2.5 Méthodes de classification

Bibliographie : Lebart, L., Morineau, A., Piron M., Analyse exploratoire multidimensionnelle, Dunod, Paris, 2000.

2.5.1 Introduction

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère. Les diverses techniques de classification (ou d'"analyse typologique", de "taxonomie", ou "taxinomie" ou encore "analyse en clusters" (amas)) visent toutes à répartir *n* individus, caractérisés par *p* variables X_1, X_2, ..., X_p en un certain nombre m de sous-groupes aussi homogènes que possible.
On distingue deux grandes familles de techniques de classification :

- La classification non hiérarchique ou partitionnement, aboutissant à la décomposition de l'ensemble de tous les individus en m ensembles disjoints ou classes d'équivalence ; le nombre m de classes est fixé.
- La classification hiérarchique : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

Remarques. Ces méthodes jouent un rôle un peu à part dans l'univers des méthodes statistiques. En effet :

- L'aspect inférentiel est ici inexistant ;
- Il existe un grand nombre de variantes de ces méthodes, et on peut être amené à appliquer plusieurs de ces méthodes sur un même jeu de données, jusqu'à obtenir une classification "qui fasse sens" ;
- Au contraire des méthodes factorielles, l'accent est souvent mis sur les n individus et non sur les p variables qui les décrivent.

2.5.2 Méthodes de type "centre mobile" : K-moyennes

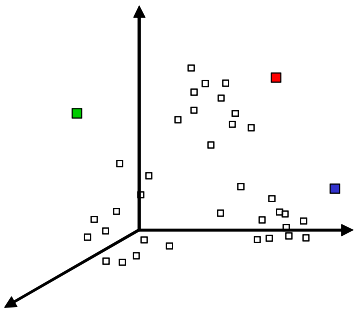
2.5.2.1 Principe de la méthode

On dispose d'un ensemble d'individus, ou observations, décrits par des variables numériques. On veut créer une partition de cet ensemble, en regroupant ces individus en un nombre déterminé K de classes : chaque individu devra appartenir à une classe et une seule. Pour cela :

On fixe de façon aléatoire K "centres de classes", ou "centres de gravité" et on exécute l'algorithme suivant :

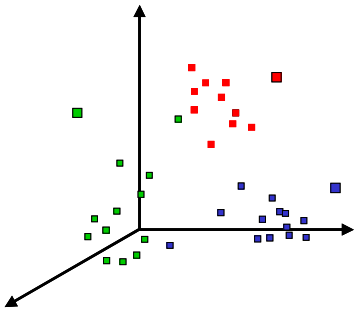
1) Chaque observation est classée en fonction de sa proximité au centre de gravité.
2) Chaque centre de gravité est déplacé de façon à être au centre du groupe correspondant.
3) On continue jusqu'à ce que les centres de gravité ne bougent plus

Méthodes de type « centres mobiles »



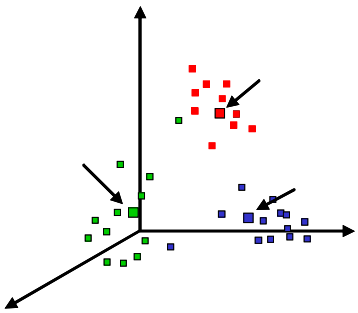
Au départ

Création aléatoire de centres de gravité.



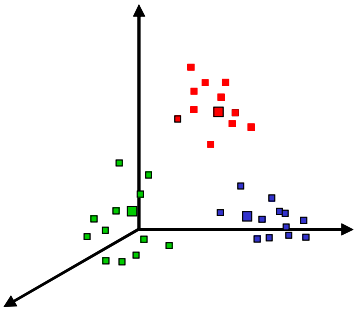
Etape 1

Chaque observation est classée en fonction de sa proximité aux centres de gravités.



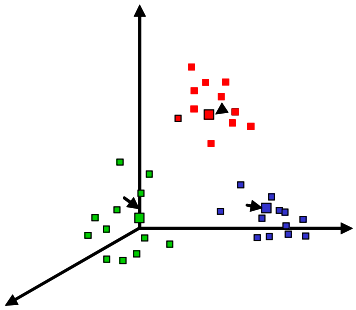
Etape 2

Chaque centre de gravité est déplacé de manière à être au centre du groupe correspondant



Etape 1’

On répète l’étape 1 avec les nouveaux centres de gravité.



Etape 2'

De nouveau, chaque centre de gravité est recalculé

On continue jusqu'à ce que les centres de gravité ne bougent plus

Choix des variables représentant les individus

Les distances étant calculées sur les valeurs observées des variables, la classification n'aura pas de sens si les variables s'expriment avec des unités différentes, et ont des plages de variation très différentes. Si c'est le cas, il faut au préalable transformer les variables (par exemple en faisant un centrage-réduction) afin d'équilibrer les "poids" des différentes variables.

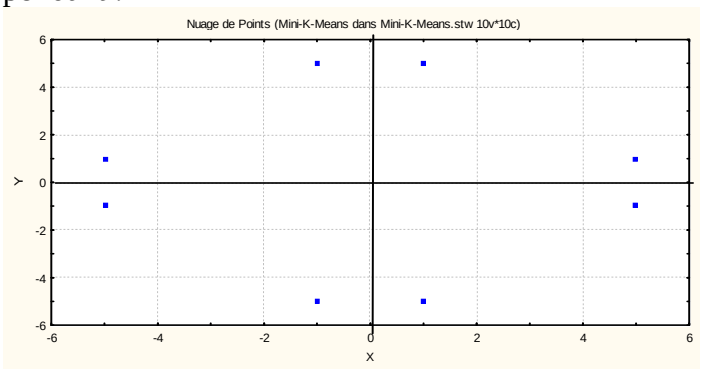
Dans le cas où les données observées sont les valeurs de p variables numériques sur n individus, on pourra choisir d'effectuer une classification des individus, ou une classification des variables. On peut choisir, par exemple, de retenir certains "traits" des individus (autrement dit certaines variables qui ont servi à les décrire) et réaliser la classification sur les individus décrits par ce choix de variables.

2.5.2.2 Mise en oeuvre avec Statistica sur un mini-exemple

On dispose de 8 individus décrits par 2 variables. Une troisième variable est constante sur l'ensemble des individus. Les données sont les suivantes :

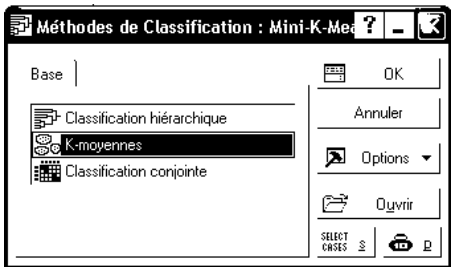
	X	Y	Z
1	5	1	10
2	5	-1	10
3	1	5	10
4	-1	5	10
5	-5	1	10
6	-5	-1	10
7	1	-5	10
8	-1	-5	10

Nuage de points correspondant :

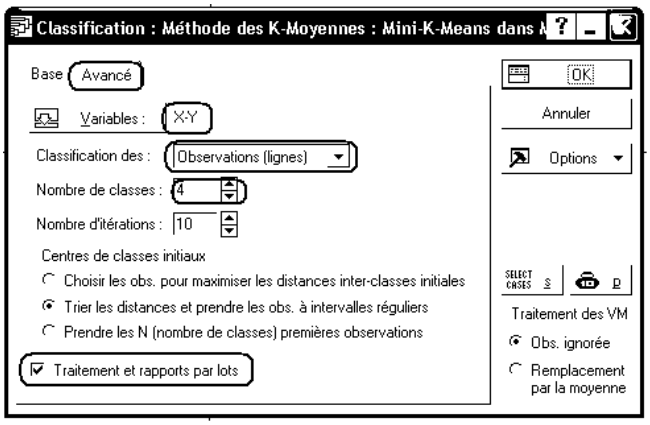


Nous souhaitons définir 4 classes à partir de ces 8 objets.

Ouvrez le classeur Mini-K-Means.stw.
Utilisez le menu Statistiques - Techniques exploratoires multivariées - Classifications et sélectionnez la méthode K-moyennes.



Sélectionnez X et Y comme variables d'analyse, et, sous l'onglet "Avancé", spécifiez une classification sur les observations, comportant 4 classes. Cochez également la case "traitements et rapports par lots", ce qui permettra de produire en une seule manipulation l'ensemble des résultats de la classification.



Comme prévu, les 4 classes formées par Statistica sont {O1, O2}, {O3, O4}, {O5, O6} et {O7, O8} (cf. les 4 feuilles de résultats "composition de la classe N° ...). Par exemple, pour la première classe :

	Composition de la Classe 1 et Distances au Centre de Classe Respecti Classe avec 2 obs.	
	Obs. # O_1	Obs. # O_2
Distance	0,707	0,707

Le centre C_1 de cette classe est évidemment le point de coordonnées (5, 0). On peut remarquer que la distance calculée par Statistica n'est pas tout à fait la distance euclidienne dans le plan, mais correspond à la formule suivante :

$$d^2(O_1, C_1) = \frac{(x_1 - \bar{x})^2 + (y_1 - \bar{y})^2}{2}$$

Le dénominateur introduit dans la formule représente le nombre de variables, comme on peut s'en rendre compte en introduisant la troisième variable (Z) dans la classification.

La même règle est appliquée pour le calcul des distances entre classes, autrement dit entre centres de classes :

Classe (Numéro)	Distances Euclidiennes Inter-Classes Dist. sous la diagonale (Dist.) ² au dessus de la diagonale			
	N°1	N°2	N°3	N°4
N°1	0,00	25,00	50,00	25,00
N°2	5,00	0,00	25,00	50,00
N°3	7,07	5,00	0,00	25,00
N°4	5,00	7,07	5,00	0,00

Les coordonnées des centres de classes sont disponibles dans la feuille de résultats "Moy. Classes" :

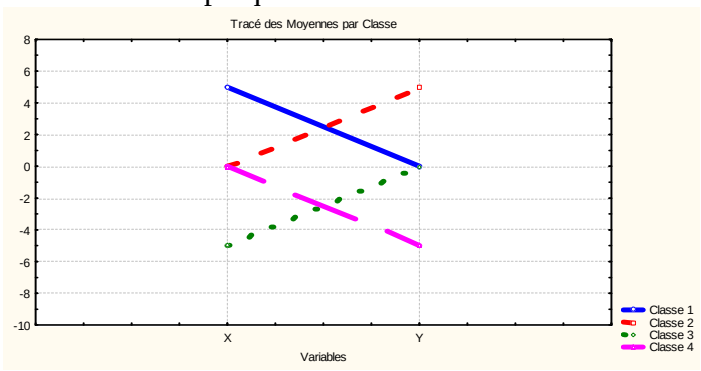
Variable	Moy. Classes (Mini-K-Means dans Mini-K-Means.stw)			
	Classe N°1	Classe N°2	Classe N°3	Classe N°4
X	5,00	0,00	-5,00	0,00
Y	0,00	5,00	0,00	-5,00

Statistica effectue également une analyse de variance à un facteur sur chacune des variables. Le facteur pris en compte ici est l'appartenance de l'observation à l'une des classes :

Variable	Analyse de Variance					
	SC Inter	dl	SC Intra	dl	F	signif. p
X	100,00	3	4,00	4	33,33	0,0027
Y	100,00	3	4,00	4	33,33	0,0027

Ces résultats peuvent être retrouvés à l'aide du menu ANOVA. On introduit une quatrième variable, nommée "Groupe", contenant le numéro de la classe à laquelle appartient l'observation. Puis, on effectue une analyse de variance à un facteur en indiquant X (par exemple) comme variable dépendante et Groupe comme variable de classement.

