Data Mining 1

Chapitre 4 : Classification (clustering)

Principe :
Ce document constitue des notes de cours illustrées sur un jeu de données. Chaque concept est complété par un exemple qui contient des commentaires de sorties obtenues avec le logiciel SAS et données en annexe. Ces sorties sont repérées par des numéros (AT1, AT2,...,AT14). Le code SAS permettant d'obtenir les résultats est aussi fourni en annexe. En fin de cours se trouvent les sorties R sur l'exemple du cours avec les principales fonctions ainsi que des commentaires.

1 Généralités

1.1 Applications

Segmentation des fichiers clients par les banques ou autres entreprises, études de marchés en Marketing, ...

1.2 Données

En fait, il existe des techniques d'analyse typologique pour n'importe quel type de tableaux de données mais nous nous restreignons aux tableaux individus/variables.

Dans la suite, on suppose donc que l'on dispose de n observations décrites par p variables quantitatives i.e. d'un tableau de type individus / variables.

Remarque 1 Lorsque les variables sont très hétérogènes (variances ou échelles très différentes), on peut être amené, comme en ACP, à centrer et réduire les données (proc standard sous SAS et fonction scale sous R).

Notations : on note $X_1, \ldots, X_n$ les n individus et $X^1, X^2, \ldots, X^p$ les p variables numériques.

Exemple 1 *En marketing, on peut chercher à mettre en évidence des groupes de clients dont les comportements d'achat sont homogènes. Les variables quantitatives à considérer peuvent être les quantités achetées pour différents produits.*

1.3 Objectifs

L'analyse typologique ou analyse classifiante ou classification (clustering en anglais) est une technique d'analyse de données permettant de construire des groupes d'individus tels que :

- chaque groupe soit homogène selon certaines caractéristiques, c'est-à-dire que les observations d'un groupe se ressemblent le plus possible,
- chaque groupe soit différent des autres selon les mêmes caractéristiques, c'est-à-dire que les observations d'un groupe sont les plus différentes possible de celles des autres groupes.

Les notions de ressemblances ou de différences entre individus sont formalisées en mesurant des **distances** entre les individus. On choisit, dans la suite, de travailler avec la distance usuelle :

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (X_i^k - X_j^k)^2}.$$

Remarque 2 *Contrairement à l'AFD, l'analyse typologique n'est pas une méthode de statistique explicative car on ne dispose pas de la variable qualitative Y qui fixe les groupes. En AT, toutes les variables jouent le même rôle et on cherche des regroupements d'individus sans aucune indication préalable.*

1.4 Présentation des techniques de classification

Pour l'essentiel, les techniques de classification font appel à une démarche algorithmique et non aux calculs formalisés usuels en Analyse Factorielle (ACP, AFC, AFD). Alors que les valeurs des composantes principales, par exemple, sont la solution d'une équation pouvant s'écrire sous une forme très condensée (même si sa résolution est complexe), la définition des classes ne se fait qu'à partir d'une *formulation algorithmique* : une série d'opérations définies de façon répétitive.

Il existe 2 grandes familles de méthodes : les méthodes de *partitionnement* et les méthodes *hiérarchiques*.

Une classification par partitionnement consiste à rechercher directement une *partition* des individus.

Rappels : on appelle **partition** d'un ensemble d'observations $\{X_1, \ldots, X_n\}$, un ensemble de k groupes $A_1, \ldots, A_k$ tels que :

- $A_1 \bigcup A_2 \ldots \bigcup A_k = \{X_1, \ldots, X_n\}$,

– $\forall i, j \in \{1, \ldots, k\}, i \neq j, A_i \bigcap A_j = \emptyset$

Les groupes obtenus visent à maximiser l'inertie inter-classes et minimiser l'inertie intra-classes. Il existe différentes méthodes de partitionnement. Nous choisissons de présenter dans ce cours une méthode dite *d'agrégation autour de centres mobiles*.

Décomposition de l'inertie :

$$\text{Inertie} = \sum_{j=1}^{p} Var(X^j) = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} (X_i^j - \overline{X^j})^2 = \frac{1}{n} \sum_{i=1}^{n} d^2(X_i, \overline{X})$$

avec $\overline{X} = (\overline{X^1}, \overline{X^2}, \cdots, \overline{X^p})$

Pour une partition en k groupes $C_1, C_2, \cdots C_k$, d'effectifs $n_1, n_2, \cdots, n_k$ ($\sum_{r=1}^{k} n_r = n$) on peut écrire l'inertie de la façon suivante :

$$\text{Inertie} = \sum_{r=1}^{k} \sum_{i \in C_r} \frac{1}{n} d^2(X_i, \overline{X})$$

On prouve que l'on peut décomposer l'inertie en :

Inertie = Inertie intra-groupes + Inertie inter-groupes

$$\text{Inertie intra-groupes} = \sum_{r=1}^{k} \frac{1}{n} \sum_{i \in C_r} d^2(X_i, \overline{X}^{(r)}) = \sum_{r=1}^{k} \frac{n_r}{n} I_r$$

avec $I_r = \frac{1}{n_r} \sum_{i \in C_r} d^2(X_i, \overline{X}^{(r)})$ inertie pour le groupe C_r

$$\text{Inertie inter-groupes} = \sum_{r=1}^{k} \frac{n_r}{n} d^2(\overline{X}^{(r)}, \overline{X})$$

avec $\overline{X}^{(r)}$ point moyen du groupe C_r.

2 Méthode d'agrégation autour de moyennes mobiles

Dans la suite, on s'intéresse à une méthode de type partitionnement qui est particulièrement rapide (appelée FASTCLUS par SAS et fonction **kmeans** sous R) et utilisable même pour de très grands tableaux.

