philippe.longchamp@hesav.ch

La notion de champ et l'analyse des correspondances multiples (ACM)

Programme Doctoral Romand en Sociologie

Séminaire résidentiel méthodologique, 22-23 septembre 2011, Montreux

Organisé par Félix Bühlmann (UNIL – FORS)

Animation : Brigitte Le Roux (brigitte.leroux@mi.parisdescartes.fr), Philippe Bonnet (philippe.bonnet@parisdescartes.fr) et Frédéric Lebaron (flebaron@yahoo.fr)

Sommaire

Présentation de Brigitte Leroux

I. Introduction

1. Généralités sur l'analyse géométrique des données (AGD)

1.1. Trois idées clés

Première idée :

L'AGD peut être présentée comme une analyse en soi, et non comme un avatar de l'ACM.

Années 1960, Benzécri crée une nouvelle approche statistique. Notamment pour traiter des données linguistiques. On l'a appelé AGD. Les gens peuvent l'appeler « French Data Analysis », très mal perçu. Le terme d'AGD est mieux adapté, adopté en 1996.

Pourquoi pas « Data Analysis » ? Car cela renvoie à des méthodes spécifiques aux Etats-Unis.

Exemple : à partir d'un tableau à deux entrées (individus et variables), comment construire deux nuages, celui des catégories (ou modalités de réponses) et un nuage représentant les individus ? C'est l'essence même de la méthode : à partir de données, on créer des nuages. Ensuite, il s'agit d'interpréter.

Deuxième idée :

Il s'agit d'une approche formelle. Il faut rechercher les valeurs propres d'un certain isomorphisme. Permet de mettre toutes les méthodes sous la même étiquette.

Troisième idée : La description d'abord

Le modèle doit suivre les données, et non l'inverse. On a des données, et on crée le modèle à partir des données. Ce qui ne veut pas dire qu'il n'y a pas d'hypothèses.

On est en statistique descriptive, et non inductive. Ce qui ne veut pas dire qu'on ne va pas faire d'inférence (inévitable dans toute démarche statistique).

1.2. Trois paradigmes

Premier paradigme : le tableau de contingence. Correspond à l'analyse des correspondances (AC)

Deuxième paradigme : le tableau individus x Variables

1. Analyse en composantes principales (ACP) : variables numériques
2. Analyse des correspondances multiples (ACM) : variables catégorielles

1.3. Bref historique

1982 : Benzécri publie L'analyse des données.

Les précurseurs : Karl Person (1901 et Hirschfeld (1935).

Benzécri publie « L'analyse des données » en 1973, ***la bible*** de l'AC.

Le cœur de l'AC est alors établi.

Texte central de Benzécri, que l'on appelle « Honolulu » (**disponible sur le site de Le Roux**).

1973 à 1980 : c'est l'âge d'or de l'AC. Les mathématiciens n'arrivent plus à suivre la demande (tout le monde vient avec ses données).

A cette époque, les software publiés par le laboratoire Benzécri sont diffusés gratuitement.

Il y a complet isolement vis-à-vis de la littérature anglo-saxonne : très peu de réactions par rapport au travail qui se fait en France.

Années 1970, l'analyse des correspondances est complètement ignorée par Shepard (qui connaît pourtant Benzécri). Kendall&Stuart (1976), bible de la statistique, l'ignore également.

1981, reconnaissance internationale, grâce à des chercheurs proches de Benzécri qui traduisent ses travaux en anglais.

Aujourd'hui, de nombreux textes existent en anglais. L'AC est reconnue et utilisée, mais l'AGD et sa méthodologie, et l'ACM en particulier, reste à découvrir par une large audience.

1.4. Quelques points méthodologiques

Il n'y a pas d'hypothèses invérifiables. Mais il y a des hypothèses. Notamment celle d'homogénéité et exhaustivité des données.

- Exhaustivité: le thème de l'étude doit délimiter le domaine de recueil des données : individus et propriétés des individus. Point sociologique.
- Homogénéité : variables qualitatives et quantitatives, il faut alors procéder à un codage pour les rendre homogènes. Les variables doivent baliser de manière correcte le domaine étudié.

Sens de la démarche : Problème sociologique ➔ Données pertinentes ➔ AGD ➔ Interprétation statistique ➔ Interprétation sociologique

Données manquantes : les individus qui ne répondent pas doivent être éjectés de l'analyse. Mais on ne peut pas éjecter ceux qui n'ont pas répondu qu'à une réponse (des variantes permettent de les inclure). Un nombre élevé de non-réponses dans un questionnaire affecte l'exhaustivité et l'homogénéité. Cela montre qu'il y a tout de même un modèle derrière l'AC, même s'il est faible (pas de contrainte de distribution normale ou de choses comme ça).

Remarque : Dans l'analyse factorielle classique, (« analyse factorielle des psychologues »), il y a un modèle théorique en psychologie derrière. Le modèle était premier : l'intelligence pourrait être un facteur général auquel s'ajoutent des facteurs spécifiques. On part de la matrice des corrélations, et on se demande ce qu'on met dans la diagonale, ce qui relève d'un choix apriori. On veut que le premier facteur sature sur toutes les variables, et que les facteurs suivants soient plus spécifiques. Il y a ensuite d'autres choix, celui des rotations. On oublie que l'analyse factorielle classique importe un modèle (psychologique). Mais tout le monde a oublié le modèle sous-jacent. On est donc dans une démarche rigoureusement opposée à celle de l'AC.

Dans l'AC, il n'y a pas de théorie apriori. Conséquence : la solution est unique.

2. Axes principaux d'un nuage de points

2.1. Notions géométriques de base

Eléments d'un espace géométrique : points, droite, plan. Une géométrie à 2 ou 3 dimensions.

On va s'intéresser au couple de points (P,M), ou *dipôle*. On s'intéresse à la distance entre les points, c'est-à-dire le vecteur qui relie P à M.

Notions affines : alignement, direction et barycentre (moyenne pondérée de points)

Les écarts s'ajoutent vectoriellement : la somme de vecteurs se fait par la règle du parallélogramme.

Notions métriques : distances et angles.

Le théorème de Pythagore est fondamental pour l'AC.

2.2. Nuage de points

Explications théoriques…

2.3. Axes principaux

Y a-t-il une droite qui récupérerait le plus de variance possible ? La variance maximale correspond à un angle qui définit le premier axe du modèle. Le nuage projeté sur cette droite est celui qui représente le mieux les données.

On peut essayer de trouver un deuxième axe. Cela donne la représentation du nuage rapporté à ses axes principaux (présentation usuelle de l'AC).

Les cosinus carré correspondent à la qualité de la représentation : le point est-il sur la droite ou non ? S'il est égal à 1, c'est qu'il est sur la droite. Cet indice permet de savoir si le point qui nous intéresse est proche ou non du plan étudié.

2.4. D'un nuage plan à un nuage de grande dimensionnalité

Un nuage dans un espace à 3 dimensions. La droite qui définit le plus de variance est l'axe principal. L'orthogonale est le deuxième axe. Comme les deux axes ne sont pas corrélés, leur interprétation renvoie forcément à des réalités différentes.

La **propriété d'hérédité** est le plan qui résume les deux axes. C'est important, car toutes les méthodes n'ont pas ces propriétés (la régression, etc.). Le nuage plan qui ajuste le mieux le nuage initial est le nuage projeté sur le plan déterminé par les 2 premiers axes principaux.

On veut projeter un groupe d'individus supplémentaires (dans l'exemple présenté ici, Indiens et Pakistanais) dans le plan crée par l'ensemble de l'échantillon. On voit ainsi où ils se situent par rapport à la population de base.

On peut projeter des variables supplémentaires (âge, sexe…). On peut le faire dans le nuage des modalités ou dans celui des individus. Ici, on fait dans le nuage des modalités. Il faut toujours regarder les écarts des coordonnées sur les axes, pour voir si une variable est associée à un axe, et voir l'écart entre la coordonnée la plus grande et la plus petite. Comme règle, on dit qu'un écart est « notable » à 0.5, et « grand » à 1. Cela permet de ne pas faire de la surinterprétation.

3. L'analyse des correspondances multiples spécifiques

Permet de tenir compte des catégories peu fréquentes (moins de 5% des réponses) : loin du centre du nuage, elles contribuent fortement à sa question et risquent donc d'être trop influentes pour la détermination des axes. Ces catégories posent donc problème, car elles tirent trop les axes (même problème que pour les très hauts revenus lorsqu'on cherche à déterminer une moyenne). Il y a aussi les « catégories poubelle », les catégories de non intérêt. Genre la catégorie « autre » dans un questionnaire. Ça recouvre plein de choses très différentes, qui ne peuvent pas être représentées par un seul point. Il faut donc leur donner un traitement particulier. Il y a également les non-réponses qui peuvent être prises en compte dans l'analyse spécifique.

Il faut réfléchir sur la distance. On ne change pas le nuage des catégories.

Cette méthode a été utilisée la première fois dans l'article de Bourdieu sur les éditeurs.

4. Analyse des données structurées

Sous-jacent déjà chez Bourdieu. Mais on lui donne ici un statut systématique. Idée : j'ai fait mon analyse sur les styles de vie, j'aimerais savoir s'il y a un rôle du sexe, etc. On appelle ça un **facteur structurant**. J'ai deux groupes : hommes et femmes, et on se demande, comme dans l'analyse de variance, s'il y a une différence entre ces groupes. Les techniques conventionnelles sont Analysis of variance : ANOVA, MANOVA, régression…

4.1. Des variables supplémentaires aux facteurs structurants

Je vais m'intéresser au groupe des individus hommes et des individus femmes. Je change de problématique en passant dans le nuage des individus.

4.2. Des données expérimentales aux données d'observation

Contrairement à l'expérimentation, les données d'observation ne sont pas contrôlées (sauf si on a des échantillons par quotas). Dans l'expérimentation, on essaye de mesurer l'effet des facteurs sur la ou les variables dépendantes.

Il faut donc manipuler l'interprétation de façon prudente lorsqu'il s'agit de données d'observation.

4.3. Ellipses de concentration

Les ellipses de concentration sont des résumés géométriques de sous-nuages dans un plan. C'est l'idée déjà développée par Bourdieu dans La Distinction, lorsqu'il traçait des contours dans ses graphiques. On remplace les nuages de points par un résumé géométrique, l'ellipse de concentration. L'ellipse se fait généralement sur les variables supplémentaires (facteurs structurants). Elle peut aussi se faire sur les variables actives pour se donner une idée, mais généralement on ne le fait pas vraiment.

Pour un nuage de forme normale, l'ellipse de concentration contient environ 86% des points du nuage. S'il y a plus de 14% des points hors du nuage, c'est donc que l'ellipse ne résume que mal l'information.

II Analyse des Correspondances Multiples (ACM)

Données de bases : tableau IndividusXQuestions

Questions : variables catégorisées, par exemple variables avec un nombre fini de catégories ou *modalités*. Les catégories ou modalités peuvent être qualitatives ou le résultat de la mise en classe d'une variable continue.

Les individus sont des personnes ou individus statistiques (entreprises, pays, année, etc.)

Questionnaire sous forme standard : pour chaque question, chaque individu choisit une et une seule modalité de réponse. Sinon, phase de codage fastidieuse.

1. Principes de l'ACM

Tableau de données de base pour l'ACM : IndividuXQuestions

L'ACM produit 2 nuages de points : le nuage des individus et le nuage des catégories.

2. Exemple Taste

4 questions, 29 modalités 8+8+7+6 de réponse au total. N=1215 individus (ceux qui ont répondu à toutes les questions).

Création du nuage des individus. Si deux individus sont d'accord sur une question, cela définit une distance nulle pour cette question. S'il y a désaccord sur une question, il faut créer de la distance liée à la différence. Cette distance dépend de la fréquence de cette réponse : si un individu a choisi une réponse rare, on va vouloir que cela l'écarte d'un individu qui a choisi une réponse courante. La distance est modulée par l'importance de la fréquence de choix de la modalité. La distance entre deux individus ayant choisi 2 modalités différentes mais courantes va être petite. En revanche, si l'un a choisi une réponse rare, alors la distance sera grande. On fait ensuite la moyenne du total des distances. Les individus qui choisissent des modalités rares vont se retrouver à la périphérie du graphique.

G est le point moyen (centre) du nuage.

Création du nuage des modalités. C'est le recto du verso qu'est le nuage des individus. On définit la distance entre deux modalités k et k'. Deux modalités sont d'autant plus proches que ce sont les mêmes individus qui les ont choisi et inversement. Moins la modalité est fréquemment choisie, plus elle est située à la périphérie du nuage. Les modalités rares se trouvent donc à la périphérie comme les individus qui ont choisi les modalités rares.

(La contribution est le poids relatif X la distance au centre. Une modalité peut donc avoir peu de poids mais beaucoup de contribution).

La contribution d'une question est la somme des contributions de chacune de ses modalités. Plus une question a de modalité, plus elle contribue à la variance totale. Ce qui est une propriété discutable : une question à 20 modalités écrase forcément une question à 2 modalités. **On cherche donc à équilibrer les modalités des questions dans le questionnaire.** Pratiquement, c'est difficilement réalisable. Donc on définit des thèmes dans le questionnaire, en essayant d'équilibrer les contributions des thèmes (les thèmes étant des regroupements de questions). Si on

a un thème « éducation » et un thème « économie », il faut veiller à ce que les contributions soient à peu près équivalentes si on veut qu'il y ait des chances que chacun puisse contribuer à un axe. C'est important : quand on regarde la contribution à un axe, il faut toujours vérifier qu'il ne s'agit pas d'un effet purement mécanique lié à un nombre très élevé de modalités de réponse sur un thème.