Le seul résultat qui n'est pas automatiquement produit par le traitement par lots est le graphique des moyennes. Pour l'obtenir, ré-affichez la fenêtre du traitement en cours, désactivez la case "traitement par lots" et cliquez sur OK. Dans la fenêtre de dialogue "Résultats de l'analyse par les k-moyennes", cliquez sur le bouton "Graphique" :



2.5.2.3 Mise en oeuvre sur les exemples traités dans les paragraphes ACP et AFC

Classification des variables du cas "Représentations sociales de l'homosexualité"

On reprend l'exemple "Représentations sociales de l'homosexualité" que nous avions traité par une ACP (classeur Statistica Rep-Soc-Homo.stw). Rappelons que les variables sont ici homogènes, puisque chaque variable est un protocole de rangs observés sur les 15 traits étudiés.

Une classification en 3 classes, portant sur les variables va-t-elle confirmer les résultats que nous avions obtenus en analysant les résultats de l'ACP ?

	Repr-Soc-Homo dans Rep-Soc-Homo-correction.st		
	1 VARIABL	2 CLASSE	3 DISTANC
He:H	1	3	1,79
Ho:H	2	3	1,43
He:Soi	3	3	2,36
Ho:Soi	4	1	1,21
Ho:Ho	5	1	1,21
Ho:F	6	2	1,48
He:Ho	7	2	1,10
He:F	8	2	1,42

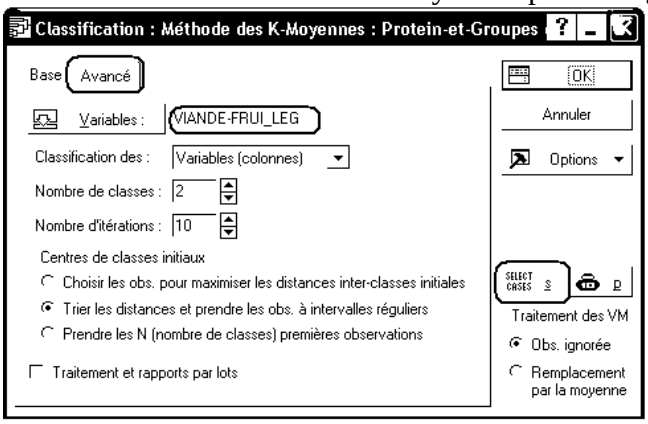
On constate que la classe 3 regroupe les variables correspondant à une cible masculine, la classe 1 regroupe les jugements portés par les homosexuels sur eux-mêmes et sur leur stéréotype, tandis que la classe 2 rassemble non seulement les variables correspondant à une cible féminine mais aussi He:Ho, c'est-à-dire la description de la cible "homosexuels" faite par les hétérosexuels.

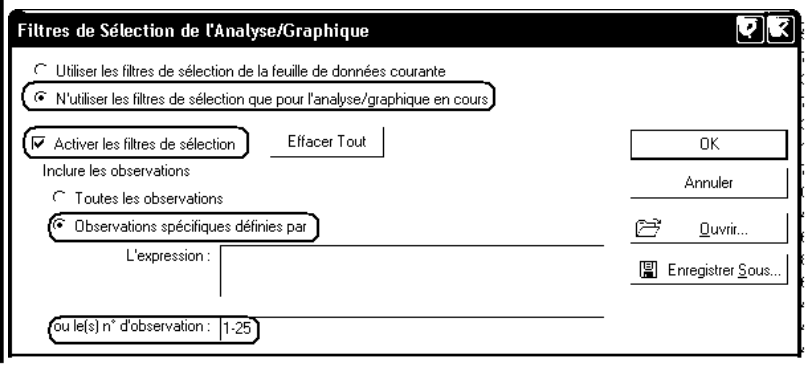
Classifications sur le cas "Protéines"

On reprend le fichier Proteines-2008.stw.

La répartition en 2 groupes "protéines animales v/s protéines végétales" apparaît-elle naturellement dans les données étudiées ?

Effectuez une classification de type K-moyennes, portant sur les variables 1 à 9 de la feuille "Protein et Groupes" en indiquant deux classes. Faites une sélection des observations, de manière à éliminer de l'étude les moyennes par zone géographique :





On voit que l'une des classes est constituée de la seule variable "céréales" pendant que l'autre classe rassemble les 8 autres variables. En effet, l'étendue de la variable "Céréales" est très différente de celle des autres variables, et le résultat produit ne fait que l'illustrer.

On peut résoudre ce problème soit en travaillant sur des données centrées réduites, soit en utilisant les coordonnées des variables selon les axes factoriels produites par une ACP normée. Par exemple, activez la feuille "Proteines-Centre-Reduit". Reprenez une classification analogue, mais portant sur les variables centrées-réduites. Cette fois, la classification recouvre assez bien l'origine (animale v/s végétale) des protéines, mais les féculents restent regroupés avec les protéines animales :

	Proteines-Centre-Reduit dans Proteines-2008.st\		
	1 VARIABL	2 CLASSE	3 DISTANC
VIANDE	1	2	0,75
PORC_VOL	2	2	0,82
OEUFS	3	2	0,54
LAIT	4	2	0,69
POISSON	5	2	0,94
CEREALES	6	1	0,65
FECULENT	7	2	0,73
NOIX	8	1	0,46
FRUI_LEG	9	1	0,77

Classification des lignes dans le cas "Régions-2001"

On reprend le classeur Statistica Regions-2001.stw.

Une classification basée sur le tableau de contingence n'aurait pas grand sens. En revanche, on peut utiliser les résultats de l'AFC comme données de base pour essayer de faire une classification des régions en 3 ou 4 ensembles.

Refaites au besoin une AFC sur ce tableau de contingence et rendez active la feuille contenant les résultats relatifs aux individus lignes (les régions). Faites ensuite une classification de type "K-moyennes", en utilisant les variables "Coord." de cette feuille et en spécifiant 3 ou 4 classes. Vous devriez retrouver en grande partie la typologie que nous avions obtenue en analysant les résultats de l'AFC.

Remarque. Les résultats de la classification dépendent-ils du nombre d'axes factoriels représentés dans la feuille de résultats de l'AFC ? On pourra essayer de refaire la classification sur les coordonnées factorielles d'un plus grand nombre d'axes, et constater qu'il en résulte peu de modifications des résultats produits : l'essentiel de la variation est représenté par les premiers axes.

2.5.2.4 Remarques et conclusion

Cette méthode produit des résultats qui peuvent être facilement exploitables. On notera cependant que l'on doit indiquer a priori le nombre de classes, ce qui nuit à l'aspect véritablement "exploratoire" de la méthode. D'autre part, les variables traitées doivent être homogènes (s'exprimer avec la même unité, ou au moins avoir la même plage de variation) et c'est toujours la distance euclidienne qui est utilisée pour évaluer les distances entre objets.

2.5.3 Classification Ascendante Hiérarchique

2.5.3.1 Les 4 étapes de la méthode

Choix des variables représentant les individus

Les distances étant calculées sur les valeurs observées des variables, la classification n'aura pas de sens si les variables s'expriment avec des unités différentes, et ont des plages de variation très différentes. Si c'est le cas, il faut au préalable transformer les variables (par exemple en faisant un centrage-réduction) afin d'équilibrer les "poids" des différentes variables.

Dans le cas où les données observées sont les valeurs de p variables numériques sur n individus, on pourra choisir d'effectuer une classification des individus, ou une classification des variables. On peut choisir, par exemple, de retenir certains "traits" des individus (autrement dit certaines variables

qui ont servi à les décrire) et réaliser la classification sur les individus décrits par ce choix de variables.
On peut noter qu'il revient au même par exemple :

- de réaliser la CAH des individus à partir de p variables centrées réduites ;
- de réaliser la CAH des individus à partir des p facteurs obtenus à l'aide d'une ACP normée sur les variables précédentes.

Toutefois, il peut être intéressant de réaliser la CAH à partir des q premiers facteurs (q<p). Cela a pour effet d'éliminer une partie des variations entre individus, qui correspond en général à des fluctuations aléatoires, c'est-à-dire à un "bruit statistique".

Dans le cas où les données observées sont représentées par un tableau de contingence, c'est-à-dire sont les valeurs de 2 variables nominales sur n individus, on pourra effectuer une CAH des modalités-lignes par exemple, à partir des coordonnées lignes obtenues par une AFC. On pourra, de même, réaliser une CAH des modalités-colonnes.

Enfin, si les données observées sont les valeurs de p variables nominales sur n individus, on pourra effectuer une CAH des individus en partant du tableau disjonctif complet, ou en utilisant les coordonnées des individus obtenues par une ACM. On pourra également traiter les modalités comme dans le cas d'une AFC.

Choix d'un indice de dissimilarité

De nombreuses mesures de la "distance" entre individus ont été proposées. Le choix d'une (ou plusieurs) d'entre elles dépend des données étudiées. Statistica nous propose les mesures suivantes :

- Distance Euclidienne. C'est probablement le type de distance le plus couramment utilisé. Il s'agit simplement d'une distance géométrique dans un espace multidimensionnel.
$$d(I_i,I_j)=\sqrt{\sum_k (x_{ik}-x_{jk})^2}$$
- Distance Euclidienne au carré. On peut élever la distance euclidienne standard au carré afin de "sur-pondérer" les objets atypiques (éloignés).
$$d(I_i,I_j)=\sum_k (x_{ik}-x_{jk})^2$$
- Distance du City-block (Manhattan) :
$$d(I_i,I_j)=\sum_k \left|x_{ik}-x_{jk}\right|$$
- Distance de Tchebychev :
$$d(I_i,I_j)=Max\ \left|x_{ik}-x_{jk}\right|$$
- Distance à la puissance.
$$d(I_i,I_j)=\left(\sum_k \left|x_{ik}-x_{jk}\right|^p\right)^{1/r}$$
- Percent disagreement. Cette mesure est particulièrement utile si les données des dimensions utilisées dans l'analyse sont de nature catégorielle.
$$d(I_i,I_j)=\frac{Nombre\ \ de\ \ x_{ik}\neq x_{jk}}{K}$$
- 1- r de Pearson : calculée à partir du coefficient de corrélation, à l'aide de la formule :
$$d(I_i,I_j)=1-r_{ij}$$

Indices de dissimilarité et distances

On peut également utiliser d'autres indices de dissimilarité puisque Statistica permet d'effectuer la classification à partir du tableau des scores de dissimilarités entre individus. En fait, un indice de dissimilarité doit simplement satisfaire les conditions suivantes :

- non-négativité : $d(I_i,I_j) \geq 0$
- symétrie : $d(I_i,I_j) = d(I_j,I_i)$
- normalisation : $d(I_i,I_i) = 0$

Un indice de dissimilarité est une "vraie" distance, s'il vérifie également l'inégalité triangulaire :

$$d(I_i,I_j) \leq d(I_i,I_k) + d(I_k,I_j).$$

La plupart des "distances" proposées par Statistica sont de véritables distances.

De nombreux indices de dissimilarité (ou au contraire de similarité) ont été proposés dans le cas de variables qualitatives (à deux modalités, ou après codage disjonctif). Par exemple, si les individus sont décrits par K variables dichotomiques (oui/non), on peut introduire :

$a_{ij} =$ *Nombre co−occurrences entre les individus i et j*

$d_{ij} =$ *Nombre co−absences entre les individus i et j*

$b_{ij} =$ *Nombre d'attributs présents chez i et absents chez j*

$c_{ij} =$ *Nombre d'attributs absents chez i et présents chez j*

On peut proposer par exemple, comme indice de dissimilarité :

$$d(I_i,I_j) = \sqrt{b_{ij} + c_{ij}}$$

ou au contraire, comme indice de similarité :

$$s(I_i,I_j) = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

Un indice de similarité peut être converti en distance par la relation :

$$d(I_i,I_j) = s_{\max} - s(I_i,I_j)$$

Choix d'un indice d'agrégation

L'application de la méthode suppose également que nous fassions le choix d'une "distance" entre classes. Là encore, de nombreuses solutions existent. Il faut noter que ces solutions permettent toutes de calculer la distance entre deux classes quelconques sans avoir à recalculer celles qui existent entre les individus composant chaque classe.