Algorithme d'agrégation autour de centres mobiles :

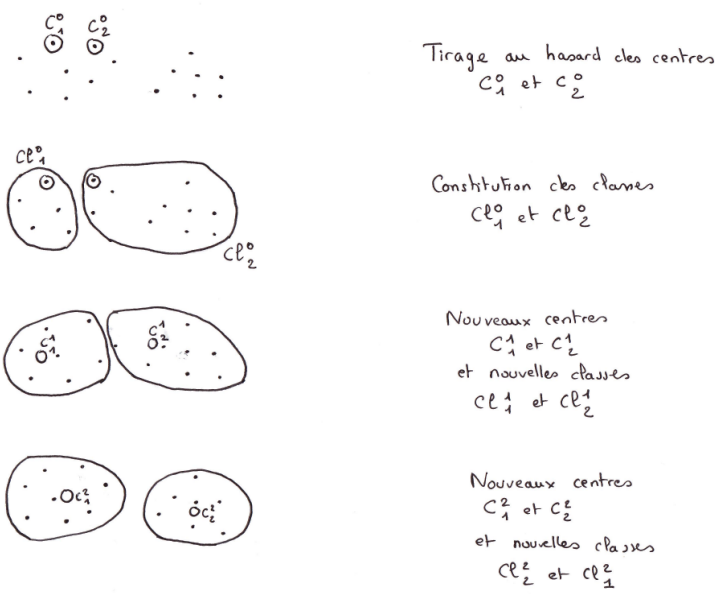


FIGURE 1 – Graphique 1

2.1 Présentation générale

La méthode consiste tout d'abord à fixer à priori un nombre k de groupes et à représenter chaque groupe par un individu. On affecte chacun des individus restants au groupe le plus proche. On calcule alors les centres (ou moyennes) de chaque groupe et les individus sont éventuellement réaffectés au groupe dont ils sont le plus proche. Cette dernière étape est répétée (on parle d'itérations) juqu'à ce que les centres de groupes soient peu ou pas modifiés (Cf. graphique 1 dans le cas $k = 2$ groupes)

On montre d'un point de vue théorique que l'algorithme précédent converge sous des hypothèses peu restrictives. De plus, à chaque itération, la variance intra-classes diminue.

Cet algorithme recouvre de nombreuses variantes selon :

- la façon de choisir les individus représentant les groupes au départ (les "centres" initiaux),
- la fréquence de calcul des centres (après la réaffectation de chaque individu ou après toutes les réaffectations),
- le test d'arrêt de la procédure.

2.2 Choix des centres de classes initiaux

Il existe de nombreuses façons de choisir les k individus représentant chacun des groupes pour l'initialisation de l'algorithme. On peut choisir entre autres :

- les k premiers individus (option REPLACE=NONE avec SAS),
- k individus tirés aléatoirement parmi les n (REPLACE=RANDOM),
- k individus "suffisamment" éloignés les uns des autres (REPLACE=PART ou FULL)...

Or, ce choix n'est pas neutre. Il peut conditionner le résultat final. On propose donc d'utiliser la dernière méthode (REPLACE=PART) qui est parmi les plus efficaces. Détaillons la procédure. Au départ, on choisit les k premiers individus comme centres. Puis, on examine un à un les $(n - k)$ individus restants pour vérifier qu'ils ne sont pas de meilleurs candidats. Un individu est meilleur candidat et va donc remplacer un des centres précédemment choisi si sa distance au centre le plus proche est plus grande que la distance minimale entre les centres. Cet individu remplace alors le centre qui lui est le plus proche parmi les 2 centres à distance minimale.

Exemple 2 *Exemple :*

OBS	*REV*	*EDUC*
1	*5*	*5*
2	*6*	*6*
3	*15*	*14*
4	*16*	*15*
5	*25*	*20*
6	*30*	*19*

3 groupes très nettement différenciés avec bas-moyens-hauts revenus et niveaux d'études. Cet exemple n'a d'autre intérêt que de nous faire comprendre l'algorithme.

Cf. FASTCLUS procedure REPLACE=PARTIAL Maxclusters=3 et AT 2 : choix des centres initiaux (initial seeds)*.*

Explication : SAS choisit d'abord les obs. 1,2 et 3 comme centres pour chacun des 3 groupes. La distance minimale entre ces centres est celle entre 1 et 2 ($= \sqrt{2}$*). Considérons l'observation 4. Sa distance au centre le plus proche est la distance à 3 soit* $\sqrt{2}$ *qui n'est pas inférieure à d(1,2) donc 4 ne remplace pas de centre.*

Par contre l'observation 5 est à une distance $> \sqrt{2}$ de tous les centres donc 5 va remplacer un centre. Lequel ? ou bien 1, ou bien 2 (les 2 centres les plus proches). 5 remplace 2 car 5 est plus proche de 2 que de 1. Quant à l'obsercation 6, sa distance au centre le plus proche (5) est plus petite que la distance minimale entre centres (d(3,5)) donc il ne remplace aucun centre. Finalement, SAS choisit les observations 1, 3 et 5 comme centres initiaux . SAS précise aussi la distance minimale entre ces centres (d(3,5)=$\sqrt{10^2 + 6^2} = \sqrt{136} = 11.66$).

Ce choix des centres initiaux influence la partition finale. Ainsi, dans l'exemple, si on choisit les observations 1, 2 et 3 comme centres initiaux, on ne trouve pas la partition $\{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$ mais la partition $\{\{1\}, \{2\}, \{3, 4, 5, 6\}\}$. Il est donc conseillé d'utiliser cette option.

A partir des centres initiaux, SAS affecte les observations et recalcule des centres de groupes.

2.3 Fréquence de calcul des centres

On peut décider d'affecter toutes les observations au groupe dont le centre est le plus proche avant de recalculer des nouveaux centres qui seront les moyennes des groupes. Mais SAS offre aussi la possibilité (option DRIFT) de recalculer les centres après l'affectation de chacune des obs. à un groupe. Il est conseillé d'utiliser cette dernière option.

2.4 Test d'arrêt de la procédure

Les itérations qui englobent reaffectations d'individus et recalculs des centres de classes s'arrêtent lorsque les centres sont peu ou pas modifiés. En fait, SAS quantifie ce "peu" en fixant un seuil de convergence qui vaut 0.02 par défaut mais qui peut être modifié. L'algorithme s'arrête donc lorsque les *changements relatifs* des centres *(relative changes in clusters seeds)* sont inférieurs à 0.02. Ce changement relatif correspond, pour un centre, au quotient de :

- la distance entre le centre à l'itération i et le centre à l'itération $i + 1$ sur
- la distance minimale entre les centres initiaux.

