	
5	0	83	41.92	28350	19	13.00	5	0	1	
6	0	80	29.50	27250	18	2.42	4	0	1	
7	0	79	28.00	16080	15	3.17	1	0	1	
8	0	67	28.75	14100	15	.50	1	0	1	
9	0	96	27.42	12420	15	1.17	1	0	1	
10	0	77	52.92	12300	12	26.42	3	0	1	
11	0	84	33.50	15720	15	6.00	1	0	1	

1.2 La spécification des paramètres de l'analyse

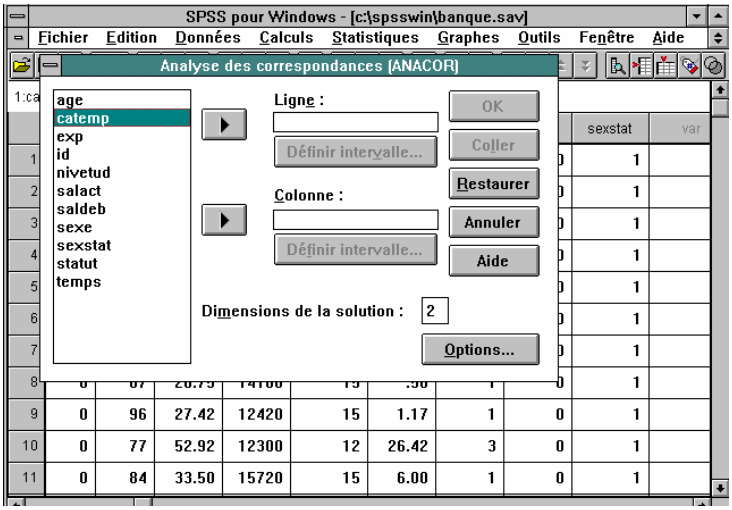
Définition des variables ligne et colonne du tableau de contingence

1.2.1 Afin d'afficher la boîte de dialogue principale de la procédure ANACOR, sélectionnez à partir du menu principal les options suivantes :

Statistiques

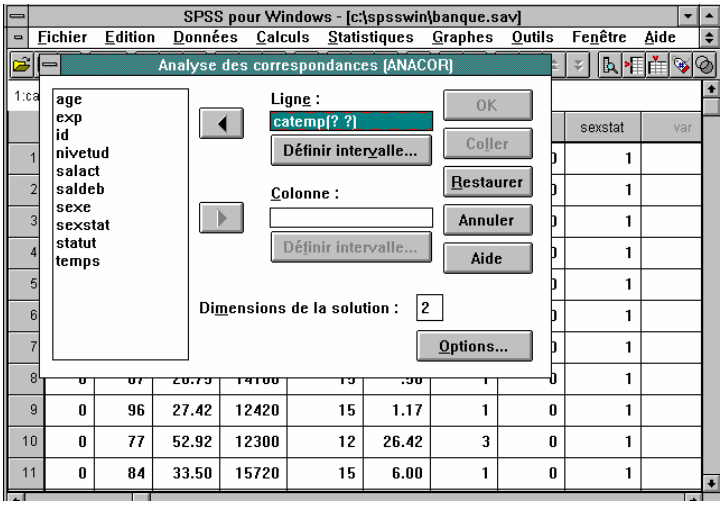
Factorisation

Analyse des Correspondances ...



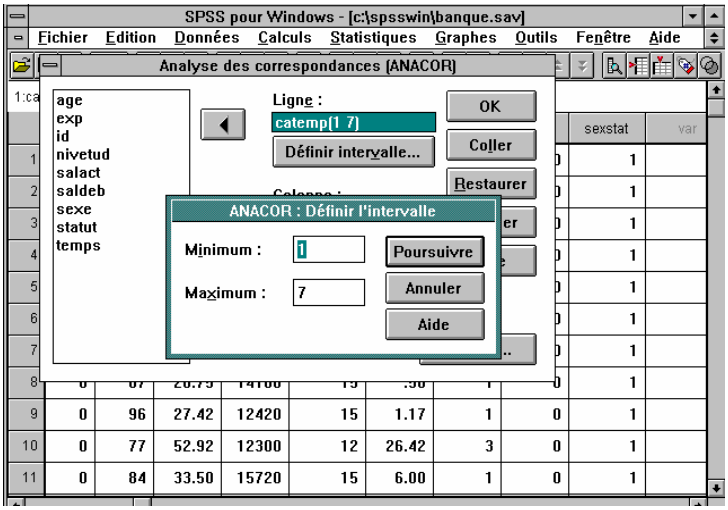
1.2.2 Sélectionnez la variable-ligne catemp dont les catégories constitueront les lignes du tableau de contingence ainsi que la plage de valeurs définissant les catégories soumises à l'analyse. Pour ce faire, il faut :

i. sélectionner la variable catemp dans la liste de variable et la transférer dans la sélection **Ligne** à l'aide du bouton de sélection;

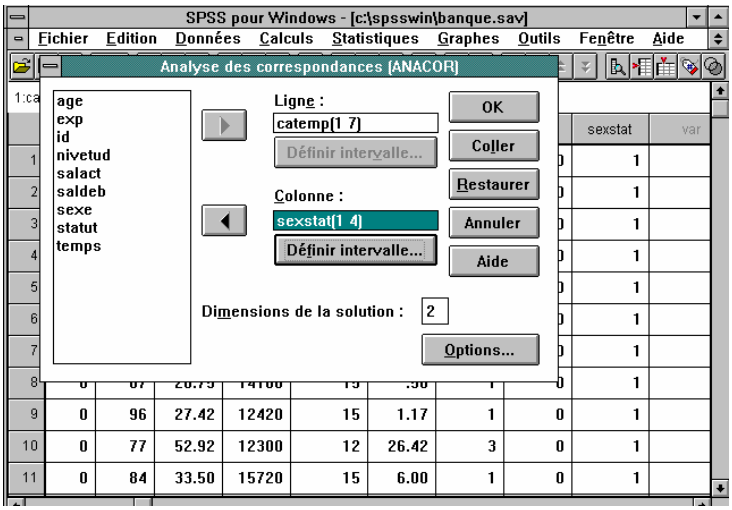


Définition de la plage des catégories

ii. spécifier la plage de valeurs définissant les catégories soumises à l'analyse en appelant la boîte de dialogue secondaire par l'intermédiaire du bouton **Définir intervalle ...** puis en donnant la valeur 1 pour le **Minimum** et la valeur 7 pour le **Maximum**.