Les choix proposés par Statistica sont les suivants :

- Saut minimum ou "single linkage" (distance minimum). C'est celle que nous avons utilisée ci-dessus.

- Diamètre ou "complete linkage" (distance maximum). Dans cette méthode, les distances entre classes sont déterminées par la plus grande distance existant entre deux objets de classes différentes (c'est-à-dire les "voisins les plus éloignés").

$$D(A,B) = \max_{I \in A} \max_{J \in B} d(I,J)$$

- Moyenne non pondérée des groupes associés. Ici, la distance entre deux classes est calculée comme la moyenne des distances entre tous les objets pris dans l'une et l'autre des deux classes différentes.

$$D(A,B) = \frac{1}{n_A n_B} \sum_{I \in A, J \in B} d(I,J)$$

- Moyenne pondérée des groupes associés. La moyenne précédente est étendue à l'ensemble des paires d'objets trouvées dans la réunion des deux classes.

$$D(A,B) = \frac{1}{(n_A + n_B)(n_A + n_B - 1)} \sum_{I,J \in A \cup B} d(I,J)$$

- Centroïde non pondéré des groupes associés. Le centroïde d'une classe est le point moyen d'un espace multidimensionnel, défini par les dimensions. Dans cette méthode, la distance entre deux classes est déterminée par la distance entre les centroïdes respectifs.

- Centroïde pondéré des groupes associés (médiane). Cette méthode est identique à la précédente, à la différence près qu'une pondération est introduite dans les calculs afin de prendre en compte les tailles des classes (c'est-à-dire le nombre d'objets contenu dans chacune).

- Méthode de Ward (méthode du moment d'ordre 2). Cette méthode se distingue de toutes les autres en ce sens qu'elle utilise une analyse de la variance approchée afin d'évaluer les distances entre classes. En résumé, cette méthode tente de minimiser la Somme des Carrés (SC) de tous les couples (hypothétiques) de classes pouvant être formés à chaque étape. Les indices d'agrégation sont recalculés à chaque étape à l'aide de la règle suivante : si une classe M est obtenue en regroupant les classes K et L, sa distance à la classe J est donnée par :

$$D(M,J) = \frac{(N_J + N_K)D(K,J) + (N_J + N_L)D(L,J) - N_J D(K,L)}{N_J + N_K + N_L}$$

La méthode de Ward se justifie bien lorsque la "distance" entre les individus est le carré de la distance euclidienne. Choisir de regrouper les deux individus les plus proches revient alors à choisir la paire de points dont l'agrégation entraîne la diminution minimale de l'inertie du nuage. Le calcul des nouveaux indices entre la paire regroupée et les points restants revient alors à remplacer les deux points formant la paire par leur point moyen, affecté du poids 2.

Algorithme de classification et résultat produit

L'algorithme de classification

La classification proprement dite peut être décrite de la manière suivante : :

Étape 1 : il y a n éléments à classer (qui sont les n individus);

Étape 2 : on construit la matrice de distances entre les n éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à n-1 classes;

Étape 3 : on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement (n-1) éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec n-2 classes et qui englobe la première;

Étape m : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

Hiérarchie de classes et partition de l'ensemble des individus

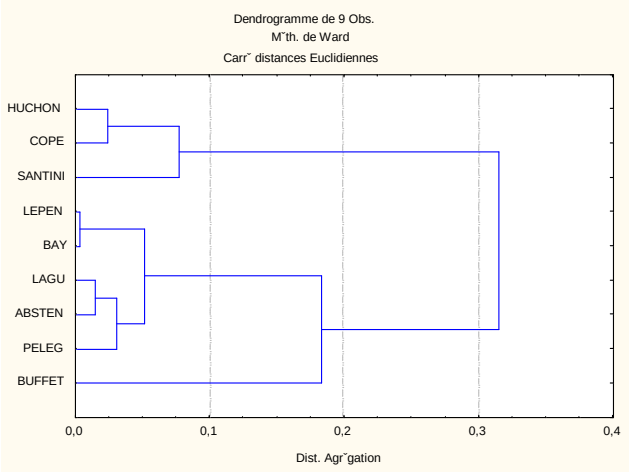
Opérer une classification, c'est définir une partition de l'ensemble des individus, c'est-à-dire, définir un ensemble de parties, ou classes de l'ensemble I des individus telles que :

- toute classe soit non vide
- deux classes distinctes sont disjointes
- tout individu appartient à une classe.

Le résultat d'une CAH n'est pas une partition de l'ensemble des individus. C'est une hiérarchie de classes telles que :

- toute classe est non vide
- tout individu appartient à une (et même plusieurs) classes
- deux classes distinctes sont disjointes, ou vérifient une relation d'inclusion (l'une d'elles est incluse dans l'autre)
- toute classe est la réunion des classes qui sont incluses dans elle.

Ce résultat est souvent représenté sous forme de dendrogramme :



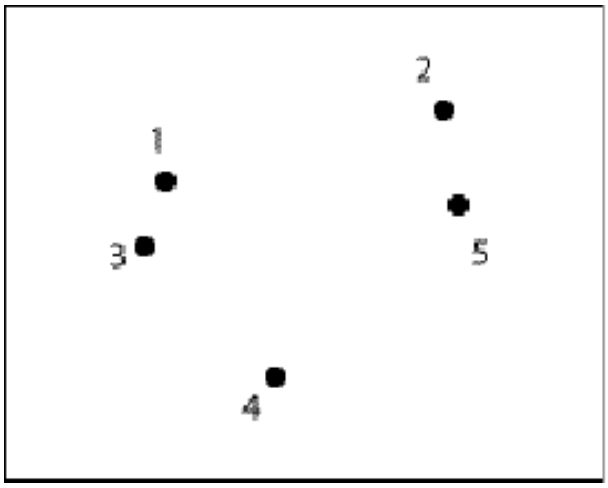
Sur la figure ci-dessus, l'axe vertical indique les individus statistiques qui ont été rassemblés pour former les classes, tandis que la graduation de l'axe horizontal indique la distance séparant les deux classes qui ont été rassemblées une étape donnée.

Choix d'une partition à partir de la hiérarchie des classes

Le dendrogramme nous indique l'ordre dans lequel les agrégations successives ont été opérées. Il nous indique également la valeur de l'indice d'agrégation à chaque niveau d'agrégation. Il est généralement pertinent d'effectuer la coupure après les agrégations correspondant à des valeurs peu élevées de l'indice et avant les agrégations correspondant à des valeurs élevées. En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité car les individus regroupés en-dessous de la coupure étaient proches, et ceux regroupés après la coupure sont éloignés.

2.5.3.2 CAH "à la main"

Le dessin suivant représente 5 objets "en vraie grandeur". La distance utilisée entre les objets est la distance euclidienne (mesurée au double-décimètre). L'indice d'agrégation est celui du "saut minimal".
Réalisez une CAH sur ces données :

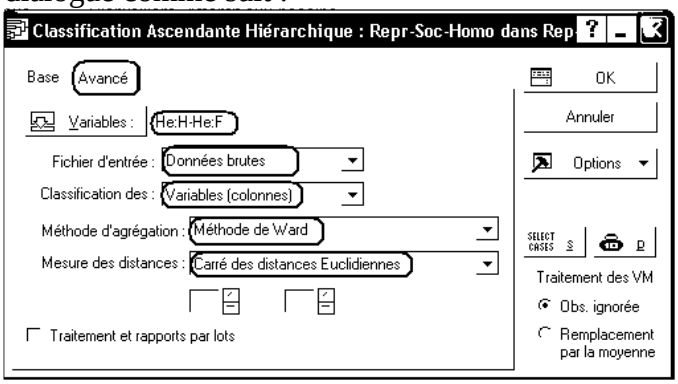


2.5.3.3 La CAH avec Statistica

CAH sur le cas "Représentations sociales de l'homosexualité

On reprend le classeur Rep-Soc-Homo.stw.

Faites une classification ascendante hiérarchique des variables, en utilisant le carré de la distance euclidienne et la méthode de Ward. Pour cela, utilisez le menu Statistiques - Techniques exploratoires multivariées - Classifications. Sélectionnez l'item "Classification hiérarchique" et complétez la fenêtre de dialogue comme suit :

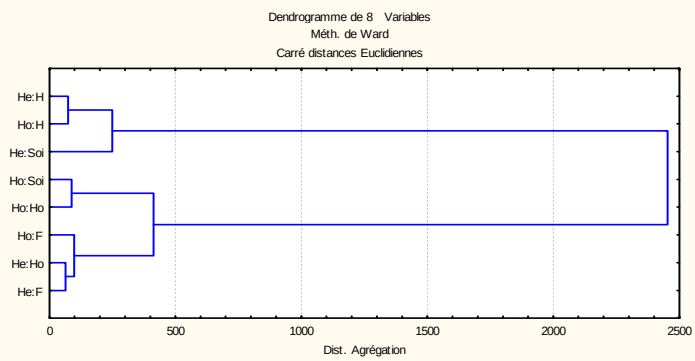


Pour l'essentiel, les résultats de la CAH rejoignent ceux de la classification précédente. Les principaux résultats fournis par Statistica sont les suivants :

La matrice des distances initiales entre les différents objets :

	Carré distances Euclidiennes							
Variable	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F
He:H	0	74	232	788	736	898	1018	1048
Ho:H	74	0	180	694	594	830	920	1028
He:Soi	232	180	0	466	438	634	718	810
Ho:Soi	788	694	466	0	88	246	160	278
Ho:Ho	736	594	438	88	0	182	178	284
Ho:F	898	830	634	246	182	0	72	108
He:Ho	1018	920	718	160	178	72	0	64
He:F	1048	1028	810	278	284	108	64	0

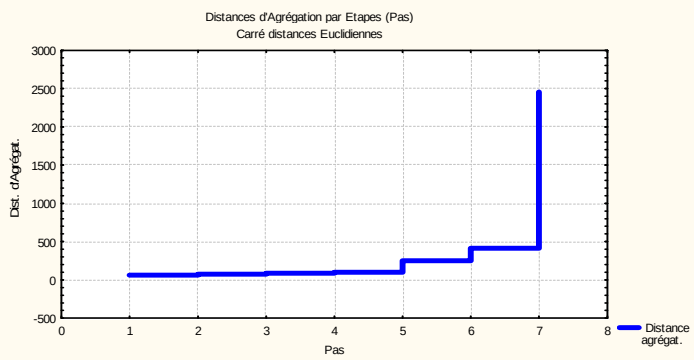
Le dendrogramme de la classification :



Le tableau donnant les différentes étapes de la classification :

distance agrégat.	Agrégation Finale (Repr-Soc-Homo dans Rep-Soc-Homo-correction.stw Méth. de Ward Carré distances Euclidiennes Objet # 1	Objet # 2	Objet # 3	Objet # 4	Objet # 5	Objet # 6	Objet # 7	Objet # 8
64,00	He:Ho	He:F						
74,00	He:H	Ho:H						
88,00	Ho:Soi	Ho:Ho						
98,66	Ho:F	He:Ho	He:F					
250,00	He:H	Ho:H	He:Soi					
413,33	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F			
2453,5	He:H	Ho:H	He:Soi	Ho:Soi	Ho:Ho	Ho:F	He:Ho	He:F

Le graphique de l'agrégation finale, donnant à chaque étape l'indice d'agrégation des deux classes que l'on réunit :



Remarques.

1. Contrairement à d'autres logiciels de traitement statistique, Statistica ne propose pas de faire un centrage réduction des variables avant de faire la CAH. Si les variables retenues pour décrire les individus s'expriment avec des unités différentes, ou ont des plages de variation très différentes, une telle transformation des variables est pourtant indispensable.