Exemple 3 *Voir le tableau AT3.*

Si on considère le premier groupe. Son centre initial a pour coordonnées $(5, 5)$. *A l'itération suivante, on a affecté les observations 1 et 2 au groupe 1 et donc, le nouveau centre vaut la moyenne des 2 observations soit* $(5.5, 5.5)$. *La distance entre les 2 centres successifs du groupe 1 est donc :* $\sqrt{(0.5)^2 + (0.5)^2} = 0.7071$ *et*

le changement relatif est donc :

$$\frac{0.7071}{11.6619} = 0.0606$$

3 Étapes de l'analyse typologique par la méthode d'agrégation autour des moyennes mobiles

3.1 Choix du du nombre de groupes

Pour utiliser une méthode de partitionnement, il faut se fixer un nombre k de groupes (défaut de ce type de méthodes). Si on n'a pas d'indication préalable, il est conseillé de faire plusieurs essais avec différentes valeurs de k.

3.2 Interprétation des groupes ou typologie

L'objectif d'une Analyse typologique est d'obtenir une **typologie** c'est-à-dire des groupes ou types d'individus avec une description de chacun de ces groupes. Pour obtenir cette interprétation, on doit revenir aux variables initiales (comme en analyse factorielle mais ici, on ne dispose plus de facteurs).

3.2.1 Position relative des groupes

On peut commencer à repérer comment les groupes se situent les uns par rapport aux autres. Pour cela, SAS fournit le groupe le plus proche de chaque groupe ("Nearest cluster")

Exemple 4 *Voir le tableau AT9. Le groupe 3 se trouve "entre" les groupes 1 et 2.*

3.2.2 Caractérisation par les moyennes

Pour repérer quelles sont les variables initiales qui jouent un rôle important dans la classification, on regarde, pour chaque variable, les différences entre les moyennes des groupes.

Exemple 5 *Cf. A.T. 7 : dans cet exemple, les moyennes entre les 3 groupes différent de la même façon (ou presque) pour la variable Rev que pour Educ. Donc, les 2 variables interviennent et dans le même sens, ce qui montre qu'il existe un groupe faible (1), un groupe moyen (2) et un groupe fort (3). Il est également intéressant de commenter les écarts-types (voir le tableau AT8).*

3.2.3 Graphiques

Un fois repérées les variables importantes et si elles ne sont pas en nombre trop important, on peut représenter les individus sur ces variables en tenant compte du groupe d'affectation.

Exemple 6 *Cf. AT 10 : idem que le premier graphique avec le label "CLUSTER" (variable SAS) en plus. Évidemment, dans ce cas, avec seulement 2 variables, il est inutile de faire une classification...*

3.2.4 Cas des très petits groupes

Souvent, pour de gros fichiers, on trouve des classes regroupant beaucoup d'individus et d'autres classes très petites (parfois un seul individu par classe). C'est le cas en particulier lorqu'un individu est très atypique. Une fois repérés de tels individus, il peut être intéressant de les enlever ou de refaire l'analyse avec d'avantage de groupes.

3.3 Validation de la typologie

En pratique, on cherche à valider la classification obtenue.

3.3.1 Facilité de description

Pour que la classification ait un intérêt, il faut que l'utilisateur soit en mesure de décrire facilement les groupes. C'est le cas pour notre exemple mais les choses se compliquent si l'on dispose de beaucoup de variables. Des spécialistes du domaine d'études peuvent être d'une aide précieuse à cette étape de l'étude.

3.3.2 Analyse des R^2

Pour une partition donnée, on peut s'intéresser (comme en AFD) au quotient entre variance inter-classes et variance intra-classes (variable par variable) ou global. Cette mesure est un R^2 et s'interprète comme en régression. SAS fournit les R^2 (*R-squared* dans *Statistics for Variables*) par variable et global (*over-all*).

Exemple 7 *Voir le tableau AT6*

	R-Squared
REV	*0.97*
EDUC	*0.99*
OVER-ALL	*0.98*

Les R^2 sont très proches de 1 donc, que ce soit globalement ou pour chaque variable, les groupes sont bien homogènes à l'intérieur (il n'y a que 2 ind. !) et hétérogènes entre eux, ce qui est l'objectif recherché.

3.3.3 Compararaison avec l'AFD

A la suite d'une classification, on dispose d'une variable groupe (appelée CLUSTER par SAS) que l'on peut utiliser avec les var. quantitatives initiales pour réaliser une AFD. Si cette AFD donne de bons résultats (axes bien discriminants), elle valide la classification précédente.

4 Classification ascendante hiérarchique (CAH)

4.1 Principe

La méthode consiste à considérer comme typologie initiale autant de classes que d'individus (n) et à regrouper les classes par étapes successives jusqu'à l'obtention d'une seule classe contenant tous les individus.

4.1.1 Algorithme

Initialisation : on calcule toutes les distances entre individus (tableau de taille $n \times n$). Chaque individu représente une classe.

Itérations :

- on agrège les 2 classes les plus proches en une nouvelle classe,
- on met à jour le tableau des distances en tenant compte de la nouvelle classe (calcul des distances de cette nouvelle classe avec les autres classes).

Les itérations s'arrêtent lorsque l'on ne dispose plus que d'une seule classe.

La procédure SAS qui permet de réaliser une CAH est la procédure CLUSTER (fonction **hclust** sous R).

4.2 Calcul des distances

Dans la présentation de l'algorithme ci-dessus, on utilise la notion de "distance" entre classes, c'est-à-dire entre groupes d'individus. Cette notion doit être précisée et, selon la distance choisie, on obtient différents algorithmes.

Dans la suite, on propose de se concentrer sur un choix particulier qui consiste à utiliser une distance euclidienne entre les centres de classes.