1.2.3 Sélectionnez la variable-colonne sexstat dont les catégories constitueront les colonnes du tableau de contingence ainsi que la plage de valeurs définissant les catégories soumises à l'analyse en opérant de manière similaire pour la sélection **Colonne**.



Dimension de la solution

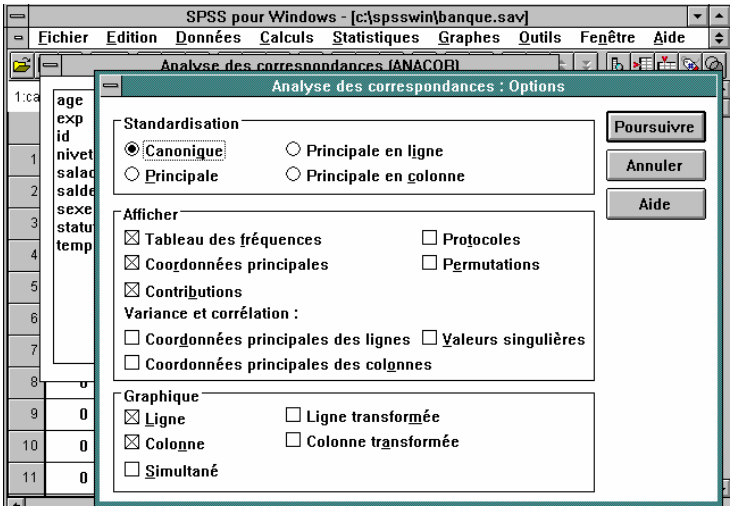
Vous pouvez spécifier le nombre de facteurs que vous voulez extraire. En AFC, la dimension de l'espace des facteurs est égale à {min(Card*I*,Card*J*)-1}, plus petite dimension du tableau (nombre minimum de catégories en ligne ou colonne) moins une unité[1]. Dans l'exemple ci-dessus, la variable-ligne sexstat comportant 4 catégories donne la plus petite dimension de tableau, la dimension de l'espace des facteurs est donc égale à 3. Dans cet exemple, vous ne pouvez donc extraire au maximum que 3 facteurs. Si vous spécifiez un nombre de facteurs à extraire supérieur à la dimension de l'espace des facteurs, c'est ce maximum qui sera retenu par la procédure. La valeur par défaut du paramètre est de 2 (on se limite à l'examen du plan factoriel $F_1 \times F_2$).

Options

[1] En raison de l'existence d'une valeur propre triviale correspondant au facteur reliant le barycentre du nuage à l'origine du repère orthonormé canonique de l'espace affine euclidien.

Le bouton Options affiche la boîte de dialogue secondaire permettant de lister le tableau de contingence soumis à l'analyse, de choisir la norme utilisée pour calculer les projections, d'imprimer les coordonnées factorielles des lignes et des colonnes, de sélectionner les graphiques factoriels que vous voulez visualiser et de consulter les indicateurs statistiques associés.

Voici la valeur des options définies par défaut dans la procédure ANACOR :



Standardisation

Vous pouvez spécifier la norme utilisée pour calculer les projections des lignes et des colonnes du tableau. Il faut choisir l'une des méthodes suivantes :

- **Canonique**. Pour chaque facteur, les lignes sont au pseudo-barycentre des colonnes (moyenne pondérée des colonnes standardisée par la valeur propre associée au facteur). On utilise cette norme pour étudier la relation entre les deux variables (écart au modèle d'indépendance). Il s'agit de l'option par défaut.
- **Principale**. Les distances entre points-lignes et entre point-colonnes approxime la distance du χ^2. Ce type de normalisation doit être utilisé si l'on souhaite étudier les différences entre modalités de l'une et/ou l'autre des variables au lieu d'examiner les différences entre les deux variables. Il s'agit de la normalisation standard utilisée par l'ensemble des présentations classiques de l'AFC.
- **Principale en ligne**. La distance entre points-lignes approxime la distance du χ^2. Cette norme maximise les distances entre points-lignes. On utilise cette norme si l'on souhaite étudier les différences ou les ressemblances entre les modalités de la variable-ligne.
- **Principale en colonne**. La distance entre points-colonnes approxime la distance du χ^2. Cette norme maximise les distances entre points-colonnes. On utilise cette norme si l'on souhaite étudier les différences ou les ressemblances entre les modalités de la variable-colonne.

Afficher

Vous pouvez spécifier une ou plusieurs options d'affichage :

- **Tableau des fréquences**. Affichage du tableau de contingence croisant les modalités-lignes et les modalités-colonnes. Cette présentation comporte le nombre d'occurrences dans chaque case ainsi que les marges en ligne et en colonne.
- **Coordonnées principales**. Coordonnées factorielles et proportions marginales pour chaque modalité-ligne et chaque modalité-colonne.
- **Contributions**. Contribution de chaque ligne et de chaque colonne à chacun des facteurs de l'analyse, ainsi que distance relative à l'origine sommée sur chacun des axes.
- **Protocoles**[2]. Profils lignes et colonnes.
- **Permutations**. Tableau de contingence avec permutation des lignes et des colonnes en fonction des projections des modalités-lignes et des modalités colonnes selon chacun des facteurs.

Variance et corrélation

Vous pouvez afficher l'une ou plusieurs des statistiques de variance et de corrélation proposées :

- **Coordonnées principales des lignes**. Variance et corrélation des coordonnées factorielles des points-lignes.
- **Coordonnées principales des colonnes**. Variance et corrélation des coordonnées factorielles des points colonnes.
- **Valeurs singulières**. Variance et corrélations des valeurs propres.