2. Statistica ne permet pas de choisir a priori un nombre déterminer de classes et, en conséquence ne fournit pas non plus de table d'appartenance du type suivant (produit par Statgraphics) :

```
Table d'appartenance
--------------------
Méthode de classification: Ward
Distance:  Euclidienne au carré

Variable                Classe
```

```
---------------------- -------
He:H                        1
Ho:H                        1
He:Soi                      1
Ho:Soi                      2
Ho:Ho                       2
Ho:F                        3
He:Ho                       3
He:F                        3
---------------------- -------
```

Ces limitations ne sont guère gênantes sur l'exemple traité ici (8 variables et 15 individus) mais le deviennent lorsque le nombre d'objets à classer est important.

Un exemple de CAH effectué à partir d'un tableau de contingence

Source : Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle.

L'exemple concerne l'analyse d'un tableau de contingence qui croise 8 professions et catégories socioprofessionnelles (PCS) et 6 types de médias pour un échantillon de 12 388 "contacts média" relatifs à 4433 personnes interrogées. L'individu statistique sera pour nous le "contact média" et non la personne interrogée dans l'enquête. Les données sont extraites de l'Enquête Budget-temps Multimédia 1991-1992 du CESP.

Afin d'interpréter plus efficacement les représentations obtenues, on projettera en éléments supplémentaires certaines autres caractéristiques de la population enquêtée telles que le sexe, l'âge, le niveau d'instruction.

Tables de contingence croisant les types de contacts-média (colonnes) avec professions, sexe, âge, niveau d'éducation (lignes).

	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV
Professions						
Agriculteur	96	118	2	71	50	17
Petit patron	122	136	11	76	49	41
Prof. Cad. S.	193	184	74	63	103	79
Prof. interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782

Nous disposons des tables de contingence suivantes (cf. tableau) : On trouve, à l'intersection de la ligne i et de la colonne j le nombre kij d'individus appartenant à la catégorie i et ayant eu la veille (un jour de semaine) au moins un contact avec le type de média j. Une personne interrogée pouvant avoir des contacts avec plusieurs médias, les valeurs en ligne représentent des "nombres de contacts".

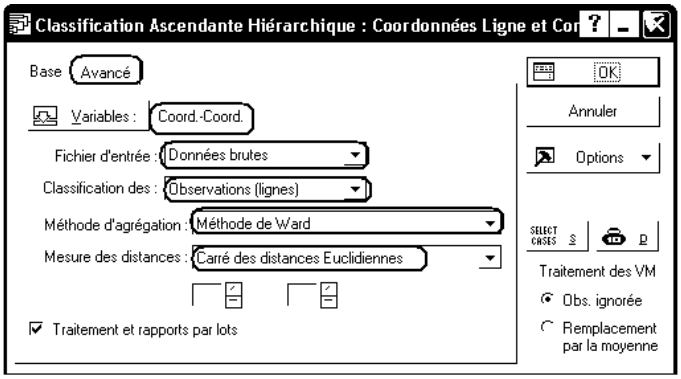
Nous nous proposons de réaliser une CAH sur les professions à partir de ce tableau. Comme nous l'avons vu dans le paragraphe sur l'AFC, la "distance" pertinente entre deux lignes du tableau est la distance du khi-2, ou, ce qui revient au même, le carré de la distance euclidienne entre les images des modalités lignes obtenues par AFC.

Dans un premier temps, ouvrez le classeur Contacts-Medias.stw et réalisez une AFC en calculant les coordonnées lignes et colonnes sur tous les facteurs.

Rendez ensuite active la feuille de données contenant les résultats relatifs aux lignes.

Utilisez ensuite le menu Statistiques - Techniques Exploratoires Multivariées - Classifications .

On choisit ici comme mesure des distances, le carré des distances euclidiennes. Cela revient à mesurer la distance entre deux lignes à l'aide de la distance du khi-2 (propriété de l'AFC). L'indice d'agrégation choisi est celui calculé par la méthode de Ward.



On obtient ainsi les résultats suivants :

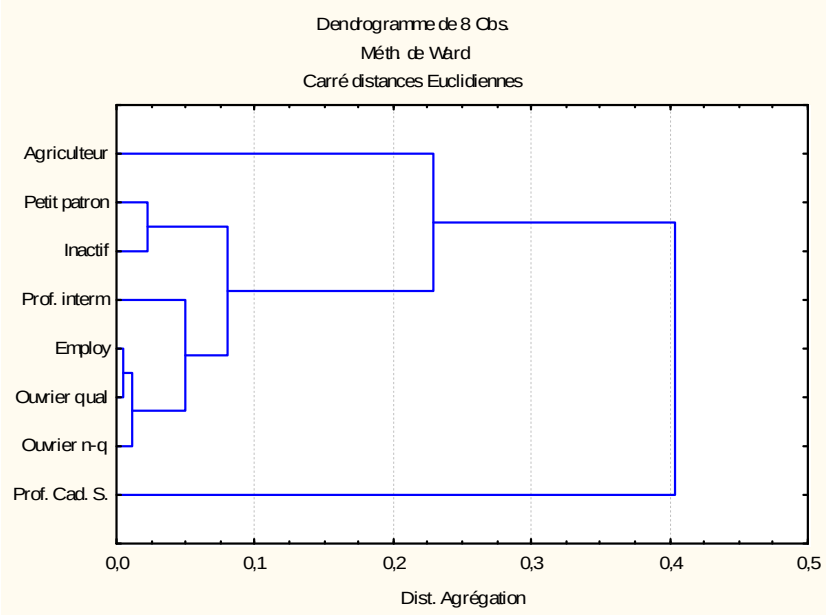
- un tableau donnant les étapes de la classification :

distance agrégat.	Agrégation Finale (Coordonnées Ligne et Contributions à l'Inertie (Contacts-medias.sta dans Classeur4) dans Classeur Méth. de Ward Carré distances Euclidiennes							
	Objet # 1	Objet # 2	Objet # 3	Objet # 4	Objet # 5	Objet # 6	Objet # 7	Objet # 8
,0041508	Employé	Ouvrier qual						
,0120153	Employé	Ouvrier qual	Ouvrier n-q					
,0225031	Petit patron	Inactif						
,0496726	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q				
,0805143	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q		
,2296203	Agriculteur	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	
,4046085	Agriculteur	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	Prof. Cad. S.

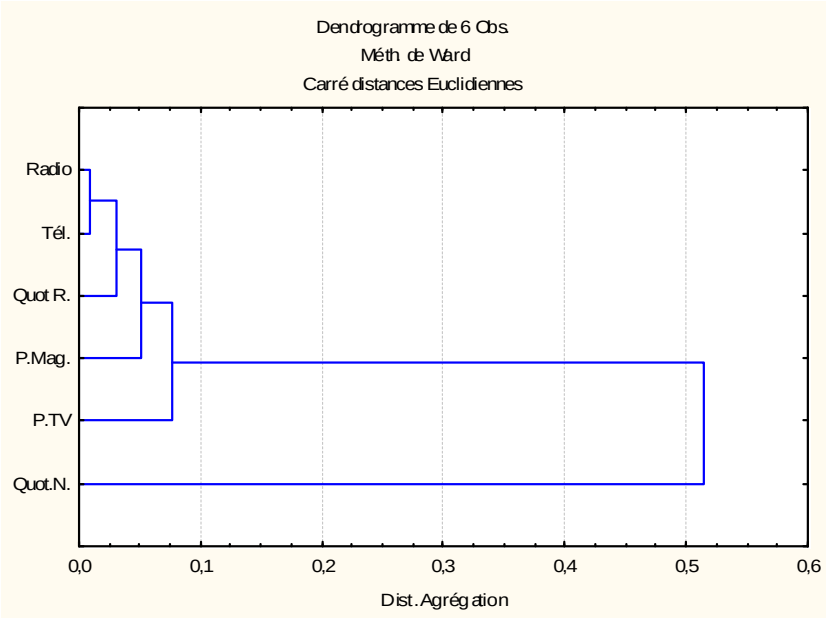
- Le tableau des distances entre individus :

N°Obs.	griculteu	Petit patron	Prof. Cad. S.	Prof. interm	mploye	Ouvrier qual	Ouvrier n-q	Inactif
Agriculteur	0,00	0,04	0,42	0,19	0,19	0,19	0,17	0,10
Petit patron	0,04	0,00	0,26	0,06	0,07	0,06	0,06	0,02
Prof. Cad. S.	0,42	0,26	0,00	0,12	0,22	0,25	0,33	0,22
Prof. interm	0,19	0,06	0,12	0,00	0,02	0,03	0,06	0,03
Employé	0,19	0,07	0,22	0,02	0,00	0,00	0,01	0,02
Ouvrier qual	0,19	0,06	0,25	0,03	0,00	0,00	0,01	0,02
Ouvrier n-q	0,17	0,06	0,33	0,06	0,01	0,01	0,00	0,03
Inactif	0,10	0,02	0,22	0,03	0,02	0,02	0,03	0,00

- Le dendrogramme correspondant à la CAH :



Une CAH analogue, réalisée à partir des individus colonnes, conduit au dendrogramme suivant :

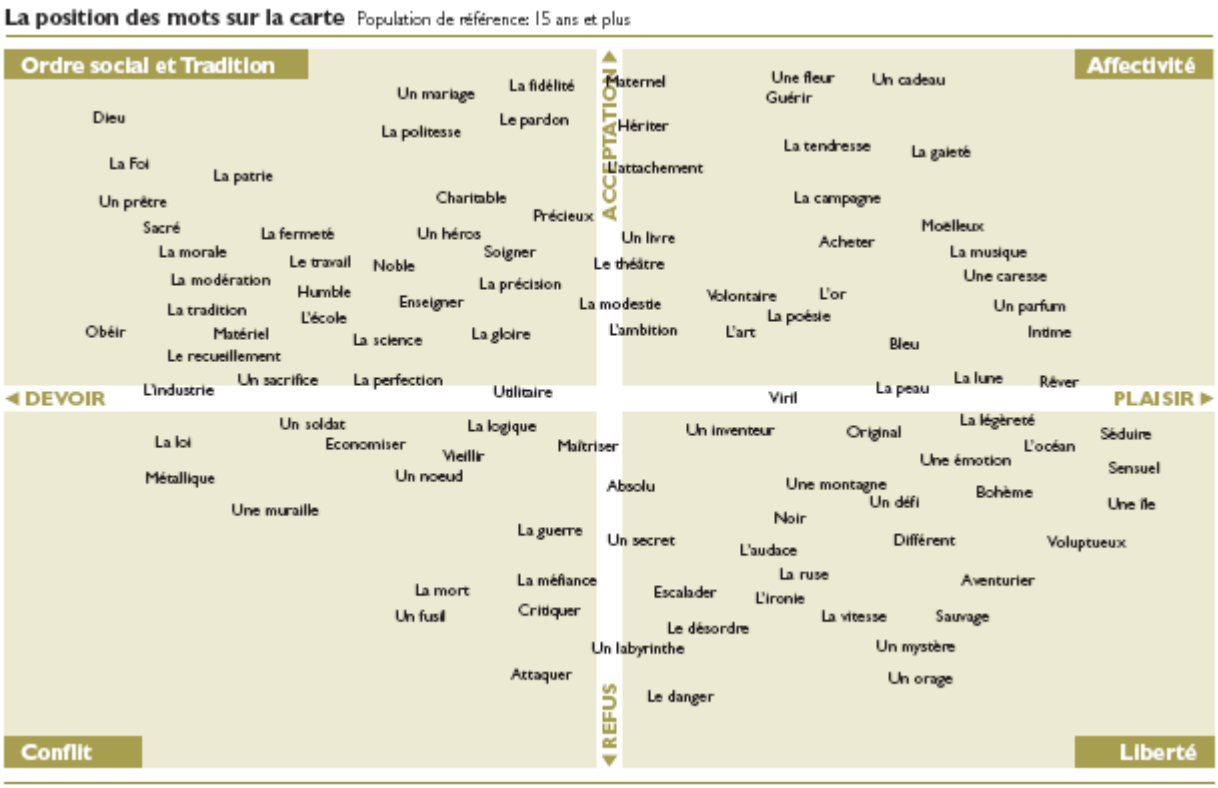


Une telle classification, dans laquelle chaque nouvelle classe est obtenue en agrégeant un unique individu à la classe formée à l'étape précédente revient en fait à définir une relation d'ordre sur les individus, et ne présente qu'un intérêt fort limité.