Si on dénote par c_a (respectivement c_b) le centre de la classe a de taille n_a (respectivement b de taille n_b), la distance utilisée est le **saut de Ward** définie par :

$$D^2_{ward}(a,b) = \frac{n_a n_b}{n_a + n_b} d^2(c_a, c_b).$$

Le saut de Ward est la distance la plus utilisée en pratique. Il permet de minimiser, à chaque itération, la décroissance de l'inertie [1] inter-groupes. En effet, lorsque l'on agrège 2 classes, l'inertie inter-classe diminue nécessairement mais on souhaite qu'elle diminue le moins possible car on veut séparer au mieux les groupes.

Pour utiliser ce critère dans la proc CLUSTER de SAS, il faut mettre l'option *method=ward*.

Remarque 3 *Autres méthodes d'agrégation*

- *Méthode du **saut minimal***

$$D_{min}(a,b) = \inf\{d(i,j), i \in a, j \in b\}$$

Cette méthode peut entraîner des classes trop larges.

- *Méthode du **diamètre***

$$D_{diam}(a,b) = \sup\{d(i,j), i \in a, j \in b\}$$

Cette méthode est très restrictive.

- *Méthode de la **distance moyenne***

$$D_{moy}(a,b) = \frac{1}{n_a n_b} \sum_{i \in a, j \in b} d(i,j)$$

Cette méthode est un bon compromis entre les deux méthodes précédentes.

La procédure CLUSTER de SAS donne en sortie l'historique de la création des groupes ainsi que les R^2 correspondant à chaque partition. Ces R^2 qui, rappelons-le sont des ratios entre l'inertie inter-groupe et l'inertie totale de l'ensemble des points, décroissent de 1 à 0. En effet, à l'initialisation, puisque chaque observation correspond à un groupe, inertie inter-groupe et inertie totale sont confondues. Au contraire, puisque la partition finale consiste en un seul groupe, il n'y a plus aucune variation inter-groupe (un seul point) et le R^2 final est nul.

1. on rappelle que l'*inertie* est la somme des variances d'un tableau de données individus $\times$ variables quantitatives

SAS fournit aussi la différence entre 2 R^2 successifs dans une colonne intitulée SPRSQ pour *Semi Partial R SQuare*.

Exemple 8 *Voir le tableau AT11 pour les regroupements successifs.*

Le graphe AT12 représente les SPRSQ en fonction du nombre de groupes. Ce graphe aide au choix du nombre de groupes. On voit qu'il y a un "saut" du SPRSQ entre la partition en 3 groupes et la partition en 2 groupes. Cela signifie qu'i l y une décroissance importante du R^2 de la partition en 2 groupes par rapport à celui de la partition en 3 groupes. On retient donc $k = 3$ groupes.

Remarque 4

- *pour réaliser une CAH sur des variables standardisées (centrées réduites), il suffit de préciser l'option* std.
- *si on ne souhaite pas afficher les valeurs propres de la matrice des corrélations (si les variables sont standardisées), il suffit de préciser l'option* noeigen.

4.3 Le dendogramme

Il s'agit d'une représentation sous forme d'arbre des regroupements successifs obtenus par CAH. Pour le représenter avec SAS, il suffit de sauvegarder un fichier résultat en utilisant l'option *out*= dans la proc CLUSTER et d'utiliser ce fichier en entrée de la procédure TREE de SAS (*data*=).

L'intéret de cette représentation est qu'elle permet également de visualiser l'évolution des R^2 d'une étape d'agrégation à la suivante puisque les SPRSQ se trouvent en ordonnée. Cette visualisation de la décroissance des R^2 va nous guider dans le choix du nombre de groupes.

En effet, en général, notre objectif est d'obtenir un nombre assez restreint de groupes mais avec un R^2 suffisamment grand.

Un fois choisi le nombre k de groupes, on obtient une partition particulière et si on veut récupérer une variable d'affectation aux groupes pour cette partition, il suffit de préciser l'option *nclusters*=k et d'enregistrer le résultat en utilisant l'option *out*=.

Exemple 9 *Voir le graphe AT14.*

4.4 Interprétation des groupes

Pour pouvoir réaliser la typologie de la partition choisie, en comparant par exemple les moyennes (en tenant compte des écarts-type) des groupes comme

avec la méthode des nuées dynamiques, il faut utiliser la proc MEANS avec la commande BY.

Exemple 10 *Voir le tableau AT13. On retrouve la classification obtenue à l'aide de la méthode des moyennes mobiles.*

5 Un exemple de classification mixte

L'avantage de la CAH comparativement à la méthode des nuées dynamiques est qu'elle ne nécessite pas de connaître le nombre de groupes à l'avance. Le dendogramme permet de nous guider dans le choix du nombre de groupes. Toutefois, cette méthode, contrairement à la méthode des nuées dynamiques, est très coûteuse en temps de calcul et ne peut-être utilisée si le nombre d'observations est grand.

Pour pouvoir cumuler les avantages des deux méthodes, on propose la stratégie suivante. Notons n le nombre d'observations du fichier de données (par exemple $n = 10000$). On ne peut pas utiliser la procédure CLUSTER mais on peut utiliser la procédure FASTCLUS en choisissant un grand nombre k_{FAST} de groupes (par exemple $k_{\mathrm{FAST}} = 500$). L'intérêt de choisir ce grand nombre de groupes est que la partition que l'on obtient n'a agrégé que les individus vraiment proches. Maintenant, il s'agit de considérer les 500 moyennes associées aux 500 groupes et d'exécuter une proc CLUSTER sur ces moyennes. Avec 500 observations (qui sont en réalité des moyennes), la procédure peut maintenant fonctionner et nous pouvons choisir grâce au dendogramme, un nombre k groupes.

6 Conclusion

Compararaison avec l'AFD

- A la suite d'une classification, on dispose d'une variable groupe que l'on peut utiliser avec les variables quantitatives initiales pour réaliser une AFD.
- Si cette AFD donne de bons résultats (axes bien discriminants), elle valide la classification précédente.