Graphique

Vous pouvez sélectionner un ou plusieurs des graphiques proposés :

- **Ligne**. graphique-plan des projections des points-lignes.
- **Colonne**. Graphique-plan des projections factorielles des points-colonnes.
- **Simultané**. Graphique-plan simultané des projections factorielles des points lignes et des points-colonnes.
- **Ligne transformée**. Édition pour chaque facteur des projections des points-colonnes.
- **Colonne transformée**. Édition pour chaque facteur des projections des points-colonnes.

Spécifications optionnelles du langage de commande

Vous pouvez adapter la procédure ANACOR à votre usage personnel en collant les paramètres de votre sélection effectuée par l'intermédiaire des boîtes de dialogue dans la fenêtre de syntaxe. Vous pourrez alors modifiez le langage de commande résultant de ces choix pour :

• choisir les plans factoriels à représenter graphiquement (utiliser le mot-clé NDIM de la sous-commande PLOT);

• spécifier le nombre de caractères des labels de valeur utilisés pour étiqueter les points dans les graphiques (avec la sous-commande PLOT).

---

[2] Il s'agit bien évidemment d'une coquille de traduction du mot "*profiles*" en anglais, qui a été pris dans un de ses sens premiers et non dans sa signification mathématique.

• sauvegarder la matrice de variance-covariance ou le tableau des coordonnées factorielles lignes et colonnes dans un fichier SPSS au format matriciel (avec la sous-commande MATRIX);

• spécifier une méthode de standardisation particulière (avec la sous-commande NORMALIZATION).

Pour une description complète de la commande ANACOR et des règles de syntaxe, consultez la section correspondante du manuel (*Syntax Reference, SPSS 6.1 Categories*).

2. Un exemple d'analyse sur données agrégées

Bien souvent le chargé d'études qui souhaite analyser un tableau de contingence publié dans un annuaire statistique ne dispose pas des questionnaires individuels qui ont permis d'élaborer ce tableau. mais simplement du comptage des occurrences généré par un tri croisé entre deux variables d'une enquête. L'utilisation du langage de commandes SPSS permet de lire directement les données agrégées du tableau de contingence puis de spécifier les paramètres de l'analyse grâce aux sous-commandes de la procédure ANACOR. Les exemples suivants permettent de s'initier aux différentes mises en oeuvre de l'AFC sur tableaux de données agrégées, dans SPSS pour Windows.

2.1 Les données

Les données agrégées sont constituées par des statistiques sur l'évolution du cheptel dans le Loiret de 1970 à 1990 (source : DDAF-Loiret 1991 d'après *Tomassone, Dervin & Masson 1993*). Au début de chaque décennie (1970, 1980, 1990), on a dénombré le bétail présent dans le Loiret selon les 5 catégories suivantes : Bovins non laitiers (BOV), Vaches (VAC), Ovins (OVI), Brebis mères (BRE), Porcins (POR). Les effectifs sont exprimés en milliers de têtes.

PRODUIT ANIMAL	SIGLE	ANNEE 1970	ANNEE 1980	ANNEE 1990
Bovins non laitiers	BOV	82	61	42
Vaches	VAC	48	41	27
Ovins	OVI	39	19	13
Brebis mères	BRE	47	36	20
Porcins	POR	39	26	25

2.2 La syntaxe des commandes

2.2.1 La spécification TABLE=ALL

La syntaxe la plus simple pour effectuer une AFC sur données agrégées consiste à lire le tableau de contingence en utilisant l'instruction DATA LIST puis analyser ce tableau en tant que données agrégées en utilisant le **mot-clé** ALL dans la sous-commande TABLE (cf. l'exemple suivant).

```
DATA LIST / AN1 TO AN3 1-9.
BEGIN DATA
 82 61 42
 48 41 27
 39 19 13
 47 36 20
 39 26 25
END DATA.
ANACOR TABLE=ALL(5,3).
```

Voici quelques règles qui vous permettront d'écrire correctement votre propre programme de commandes :

- Le mot-clé ALL de la sous-commande TABLE permet de lire et d'analyser directement le contenu des cases du tableau.
- Les colonnes du tableau en entrée doivent être spécifiées en tant que variable dans l'instruction DATA LIST. Par contre, il n'y pas spécification des lignes.
- Après le mot-clé ALL figurent, entre parenthèses, le nombre de lignes du tableau suivi par le nombre de colonnes et séparés par une virgule.
- Le nombre de lignes et colonnes spécifié peut être inférieur au nombre réel de lignes et de colonnes si l'on souhaite analyser seulement un sous-ensemble du tableau.
- Le mot-clé ALL de la sous-commande TABLE permet de lire et d'analyser directement le contenu des cases du tableau.
- Les variables (colonnes du tableau) sont traitées en tant que modalités-colonnes, et les enregistrements (lignes du tableau) en tant que modalités-lignes.
- Avec la spécification TABLE=ALL, les lignes ne peuvent être étiquetées. Si l'affichage des étiquettes dans les résultats est nécessaire, vous pouvez pour entrer vos données utiliser la méthode basée sur l'instruction WEIGHT (cf. infra).

2.2.2 La commande WEIGTH

Si l'on veut pouvoir désigner explicitement les lignes et les colonnes du tableau, l'instruction WEIGHT fournit une alternative commode pour lire le tableau des données agrégées. Ce mode de spécification convient tout particulièrement aux tableaux de petite dimension (cf. l'exemple suivant).

```
DATA LIST FREE / CHEPTEL ANNEE NOMBRE.
BEGIN DATA
1 1 82
1 2 61
1 3 42
2 1 48
2 2 41
2 3 27
3 1 39
3 2 19
3 3 13
4 1 47
4 2 36
4 3 20
5 1 39
5 2 26
5 3 25
END DATA.
WEIGHT BY NOMBRE.
ANACOR TABLE=CHEPTEL(1,5) BY ANNEE(1,3).
```

Voici les principales règles gouvernant la spécification de ce type de syntaxe :

- La commande WEIGHT pondère chaque enregistrement par la valeur de la variable NOMBRE considérant que l'on a recensé 82 milliers de têtes de bétail ayant les caractéristiques CHEPTEL=1 lors de l'année ANNEE=1, puis 61 milliers de têtes de bétail du type CHEPTEL=1 lors de l'année ANNEE=2, et ainsi de suite jusqu'au type CHEPTEL=5 qui compte 25 milliers de têtes de bétail en l'année ANNEE=3.
- Si l'une des cases du tableau s'avère nulle, alors l'exécution de la commande WEIGHT produira un message d'avertissement mais les résultats fournis par la procédure ANACOR seront corrects.
- Les cases du tableau ne peuvent pas contenir de valeurs négatives. Les valeurs négatives ou les valeurs manquantes du système sont mises à zéro lors de l'exécution de la commande WEIGHT.
- Pour de grands tableaux de données agrégées, il vaut mieux utiliser le mode de spécification TABLE=ALL ou bien le langage de commande pour lire le tableau.