Classification à partir d'un tableau Individus x Variables Numériques

Réf. Lebart L., Piron M., Steiner J.-F., La Sémiométrie, Dunod, Paris, 2003.

Dans l'ouvrage cité en référence, les auteurs ont fait le choix de 210 mots. Il est ensuite demandé aux personnes interviewées de noter les mots en fonction de la sensation, agréable ou désagréable, que provoque leur lecture. L'échelle de notation comporte 7 modalités variant de -3 à 3. Pour les traitements statistiques ultérieurs, cette échelle est ramenée à une échelle variant de 1 à 7. L'échantillon interrogé entre 1990 et 2002 s'élève à 11055 personnes. Une enquête analogue, menée pour la Belgique, a conduit au résultat suivant (deux premiers axes d'une ACP) :



On mesure la proximité entre deux mots à l'aide du coefficient de corrélation des séries statistiques obtenues pour les deux mots. Plus précisément, le carré de la distance entre deux mots a et b est égal à (1-r(a, b)), où r(a,b) désigne le coefficient de corrélation des deux séries. Pour chaque mot, les autres mots qui lui sont le mieux corrélés constituent son champ sémantique interne. Cependant, un même mot peut être corrélé avec des mots non corrélés entre eux.

Une classification ascendante hiérarchique est effectuée à partir de la distance définie précédemment. Il n'est pas évident a priori que des notes fondées seulement sur l'agrément ou le désagrément engendrent des proximités sémantiques. On constate cependant que les classes obtenues regroupent des mots qui ne sont pas de vrais synonymes (la liste de mots excluait a priori la présence de synonymes) mais appartiennent au même halo sémantique. Dans une partition en 12 classes, par exemple, on trouvera rassemblés des mots ayant trait au concept de "sublimation" tels que :

absolu, immense, infini, admirer, adorer, éternel, précieux, secret, sublime.

Exercice :
A partir de la liste de 7 mots suivants :

efficace, courage, sensuel, montagne, magie, douceur, campagne

imaginez les réponses fournies par dix interviewés et traitez-les à l'aide d'une CAH en utilisant, évidemment , la "distance" 1-r de Pearson.

Représentation des similitudes par l'arbre de longueur minimale

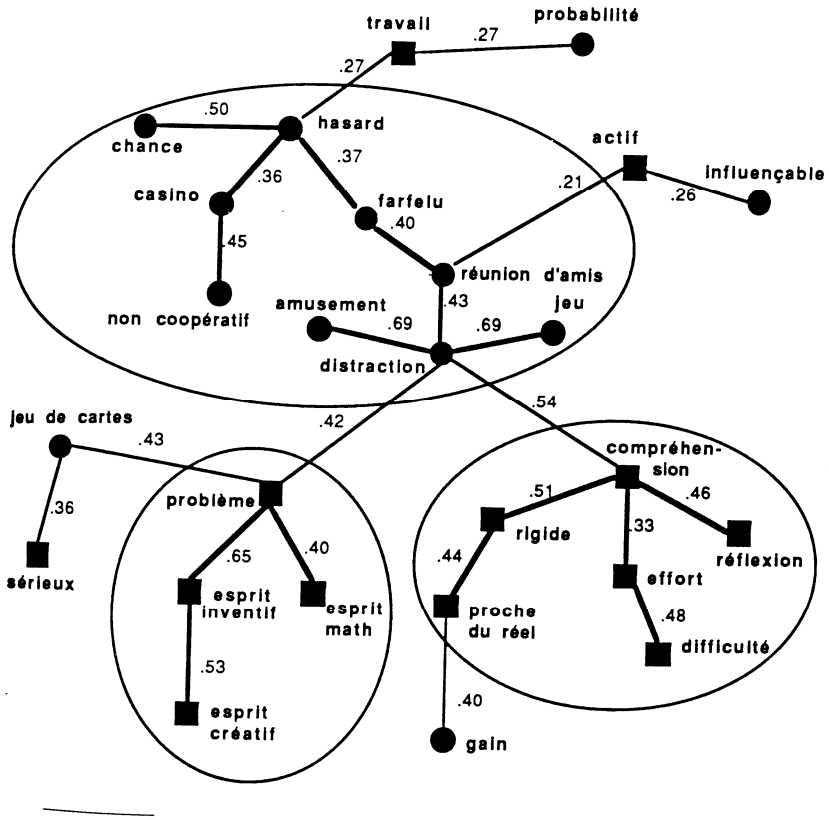
L'ensemble des n objets à classer peut être considéré comme un ensemble de points d'un espace. Si l'on ne dispose que des valeurs d'un indice de dissimilarité , on peut représenter les objets par des points (d'un plan par exemple), chaque couple d'objets étant joint par une ligne continue, à laquelle est attachée la valeur de l'indice de dissimilarité. On représente ainsi l'ensemble des objets et des valeurs de l'indice par un graphe complet valué. On cherchera ensuite à extraire de ce graphe un graphe partiel (ayant les mêmes sommets, mais moins d'arêtes) plus aisé à représenter, et

permettant néanmoins de bien résumer les valeurs de l'indice. Parmi tous les graphes partiels, ceux qui ont une structure d'arbre sont particulièrement intéressants, car ils peuvent faire l'objet d'une représentation plane. La longueur d'un arbre sera la somme des "longueurs" (valeurs de l'indice) de ses arêtes. Parmi tous les graphes partiels qui sont des arbres, l'arbre de longueur minimale a retenu depuis longtemps l'attention des statisticiens en raison de ses bonnes qualités descriptives, qui ne sont pas étrangères à sa parenté avec les classifications hiérarchiques. On peut, par exemple, montrer l'équivalence avec la classification selon le saut minimal.

Dans la procédure de Kruskal, par exemple, on range les n(n - 1)/2 arêtes dans l'ordre des valeurs croissantes de l'indice. On part des deux premières arêtes, puis on sélectionne successivement toutes les arêtes qui ne font pas de cycle avec les arêtes déjà choisies. On interrompt la procédure dès que l'on a n-1 arêtes. De cette façon, on est sûr d'avoir obtenu un arbre (graphe sans cycle ayant n-1 arêtes).

Exemple : Dans l'ouvrage "Représentations sociales et analyse des données", Doise et al. donnent l'exemple suivant :

Donnons un exemple que Flament emprunte à Abric et Vacherot (1976). Il s'agit d'une recherche effectuée sur la représentation d'une tâche de type «dilemme du prisonnier», tâche qui peut être perçue comme une situation de jeu ou une situation de résolution de problèmes. Les auteurs retiennent 26 termes d'une pré-enquête permettant de traduire l'une ou l'autre de ces situations. Ils demandent ensuite à des sujets ayant effectué une tâche de type dilemme du prisonnier de choisir parmi les 26 termes ceux qui évoquent la situation dans laquelle ils se trouvaient. L'arbre maximum du système de similitude (comprenant 325 corrélations) est de la forme suivante :



Dans cette figure, chaque terme représente un sommet. Le long des liaisons entre sommets (ou arêtes) sont indiqués les indices de similitude. Pour construire un tel arbre, la procédure est la suivante. Il s'agit d'abord d'ordonner les arêtes selon la valeur décroissante de l'indice de similitude qui leur est associé. On retient ensuite les deux premières arêtes qui appartiendront forcément à l'arbre maximum du fait qu'elles ne peuvent être les plus petites dans aucun cycle. Enfin, on ajoute à

ces deux premières arêtes, toute arête qui ne forme pas de cycle avec celles déjà retenues. Les arêtes qui sont donc retenues dans l'arbre maximum sont celles qui ne sont minimum dans aucun cycle (voir Degenne et Verges, 1973). Pour illustrer ce propos, prenons l'exemple des éléments Chance, Hasard et Casino qui figurent dans l'arbre maximum ci-dessus. Les arêtes (Chance, Hasard) et (Casino, Hasard) sont inscrites sur le graphe et valent respectivement .50 et .36. On en déduit par conséquent que l'arête (Chance, Casino) est inférieure à .36 ; si tel n'était pas le cas, l'arête (Casino, Hasard) serait supprimée au profit de l'arête (Chance, Casino). En termes de similitude, on peut dire que Chance et Hasard, d'une part, et Hasard et Casino, d'autre part, sont plus proches l'un de l'autre que Chance et Casino.

Sur la base de l'arbre maximum, il est possible de répondre à la question posée par Abric et Vacherot qui est d'identifier les termes associés à jeu ou à résolution de problème comme représentation de la tâche. Flament (1986, 144) en propose la lecture suivante: «Supprimons de l'arbre maximum les arêtes se trouvant entre items de catégories initiales différentes (voir figure) ; les sous-graphes ainsi obtenus sont alors de composition homogène (soit tout jeu, soit tout problème) ; on observe des items isolés (Travail, Probabilité, Actif, etc.), dont la signification initiale est fortement remise en cause (puisque chacun ressemble plus à des items de catégorie opposée qu'aux items de sa propre catégorie). Restent trois sous-graphes importants (indiqués dans la figure) - un pour jeu, deux pour problème -, dont les items voient leur signification initiale confirmée dans la représentation par le voisinage d'items de même catégorie.»

3 Méthodes prédictives

3.1 Régression linéaire

Bibliographie :
Bry, X., Analyses factorielles multiples, Economica, Paris, 1996.

3.1.1 Régression linéaire multiple

Sur un échantillon de n individus statistiques, on a observé :

- p variables numériques X_1, X_2, ..., X_p (variables indépendantes ou explicatives)
- une variable numérique Y (variable dépendante, ou "à expliquer").

Exemple (source : fichiers d'exemples fournis avec Statistica) :

On dispose, pour 30 comtés américains, des données suivantes :

VARI_POP	Variation de la Population (1960-1970)
N_AGRIC	Nb. de personnes travaillant dans le secteur primaire (agriculture)
TX_IMPOS	Taux d'imposition des propriétés résidentielles et fermes
PT_PHONE	Pourcentage de résidences équipés d'une installation téléphonique
PT_RURAL	Pourcentage de la population vivant en milieu rural
AGE	Age médian
PT_PAUVR	Pourcentage de familles en dessous du seuil de pauvreté

L'objectif est d'identifier les facteurs liés au pourcentage de familles en deçà du seuil de pauvreté dans ces comtés (Pt_Pauvr), et de construire un modèle prédictif pour cette variable. Nous allons donc traiter la variable Pt_Pauvr comme la variable dépendante (réponse), et les 6 autres variables comme des prédicteurs continus.

Les données sont les suivantes :

	VARI_POP	N_AGRIC	PT_PAUVR	TX_IMPOS	PT_PHONE	PT_RURAL	AGE
Benton	13,7	400	19,0	1,09	82	74,8	33,5
Cannon	-0,8	710	26,2	1,01	66	100,0	32,8
Carrol	9,6	1610	18,1	0,40	80	69,7	33,4
Cheatheam	40,0	500	15,4	0,93	74	100,0	27,8
Cumberland	8,4	640	29,0	0,92	65	74,0	27,9
DeKalb	3,5	920	21,6	0,59	64	73,1	33,2
Dyer	3,0	1890	21,9	0,63	82	52,3	30,8
Gibson	7,1	3040	18,9	0,49	85	49,6	32,4
Greene	13,0	2730	21,1	0,71	78	71,2	29,2
Hawkins	10,7	1850	23,8	0,93	74	70,6	28,7
Haywood	-16,2	2920	40,5	0,51	69	64,2	25,1
Henry	6,6	1070	21,6	0,80	85	58,3	35,9
Houston	21,9	160	25,4	0,74	69	100,0	31,4
Humphreys	17,8	380	19,7	0,44	83	72,0	30,1
Jackson	-11,8	1140	38,0	0,81	54	100,0	34,1
Johnson	7,5	690	30,1	1,05	65	100,0	30,5
Lawrence	3,7	1170	24,8	0,73	76	69,5	30,0
McNairy	1,6	1280	30,3	0,65	67	81,0	32,4
Madison	8,4	2270	19,5	0,48	85	39,1	28,7

Marshall	2,7	960	15,6	0,72	84	58,4	33,4
Maury	5,6	1710	17,2	0,62	84	42,4	29,9
Montgomery	12,7	1410	18,4	0,84	86	36,4	23,3
Morgan	-4,8	200	27,3	0,73	66	99,8	27,5
Sevier	16,5	960	19,2	0,45	74	90,6	29,5
Shelby	15,2	11500	16,8	1,00	87	5,9	25,4
Sullivan	11,6	1380	13,2	0,63	85	44,2	28,8
Trousdale	4,9	530	29,7	0,54	70	100,0	33,1
Unicoi	1,1	370	19,8	0,98	75	52,6	30,8
Wayne	3,8	440	27,7	0,46	48	100,0	28,4
Weakley	19,0	1630	20,5	0,68	83	72,1	30,4

3.1.1.1 Formulation explicative : le modèle linéaire

On cherche à exprimer Y sous la forme :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p + E$$

où E (erreur commise en remplaçant Y par la valeur estimée) est nulle en moyenne, de variance minimale et indépendante des X_i..