Le premier axe (axe principal) est celui qui a le plus de variance. On en trouve deux, un pour le nuage des individus, un pour le nuage des modalités. Les deux nuages ont les mêmes valeurs propres (les axes ont les mêmes variances).

Une fois qu'on a les axes principaux, on définit les coordonnées des individus et des modalités sur ces axes.

3. Etapes d'une analyse

1. Préparation du tableau à analyser : choix des individus actifs, des questions actives et codage des individus.

2. Etude des statistiques élémentaires

3. Effectuer l'ACM et recherche des points à contributions prédominantes

4. Inspection informelle des deux nuages : recherche d'indices qui montrent un manque d'homogénéité

5. Interprétation. Nombre d'axes à interpréter, interprétation des axes par la méthode des contributions des points et des écarts à partir du nuage des modalités.

6. Etudier le nuage des individus (patrons repères, forme, etc.).

7. Etude des éléments supplémentaires

8. Examen approfondi du nuage des individus (facteurs structurants, ellipses de concentration, variance inter et variance intra, etc.)

9. Inférence statistique. Elle doit être adaptée : quand on fait de l'inférence sur un axe, comme ils sont construits, on ne peut pas faire le même type d'inférence qu'avec des corrélations où les variables sont extérieures et non « construites ».

4. ACM de l'exemple « Taste »

Le nombre de dimensionnalité du nuage est le nombre de modalités actives moins le nombre de questions actives. Ici : 29 – 4 =25.

On regarde les contributions de chaque axe à la variance. Lorsqu'on a un décrochage, on coupe. Ici, on retient donc 3 axes. Il ne faut pas couper entre deux axes qui ont à peu près la même variance. Cependant, les arguments statistiques sont nécessaires mais pas suffisants : il faut sortir les graphiques, et voir si les axes sont interprétables. Il y a parfois de l'information au $4^{ème}$ ou $5^{ème}$ axe… Le nombre d'axes à retenir est de toute façon l'argument le plus difficile à défendre, notamment dans un article.

Nombre de questions divisé par 100 : donne la moyenne des contributions. Tout ce qui est en-dessous de cette moyenne, on ne l'interprète pas. On ne retient que ce qui est au-dessus. Ça simplifie la vie.

On peut intégrer, comme individu supplémentaire, des « individus construits ». Par exemple celui qui, sur un questionnaire sur le racisme, aurait répondu « racistement » à toutes les questions ou, au contraire, « non-racistement » à toutes

les questions. On les projette dans l'espace, ce qui donne des points de repère. On peut toujours enrichir les nuages avec des éléments extérieurs qui ne modifient pas les axes.

III. Méthode de classification

La ressemblance/dissemblance est mesurée à partir d'un ensemble de variables décrivant les objets (variables numériques ou catégorisées). En classification euclidienne, la ressemblance est mesurée par une distance euclidienne. La distance euclidienne est une méthode parmi plein d'autre possibles.

Distances (dissimilarités) :

- Tableau IndividusXVariables numériques : distance euclidienne sur données brutes ou sur données centrées réduites

- Tableau de contingence : distance du X2 (distance de l'AC)

- Tableau IndividusXQuestions : distance de l'ACM

Trois familles de méthodes :

1. Les méthodes de partitionnement qui permettent d'obtenir une partition optimale en un nombre de classes fixé à priori. Peu intéressante si on veut des analyses fines.
2. Les méthodes hiérarchiques (ascendantes ou descendantes) qui produisent une succession de partitions emboîtées. Souvent utilisées.
3. Les méthodes de segmentation qui cherchent à résoudre les problèmes de discrimination et de régression en segmentant de façon progressive un échantillon pour obtenir un arbre de décision binaire (AID, CART). De plus en plus à la mode. Plus adaptées au raisonnement sociologique que les régressions classiques.

On cherche un optimum local (on est sûr que ça n'est pas le meilleurs, le nombre de partitions étant bien trop grand pour qu'il soit possible de les tester toutes).

Partition d'un nuage : variance inter et variance intra. On classe les points d'un nuage. On souhaite que la variance intra soit la plus petite possible, la variance inter la plus grande.

Admettons qu'on ait trois classes de nuages. On définit le point moyen de ces trois nuages. Chacun est pondéré par le nombre de points qu'il représente. On obtient un nouveau nuage : le nuage inter composé de trois points.

La **contribution d'un sous-nuage** est la somme des contributions de ses points. La contribution intra d'un sous-nuage est le produit de son poids par sa variance divisée par la variance totale.

Méthode ascendante hiérarchique

1. Distance euclidienne entre individus. Variables numériques (distance de l'ACP) ou variables catégorisées (distance de l'ACM).

A chaque étape, on agrège deux à deux les éléments les plus proches (arbre hiérarchique). Où faut-il couper ? On coupe lorsqu'on a une chute des indices de niveau. Lorsqu'on a plus de variance inter que de variance intra, on peut s'arrêter.

Philippe Bonnet : l'utilisation de SPAD

Remarques préliminaires :

- Notre version SPAD est valable jusqu'au 31 octobre 2011.

- Les usagers de SPAD utilisent souvent SPSS pour le nettoyage de la base de donnée, les recodages, etc. (possibilité de le faire sur SPAD, mais un peu complexe).

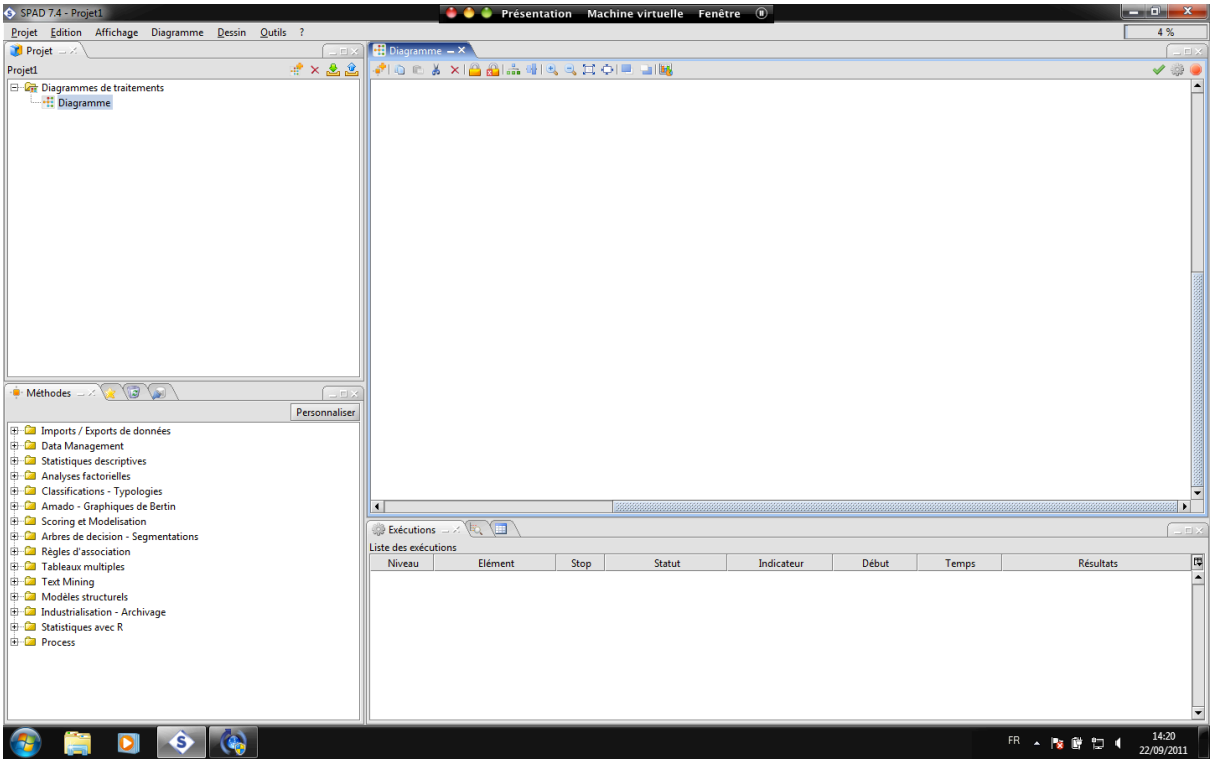
- Un guide pour l'utilisation de SPAD est disponible sous http://www.math-info.univ-paris5.fr/~lerb/Logiciels/SPAD/SPAD_Guide-English-2010.pdf. Ce guide reprend l'analyse sur la base de données « Taste ».

On ne fait jamais qu'une seule analyse, mais une série d'analyses successives afin de trouver celle qui rend le mieux compte de nos données. A la première fois, on peut voir certains individus qu'on préfère ne pas mettre en actif. On les met alors en supplémentaire. Dans le cas présent, les individus en supplémentaire sont des catégories d'individus qui ne faisaient pas partie de l'échantillon aléatoire, mais qui avaient été ciblés (indiens, pakistanais…).

Conseil : le nombre de modalités doit être ni trop élevé, ni trop faible. Ici, il n'y a pas plus de 8 modalités par question.

1. Généralités

Réglage : **Outils – Préférences – Répertoire des préférences et des projets**. Il a décidé d'emblée de placer ces répertoires à certains endroits. C'est important de savoir où c'est. Rien à changer pour l'instant.



On a 4 fenêtres.

1. Fenêtre des diagrammes (en haut à gauche)
2. Fenêtre des méthodes
3. Fenêtre qui nous donne l'historique de ce qu'on fait (en haut à droite). On va sélectionner une méthode en bas à gauche (glisser-déposer), la déposer en haut à droite, paramétrer la méthode, puis exécuter. Tout cela figurera dans cette fenêtre.
4. On retrouve tout ce qui a été fait dans la fenêtre en bas à droite avec le statut (si ça a fonctionné ou pas, etc.). Intérêt : si ça n'a pas marché, on peut passer dans les autres onglets, où il y a les messages d'erreur (explication du problème), puis troisième onglet où se trouvent les données.

2. Démarrer le projet :

On va importer le fichier excel. Dans **méthode, ligne import/export, Imports, feuille excel**. Prendre feuille excel, la glisser dans la fenêtre 3. L'icône apparaît. Faire **clic droit, paramétrer**. Deux onglets : Import et Métadonnée. **Dans premier onglet, sélectionner le fichier excel et préciser la feuille Excel (Data)**. Il affiche les premières lignes.

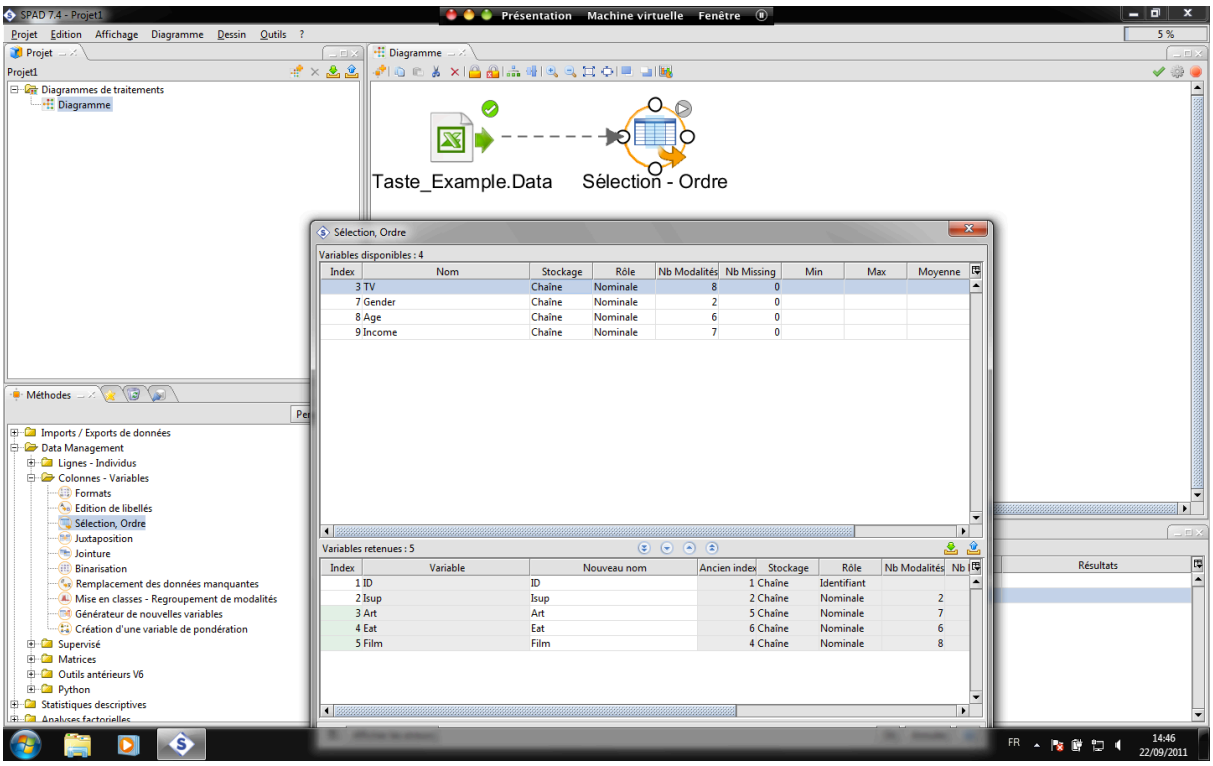
On va dans l'onglet Métadonnées. **Les 2 éléments en bas doivent être cochés.**

SPAD a décrété que les variables sont nominales. Typage automatique des variables. On peut le changer, mais c'est très délicat. **On se met sur la première variable et clic droit. Sous Rôle, pour la première variable, on sélectionne Identifiant.** Parce que cette première variable est l'identifiant des individus. La détermination de la colonne Rôle est déterminante. Spad le fait par défaut, mais il faut vérifier.