2.2.3 L'étiquetage des modalités lignes et colonnes

Les instructions suivantes montrent comment on peut compléter progressivement la spécification de la procédure ANACOR :

```
DATA LIST FREE / CHEPTEL ANNEE NOMBRE.
WEIGHT BY NOMBRE.
VALUE LABELS
CHEPTEL 1 'bov' 2 'vac' 3 'ovi' 4 'bre' 5 'por'/
ANNEE   1 'AN70' 2 'AN80' 3 'AN90'.
BEGIN DATA
1 1 82 1 2 61 1 3 42
2 1 48 2 2 41 2 3 27
3 1 39 3 2 19 3 3 13
4 1 47 4 2 36 4 3 20
5 1 39 5 2 26 5 3 25
END DATA.
ANACOR TABLE=CHEPTEL(1,5) BY ANNEE(1,3)
 / DIMENSION=2
 / NORMALIZATION=PRINCIPAL.
```

- Le format libre (mot-clé FREE de la commande DATA LIST) permet également d'entrer les données sous forme de matrice (chaque enregistrement correspond à une ligne du tableau).
- La commande VALUE LABELS permet d'affecter une étiquette à caractère descriptif à chaque valeur associée à une modalité ligne ou colonne. Cette étiquette de valeur étant reproduite dans les graphiques, il est préférable de prévoir des

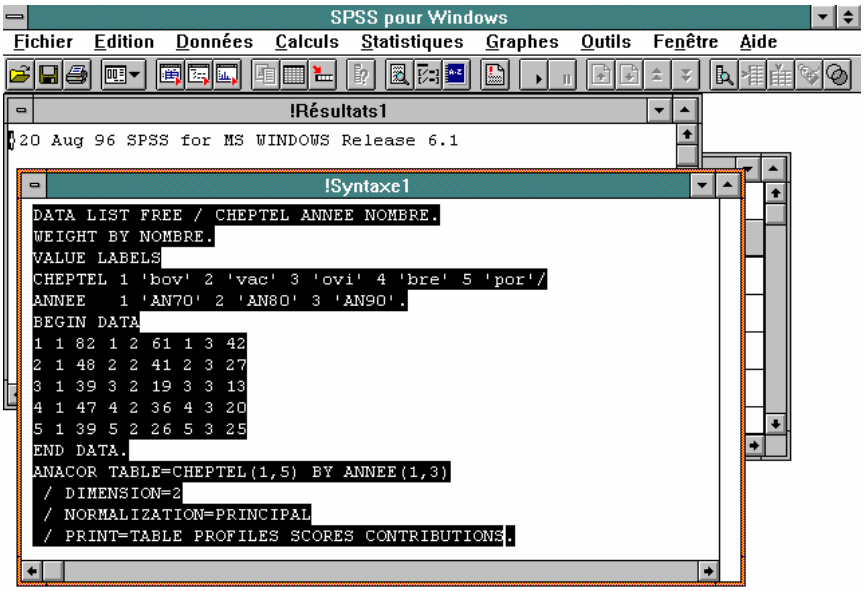
étiquettes ne dépassant pas 3 ou 4 caractères. Une pratique courante est de distinguer les projections des modalités lignes et colonnes pour les graphiques conjoints en réservant, par exemple, les minuscules pour les étiquettes des modalités-lignes et les majuscules pour celles des modalités-colonnes.

• La spécification (=2) de la sous-commande DIMENSION permet de limiter l'extraction des facteurs aux deux premiers axes factoriels et l'impression des graphiques au plan factoriel $F_1 \times F_2$.

• La spécification (=PRINCIPAL) de la sous-commande NORMALIZATION indique que les normes choisies pour les distances entre modalités-lignes et entre modalités colonnes approxime la métrique du χ^2 .

2.3 L'exécution du programme

Pour exécuter la procédure ANACOR selon les options spécifiées en mode commande, il suffit de saisir le texte du programme de commandes correspondant dans la fenêtre de syntaxe, de le sélectionner et d'en lancer l'exécution au moyen du bouton prévu à cet effet :



2.4 Listage des résultats et formulaire d'AFC

2.4.1 La procédure ANACOR

La bannière du programme nous informe qu'il s'agit de la version 0.4 de la procédure SPSS ANACOR conçue et réalisée par le Département d'Analyse des Données de l'Université de Leiden, aux Pays-Bas. Pour l'essentiel, la présentation des résultats ci-après est inspirée du formalisme utilisé par Jean-Paul Benzécri.

```
                    A N A C O R - VERSION 0.4
                              BY
                    DEPARTMENT OF DATA THEORY
                UNIVERSITY OF LEIDEN, THE NETHERLANDS
```

2.4.2 Le tableau de contingence

Le premier résultat concerne le tableau de contingence lui-même, imprimé par défaut. Par la suite, on notera ***I*** l'ensemble des lignes et ***J*** l'ensemble des colonnes de ce tableau. L'impression du tableau de contingence permet de valider le résultat de la lecture des données en vérifiant le contenu des cases. On peut également consulter les marges de ce tableau pour prendre connaissance des **distributions marginales ligne** k_I et **colonne** k_J.

```
The table to be analyzed:

                    1        2        3
                  AN70     AN80     AN90   Margin

  1 bov             82       61       42      185
  2 vac             48       41       27      116
  3 ovi             39       19       13       71
  4 bre             47       36       20      103
  5 por             39       26       25       90
              -------- -------- -------- --------
    Margin         255      183      127      565
```

La marge-ligne k_I est définie par le poids $\boldsymbol{k_{i.}}$ de la ligne i : $k_{i.} = \sum_{i \in I} k_{ij}$, soit pour la première ligne du tableau (`bov`) $k_{1.} = 185$. La marge-colonne k_J est définie par le poids $\boldsymbol{k_{.j}}$ de la colonne j : $k_{.j} = \sum_{j \in J} k_{ij}$, soit pour la première colonne du tableau (année 1970) $k_{1} = 255$. La marge-colonne représente l'effectif global du cheptel dans le Loiret : celui-ci atteignait 255 000 têtes de bétail en 1970.