La solution à ce problème est obtenue en prenant pour b_0 :

$$b_0 = \overline{Y} - b_1 \overline{X_1} - b_2 \overline{X_2} - ... - b_p \overline{X_p}$$

et pour les autres coefficients b_i, les solutions du système d'équations linéaires :

$$\begin{cases} Cov(X_1, X_1) b_1 + Cov(X_1, X_2) b_2 + ... + Cov(X_1, X_p) b_p = Cov(X_1, Y) \\ Cov(X_2, X_1) b_1 + Cov(X_2, X_2) b_2 + ... + Cov(X_2, X_p) b_p = Cov(X_2, Y) \\ .... \\ Cov(X_p, X_1) b_1 + Cov(X_p, X_2) b_2 + ... + Cov(X_p, X_p) b_p = Cov(X_p, Y) \end{cases}$$

Pour les données citées en introduction, on obtient :

PT_PAUVR = 31,2660 - 0,3923 VARI_POP + 0,0008 N_AGRIC+ 1,2301 TX_IMPOS - 0,0832 PT_PHONE + 0,1655 PT_RURAL - 0,4193 AGE

Interprétation des coefficients b_i : d'une manière générale, chaque coefficient b_i représente la variation relative de la variable Y rapportée à celle de la variable X_i, toutes les autres variables restant constantes.

Problème : les coefficients b_i dépendent des unités choisies pour mesurer les X_i. C'est pourquoi, on donne aussi les coefficients β_i, liés aux b_i par la relation :

$$\beta_i = \frac{\sigma(X_i)}{\sigma(Y)} b_i$$

Dans l'exemple, les coefficients β_i sont donnés par :

VARI_POP	N_AGRIC	TX_IMPOS	PT_PHONE	PT_RURAL	AGE
-0,630788	0,238314	0,038799	-0,129627	0,618746	-0,188205

Même ainsi "normés", les coefficients de la régression restent d'interprétation délicate. En effet, il est impossible de faire varier l'une des X_i en laissant les autres constantes, car ces variables sont elles-mêmes corrélées entre elles.

Sur notre exemple, on constate que le coefficient correspondant à la variable N_AGRIC est positif (PT_PAUVR et N_AGRIC semblent varier dans le même sens), alors que le coefficient de corrélation entre les deux variables PT_PAUVR et N_AGRIC est négatif (r=-0,17), ce qui indiquerait plutôt une variation en sens contraires. Les corrélations entre les variables sont en effet données par :

	VARI_POP	N_AGRIC	PT_PAUVR	TX_IMPOS	PT_PHONE	PT_RURAL	AGE
VARI_POP	1,00	0,04	**-0,65**	0,13	0,38	-0,02	-0,15
N_AGRIC	0,04	1,00	**-0,17**	0,10	0,36	-0,66	-0,36
PT_PAUVR	**-0,65**	**-0,17**	**1,00**	**0,01**	**-0,73**	**0,51**	**0,02**
TX_IMPOS	0,13	0,10	**0,01**	1,00	-0,04	0,02	-0,05
PT_PHONE	0,38	0,36	**-0,73**	-0,04	1,00	-0,75	-0,08
PT_RURAL	-0,02	-0,66	**0,51**	0,02	-0,75	1,00	0,31
AGE	-0,15	-0,36	**0,02**	-0,05	-0,08	0,31	1,00

Les coefficients b_i et β_i étant des valeurs "théoriques" estimées à partir des valeurs prises par les variables X_i sur l'échantillon de n individus statistiques, il est possible :
- de donner pour chaque b_i un intervalle de confiance à un degré de confiance donné ;
- de tester si chacun des coefficients est significativement différent de 0.

Dans l'exemple traité, on obtient pour les b_i :

	PT_PAUVR (param.)	PT_PAUVR Err-Type	PT_PAUVR t	PT_PAUVR p	-95,00% Lim.Conf	+95,00% Lim.Conf
Ord.Orig.	31,2660	13,2651	2,3570	0,0273	3,8251	58,7070
VARI_POP	-0,3923	0,0805	-4,8742	0,0001	-0,5589	-0,2258
N_AGRIC	0,0008	0,0004	1,6903	0,1045	-0,0002	0,0017
TX_IMPOS	1,2301	3,1899	0,3856	0,7033	-5,3686	7,8288
PT_PHONE	-0,0832	0,1306	-0,6376	0,5300	-0,3533	0,1868
PT_RURAL	0,1655	0,0618	2,6766	0,0135	0,0376	0,2935
AGE	-0,4193	0,2554	-1,6415	0,1143	-0,9476	0,1091

N.B. : Tableau obtenu sous Statistica, à l'aide du menu Statistiques - Modèles généraux de régression - Régression Multiple puis l'onglet Synthèse et le bouton Coefficients.

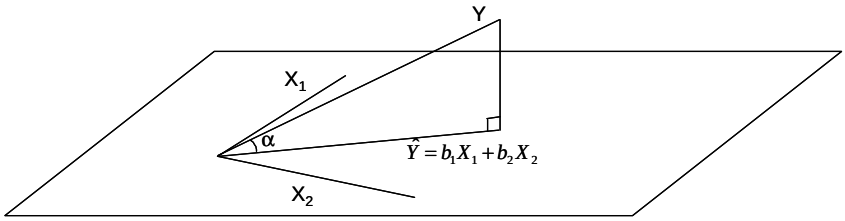
On voit que seuls b_0, b_1 et b_5 sont significativement différents de 0.

En raison des difficultés d'interprétation des résultats d'une régression multiple, différentes alternatives à la régression linéaire "ordinaire" ont été proposées. En particulier, l'interprétation d'une régression est nettement plus simple lorsque les prédicteurs sont non corrélés entre eux. C'est pourquoi il peut être intéressant de réaliser une ACP sur les prédicteurs, puis une régression de Y sur les facteurs de l'ACP. Cette méthode est appelée : *régression sur les composantes principales*.

3.1.1.2 Approche factorielle de la régression

Après centrage des données, le problème de la régression linéaire se ramène au suivant :

On cherche à expliquer la variabilité de Y à partir de celle des X_j : on cherche une combinaison linéaire des X_j qui reproduit "au mieux" la variabilité des individus selon Y. On prend donc la combinaison linéaire la plus corrélée avec Y. La solution est fournie par la combinaison linéaire des X_j qui fait avec Y un angle minimum.



A chaque valeur observée y_i de la variable Y correspond une valeur $\hat{y}_i$ estimée à l'aide de l'équation de régression. La variabilité de Y se décompose comme suit :

Variance de Y = Variance expliquée + Variance résiduelle

$$Var(Y) = Var(\hat{Y}) + Var(Y - \hat{Y})$$

L'analyse de variance permet de tester globalement si la variable régressée dépend significativement des régresseurs qui ont été considérés :

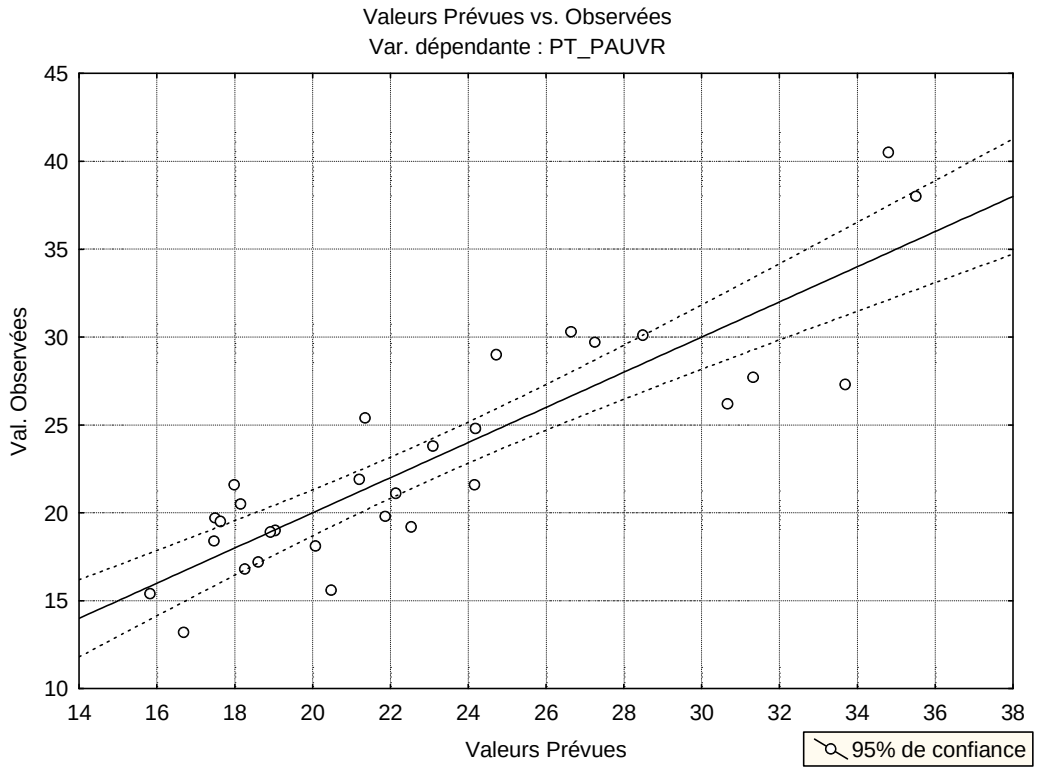
	Sommes Carrés	dl	Moyennes Carrés	F	niveau p
Régress.	932,065	6	155,3441	13,44909	0,000002
Résidus	265,662	23	11,5505		
Total	1197,727				

Le *coefficient de détermination* est le rapport : $R^2 = \dfrac{Var(\hat{Y})}{Var(Y)}$.

Cette valeur est aussi le carré du coefficient de corrélation $r(Y,\hat{Y})$, appelé *coefficient de corrélation multiple*. Sur l'exemple traité, on obtient :

$R = 0{,}8822$; $R^2 = 0{,}7782$

Le graphique suivant compare les valeurs observées de Y (les yi) avec les valeurs estimées par la régression (les y_i) :



3.1.2 Une application de la régression linéaire : analyse de médiation

Réf. http://www.psychologie-sociale.org/reps2.php?article=7

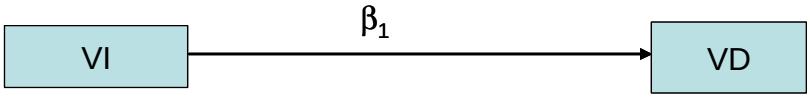
L'analyse de médiation est une technique statistique très utile pour identifier les processus responsables de l'effet d'une variable indépendante sur une variable dépendante. Ainsi, la médiation permet de distinguer, dans l'effet à expliquer, ce qui est directement imputable à la variable indépendante (effet direct de la VI sur la VD) et ce qui relève plutôt de l'intervention d'un facteur intermédiaire (effet indirect de la VI sur la VD via une variable de médiation M).