Remarque 5 *Lorsque que l'on dispose de grands tableaux, on peut parfois commencer par une ACP et utiliser les premières composantes principales dans la classification (De même, pour une table de contingence, on peut utiliser les axes définis par une AFC).*

Justification de l'AT :

– tableau de type individus / variables avec des variables quantitatives,
– recherche d'une classification des individus en vue de définir quelques typologies caractéristiques.

Annexe 1 : sorties SAS

Graphe AT1

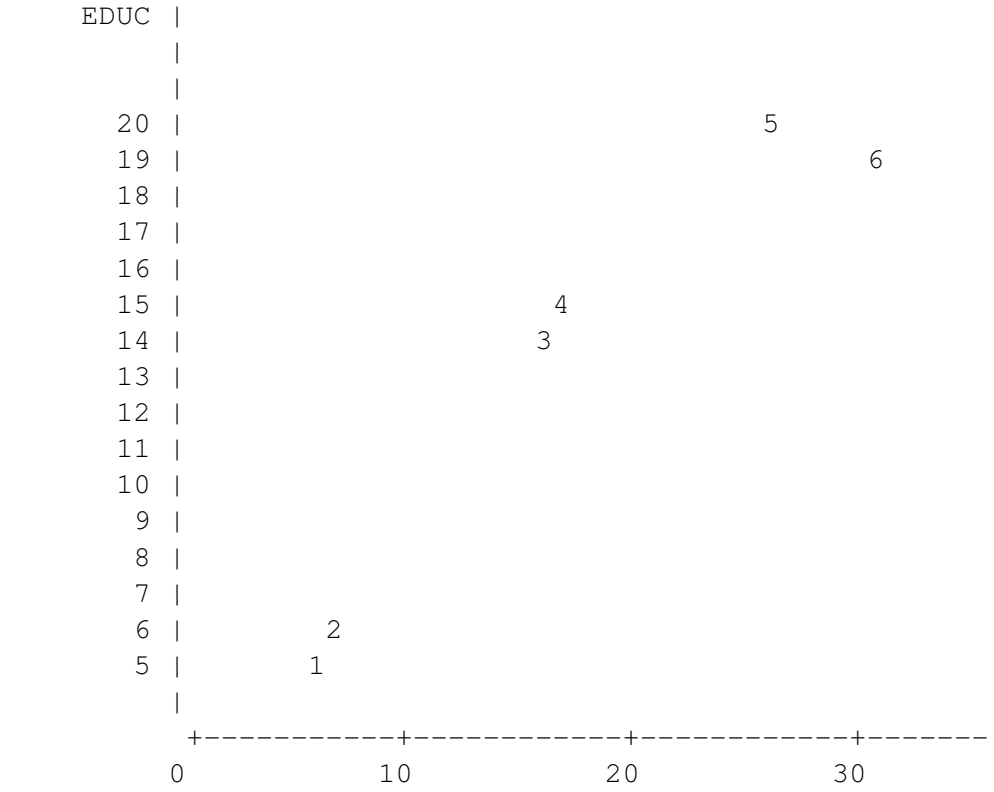


Tableau AT2

The FASTCLUS Procedure

Replace=PARTIAL Radius=0 Maxclusters=3 Maxiter=20 Converge=0.02

Initial Seeds

Cluster	REV	EDUC
1	5.00000000	5.00000000
2	25.00000000	20.00000000
3	15.00000000	14.00000000

Tableau AT3

Minimum Distance Between Initial Seeds = 11.6619

Historique des itérations

Itération	Critère	Changement relatif dans les valeurs initiales de classification 1	2	3
1	1.5811	0.0606	0.2186	0.0606
2	1.1180	0	0	0

Convergence criterion is satisfied.

Tableau AT4

Liste des classifications

Obs.	Classification	Distance par rapport à la valeur initiale
1	1	0.7071
2	1	0.7071
3	3	0.7071
4	3	0.7071
5	2	2.5495
6	2	2.5495

Criterion Based on Final Seeds = 1.1180

Tableau AT5

Récapitulatif sur la classification

Classification	Fréquence	RMS Std Deviation	Distance max. de la valeur initiale à l'obs.	Classification la plus proche
1	2	0.7071	0.7071	3
2	2	2.5495	2.5495	3
3	2	0.7071	0.7071	2

Tableau AT6

Statistiques pour variables

Variable	Total STD	Dans STD	R-carré	RSQ/(1-RSQ)
REV	9.98833	2.12132	0.972937	35.950617
EDUC	6.36920	0.70711	0.992605	134.222222
OVER-ALL	8.37655	1.58114	0.978622	45.777778

Tableau AT7 Cluster Means

Cluster	REV	EDUC
1	5.50000000	5.50000000
2	27.50000000	19.50000000
3	15.50000000	14.50000000

Tableau AT8 Cluster Standard Deviations

Cluster	REV	EDUC
1	0.707106781	0.707106781
2	3.535533906	0.707106781
3	0.707106781	0.707106781

Tableau AT9 Distance Between Cluster Centroids

Nearest Cluster	1	2	3
1	.	26.07680962	13.45362405
2	26.07680962	.	13.00000000
3	13.45362405	13.00000000	.

Graphe AT10 Graphe de REV*EDUC.