Le poids total du tableau de contingence $\boldsymbol{k_{..}}$ est défini par la sommation selon les lignes ou les colonnes des marges du tableau : $k_{..} = \sum_{i \in I} k_{i.} = \sum_{j \in J} k_{.j} = \sum_{(i,j) \in IxJ} k_{ij}$

Ces éléments permettent de définir la distribution de fréquences $f_{IJ} = \{ f_{ij}, (i, j) \in I \times J \}$, avec $f_{ij} = \frac{k_{ij}}{k_{..}}$, comme une loi de probabilité sur le couple $I \times J$.

2.4.3 Les tableaux de profils

Les tableaux de profils lignes et colonnes sont des statistiques optionnelles de la procédure qui peuvent être obtenues soit en cochant la sélection PROTOCOLES dans la liste AFFICHER de la boîte de dialogue secondaire OPTIONS de la procédure, soit en spécifiant le mot-clé PROFILES dans la sous-commande PRINT.

Le **profil-ligne** $f_J^{\,i} = \{ f_j^i, \ j \in J \}$, profil de l'élément *i* sur l'ensemble *J*, est la loi conditionnelle de l'ensemble *J* connaissant l'événement *i* de l'ensemble *I*. Le profil-ligne des bovins non laitiers ($i = 1$) est {`0,443;0,330;0,227`}.

```
The Rowprofiles:

                  1        2        3
                 AN70     AN80     AN90   Margin

  1 bov           .443     .330     .227    1.000
  2 vac           .414     .353     .233    1.000
  3 ovi           .549     .268     .183    1.000
  4 bre           .456     .350     .194    1.000
  5 por           .433     .289     .278    1.000
              -------- -------- --------
    Margin        .451     .324     .225
```

Son j^{e} élément est défini par le rapport :

$$f_j^i = \frac{k_{ij}}{k_{i.}}$$

Ce rapport est égal au pourcentage que représente le poids de la case *(i,j)* du tableau de contingence par rapport au poids de la ligne *i*; il donne une estimation de la fréquence conditionnelle de l'événement $i \cap j$ sachant l'événement *i*.

Le profil-ligne moyen $f_J = \{ f_{.j},\ j \in J \}$ (marge du tableau des profils-lignes) est la loi marginale sur *J* de la distribution de fréquences f_{IJ}, défini par sa j^{e} coordonnée : $f_{.j} = \frac{k_{.j}}{k_{..}}$, fréquence marginale de la colonne *j*. Le profil-ligne moyen f_J donne la pondération définie par la fréquence marginale de la colonne *j*, $f_{.j}$ masse affectée à chaque profil-colonne.

La définition du **nuage** $N_J(I) = \{ i \in I;\ (f_j^i,\ f_{i.}),\ j \in J \} \subset R_J$, des éléments *i* affectés des masses $f_{i.}$, permet d'associer à chaque profil-ligne une représentation euclidienne en termes de **point-ligne**.

Le profil-ligne moyen f_J, utilisé comme vecteur de pondération, permet de définir la **métrique du χ^2 de centre** f_J comme distance distributionnelle entre les points-lignes *i* et *i'* :

$$d(i,i') = \left\| f_J^i - f_J^{i'} \right\|_{f_J} = \sum_{j \in J} \left[f_j^i - f_j^{i'} \right]^2 \Big/ f_{.j}$$

Le **profil-colonne** $f_I^j = \{ f_i^j,\ i \in I \}$, profil de l'élément *j* sur l'ensemble *I*, est la loi conditionnelle de l'ensemble *I* connaissant l'événement *j* de l'ensemble *J*. Le profil-colonne de l'année 1970 ($j = 1$) est {0,322;0,188;0,153;0,184;0,153}.

```
 The Columnprofiles:

                  1        2        3
                 AN70     AN80     AN90   Margin

  1 bov           .322     .333     .331     .327
  2 vac           .188     .224     .213     .205
  3 ovi           .153     .104     .102     .126
  4 bre           .184     .197     .157     .182
  5 por           .153     .142     .197     .159
              -------- -------- --------
    Margin      1.000    1.000    1.000
```

Son i^{e} élément est défini par le rapport :

$$f_i^{\ j} = \frac{k_{ij}}{k_{.j}}$$

Ce rapport est égal au pourcentage que représente le poids de la case *(i,j)* du tableau de contingence par rapport au poids de la colonne *j*. Il donne une estimation de la fréquence conditionnelle de l'événement $i \cap j$ sachant l'événement *j*. Ainsi, la comparaison des trois profils-

colonnes correspondant aux années nous informe que les vaches (vac, $i = 2$) représentaient successivement dans le Loiret 18,8% du cheptel en 1970 ($f_2^1 = 0{,}188$), 22,4% du cheptel en 1980 ($f_2^2 = 0{,}224$), puis 21,3% ($f_2^3 = 0{,}213$) du cheptel en 1990.

Le profil-colonne moyen $f_I = \{ f_{i.},\ i \in I \}$ (marge du tableau des profils-colonnes) est la loi marginale sur I de la distribution de fréquences f_{IJ} , défini par sa i^{e} coordonnée : $f_{i.} = \frac{k_{i.}}{k..}$, fréquence marginale de la ligne *i*. Le profil-colonne moyen f_I donne la pondération définie par la fréquence marginale de la ligne *i*, $f_{i.}$ masse affectée à chaque profil-ligne.