On clique Ok. Le signe vert s'affiche sur l'icône Excel.

3. Sélection de l'ordre des variables

On ferme l'Import et on va dans Data management, Selection, Ordre. **On le glisse par dessus l'icône Excel**. Il établit une liaison (flèche) entre les deux éléments. **Se mettre sur Sélection-Ordre bouton droit, Paramétrer**. Un tableau s'ouvre. Fenêtre en haut avec variables. **Dans fenêtre d'en bas, on fait glisser les variables dans un ordre choisi par nous**.



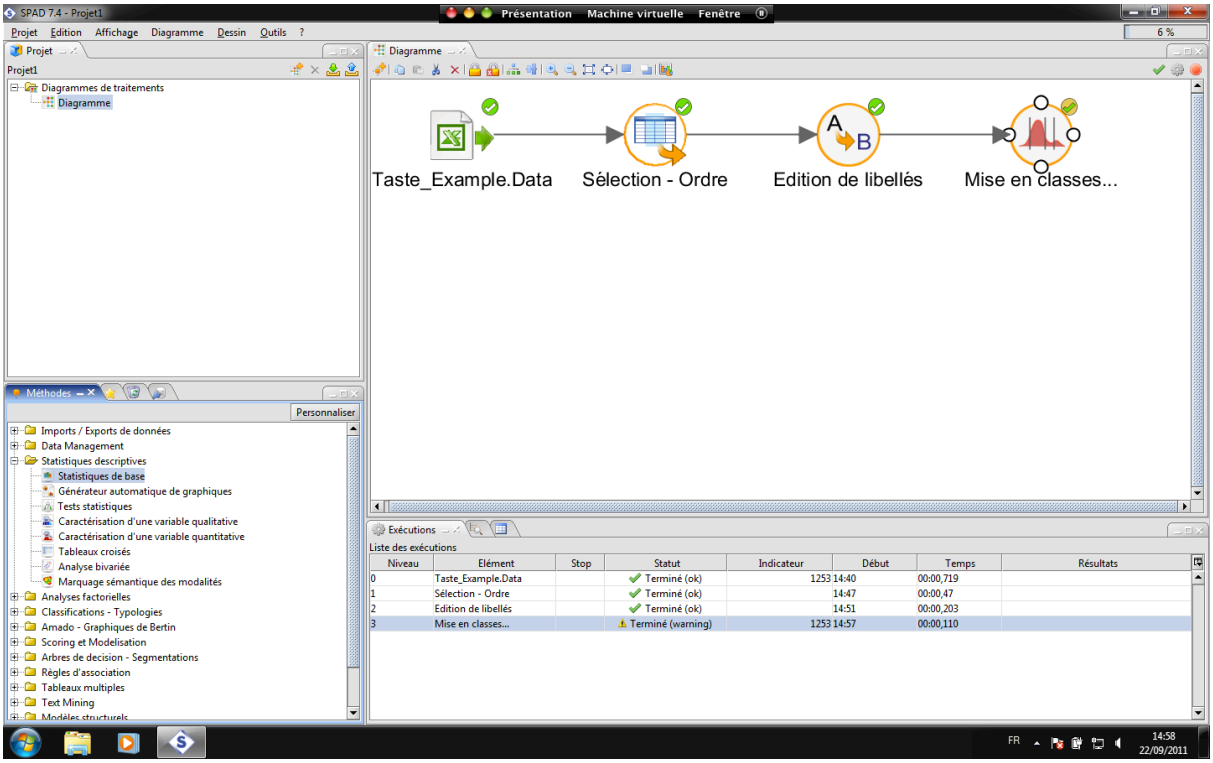
Ça peut être très important si on a plein de variables, pour la gestion par la suite. **On fait OK.**

4. Edition de libellés

On va dans **Colonnes-Variables – Edition de libellés. On le fait glisser par dessus Sélection-Ordre. Clic droit, Paramétrer.** Si je clique sur « Art », les modalités apparaissent à droite par ordre alphabétique. Il est possible que l'on veuille un autre ordre (par exemple l'ordre du questionnaire). On peut le faire ici. On le fait en cliquant-glissant la modalité concernée. On peut aussi changer le nom de la variable (à gauche) ou de la modalité. Dans les graphiques, les variables apparaissent avec le nouveau nom (donc attention à être précis pour s'y retrouver avec les output). Toujours penser que ces output vont apparaître dans le nuage du tableau. Il est donc important de pouvoir les identifier précisément. **On clique OK.**

5. Mise en classes

On est toujours dans Colonnes-Variables. **On prend Mise en classe, on le glisse par dessus Edition de libellés. Clic droit, Paramétrer. Clic droit sur Age – Regroupement de modalités. Tableau, avec à gauche les modalités de l'âge. Si on veut regrouper les vieux ensemble. Je prends les 18-24 et je les fais glisser avec les deux flèches parallèles (ce qui ne les regroupe pas). On peut en prendre plusieurs à la fois. Puis on prend les 55-64 et 65+ et on les passe à droite avec la flèche qui regroupe. On peut ensuite changer le libellé de cette nouvelle variable (55+). On fait OK.** Dans le tableau « derrière », on peut annuler la dernière manœuvre (en appuyant sur la croix rouge à droite). Attention : si on a regroupé une variable, il faut contrôler le Stockage, pour voir s'il faut le changer (ici, non). **On fait OK.**



6. Statistiques de base

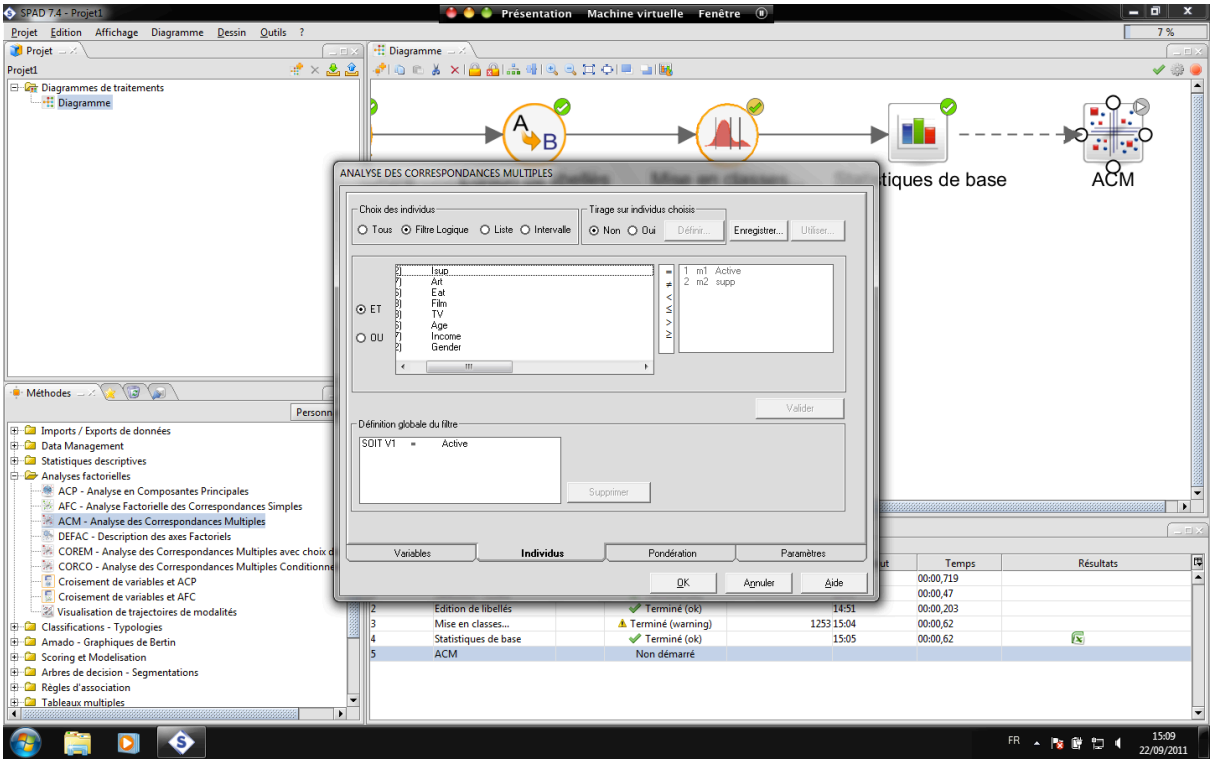
On ouvre Statistiques descriptives, Statistiques de base. On le glisse (toujours selon le même principe). Clic droit, Paramètres. On sélectionne Individu Pondération - Filtre logique. On clique sur Isup, les deux modalités apparaissent, on sélectionne Active. On sélectionne = et Valider (Ne pas oublier). Il sait à présent qu'il ne travaille que sur les individus qui sont en Active. Cela apparaît dans la fenêtre d'en bas. **On sélectionne Tri à plats, et on sélectionne ce qu'on veut. On fait OK.**

7. Analyse des Correspondances Multiples

7.1. Réglages de base

On sélectionne Analyse factorielle, ACM, on le glisse. Paramètre. On prend les 4 variables actives, on les glisse en bas comme Active. On prend les 3 Illustrative (Supplémentaires).

On va dans **l'onglet individus. On met Filtre Logique. On sélectionne Isup, Active. On Valide. Il ne travaillera donc que sur les individus Actif.**



On va sur **l'onglet Pondération**. On laisse comme c'est.

Onglet Paramètres. Paramètres de fonctionnement. Par défaut, il propose les 10 premières. Ok.

Ventilation des modalités actives… Il met 2.000 par défaut. Ca veut dire que s'il tombe sur des modalités qui font moins de 2%, il les supprime et les ventile dans le reste. PAS QUESTION. **On change et on met 0%.** C'est un parti pris de notre part. On verra plus tard que dans le cas de très petits effectifs, on peut prendre une autre solution.

En-dessous. **Coordonnées éditées. Il met les 5 premières. Ok.**

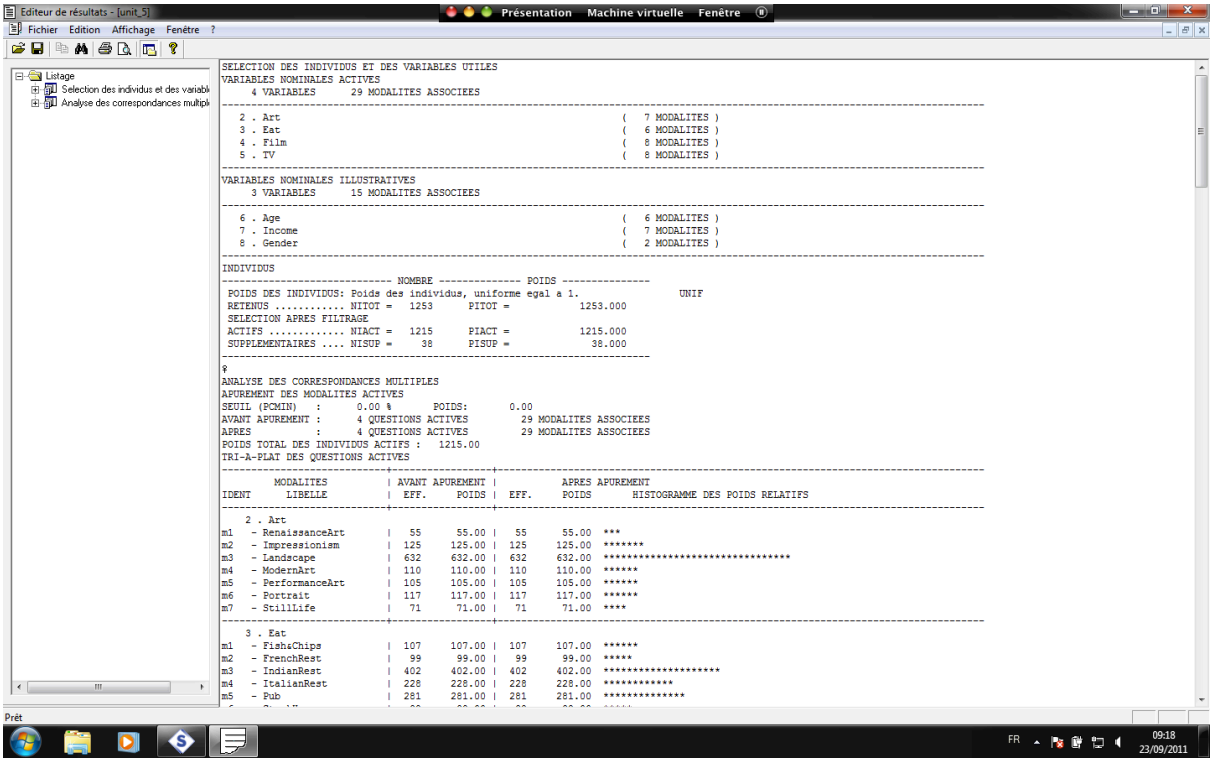
On fait OK.

Récapitulation de l'ACM :

1. Analyse factorielle – ACM – glisser dans tableau - Paramétrer
2. Variables – sélectionner les 4 actives et les 3 illustratives
3. Individus – Isup – sélectionner les actifs – Valider
4. Pondération : on laisse comme ça.
5. Paramètres : on met 0%.
6. Bouton droit sur ACM, Résultats – Editeur graphique de plans factoriels
7. Ouvrir nouveau graphique
8. Sélectionner ce qu'on veut

7.2 Editeur de résultats

Dans ACM, si on fait bouton droit, Résultats, Editeur de résultats, On se met sur Listage et on fait Enter. Les résultats s'affichent sous forme écrite avec tous les résultats.