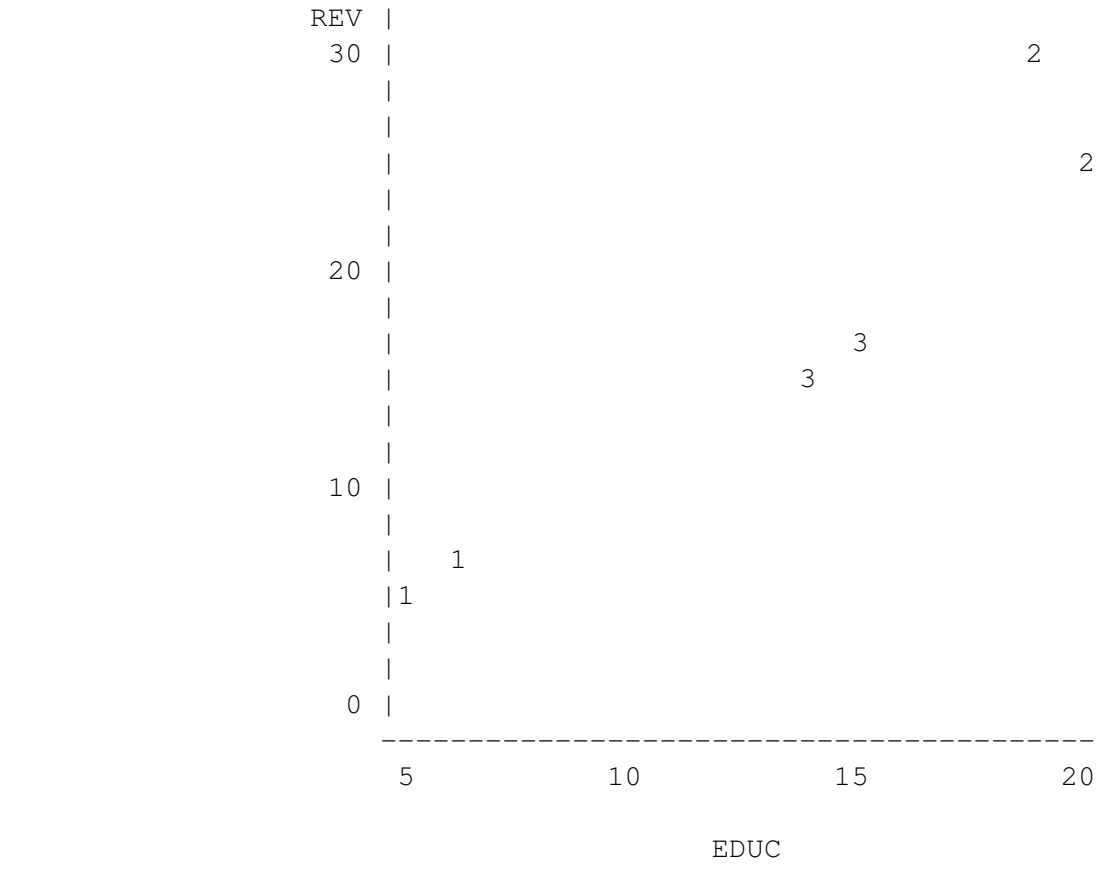


Tableau AT11

Historique des classifications

NCL	--Classifications jointes---		FREQ	SPRSQ	RSQ	Tie
5	3	4	2	0.0017	.998	T
4	1	2	2	0.0017	.997	
3	5	6	2	0.0138	.983	
2	CL5	CL3	4	0.2060	.777	
1	CL4	CL2	6	0.7768	.000	

Graphe AT 12

Graphe de _SPRSQ_*_NCL_. Légende : A = 1 obs, B = 2 obs, etc.