La définition du **nuage** $N_I(J) = \{ j \in J\ ;\ (f_i^j,\ f_{.j}),\ i \in I \} \subset R_I$, des éléments *j* affectés des masses $f_{.j}$, permet d'associer à chaque profil-colonne une représentation euclidienne en termes de **point-colonne**.

Le profil-colonne moyen f_I, comme vecteur de pondération, permet de définir la **métrique du χ^2 de centre** f_I comme distance distributionnelle entre les points-colonnes *j* et *j'* :

$$d(j,j') = \left\| f_I^{\,j} - f_I^{\,j'} \right\|_{f_I} = \sum_{i \in I} \left[f_i^{\,j} - f_i^{\,j'} \right]^2 \Big/ f_{i.}$$

2.4.4 Les valeurs propres et leur pourcentage d'inertie

Le tableau des **valeurs propres** λ_α, où α indique le **rang** du facteur est le second résultat imprimé par défaut dans la procédure ANACOR.

Dimension	Singular Value	Inertia	Proportion Explained	Cumulative Proportion
1	.08061	.00650	.646	.646
2	.05969	.00356	.354	1.000
Total		.01006	1.000	1.000

Dans cet exemple, la plus petite dimension du tableau de contingence étant 3 (nombre d'années), le nombre maximum de facteurs non triviaux pouvant être extraits en AFC est de 2. Ce tableau de résultats donne pour chaque facteur de rang α la valeur propre correspondante λ_α (colonne "`Singular Value`", ainsi on a la première valeur propre $\lambda_1 = 0{,}08061$), l'inertie relative à ce facteur (colonne "`Inertia`"), le pourcentages d'inertie expliquée par chaque facteur (colonne "`Proportion explained`") ainsi que le pourcentage d'inertie cumulée au rang α (colonne "`Cumulative Proportion`").

L'interprétation des axes

Puis sont imprimés également par défaut pour chaque ensemble de modalités (lignes et colonnes), le **tableau des facteurs** ("`Row Scores`", respectivement "`Column Scores`"), celui des **contributions aux facteurs** ("`Contribution of Row points to the inertia of each dimension`", respectivement "`Contribution of column points to the inertia of each dimension`") et celui des **cosinus**

carrés ("Contribution of dimensions to the inertia of each row point", respectivement "Contribution of dimensions to the inertia of each column point").

2.4.5 Les facteurs

Pour les lignes, nous pouvons consulter le **tableau des facteurs** sur *I*, $F_\alpha = \{F_\alpha(i),\ i \in I\}$.

```
Row Scores:

CHEPTEL       Marginal    Dim
               Profile       1        2

   1 bov          .327    -.015     .006
   2 vac          .205    -.069     .037
   3 ovi          .126     .193    -.040
   4 bre          .182     .027     .074
   5 por          .159    -.063    -.114
```

$F_\alpha(i)$ est la valeur du facteur de rang α au point *i* de l'ensemble *I*, muni du système de masses f_I. Calculons la projection sur le premier facteur du barycentre des points-lignes :

$$f_{1.} \times F_1(1) + f_{2.} \times F_1(2) + f_{3.} \times F_1(3) + f_{4.} \times F_1(4) + f_{5.} \times F_1(5) =$$
$$0,327 \times (-0,015) + 0,205 \times (-0,069) + 0,126 \times 0,193 + 0,182 \times 0,027 + 0,159 \times (-0,063) \approx 0$$

On peut vérifier les propriétés suivantes des facteurs F_α :

- $\sum_{i \in I} f_{i.} \times F_\alpha(i) = 0$ Les axes factoriels sont centrés relativement à f_I la loi marginale sur *I*.
- $\sum_{i \in I} f_{i.} \times F_\alpha^{\,2}(i) = \lambda_\alpha$ L'inertie du nuage de points $N_J(I)$, muni du système de masses f_I, relative à chacun des axes α est égale à la valeur propre associée λ_α.
- $\sum_{i \in I} f_{i.} \times F_\alpha(i) \times F_\beta(i) = 0$ pour $\alpha \neq \beta$. Les axes factoriels sont othogonaux deux à deux.

Une normalisation des facteurs F_α par la constante $\lambda_\alpha^{1/2}$ permet de définir des fonctions $\varphi_\alpha^{\,I}$ de moyenne nulle, de variance unité, non corrélées deux à deux sur *I* muni du système de masses f_I.

Un tableau similaire $G_\alpha = \{G_\alpha(j),\ j \in J\}$ est imprimé pour l'ensemble *J* des points-colonnes.

```
Column Scores:

ANNEE         Marginal    Dim
               Profile       1        2

   1 AN70         .451     .087    -.015
   2 AN80         .324    -.051     .078
   3 AN90         .225    -.100    -.082
```

$G_\alpha(j)$ est la valeur du facteur de rang α au point *j* de l'ensemble *I*, muni du système de masses f_J. Calculons la projection sur le premier facteur du barycentre des points-colonnes :

$$f_{.1} \times G_1(1) + f_{.2} \times G_1(2) + f_{.3} \times G_1(3) = 0,457 \times$$

On peut vérifier sur ce tableau les propriétés suivantes des facteurs G_α :

- $\sum_{j\in J} f_{.j} \times G_\alpha(j) = 0$ Les axes factoriels sont centrés relativement à f_J la loi marginale sur *J*.
- $\sum_{j\in J} f_{.j} \times G_\alpha{}^2(j) = \lambda_\alpha$ L'inertie du nuage de points $N_I(J)$, muni du système de masses f_J, relative à chacun des axes α est égale à la valeur propre associée λ_α.
- $\sum_{j\in J} f_{.j} \times G_\alpha(j) \times G_\beta(j) = 0$ pour $\alpha \neq \beta$. Les axes factoriels sont othogonaux deux à deux.

Une normalisation des facteurs G_α par la constante $\lambda_\alpha{}^{1/2}$ permet de définir des fonctions $\psi_\alpha{}^J$ de moyenne nulle, de variance unité, non corrélées deux à deux sur *J* muni du système de masses f_J.