C'est ici qu'on trouve les taux modifiés (pas présents dans Excel). On regarde d'abord les taux non modifiés. Il s'agit des valeurs propres, qui permettent de voir la variance de chaque axe, ce qui permet de décider combien d'axes on retient. On va ensuite voir les « Valeurs propres avec correction de Benzécri » (bien plus – trop – optimistes que les valeurs propres). Il ne s'agit plus de valeurs de variance, mais de **taux d'importance**. On regarde le pourcentage cumulé, qui permet de voir ce que représentent les axes que l'on retient.

7.3. Sorties Excel

On fait bouton droit ACM, Sorties Excel. Il sort des tableaux.

Tris à plat des 4 questions actives. On en a généralement pris connaissance avant l'analyse.

On retrouve les valeurs propres (mais pas les taux modifiés de Benzécri, qui ne sont pas dans Excel).

Dans les feuilles suivantes : Cornu-5. On a « Coordonnées des modalités actives ». Ici, on a les 5 premiers axes (on lui a dit d'éditer les 5 premiers). Dans la feuille Cornu-6, on a les contributions aux axes. On a la ligne TOTAL : contribution totale de toutes les modalités des questions.

Nous savons qu'il y a 29 modalités. Nous allons identifier les catégories qui ont une forte contribution. Critère empirique : si toutes les catégories contribuaient également, chacune contribuerait pour 100/29. Contribution moyenne virtuelle. On considère

qu'une catégorie contribue fortement si elle contribue plus que ce résultat (3.34). On voit que ModernArt et Portrait contribuent « fortement » au premier axe, mais pas les autres.

Mais si on a beaucoup plus de questions ? **On sélectionne les trois premiers axes (seulement les contributions des modalités, pas le total). On fait une « mise en forme conditionnelle ». Format – Mise en forme conditionnelle… La valeur de la cellule est – supérieure ou égale à – 3.34. Format… On lui demande de le mettre en jaune. Ok**. Et voilà que dans le tableau, sont coloriées en jaune les catégories qui contribuent fortement aux différents axes selon le critère que nous avons retenu. On peut faire la même manœuvre en sélectionnant toutes les questions à la fois (mais jamais les lignes TOTAL !). Ainsi, on a une vue beaucoup plus simple du tableau. On se sert des fonctionnalités d'Excel.

On peut également compter le nombre de cellules qui ont une valeur supérieure à 3.34. **Dans formule : =NB.SI(D4 :D10 etc. (on sélectionne les colonnes des valeurs pour le 1er axe) ; « >3.34 »).** On refait la manœuvre pour les autres axes. Normalement (ici ça n'a pas marché), il affiche le nombre de cellules concernées. Quand on fait un graphique avec les modalités qui contribuent fortement aux différents axes, on sait le nombre de modalités concernées (ici 13 sur 29).

On voit que Portrait et ModernArt contribuent fortement au premier axe. On revient à la feuille Cornu-5, et on regarde le signe des coordonnées de ces deux modalités.

Ce classeur Excel, il faut penser à l'enregistrer sous un autre nom que « Classeur 1 ».

On quitte Excel.

7.4. Editeur de graphiques

7.4.1. Généralités

Résultats – Editeur de graphiques – Dans graphique – Nouveau. Se présente ce qui existe à l'issue de l'analyse (individus actifs, etc.). On sélectionne les variables nominales actives (pour dégager l'espace des styles de vie). On obtient un graphique « muet ».

Dans **Préférences – Style pour la page. On coche ECHELLES identiques (les deux axes doivent avoir la même échelle !), TITRES D'AXES avec % inertie (c'est le pourcentage de variance). Ce sont les deux choses importantes. OK. Enregistrement des préférences : on dit OK**. Normalement, au prochain graphique, il aura retenu ces options.

On quitte le graphique et on recommence, pour qu'il applique nos préférences. On voit qu'il affiche les taux de variances.

7.4.2. Ajuster la taille des points au poids

Préférences – Style pour les catégories – TAILLE DES SYMBOLES proportionnelle – poids (pour les modalités on prend le poids. Mais pour les individus on prendra la « superposition »). Ainsi, le poids des modalités sera représenté par la taille du point.

« Fond » « Transparent » (aide à la lecture du graphique).

Autre façon d'ajuster la taille des points au poids. Sélection – De tous les points – ils deviennent fushia. Icône abc fait apparaître les étiquettes des points. Il faut ensuite désélectionner les points : Sélection – désélection totale.

On peut déplacer les étiquettes en les glissant. Avec clic droit sur étiquette, on peut changer la police etc.

7.4.3. Représentation des différents axes

Icône sous Sélection qui représente deux axes orthogonaux. On clique dessus, il nous demande quels axes on veut représenter. On demande le 1-3, il affiche les axes 1 et 3. Bon, on revient à 1 et 2.

7.4.4. Mettre en évidence les modalités de certaines thématiques

Sélection – Des variables par liste. On prend la TV, on la met en bas. OK. Les points TV apparaissent en fushia dans le graphique.

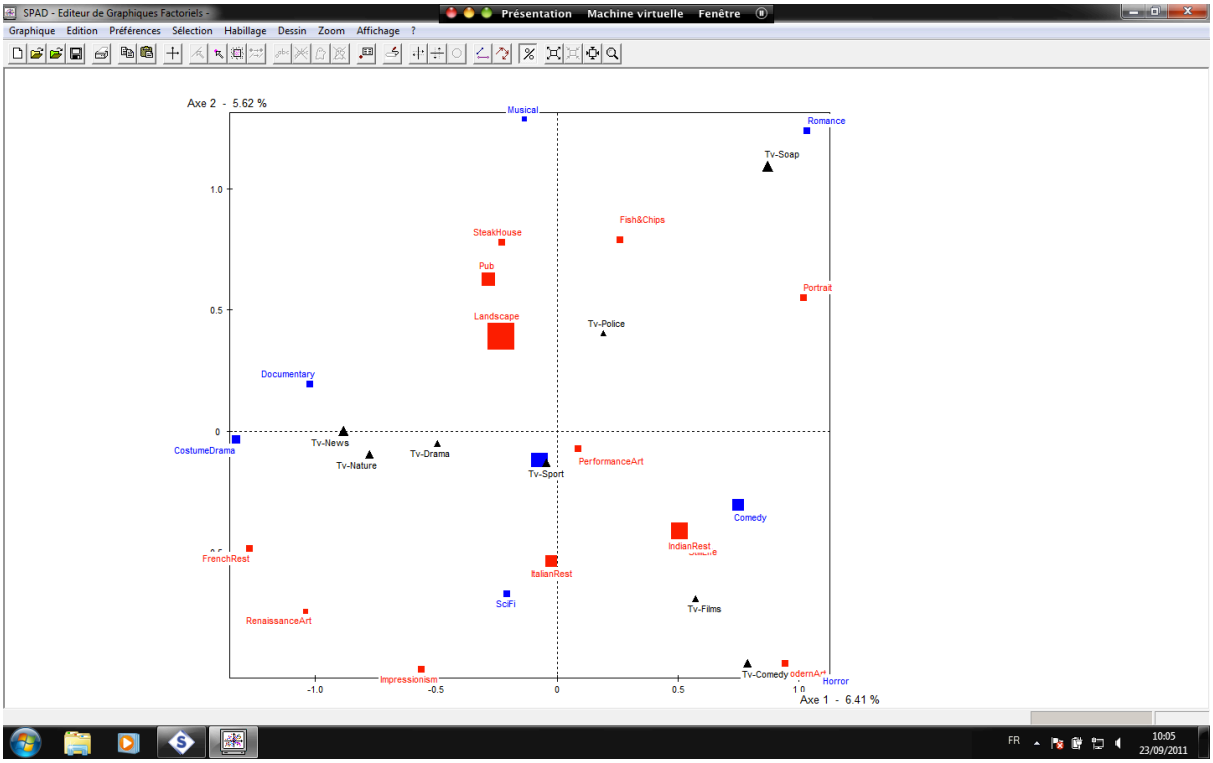
Habillage – Couleurs, symboles, …. On choisit le noir, et on peut changer le symbole. OK.

En-dessous de Dessin, il y a une **icône qui permet de désélectionner les points sélectionnés.**

La TV apparaît comme on l'a voulu.

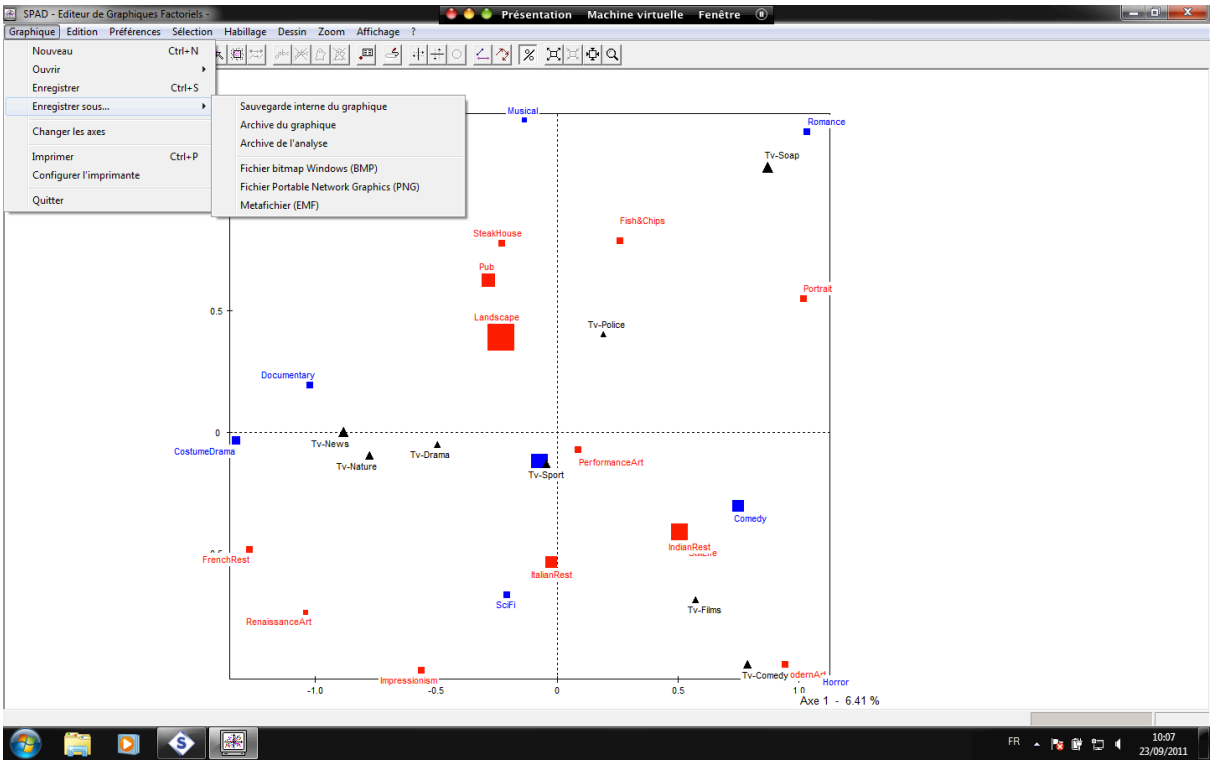
On peut faire la même chose avec d'autres variables.

Quand on fait un graphique, il faut penser à une chose : il doit être lisible. Il faut donc jouer sur les symboles, les couleurs etc. Il faut qu'il soit lisible en une demi-seconde, sinon il ne sert à rien.