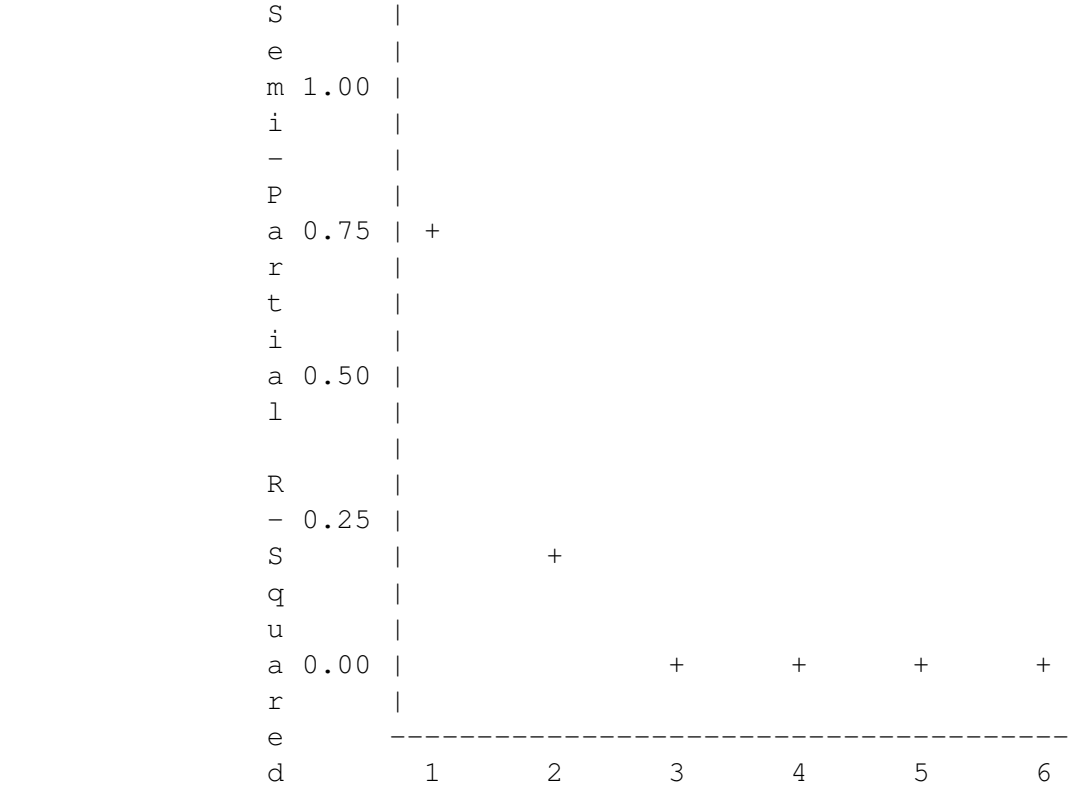


Tableau AT13

---------- CLUSTER=1 ----------

La procédure MEANS

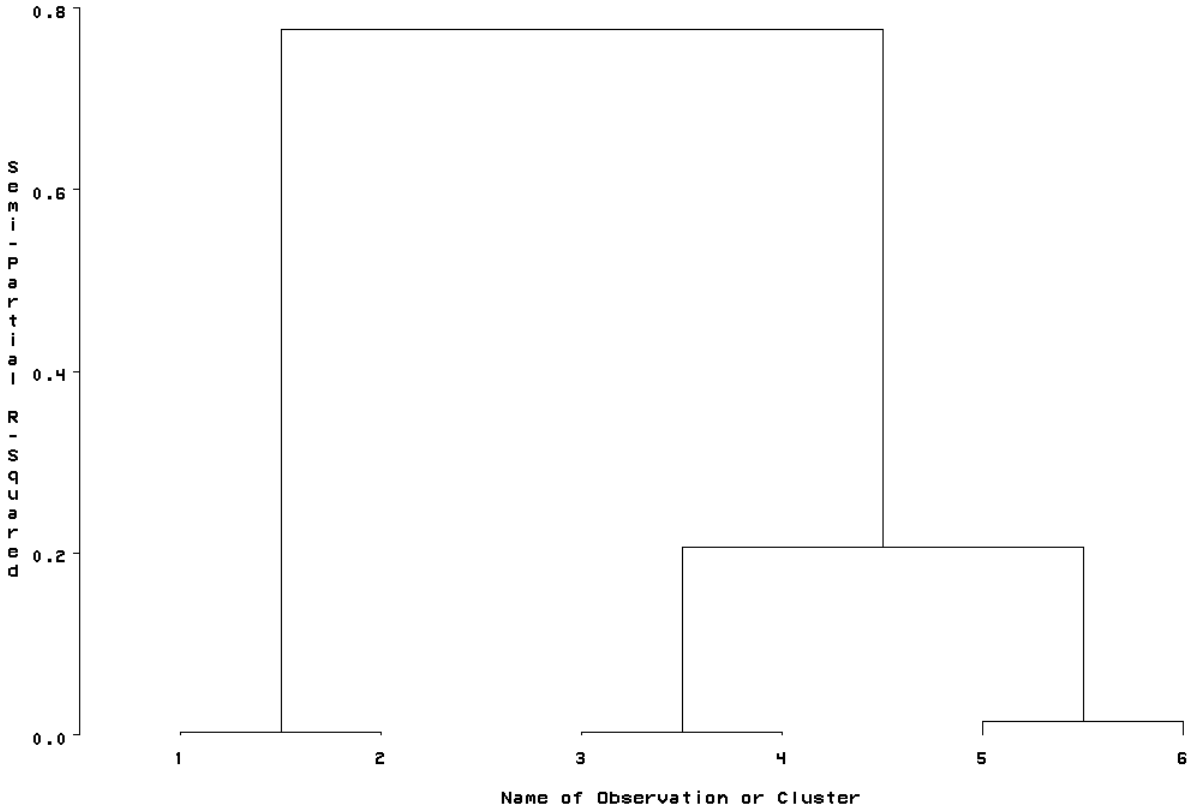
Variable	Nb	Moyenne	Écart-type	Minimum	Maximum
EDUC	2	14.5000000	0.7071068	14.0000000	15.0000000
REV	2	15.5000000	0.7071068	15.0000000	16.0000000

---------- CLUSTER=2 ----------

Variable	Nb	Moyenne	Écart-type	Minimum	Maximum
EDUC	2	5.5000000	0.7071068	5.0000000	6.0000000
REV	2	5.5000000	0.7071068	5.0000000	6.0000000

---------- CLUSTER=3 ----------

Variable	Nb	Moyenne	Écart-type	Minimum	Maximum
EDUC	2	19.5000000	0.7071068	19.0000000	20.0000000
REV	2	27.5000000	3.5355339	25.0000000	30.0000000


0.8
0.6
0.4
0.2
0.0
Semi-Partial R-Squared
1
2
3
4
5
6
Name of Observation or Cluster

Graphe AT14

Annexe 2 : code SAS

```
data tabat;
input OBS REV EDUC;
cards;
1 5 5
2 6 6
3 15 14
4 16 15
5 25 20
6 30 19
;
run;

proc plot data=tabat;
plot EDUC*REV=OBS / vpos=20 hpos=40;
run;
/*Classification par la méthode des moyennes mobiles*/
proc fastclus data=tabat out=resat
maxclusters=3 /*nombre de groupes*/
replace=part /*choix des centres initiaux les plus éloigés possible*/
maxiter =20 /*nombre maximum d'itératioons*/
/* recalcul ds centres apres chaque reaffectation d'observations*/
list /*groupe auquel appartient chaque observation et distance au
centre final du groupe*/
distance /* distance entre les centres finaux des groupes*/;
var rev educ;
run;
proc plot data=resat;
plot rev*educ=cluster / vpos=20 hpos=40;
run;
quit;

/*CAH*/


proc cluster data=tabat std method = ward noeigen out=rescah;
var REV EDUC;
id OBS;
```

```
run;

proc gplot data=rescah;
plot _SPRSQ_*_NCL_;
run;
quit;

proc tree data=rescah;
run;
quit;
proc tree data=rescah nclusters=3 out=cluster;
copy EDUC REV;
run;
quit;
proc sort data=cluster;
by cluster;
run;
proc means data=cluster;
var EDUC REV;
by cluster; run;
```

Annexe 3 : codes et sorties R

```
> reveduc   # les données
  rev educ
1   5    5
2   6    6
3  15   14
4  16   15
5  25   20
6  30   19

# Méthode d'agrégation autour des moyennes mobiles (AMM)
# on ne standardise pas car les données ont le même ordre de grandeurs
> classif=kmeans(reveduc,3)  # on choisit de faire une classification en
3 groupes : le choix du nombre de groupe
se fait en examinant les R2 par variable et le R2 global : s'ils sont
suffisamment élevés, cela valide le choix du
nombre de groupes

> names(classif)
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"
> classif$cluster # vecteur contenant les numéros du groupe auquel
chaque observation est affectée
[1] 2 2 1 1 3 3
> classif$centers # les centres des groupes dont il faut commenter
les coordonnées pour faire la typologie
   rev educ
1 15.5 14.5
2  5.5  5.5
3 27.5 19.5


> classif$size #le nombre d'individus de chaque groupe
[1] 2 2 2

> classif$betweenss #somme des carrés inter-groupes
[1] 686.6667
> classif$totss  #somme des carrés totale
[1] 701.6667
```

```
> R2=classif$betweenss/classif$totss   #R2 global : très élevé donc
classification très statisfaisante
> R2
[1] 0.9786223

> classif$tot.withinss # somme des carrés intra-groupes
[1] 15
> classif$withinss #somme des carrés intra-groupes pour chaque groupe
[1]  1  1 13
```