Principe de la méthode :

On effectue la régression linéaire de la VD sur la VI. On obtient l'équation de régression :

$VD = b0 + b1\ VI$

et un coefficient de régression standardisé : β_1 :



On effectue ensuite la régression linéaire de la variable de médiation M sur la VI. On obtient l'équation de régression :

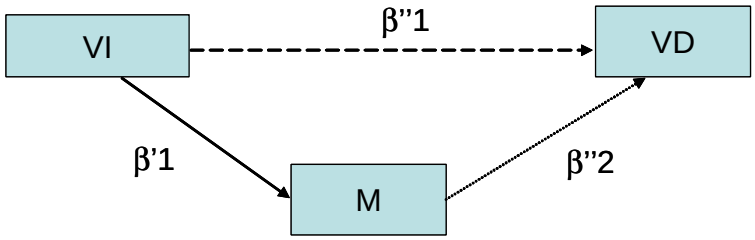
$M = b'0 + b'1\ VI$

et le coefficient de régression standardisé : β'_1

Enfin, on effectue la régression linéaire multiple de la VD sur les deux variables M et VI. On obtient l'équation de régression :

$VD = b''0 + b''1\ VI + b''2\ M$

et les coefficients de régression standardisés β''_1 et β''_2 :



Interprétation :

Si β''_2 est significativement différent de 0 et que β''_1 est nettement plus proche de 0 que β_1, en particulier si β''_1 n'est pas significativement différent de 0 alors que β_1 l'était, il y a médiation (partielle ou totale).

Exemple :

Ref. Costarelli, S., Callà, R.-M.. Self-directed negative affect: The distinct roles of ingroup identification and outgroup derogation, Current research in Social Psychology, Volume 10 No 2, 2004.

Le Sud-Tyrol est une région de l'Italie du Nord dans laquelle coexistent une population de langue italienne et une population de langue allemande. La population de langue allemande a fait l'objet d'une discrimination négative durant le régime fasciste, puis a bénéficié de dispositions favorables

ensuite. De ces événements résulte un fort sentiment d'appartenance à un groupe pour les membres de chacun de ces deux groupes ethniques.

Une enquête par questionnaire a été menée en 2002 auprès d'un échantillon de 71 lycéens italophones. En particulier, les sujets devaient se positionner sur des échelles unipolaires à 6 points (cotées de 0 à 5, 0=pas du tout, 5=extrêmement), selon leur opinion relativement aux deux communautés. Pour moitié, les adjectifs utilisés étaient à connotation positive (par exemple : les germanophones : ne sont pas du tout/sont extrêmement sympathiques), et pour moitié, les adjectifs utilisés étaient à connotation négative (par exemple : antipathique, repoussant, méprisable). En calculant un score moyen par sujet pour les échelles de même connotation, appliquées à la même cible ethnique, on obtient ainsi pour chaque sujet quatre mesures comprises dans l'intervalle de 0 à 5:

- l'évaluation positive de l'endogroupe, notée ici ENDOP
- l'évaluation positive de l'exogroupe, notée ici EXOP
- l'évaluation négative de l'endogroupe, notée ici ENDON
- l'évaluation négative de l'exogroupe, notée ici EXON.

Par ailleurs, le questionnaire comportait également des questions permettant d'évaluer deux autres variables, également dans l'intervalle de mesure de 0 à 5 :

- l'intensité de l'identification à l'endogroupe, notée ici IDENT;
- l'estime négative de soi (self-directed negative affect), notée ici SDNA.

Les paramètres descriptifs des variables observées sont donnés par :

Description	Notation	Moyenne	Ecart type
Identification à l'endogroupe	IDENT	3.61	0.57
Estime négative de soi	SDNA	3.07	0.63
Evaluation positive de l'endogroupe	ENDOP	4.39	0.71
Evaluation positive de l'exogroupe	EXOP	3.66	0.66
Evaluation négative de l'endogroupe	ENDON	0.56	0.50
Evaluation négative de l'exogroupe	EXON	1.58	0.59

On cherche à expliquer les variations de la variable "estime négative de soi" (SDNA) par celles des autres variables.

On définit une variable notée DEROG (outgroup derogation, ou partialité envers l'exogroupe) en formant la différence EXON - ENDON.

a) La régression linéaire de la variable SDNA sur la variable IDENT fournit les résultats suivants :

	Synthèse de la Régression; Variable Dép. : SDNA F(1.69)=4.0587 p<.04784 Err-Type de l'Estim.: .62106					
	Béta	Err-Type de Béta	B	Err-Type de B	t(69)	niveau p
OrdOrig.			2.129557	0.472590	4.506145	0.000026
IDENT	0.235700	0.116994	0.260511	0.129309	2.014632	0.047842

L'effet de IDENT sur SDNA est donc significatif au seuil de 5%.

La régression linéaire de IDENT sur DEROG fournit les résultats suivants :

	Synthèse de la Régression; Variable Dép. : DEROG F(1.69)=8.4320 p<.00495 Err-Type de l'Estim.: .52667

	Béta	Err-Type de Béta	B	Err-Type de B	t(69)	niveau p
OrdOrig.			-0.129500	0.400765	-0.323132	0.747572
IDENT	0.329994	0.113642	0.318421	0.109657	2.903799	0.004948

L'effet de IDENT sur DEROG est donc significatif au seuil de 5% ?

Enfin, on réalise une régression linéaire multiple de SDNA sur les variables DEROG et IDENT. Les résultats sont alors les suivants :

	Synthèse de la Régression; Variable Dép. : SDNA F(2.68)=5.1029 p<.00861 Err-Type de l'Estim.: .60028					
	Béta	Err-Type de Béta	B	Err-Type de B	t(68)	niveau p
OrdOrig.			2.172575	0.457119	4.752756	0.000011
IDENT	0.140000	0.119789	0.154737	0.132398	1.168723	0.246596
DEROG	0.290005	0.119789	0.332182	0.137210	2.420970	0.018154

Dans cette dernière régression, l'effet de IDENT sur SDNA n'est plus significatif. L'effet constaté dans la première régression s'explique donc par un effet de médiation joué par la variable DEROG.

Remarque 1. Dans l'article cité supra, les auteurs définissaient également la variable FAVO (favoritisme pour l'endogroupe) comme la différence ENDOP-ENDON et réalisaient une analyse de médiation analogue. Mais, au contraire de la variable DEROG, la variable FAVO ne joue pas de rôle de médiation significatif.

Remarque 2. Ces résultats, obtenus sur des données analogues à celles utilisées par les auteurs peuvent être retrouvés dans le classeur Statistica Analyse-mediation1.stw.

3.1.3 Régression linéaire avec Statistica

Exemple

Source : A study on significant sources of the burnout syndrome in workers at occupational centres for mentally disabled, Pedro R. Gil-Monte and José Ma Peiró, Psychology in Spain, 1997, Vol. 2. No 1, 116-123.
Page Web : http://www.psychologyinspain.com/content/full/1997/6bis.htm

Subjects
Subjects were 95 employees in occupational institutions for mentally retarded people in the Valencia Autonomous Community (...).

Description des variables.

Self-confidence levels were measured by using five items of an adaptation of the Trait Sport-Confidence Inventory" (TSCI) (Vealey, 1986), in which the word "athlete" was replaced by "workmate". Cronbach's alpha coefficient for the present study was .84.
Social support at work was estimated using 6 items of the "Organisational Stress Questionnaire" (OSQ) (Caplan, Cobb, French, Van Harrison and Pinneau, 1975). These items reflect some aspects of social support coming from *workmates* (3 items) and *supervisors* (3 items). Reliability coefficient in this study was a=.86 for the supervisors' social support scale, and a=.76 for the workmates' social support scale.

Perceived *role conflict* and *role ambiguity* levels were measured by 3 items, for each of the variables, taken from their respective OSQ scales. Reliability values were a=.69 for role ambiguity and .68 for the role conflict scale.
The burnout syndrome was estimated by MBI (Maslach and Jackson, 1986). This instrument is comprised of 22 items measuring the three dimensions in the syndrome: *personal accomplishment* (8 items), *emotional exhaustion* (9 items), and *depersonalisation* (5 items). Reliability coefficients obtained in the study were: a=.76 for the personal accomplishment subscale, a=.87 for emotional exhaustion, and a=.52 for depersonalisation.

Ouvrez le classeur Valencia-Burnout.stw.

N.B. Les données figurant dans ce classeur ont été générées à partir des indications (moyennes, écarts-types, coefficients de corrélation) figurant dans l'article. Cela explique qu'il ne s'agisse pas de valeurs entières, comme on aurait pu le penser à la lecture de la description des variables.

Affichez les statistiques descriptives concernant ces variables. Vous devriez obtenir :

	Statistiques Descriptives (Valencia-Burnout dans Valencia-Burnout.stw				
Variable	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
Self-Confidence	95	6,4800	4,1632	9,4430	1,0656
Workmates Social Support	95	3,2500	1,5810	5,0078	0,6635
Supervisor Social Support	95	2,9000	0,5435	5,2818	0,8545
Role Conflict	95	2,7300	0,9488	4,7904	0,7942
Role Ambiguity	95	2,1100	0,1915	4,3977	0,7640
Personal Accomplishment	95	36,4300	23,2596	57,2632	6,9266
Emotional Exhaustion	95	17,5600	-5,4779	42,1888	10,1737
Depersonalisation	95	4,6900	-4,6758	15,3578	4,4636

Affichez de même la matrice des corrélations :

	Corrélations (Valencia-Burnout dans Valencia-Burnout.stw) Corrélations significatives marquées à p < ,05000 N=95 (Observations à VM ignorées)							
Variable	f-Confiden	Vorkmates Social Support	Supervisor Social Support	Role Conflict	Role Ambiguity	Personal omplishme	Emotional Exhaustion	ersonalisat
Self-Confidence	1,00	0,08	0,16	-0,11	-0,33	0,35	-0,14	0,00
Workmates Soci	0,08	1,00	0,50	-0,41	-0,40	0,33	-0,45	-0,22
Supervisor Socia	0,16	0,50	1,00	-0,37	-0,43	0,22	-0,40	-0,12
Role Conflict	-0,11	-0,41	-0,37	1,00	0,38	-0,32	0,69	0,32
Role Ambiguity	-0,33	-0,40	-0,43	0,38	1,00	-0,48	0,40	0,28
Personal Accom	0,35	0,33	0,22	-0,32	-0,48	1,00	-0,40	-0,28
Emotional Exhau	-0,14	-0,45	-0,40	0,69	0,40	-0,40	1,00	0,40
Depersonalisatic	0,00	-0,22	-0,12	0,32	0,28	-0,28	0,40	1,00

Comparez avec les valeurs indiquées dans l'article :

Table 1
Descriptive Statistics, reliability values and variables correlations[2]

	M	SD	Range	1	2	3	4	5	6	7	8
1. Self-confidence	6.48	1.06	1-9	(.84)							
2. Workmates Social Support	3.25	.66	1-4	.08	(.76)						
3. Supervisor Social Support	2.90	.85	1-4	.16	.50	(.86)					
4. Role Conflict	2.73	.79	1-5	-.11	-.41	-.37	(.68)				
5. Role Ambiguity	2.11	.76	1-5	-.33	-.40	-.43	.38	(.69)			
6. Personal Accomplishment	36.43	6.89	0-48	.35	.33	.22	-.32	-.48	(.76)		
7. Emotional Exhaustion	17.56	10.12	0-54	-.14	-.45	-.40	.69	.40	-.40	(.87)	
8. Depersonalisation	4.69	4.44	0-30	-.00	-.22	-.12	.32	.28	-.28	.40	(.52)

3.1.3.1 La régression linéaire ordinaire

Effectuez ensuite une régression multiple ordinaire des 3 dernières variables sur les 5 premières :

Pour la variable Personal Accomplishment :

Le bouton "Synthèse de la régression" (onglet "Avancé") affiche les résultats suivants :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,56173332 R²= ,31554432 R² Ajusté = ,27709175 F(5,89)=8,2061 p<,00000 Err-Type de l'Estim.: 5,8892					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(89)	niveau p
OrdOrig.			32,4726	7,3143	4,4396	0,0000
Self-Confidence	0,2293	0,0932	1,4906	0,6057	2,4610	0,0158
Workmates Social Support	0,1730	0,1074	1,8063	1,1213	1,6109	0,1108
Supervisor Social Support	-0,0923	0,1071	-0,7479	0,8678	-0,8618	0,3911
Role Conflict	-0,1351	0,1006	-1,1779	0,8773	-1,3426	0,1828
Role Ambiguity	-0,3235	0,1071	-2,9325	0,9710	-3,0201	0,0033

La colonne "B" donne les coefficients de l'équation de régression linéaire. Le modèle fourni par la régression linéaire est le suivant :
Personal Accomplishment = 32,47 + 1,49 *Self-Confidence +1,81 * Workmates Social Support - 0,75 * Supervisor Social Support - 1,18 * Role Conflict - 2,93 * Role Ambiguity

La valeur de R^2 est de 0,315 : 31,5% de la variance de la variable Personal Accomplishment est expliquée par le modèle.