7.4.5. Sauvegarde du graphique

On sauvegarde le graphique. Graphique, Enregistre sous… 6 options. Les 3 options du haut concernent des sauvegardes internes au projet. **On sélectionne Archive du graphique. Si on veut sauvegarder dans un fichier word par exemple, on prend la série du bas, soit le BMP, soit EMF. L'EMF permet de sauvegarder dans word sans aucun problème.**



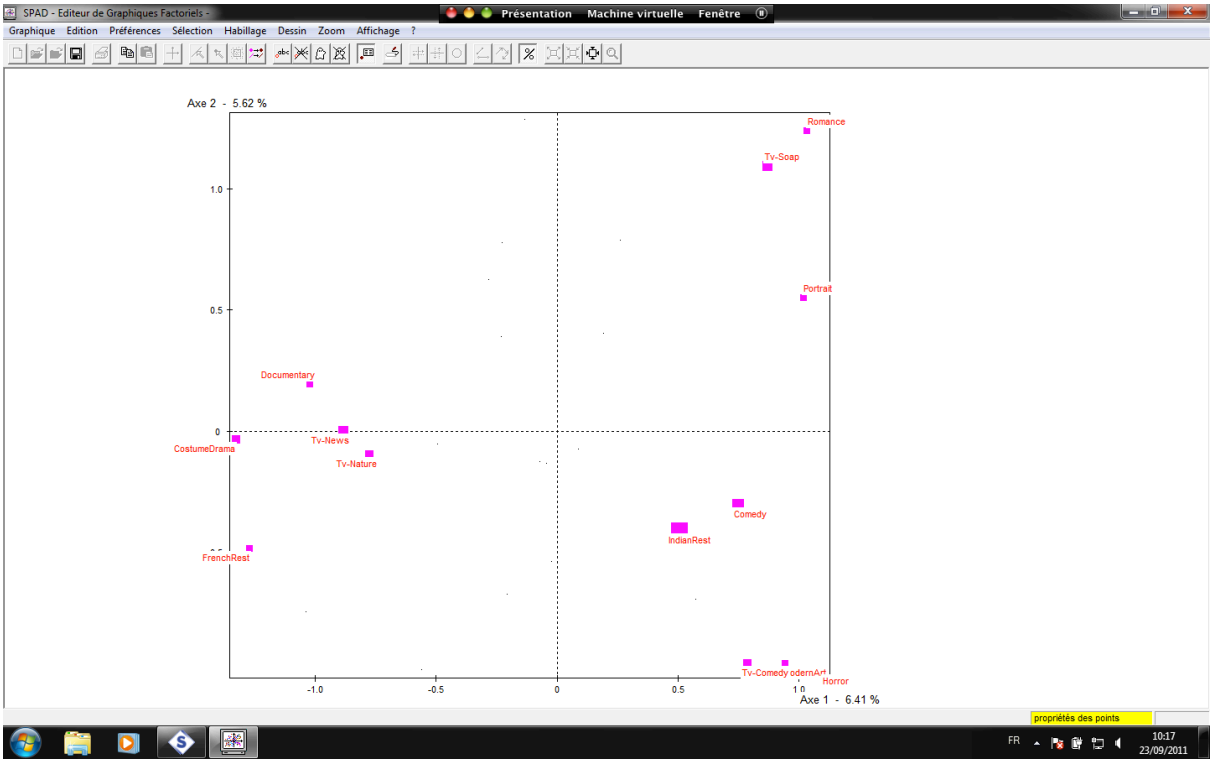
On sélectionne Metafichier (EMF). Il ouvre une fenêtre, avec un nom de fichier par défaut. Supprimer l'* et renommer. Ici, on le nomme « TOTO.emf ».

7.4.6. Editer du texte sur le graphique

Dessin – Editeur de texte. On clique où on veut sur le graphique et il édite le texte qu'on met. Ici : Figure 1 : plan des axes 1 et 2, nuage des modalités.

7.4.7. Sélectionner les modalités en fonction de leurs contributions aux axes

On veut à présent sélectionner les modalités qui contribuent fortement au 1er axe. **Sélection – De tous les points. On reprend Sélection – Filtrage statistique de la sélection. Une fenêtre s'ouvre. On sélectionne « contribution ». à droite : « Sélection exclusive ». Nombre d'éléments sélectionnée en % : 45% (car on a fixé tout à l'heure le % total des 13 variables sur 29 qui contribuent fortement au 1er axe, soit plus que 100/29=3.34%).** On désélectionne tous les points : on a des modalités à gauche, d'autres à droite, et plus rien au milieu. Ce qui est normal.



On a ici un graphique avec les modalités qui contribuent fortement (plus que la moyenne). On interprète les axes avec cela. Mais c'est basé sur un critère relativement arbitraire. Il faut ensuite intégrer le raisonnement sociologique : si une modalités qui contribue faiblement (qui n'apparaît pas sur ce graphique) semble sociologiquement intéressante, il n'y a pas de raison de se priver de la faire figurer sur le graphique.

Sur le graphique, tous les autres points sont présents, sous forme « fantôme » (un tout petit point). On met le curseur dessus, bouton droit, et on peut les faire réapparaître. (Normalement ça marche, mais apparemment sur mon ordi ça ne marche pas).

On refait la **Sélection tous les point – Filtrage par catégories – Et on sélectionne l'axe 2. Sur le graphique, les modalités sont distribuées en haut et en bas (car par rapport au 2ème axe)**.

7.5. Nuage des catégories des variables supplémentaires

7.5.1. Sorties Excel

On affiche les résultats excel. On regarde la feuille Cornu-8. Coordonnées des modalités actives et illustratives. On regarde Age, Genre et Revenu (variables supplémentaires). On regarde les coordonnées sur les axes pour le Sexe. On n'a que 2 modalités. Sur l'axe 1, -0.18 et 0.13. L'écart entre les deux peut être lu directement comme une indication de l'importance de l'« effet » du sexe. L'écart est ici de -0.31. On peut se fixer des bornes, genre en-dessous de 0.4 et 0.6, ou binaire : en-dessous de 0.5. Ici, on prend binaire. Sur l'axe 1 et 2, on voit que l'écart est inférieur. Mais sur l'axe trois, la différence est de 0.91. Donc si on veut trouver une différence liée à l'âge, c'est sur l'axe 3 qu'il faut la chercher.

On peut faire la même chose avec Age, en calculant la différence entre les extrêmes (18-24 ans et 65+). La différence la plus importante est sur le premier axe. Tout cela nous donne des indications. On voit que sur le premier axe, on a quelque chose de régulier, d'ordonné : ça décroit régulièrement, allant de droite à gauche de l'axe.

Sur le Revenu, on peut refaire la même chose.

L'observation de ces coordonnées des catégories supplémentaires est déjà une information. D'un point de vue théorique, on ne fournit pas la contribution des catégories supplémentaires, car elles ne contribuent pas à la construction des axes.

On quitte Excel.

7.5.2. Graphique

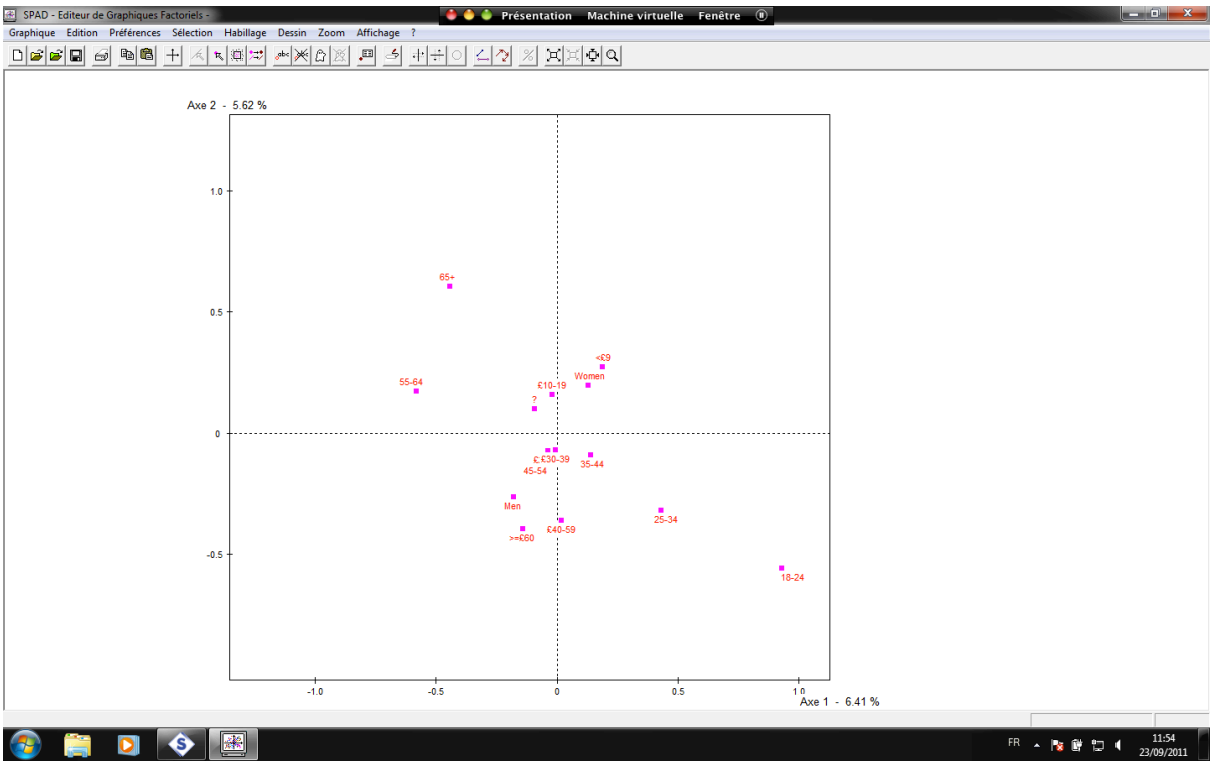
On va afficher le nuage des modalités (on affichera plus tard le nuage des individus)

On retourne dans le module graphique : Résultats, Editeur graphique…

On ouvre un nouveau graphique en **sélectionnant les Variables nominales illustratives.**

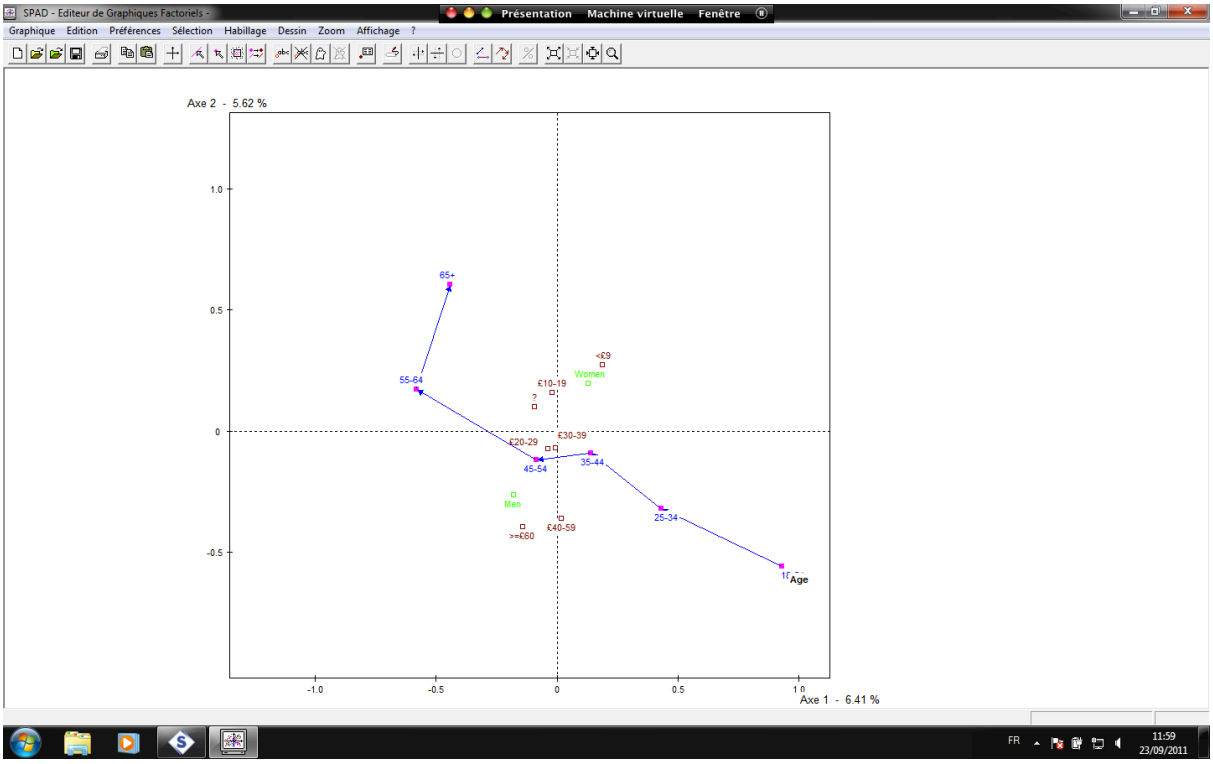
On peut aussi sélectionner les illustratives et les actives. Puis sélectionner les nominales actives pour les mettre en « fantôme » (petite icône en forme de fantôme).

On a le graphique, **on demande les étiquettes**.



On **sélectione par liste. On met Gender en bleu, etc.**

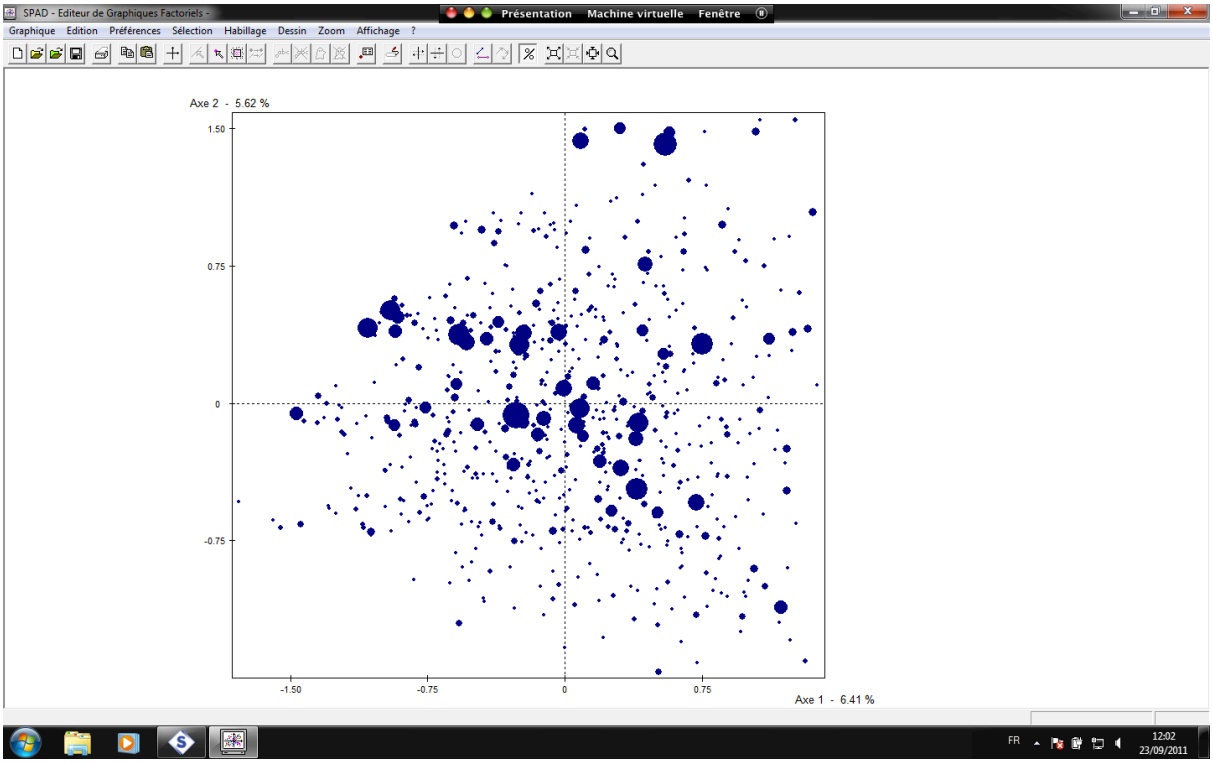
On fait **Sélection Variables par liste. On sélectionne Age. Puis on clique Dessin Trajectoires. On sélectionne Age. L'idée, c'est de tracer la variable, puisqu'elle est ordonnée. On fait OK**, et on a la ligne. On voit que c'est ordonné sur l'axe 1 : on passe des jeunes aux vieux (à l'exception des vieux, qui reviennent en arrière). Cela confirme ce qu'on avait vu dans Excel.