2.4.6 Les contributions aux facteurs

L'interprétation des axes de l'AFC ne peut être effectuée sur la base des seules coordonnées factorielles. Les **contributions des points à l'inertie des facteurs** constitue l'indice statistique permettant d'apprécier la part que les modalités correspondantes ont prise dans la constitution d'un axe.

Imprimé par la procédure ANACOR, le tableau suivant rassemble les **contributions relatives des points-lignes *i* à l'inertie des facteurs de rang α**, $CTR_\alpha^I = \{CTR_\alpha(i),\ i \in I\}$ muni du système de masses f_I, profil-moyen de ces modalités-lignes toutes années confondues. Cette contribution est définie par le rapport $CTR_\alpha(i) = \dfrac{f_{i.} \times F_\alpha^2(i)}{\lambda_\alpha}$

```
Contribution of row points to the inertia of each dimension:

CHEPTEL       Marginal    Dim
               Profile       1        2

   1 bov          .327     .012     .004
   2 vac          .205     .150     .077
   3 ovi          .126     .719     .056
   4 bre          .182     .021     .282
   5 por          .159     .098     .581
                       -------- --------
                          1.000    1.000
```

Rappelons que chaque valeur propre représente l'inertie du facteur associé : $\lambda_\alpha = \sum_{i\in I} f_{i.} \times F_\alpha^2(i)$

De façon symétrique, le tableau $CTR_\alpha^J = \{CTR_\alpha(j),\ j \in J\}$ rassemble les **contributions relatives des points-colonnes *j* à l'inertie des facteurs de rang α**, muni

du système de masses f_J, profil-moyen de ces modalités-lignes toutes années confondues. Cette contribution est définie par le rapport $CTR_\alpha(j)= \frac{f_{.j} \times F_\alpha^2(j)}{\lambda_\alpha}$

```
Contribution of column points to the inertia of each dimension:

ANNEE          Marginal    Dim
                Profile       1        2

   1 AN70          .451     .521     .027
   2 AN80          .324     .130     .546
   3 AN90          .225     .349     .426
                        -------- --------
                           1.000    1.000
```

Rappelons que chaque valeur propre représente l'inertie du facteur associé : $\lambda_\alpha = \sum_{j \in J} f_{.j} \times G_\alpha^2(j)$

2.4.7 Les corrélations profils-facteurs

L'interprétation des représentations graphiques fournies par l'AFC dépend également de la qualité des projections des profils lignes ou colonnes sur les facteurs. L'indice permettant de s'assurer de la qualité de ces projections est le **cosinus carré** de l'angle que fait le vecteur représentant le profil avec les vecteurs unitaires des axes factoriels. Ce cosinus carré est assimilable à une corrélation entre le profil et le facteur.

Dans le listage fourni par la procédure ANACOR, le tableau des **contributions des facteurs à l'inertie des points-lignes** rassemble les corrélations $COR_\alpha^I = \{COR_\alpha(i),\ i \in I\}$ muni du système de masses f_I, profil-moyen de ces modalités-lignes toutes années confondues. Cette corrélation est définie comme le cosinus carré de l'angle de projection du profil-ligne f_J^i, représenté par le point-ligne *i*, sur l'axe factoriel de rang α par le rapport $COR_\alpha(i)= \frac{F_\alpha^2(i)}{\rho^2(i)}$ où $\rho^2(i)$ est le carré de la distance d'un point ligne *i* au centre de gravité du nuage $N_J(I)$.

$\rho^2(i)= \left\| f_J^i - f_J \right\|_{f_J} = \sum_{j \in J} \left[f_j^i - f_{.j} \right]^2 / f_{.j} = \sum_{\alpha \in A} F_\alpha^2(i)$

```
Contribution of dimensions to the inertia of each row point:

CHEPTEL        Marginal    Dim                 Total
                Profile       1        2

   1 bov           .327     .846     .154    1.000
   2 vac           .205     .780     .220    1.000
   3 ovi           .126     .959     .041    1.000
   4 bre           .182     .119     .881    1.000
   5 por           .159     .235     .765    1.000
```

De façon symétrique, le tableau des **contributions des facteurs à l'inertie des points-colonnes** rassemble les corrélations $COR_\alpha^{\ J} = \{COR_\alpha(j),\ j \in J\}$ muni du système

de masses f_I, profil-moyen de ces modalités-lignes toutes années confondues. Cette corrélation est définie comme le cosinus carré de l'angle de projection du profil-colonne f_I^j, représenté par le point-colonne *j*, sur l'axe factoriel de rang α par le rapport $COR_\alpha(j) = \frac{F_\alpha^2(j)}{\rho^2(j)}$ où $\rho^2(j)$ est le carré de la distance du point-colonne *j* au centre de gravité du nuage $N_I(J)$.

$$\rho^2(j) = \left\| f_I^j - f_I \right\|_{f_I} = \sum_{i \in I} \left[f_i^j - f_{i.} \right]^2 \Big/ f_{i.} = \sum_{\alpha \in A} G_\alpha^2(j)$$

```
Contribution of dimensions to the inertia of each column point:

ANNEE          Marginal    Dim              Total
                Profile       1        2

   1 AN70          .451     .972     .028    1.000
   2 AN80          .324     .302     .698    1.000
   3 AN90          .225     .599     .401    1.000
```

2.4.8 Les vecteurs propres

Si l'on souhaite obtenir les coordonnées des vecteurs propres dans la base canonique de R_J (espace des profils-lignes), il convient alors de choisir la méthode de standardisation "Principale en ligne" (NORMALISATION=RPRINCIPAL), comme suit :

```
ANACOR TABLE=CHEPTEL(1,5) BY ANNEE(1,3)
 / DIMENSION=2
 / NORMALIZATION=RPRINCIPAL.
```