Les coefficients de la colonne "Bêta" sont les coefficients standardisés, c'est-à-dire les coefficients que l'on observerait si on utilisait des variables centrées réduites au lieu des variables observées. On peut également les interpréter comme suit : lorsque "Self-Confidence" augmente d'un écart type, la variable "Personal Accomplishment" estimée augmente de 0,23 écart type, lorsque la variable "Role Conflict" augmente d'un écart type, "Personal Accomplishment" diminue de 0,135 écart type.
Par exemple, on pourra vérifier que

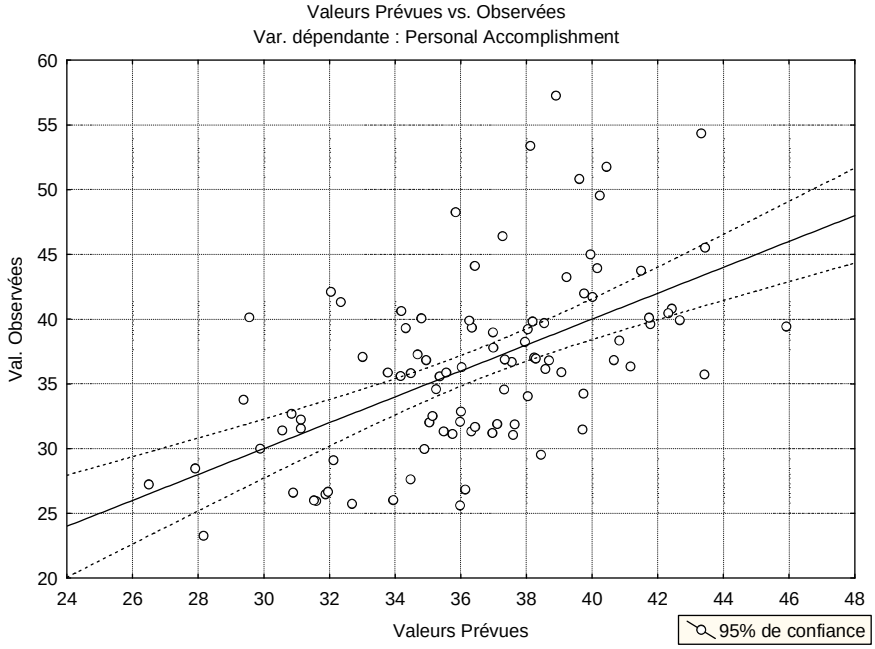
$$Beta(Self-Confidence) = \frac{Ecart\ type(Self-Confidence)}{Ecart\ type(Personal\ Accomplishment)} \times B(Self-Confidence) = \frac{1,0656}{6,9266} \times 1,4906 = 0,2293$$

Les valeurs de t sont obtenues en divisant la valeur correspondante de B par son erreur type. Autrement dit, on teste si le coefficient B est significativement différent de 0.

On peut afficher les résultats de l'ANOVA (bouton ANOVA) montrant qu'ici, le coefficient de régression multiple est significativement différent de 0, ou encore qu'il existe un lien linéaire significatif entre la variable dépendante et les autres variables :

Effet	Analyse de Variance (Valencia-Burnout dans Valencia-Burnout.stw				
	Sommes Carrés	dl	Moyennes Carrés	F	niveau p
Régress.	1423,058	5	284,6115	8,2061	0,0000
1Résidus	3086,793	89	34,6831		
Total	4509,850				

Sous l'onglet "Nuage", on pourra obtenir différentes représentations graphiques dont, par exemple, le graphique illustrant l'adéquation entre les valeurs observées et les valeurs théoriques :



3.1.3.2 La régression linéaire pas à pas

Dans l'article, les auteurs indiquent qu'ils ont fait une régression linéaire pas à pas des dimensions du MBI sur les 5 premières variables.

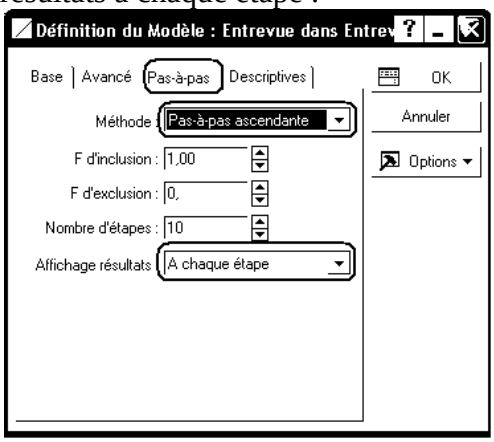
Principe de la méthode

Les données sont formées par une VD Y et plusieurs variables explicatives X1, X2, ..., Xp.
On choisit, parmi les variables explicatives, celle qui est le mieux corrélée à Y. Pour simplifier les notations, nous supposerons qu'il s'agit de la variable X1.
On calcule l'équation de régression linéaire de Y sur X1 : Y = b1 X1 + b0.
On calcule alors les résidus : R1 = Y - b1 X1.- b0
On choisit, parmi les variables explicatives restantes, celle qui est le mieux corrélée à R1. Nous supposons ici qu'il s'agit de la variable X2.
On calcule l'équation de régression linéaire de Y sur X1 et X2 : Y = b'1 X1 + b2 X2 + b'0.
On calcule les nouveaux résidus : R2 = Y - (b'1 X1 + b2 X2 + b'0) et on poursuit la méthode jusqu'à ce que les variables explicatives restantes ne soient plus significativement corrélées aux résidus.

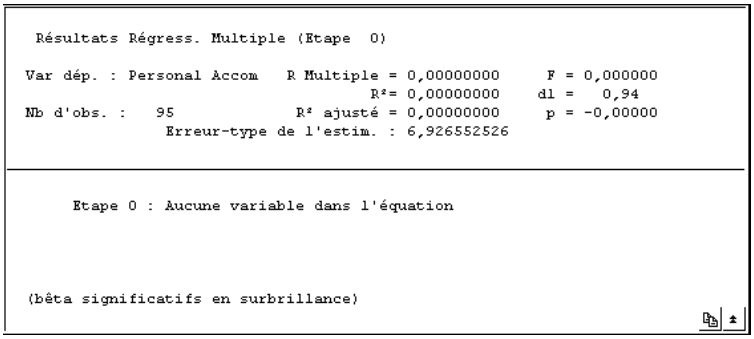
La régression linéaire pas à pas pour la variable Personal Accomplisment

Utilisez de nouveau le menu Statistiques - Régression Multiple
Sous l'onglet "Avancé", spécifiez "Personal Accomplishment" comme variable dépendante, les 5 premières variables comme variables indépendantes. Cochez l'option "régression ridge ou pas-à-pas".

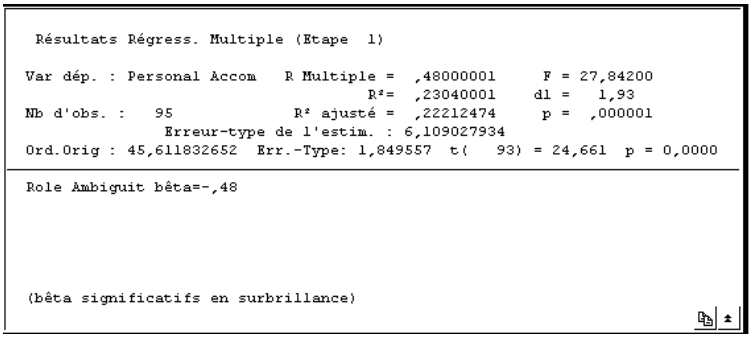
Dans le dialogue suivant, activez l'onglet "pas-à-pas" et sélectionnez la méthode "pas à pas ascendante", et l'affichage des résultats à chaque étape :



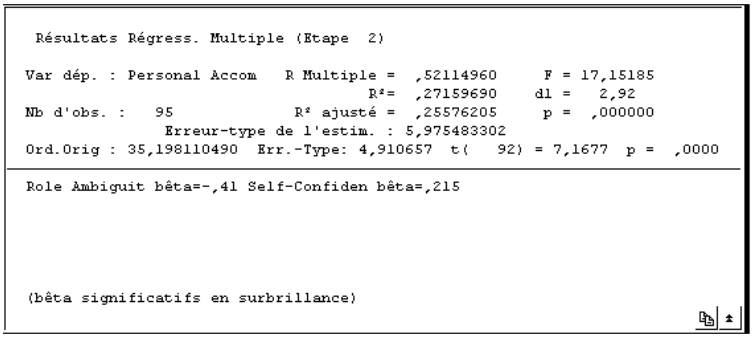
A la première étape, Statistica affiche les résultats suivants :



Cliquez sur "suivant". On obtient :



Puis :



Statistica accepte encore de faire rentrer deux autres variables dans la régression. Cependant, en affichant les résultats disponibles sous le bouton "Synthèse de la régression", on se rend compte que seules ces deux premières variables sont significativement corrélées aux résidus :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishmen R= ,55662608 R²= ,30983259 R² Ajusté = ,27915848 F(4,90)=10,101 p<,00000 Err-Type de l'Estim.: 5,8808					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(90)	niveau p
OrdOrig.			30,8772	7,0660	4,3698	0,0000
Role Ambiguity	-0,3029	0,1043	-2,7456	0,9451	-2,9050	0,0046
Self-Confidence	0,2253	0,0929	1,4647	0,6041	2,4247	0,0173
Workmates Social Support	0,1406	0,1005	1,4680	1,0489	1,3996	0,1651
Role Conflict	-0,1225	0,0994	-1,0682	0,8668	-1,2323	0,2210

On peut alors reprendre la méthode en ne spécifiant que deux étapes et retrouver les résultats indiqués par les auteurs :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,52114960 R²= ,27159690 R² Ajusté = ,25576205 F(2,92)=17,152 p<,00000 Err-Type de l'Estim.: 5,9755					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(92)	niveau p
OrdOrig.			35,1981	4,9107	7,1677	0,0000
Role Ambiguity	-0,4090	0,0943	-3,7083	0,8545	-4,3395	0,0000
Self-Confidence	0,2150	0,0943	1,3976	0,6127	2,2811	0,0249

Résultats indiqués dans l'article :

Table 2
Stepwise regression analysis for MBI dimensions

Variable Step	R2 increase	Beta	F for equation
	personal Accomplishment		
1 Role ambiguity	.23	-.41	
2 Self-confidence	.04	.22	17.27***
	Emotional Exhaustion		
1 Role conflict	.47	.60	
2 Workmates' social support	.03	-.20	47.27***
	Depersonalisation		
1 Role conflict	.10	.32	10.45***

*** p < 001