7.6. Nuage des individus

7.6.1. Editer le graphique

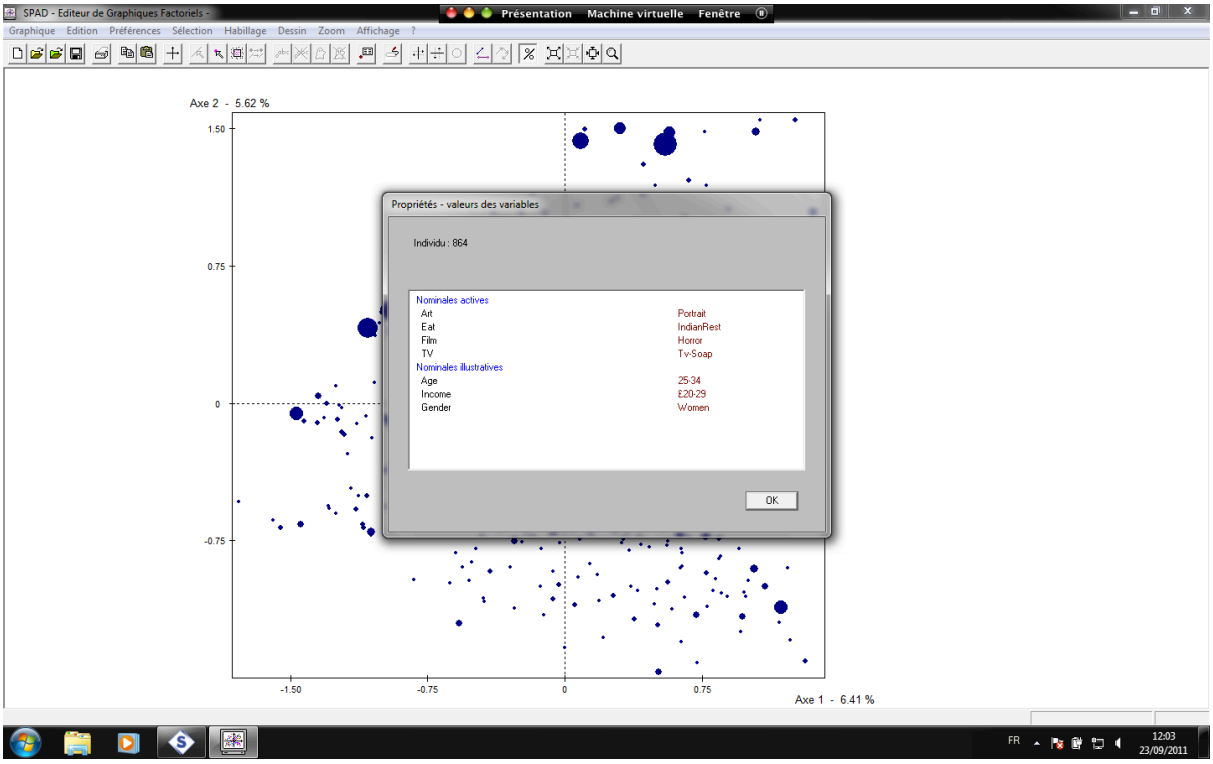
On fait le nuage des individus. **On édite un nouveau graphique, en sélectionnant les individus actifs**. Il nous affiche les gros points en raison de nos options. **Sélectionner tous les points. Habillage, Couleurs symboles, Tailles proportionnelle à la superposition.**



Les gros points sont les superpositions de points.

7.6.2. Identification des individus correspondant aux points

On voit un point à droite. On clique droit dessus et on demande ses propriétés. On demande la valeur des variables. On sait tout de lui.



On voit que le profil de cette personne « extrême » correspond à notre interprétation des axes. On en fait 4 ou 5 comme ça, pour confirmer l'interprétation des axes. Ici, l'exemple est simple car on n'a que 4 questions.

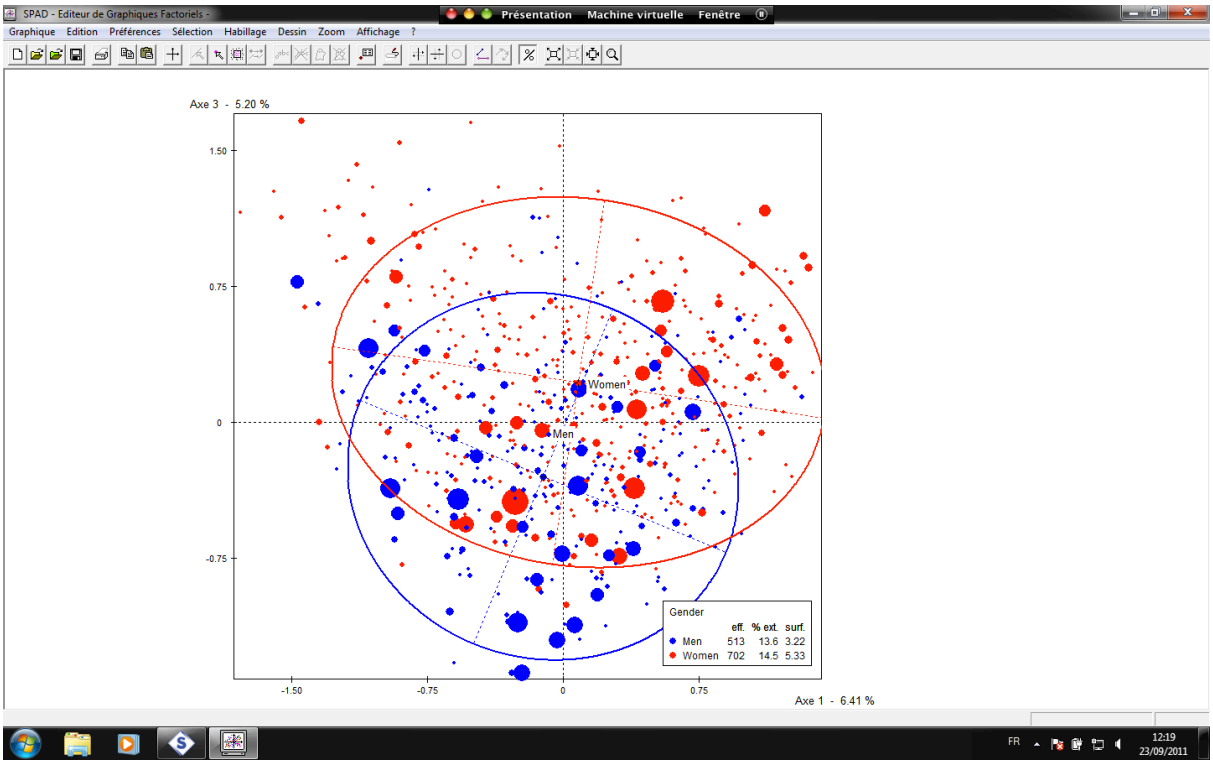
Si deux individus se trouvent au même endroit, c'est parce qu'ils ont répondu pareil sur les 4 questions. Ça ne veut pas dire qu'ils sont pareils sur le plan des variables supplémentaires. Si c'est le cas, c'est un résultat de recherche !

7.6.3. Facteurs structurants et ellipses de concentration

Prenons un facteur structurant, le sexe. On sélectionne le plan 1-3, car on sait que le sexe joue plutôt sur l'axe 3 (nous l'avions vu dans Excel).

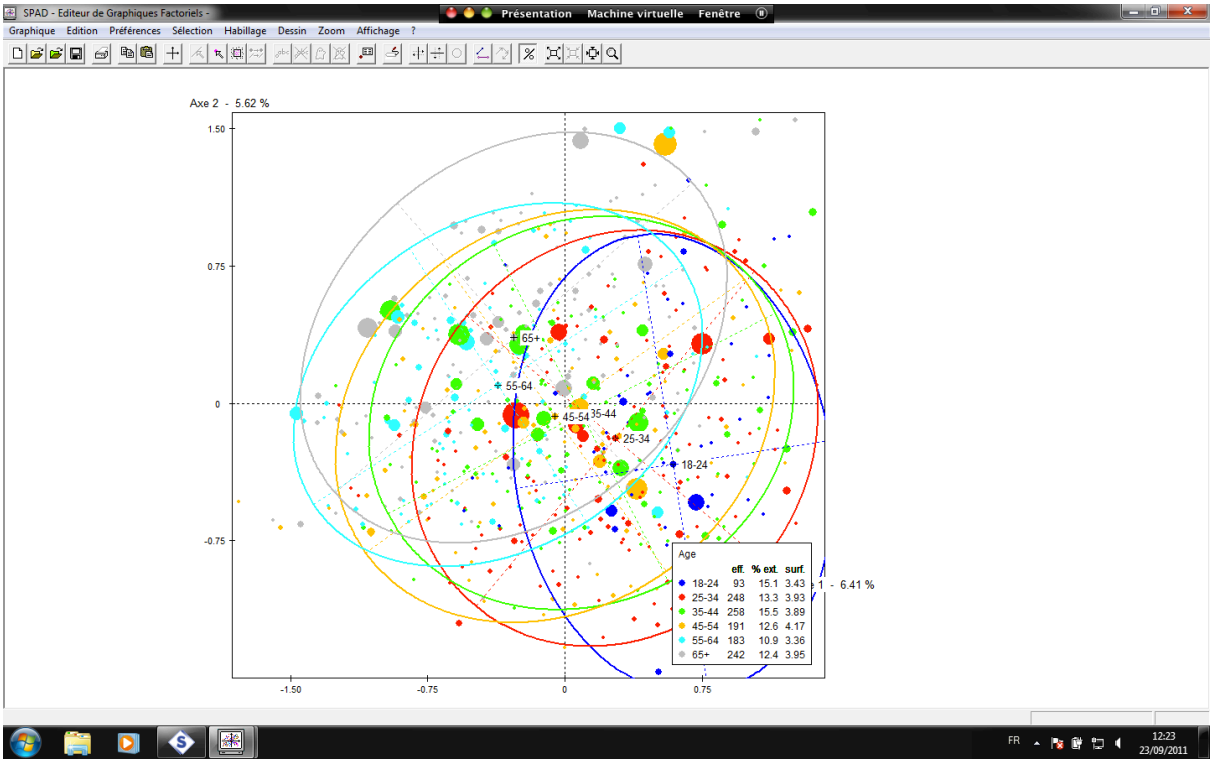
On sélectionne **Habillage – Des individus par une nominale ou une partition. On sélectionne Gender. On a une case « ellipses » que l'on coche. Il faut sélectionner « Ellipses de concentration ». Sous Tracé, on sélectionne l'Epaisseur 2. On sélectionne les axes des ellipses. OK. Spad nous demande des choses, mais on ignore. OK.**

On a le graphique avec les deux ellipses.



On voit que, effectivement, l'écart sur l'axe 3 est important : les hommes sont surreprésentés vers le bas de l'axe. **On peut changer la représentation des axes : Graphique – Changer les axes.**

On peut afficher les ellipses de l'âge.