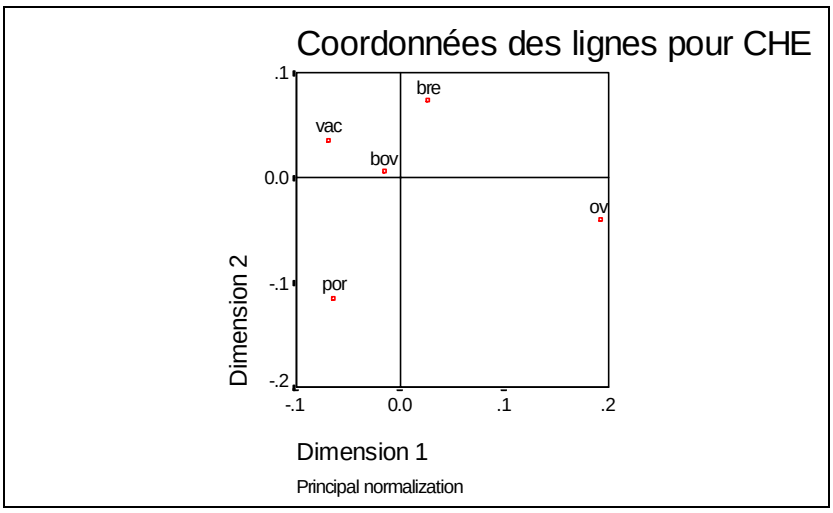
puis de consulter le tableau des facteurs des points-colonnes dont les colonnes DIM1 et DIM2 nous donnent directement dans la base canonique de R_J les coordonnées des deux vecteurs propres V_1, respectivement V_2 :

```
Column Scores:

ANNEE          Marginal    Dim
                Profile       1        2

   1 AN70          .451    1.075    -.246
   2 AN80          .324    -.633    1.299
   3 AN90          .225   -1.246   -1.377
```

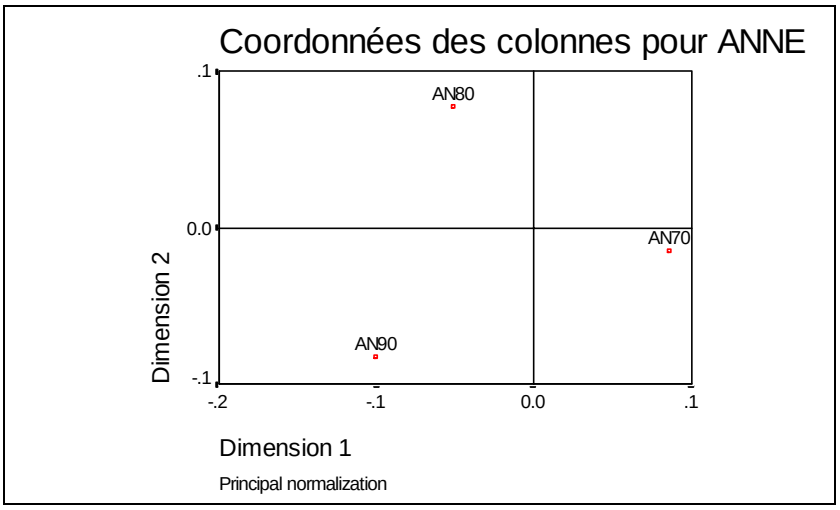
On peut de façon symétrique dans l'espace des profils-colonnes obtenir les coordonnées des vecteurs propres dans la base canonique de R_I en utilisant la méthode de standardisation "Principale en colonne" (NORMALISATION= CPRINCIPAL) et en consultant le tableau des facteurs pour les points-lignes.

2.4.9 Les représentations graphiques

Les facteurs sur l'ensemble *I* ou l'ensemble *J*, utilisés comme **coordonnées factorielles** permettent de dresser des représentations graphiques des modalités lignes, respectivement colonnes afin de visualiser la distribution des profils lignes, respectivement colonnes. La procédure ANACOR produit des graphiques factoriels par défaut pour chaque ensemble de profils lignes et colonnes dans le plan factoriel $F_1 \times F_2$.
Le graphique ci-dessous visualise les projections dans le plan factoriel $F_1 \times F_2$ des profils-lignes associés aux modalités de la variable CHEPTEL.



Le graphique ci-dessous visualise les projections dans le plan factoriel $F_1 \times F_2$ des profils-colonnes associés aux modalités de la variable ANNEE.



Soulignons que pour la méthode de normalisation utilisée (PRINCIPAL), il n'est pas possible de demander à la procédure ANACOR le tracé conjoint de l'ensemble des points-lignes et de l'ensemble des points-colonnes sur un même graphique factoriel.

3. En guise de conclusion

En tant que méthodologie pour l'analyse de données statistiques multidimensionnelles, l'AFC est susceptible de bien d'autres applications que l'analyse d'un tableau de contingence. Parmi celles-ci, citons le dépouillement multidimensionnel des réponses à un questionnaire d'enquête, l'analyse lexico-statistique de réponses à des questions ouvertes, l'analyse économique de tableaux d'entrées-sorties à trois indices, l'analyse des tableaux de distance, la discrimination de sous-populations sur tableau de descripteurs cliniques, etc. Chacune de ces applications fait référence, explicitement ou implicitement, à un modèle théorique : AFC d'un codage disjonctif complet, AFC du tableau de Burt ou Analyse des Correspondances multiples (**ACM**), Analyse factorielle multiple (**AFM**), Analyse intra ou ACM conditionnelle, Analyse factorielle de tableaux de distance, etc. À l'instar de la pratique courante de projection d'éléments supplémentaires, ces modèles théoriques ne sont pas incorporés explicitement à l'interface d'utilisation de SPSS pour Windows; il sera donc nécessaire de diffuser ultérieurement d'autres notices d'utilisation proposant une introduction simple et un accès facilité à ces techniques statistiques dans le contexte de ce logiciel. Nous espérons bien entendu pouvoir répondre prochainement à de tels besoins.

4. Références

Tomassone R., Dervin C., Masson J.-P. 1993. *BIOMÉTRIE, Modélisation des phénomènes biologiques*, Masson, Paris, 553 p.

SPSS Inc. 1994. *SPSS 6.1 Categories*, SPSS Inc., Chicago, 209 p.