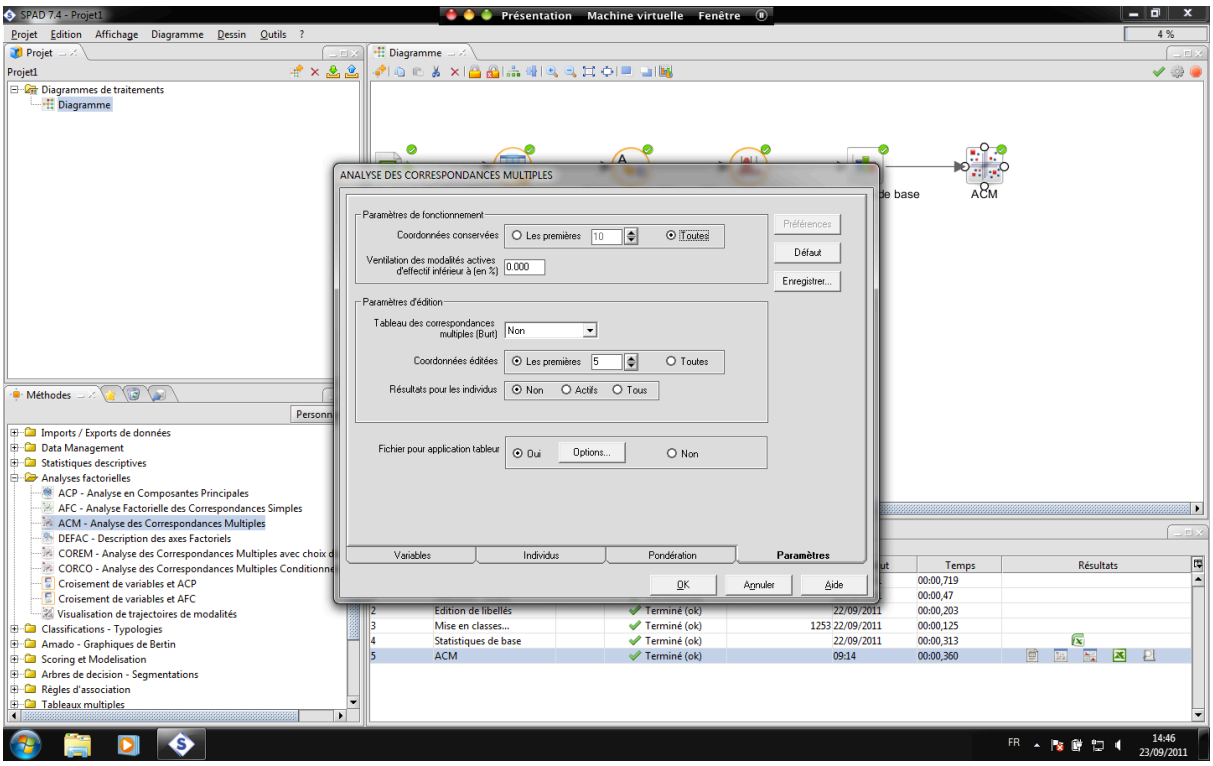


Beaucoup d'ellipses. Dans ce cas, on peut faire **bouton droit sur l'ellipse**, afin d'en supprimer certaines. On ne garde que celles que l'on commente.

8. La classification

8.1. Reparamétrer l'ACM

On revient à l'ACM. On va reparamétrer. Une classification ne se fait pas sur les données brutes dans SPAD. Elle ne peut se faire qu'après une ACM.



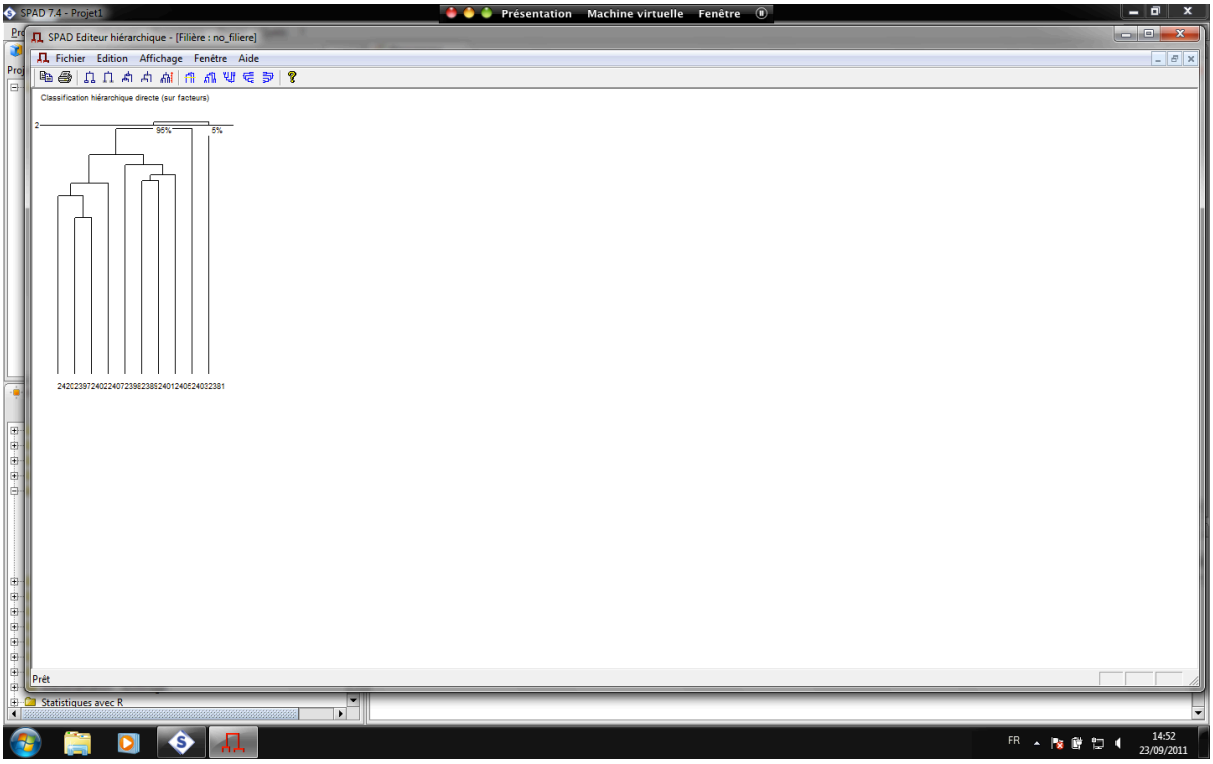
On fait clic droit, Reparamétrer, onglet Paramètres. On sélectionne Toutes (coordonnées observées). C'est pour que l'on dispose de toutes les coordonnées des individus. **On fait OK.**

8.2 Classification

On sélectionne dans méthodes **Classification-Typologies –CAH-MiXTE. On le glisse. Clic droit Paramètre. Choix de la méthode : Hiérarchique RECIP. Paramètres de fonctionnement : Toutes. Sauvegarde partielle de l'arbre : On met 10. Coordonnées des éléments terminaux : le nombre d'axes retenus (3). OK**. La classification apparaît dans le tableau.

8.2.1. Graphique des hiérarchies

Clic droit : Résultats – Editeur graphique des hiérarchies.



Cette représentation graphique permet de voir où l'on peut faire la coupure. On place la souris où on veut couper et on voit les données pour voir la pertinence. On voit que les individus de la droite du graphique ne s'agrègent que tardivement (lors de la dernière agrégation). Ce qui signifie qu'ils sont tellement différents qu'ils ne peuvent pas être agrégés aux autres. Ils représentent 5% du total (environ 1'200) des individus.

On ferme ce fichier.

8.2.2. Coupure de l'arbre et caractérisation des classes

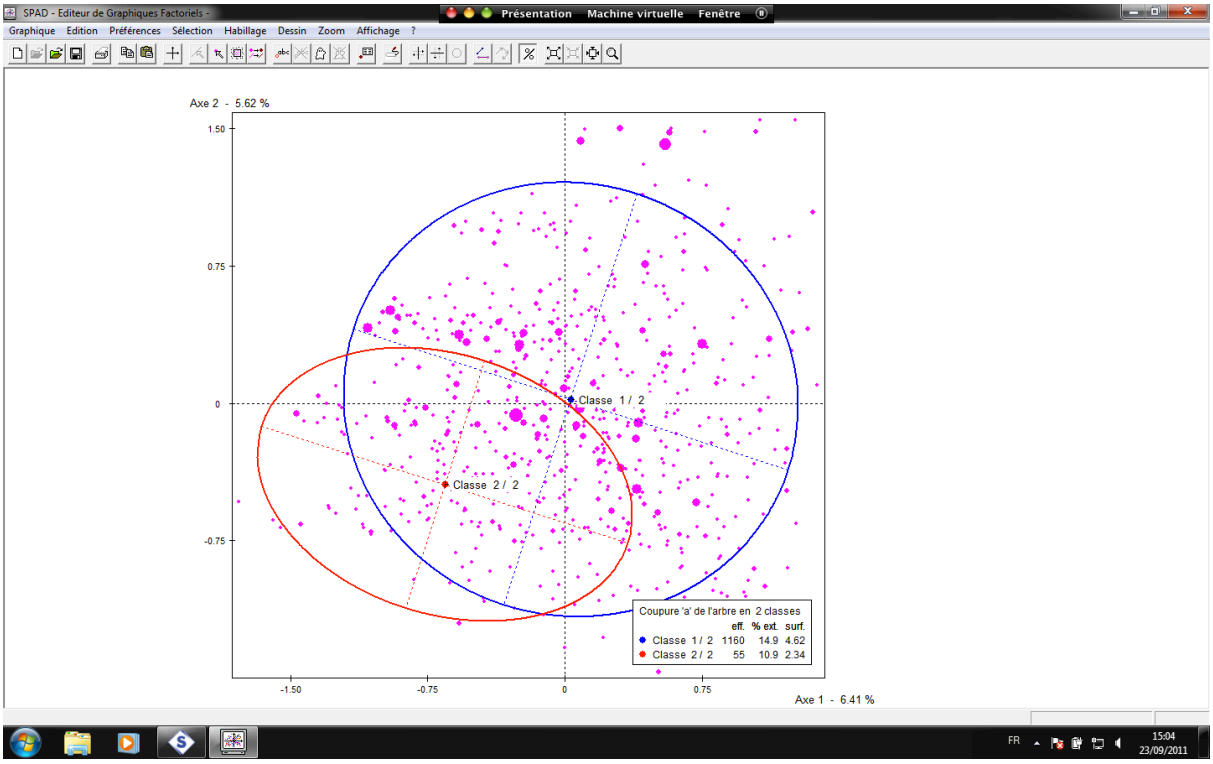
Dans méthodes : **Coupure de l'arbre et caractérisation des classes. On le glisse. Clic droit. Paramétrage. Tableau Choix des partitions par coupure de l'arbre. On sélectionne Définies par l'utilisateur, et on met 2,3 (pour avoir les partitions en 2 et 3 classes).**

Onglet Paramètres de partitionnement. Itérations de consolidation : on met 0. Coordonnées éditées pour les classes : on met 3. Edition des parangons : Non. Fichier pour application tableur : Oui. OK.

8.2.3. Présentation graphique

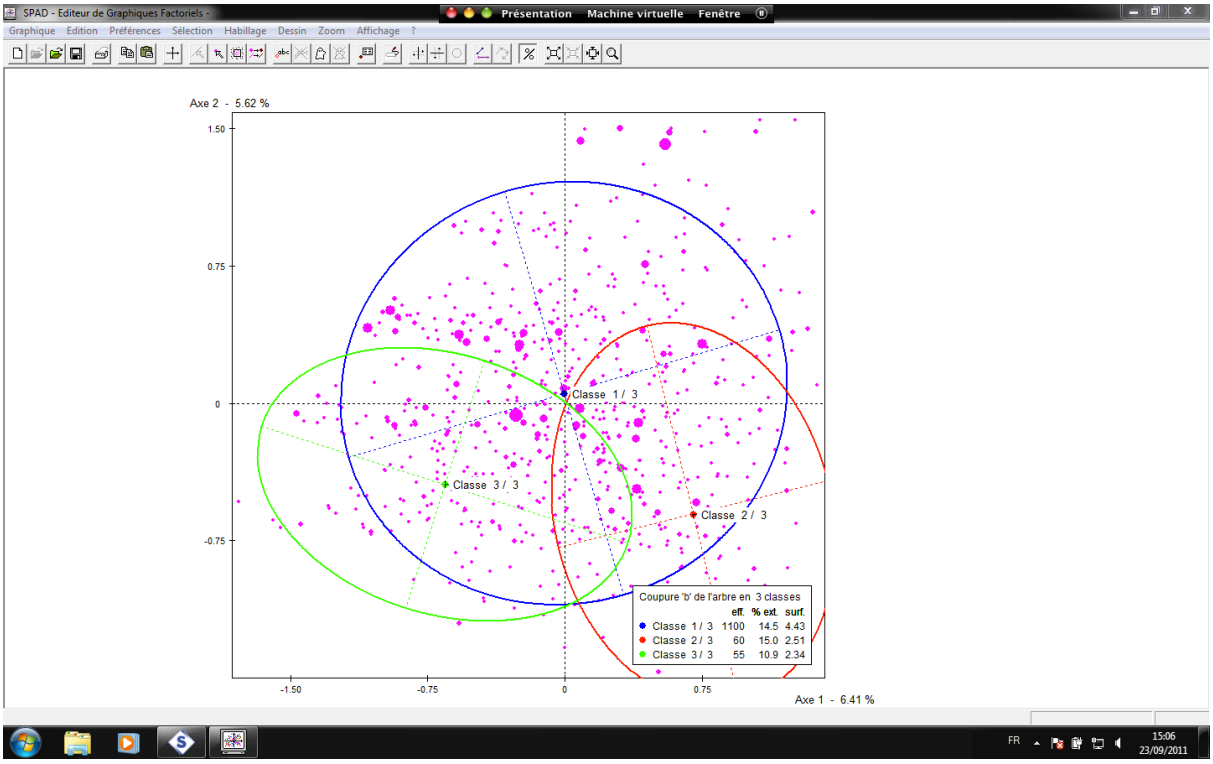
On fait clic droit sur « Coupure de l'arbre », Résultats Editeur graphique – Nouveau graphique.

Habillage – Habillage des individus par une nominale ou une partition – Dans fenêtre : Partitions. Deux partitions à disposition (celles qu'on a demandé tout à l'heure). On prend 2 classes. Couleur : bleu et rouge. On coche la case Ellipse. OK.



La classe rouge est petite : c'est la fameuse classe à 5% d'individus (55 individus). On voit la partition en 2 classes.

On va tenter en trois classes. On refait la même manœuvre, mais on sélectionne 3 classes.



La classe rouge est devenue verte, mais elle n'a pas bougé, elle comprend toujours 55 personnes. En revanche, la classe bleue a été divisée en deux, avec la classe rouge. On a donc retiré deux classes d'individus particuliers. On va essayer d'identifier ce qui les distingue des autres (pour l'instant, on ne le sait pas).

8.2.4. Caractérisation des classes de la typologie (Class Miner) relativement aux variables actives

On quitte le graphique, mais on le garde en tête.

8.2.4.1. Class Miner

Dans méthode : **Caractérisation des classes de la typologie. On le glisse, il se transforme en Class Miner. Clic droit, Paramétrer**. On va d'abord s'occuper des variables actives. **Nominales caractérisantes : on sélectionne Art, Eat, Film et TV (les 4 actives)**. On va dans **l'onglet Paramètres : Caractérisation des classes par… Les modalités. On clique Options. Critère de tri des modalités : Valeur test. Sélection des modalités : Toutes. Poids relatif minimal d'une modalité : 2.0. Modalités conservées : Si sur ou sous-représentées. OK. Les nominales on touche pas. « Les continues » sont sélectionnées. Les fréquences on ne sélectionne pas. OK**.

8.2.4.2. Résultats Excel

Clic droit. On regarde les résultats en termes de sortie Excel.

Sur Excel : **Dans la feuille Decla%, on commence par supprimer la colonne de la Valeur-Test (ici la colonne F). Dans la colonne H, on calcule la différence entre le pourcentage de la modalité dans la classe et celle dans l'échantillon. On met une formule dans la colonne H pour faire la soustraction entre C5 et D5.** Puis on tire la formule vers le bas (jusqu'à tout en bas de la feuille) pour l'avoir sur toute la colonne G.

Dans la colonne **I, on fait C5 divisé par D5**. On tire la formule vers le bas pour avoir toute la colonne I.

On voit que dans le premier tableau (première classe d'individus), les différences sont très faibles (même si elles sont significatives, cf. colonne Probabilité). Ce qui confirme ce que nous avions vu dans le graphique : ce premier groupe rassemble l'écrasante majorité de l'échantillon, donc des individus statistiquement « banals ».

Dans le deuxième tableau (deuxième classe d'individus), on voit que 100% regardent des films d'horreur. La différence par rapport à la population générale est de 94.9%. Ce qui est énorme. Ces différences sont significatives pour les 4 premières lignes. **Deux critères : Différence supérieure à 5 ET Probabilité=significativité plus petite que 5%.** On retient donc les 4 premières lignes : Horror, Tv-Comedy, IndianRest et StillLife.

Tout en bas du tableau, on voit des différences négatives : personne ne regarde de film d'action, de Comedy, etc. Mais ces différences négatives sont moins évidentes à mobiliser pour caractériser la classe.

On passe au tableau suivant, la troisième classe d'individus. Les trois premières lignes sont significatives avec des différences supérieures à 5 : RenaissanceArt, CostumeDrama et ItalianRest. **On quitte Excel.**

8.2.5. Caractérisation des classes de la typologie (Class Miner) relativement aux variables supplémentaires

8.2.5.1. Class Miner

On reprend le Class Miner. **Paramétrer. On sélectionne Age, Gender et Income (variables supplémentaires). On refait exactement le même chemin qu'avec les variables actives.**

8.2.5.2. Résultats Excel

On sort un tableau Excel et on refait la même interprétation (sur les variables supplémentaires). On voit que la modalité « jeune » (18-24 ans) est surreprésentée dans la première classe. Etc.

9. Quitter SPAD et enregistrer le projet

Note : Quand on quitte SPAD, tout ce qui a été fait, on le retrouve. Si on veut transmettre notre projet à quelqu'un, on fait **Enregistrer sous. Dans Archive du projet, on sélectionne Inclure les données, sélectionner Verrouiller les méthodes d'imports**. On peut ainsi envoyer le projet par mail. Mais ça fait des projets assez lourds.

Exemples d'applications de l'AGD en sociologie

1. Présentations de Frédéric Lebaron

1.1. L'engagement statistique de Bourdieu

Cf. le texte de Lebaron (2010).

Salah Bouhedja : il est l'informaticien qui modélise les analyses de Bourdieu, dans l'anatomie du goût et La Distinction. Méthode qui préfigure ce que sera l'ACM.

Avec l'article « Le patronat », c'est véritablement le premier usage de l'ACM.

Dans une note de l'article « Le champ économique », Bourdieu explique que l'ACM représente une formalisation de la notion de champ.

La théorie ne serait pas ce qu'elle est si Bourdieu n'avait pas utilisé l'ACM. Cependant, dire qu'il y a affinité élective entre une méthode et le concept de champ ne veut pas dire que l'usage de l'ACM soit réservé au concept de champ.

Il y a refus de l'opposition entre méthode « descriptive » (qui serait l'ACM) et méthode « explicative » (qui serait la régression). L'explication relève toujours de la théorie sociologique, et jamais de l'outil en tant que tel. Rien ne s'oppose donc à mobiliser un raisonnement explicatif dans le cadre d'une ACM.

La prosopographie se développe avec la notion de champ. Elle est basée sur des dictionnaires biographiques etc. et non plus sur le questionnaire. Bourdieu ne parle d'ailleurs plus de variables, mais de « propriétés » des individus (qui peuvent se rapporter tant à des réponses à un questionnaire qu'à des données prosopographiques).

1.2. Le champ du pouvoir norvégien en 2000.

Trois questions de recherche :

1. Quelles sont les principales dimensions du champ du pouvoir norvégien, en termes de types de capitaux ? Quels sont les principes de différenciation dans ce champ ?

2. Quelles sont les fractions les plus mobiles et celles où la reproduction intergénérationnelle est la plus forte ?

3. Existe-t-il des fractions particulièrement homogènes en termes de capital ?

Critère de sélection des individus enquêtés : la position institutionnelle dirigeante. 1710 répondants, soit un taux de retour de 87.3% (passation en face-à-face et par téléphone). 85% d'hommes, 62% de niveau de diplôme supérieur, 50% avaient un revenu annuel supérieur à 1 million de couronnes (150'000 euros), âgés en moyenne de 52 ans.

Méthode : ACM spécifique : des modalités de « non réponse » ou à très faibles effectifs sont définies comme passives. 6 grandes rubriques : capital économique, capital scolaire personnel acquis, capital scolaire hérité ou « par alliance » (diplômes du conjoint), capital social personnel, capital social hérité, capital d'expérience professionnelle.

Revenu : codé en trois catégories (haut, moyen et bas) non égales : les catégories haut et bas sont plus faiblement dotées, de manière à être cohérent avec l'ACM, où les modalités moins dotées sont situées plus « loin » du centre.

Le capital social hérité se mesure par la position des ascendants : appartenance à un conseil d'administration, à une ONG, au Parlement, etc.

Trois questions du deuxième article :

1. Comment le capital social institutionnalisé est-il distribué dans ce champ ?

2. Quelle est la relation entre ce capital et les autres formes de capitaux ?

3. Y a-t-il des zones de plus forte « endogamie » (définie ici en termes de relations sociales institutionnalisées) ?

2. Frédéric Lebaron et Philippe Bonnet : Les pratiques culturelles des Français

1. Enjeux théoriques

Même perspective que celle de la Distinction : relation entre les pratiques culturelles, les styles de vie et la distribution des différentes espèces de capital dans la société française. Touche aux questions de la bonne volonté culturelle, de la culture légitime, de l'habitus petit-bourgeois, du goût de nécessité, etc. On peut aller plus loin que Bourdieu, car on dispose de nouveaux outils.

2. Les données

Enquête permanente de l'INSEE. Données de 2003, 5'626 individus (15 ans et plus). Etude peu utilisée par les sociologues (sauf Philippe Coulangeon, qui vient de sortir quelque chose en ayant partiellement recours aux mêmes données).

3. Construction de l'espace des pratiques culturelles

Questions sur les pratiques légitimes (lecture, écoute de musique classique), les pratiques caractéristiques d'un style de vie jeune (styles de musique, types de radio), les pratiques populaires. Il n'y a pas de questions d'attitudes dans cette enquête de l'INSEE. Les questions ne portent que sur des pratiques.

Les variables qui définissent l'espace :

10 questions sur la télévision

8 questions sur les sorties (théâtre…)

5 questions sur journaux et magazines

8 questions sur les livres

2 questions sur l'écoute de musique/radio

Il y a plus de questions sur la TV que sur les autres domaines. Mais le nombre de modalités est plus équilibré que cela. Les deux questions sur la musique/radio ont par exemple beaucoup de modalités de réponses. **Il faut toujours considérer le nombre de modalités actives moins le nombre de questions par thème pour équilibrer les choses.**

Il y a 33 questions en tout, et 104 modalités.

Une ACM spécifique a été pratiquée. Elle est utilisée pour restreindre l'analyse aux catégories d'intérêt. Les catégories peu fréquentes et celles sans signification ont été

mises en « passif ». Le nombre de catégories actives est K=90. Le nombre d'individus est de 5'497 (n'ont retenu que les 18 ans et plus).

4 axes ont été interprétés. Car les valeurs propres chutent entre le 4ème et le 5ème. Le taux d'importance est de 90.45%.

Contributions des thèmes aux trois premiers axes : le 1er axe est influencé par les spectacles et les livres. La TV caractérise les 2ème et surtout 3ème axe. Etc.

Dans le graphique, 36 catégories ont été retenues. Ce sont celles qui caractérisent le plus la variance du 1er axe. Il représente le capital culturel : à droite, les pratiques légitimes, à gauche, des pratiques moins légitimes (ne pas lire, ne pas aller au spectacle, regarder TF1).

Sur l'axe 2, 29 catégories sur 90 ont été retenues (celles qui le déterminent le plus). En haut, regarder M6. Musique Pop internationale, Techno. Lire des BD. Il s'agit de la culture jeune. En bas, Aller à l'opéra, écouter la radio, lire des quotidiens nationaux, etc., bref, culture plus classique.

Sur l'axe 3, 28 catégories sur 90 ont été retenues. Cet axe oppose les pratiques culturelles « hard » (rock, techno, BD, cinéma, etc. mais pas d'art, pas de roman, etc.) opposé à quelque chose de plus calme (théâtre, lire des romans sentimentaux, regarder France2). Donc probablement lié à l'âge et au sexe.

On passe au nuage des individus sur les axes 1 et 2. Enorme concentration à gauche, puis se disperse à droite. Dans le plan 1 et 3, on a toujours cette concentration à gauche.

Exploration du nuage. On met le niveau d'éducation comme facteur structurant. Puis on met le sexe. Par rapport au sexe, on voit que les deux ellipses sont superposées : donc il n'y a pas de différences entre hommes et femmes (ce qui représente un résultat).

On prend l'âge comme facteur structurant (en liant les points entre eux, car variable ordinale, donc pratique pour la représentation graphique). Les jeunes se distinguent beaucoup des vieux, surtout sur l'axe 2. On cible les jeunes par une ellipse de concentration. On voit que, pour une bonne partie, ils se trouvent vers le haut de l'axe 2.

On prend la classe sociale comme facteur structurant. A partir des CSP françaises en 42 catégories, on regroupe en 5 classes. On fait les ellipses de concentration. L'ellipse des ouvriers est concentrée à gauche de l'axe 1 (peu de pratiques et pratiques peu légitimes). Ils ont l'ellipse qui a la surface la plus petite, ce qui peut s'interpréter comme un signe d'homogénéité sur le premier axe. On retrouve donc les résultats de La Distinction (qui ont pourtant été beaucoup critiqués…).

Ont sélectionné 180 managers pour voir comment ils se situent. Pôle économique de la classe dominante. On voit qu'ils ont une ellipse qui est pas mal à droite sur l'axe 1. Ils sont surtout très étendus sur cet axe, ce qui montre une certaine dispersion.

Classification. Classification ascendante hiérarchique. 4 classes ont été retenues. Dont une à seulement 4% de la population (217 individus). On fait les ellipses de concentration de ces 4 classes dans le plan 1-2. L'idée est d'interpréter ces 4 classes sur la base des valeurs actives et de quelques variables supplémentaires.

Première classe. 217 personnes. C'est l'élite ou l'avant-garde culturelle légitime. Questions actives : 95% vont à l'opéra, 67% vont au théâtre, etc. La UpperClass et les Managers y sont largement surreprésentés.

Bourdieu construit des habitus à partir de styles de vie : bonne volonté culturelle, distinction, etc. Ce que montre la partie droite de l'axe 1, c'est que cet habitus est en grande partie un habitus de classe.

Deuxième classe. C'est la « bonne volonté culturelle ». Le point moyen est plutôt à droite, mais les axes de l'ellipse sont parallèles aux axes généraux. L'art les intéresse, mais plutôt la télé. Les pratiques sont assez intenses, mais peu légitimes. On est dans la « bonne volonté culturelle » version années 2000 (à discuter sur le plan sociologique). Les classes sup y sont surreprésentées, mais la classe moyenne aussi.

Troisième classe : La culture jeune. Regardent des clips, des sitcoms, etc. Il y a plus de jeunes que de vieux.

Quatrième classe : populaire traditionnelle. La modalité « never » est surreprésentée dans les valeurs actives. Individus peu éduqués et âgés surreprésentés.

Conclusion : Les nuages d'individus contiennent toute l'information : l'investigation systématique des nuages et sous-nuages d'individus en utilisant largement plusieurs facteurs structurants s'avère très pertinente.