Grahique (voir figure 2)

```
groupe=classif$cluster
> reveduc2=cbind(reveduc,groupe)
> reveduc2
  rev educ groupe
1   5    5      2
2   6    6      2
3  15   14      1
4  16   15      1
5  25   20      3
6  30   19      3
plot(reveduc2$educ,reveduc2$rev,type="n",xlab="education",ylab="revenu")
text(reveduc2$educ,reveduc2$rev,as.character(reveduc2$groupe))
```

```
R2 par variable : on rappelle que pour la typologie (description des groupes),
on ne commente que les variables de fort R2 (>0.5)
> attach(reveduc2)

    groupe
> summary(lm(rev~as.factor(groupe)))

Call:
lm(formula = rev ~ as.factor(groupe))

Residuals:
   1    2    3    4    5    6
-0.5  0.5 -0.5  0.5 -2.5  2.5

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              15.500      1.500  10.333  0.00193 **
as.factor(groupe)[T.2]  -10.000      2.121  -4.714  0.01807 *
as.factor(groupe)[T.3]   12.000      2.121   5.657  0.01094 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.121 on 3 degrees of freedom
Multiple R-squared: 0.9729,     Adjusted R-squared: 0.9549     # donc R2(REV)=0.9729
F-statistic: 53.93 on 2 and 3 DF,  p-value: 0.004452

> summary(lm(educ~as.factor(groupe)))

Call:
lm(formula = educ ~ as.factor(groupe))

Residuals:
   1    2    3    4    5    6
-0.5  0.5 -0.5  0.5  0.5 -0.5

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             14.5000     0.5000  29.000    9e-05 ***
as.factor(groupe)[T.2]  -9.0000     0.7071 -12.728  0.00105 **
as.factor(groupe)[T.3]   5.0000     0.7071   7.071  0.00582 **
```

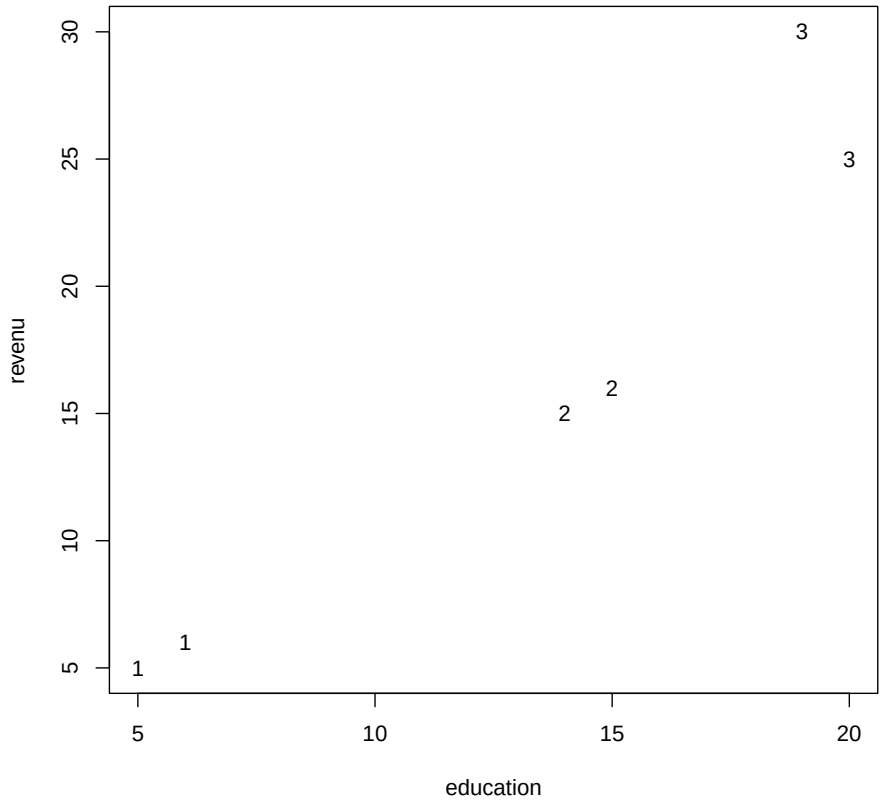


FIGURE 2 – Représentation des groupes sur les variables

```
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.7071 on 3 degrees of freedom
Multiple R-squared: 0.9926,     Adjusted R-squared: 0.9877   #donc R2(EDUC)=0.9926
F-statistic: 201.3 on 2 and 3 DF,  p-value: 0.0006362

Typologie : comme les R2 des variables educ et rev sont très élevés, on peut
commenter le niveau de ces 2 variables dans chaque groupe :
groupe 1 : rev et educ moyens
groupe 2 : rev et educ faibles
groupe 3 : rev et educ élevés

#Classification ascendante hiérarchique (CAH)
> classif2= hclust(dist(reveduc),method="ward")
> names(classif2)
[1] "merge"       "height"      "order"       "labels"      "method"
[6] "call"        "dist.method"

> plot(classif2) #pour représenter le dendogramme



> require(graphics)
> groupe2=cutree(classif2,k=3) #vecteur des numéros de
groupe auquel chaque observation
est affectée
> groupe2
[1] 1 1 2 2 3 3

> reveduc3=cbind(reveduc,groupe2)
> reveduc3
  rev educ groupe2
1   5    5       1
2   6    6       1
3  15   14       2
4  16   15       2
5  25   20       3
6  30   19       3
```

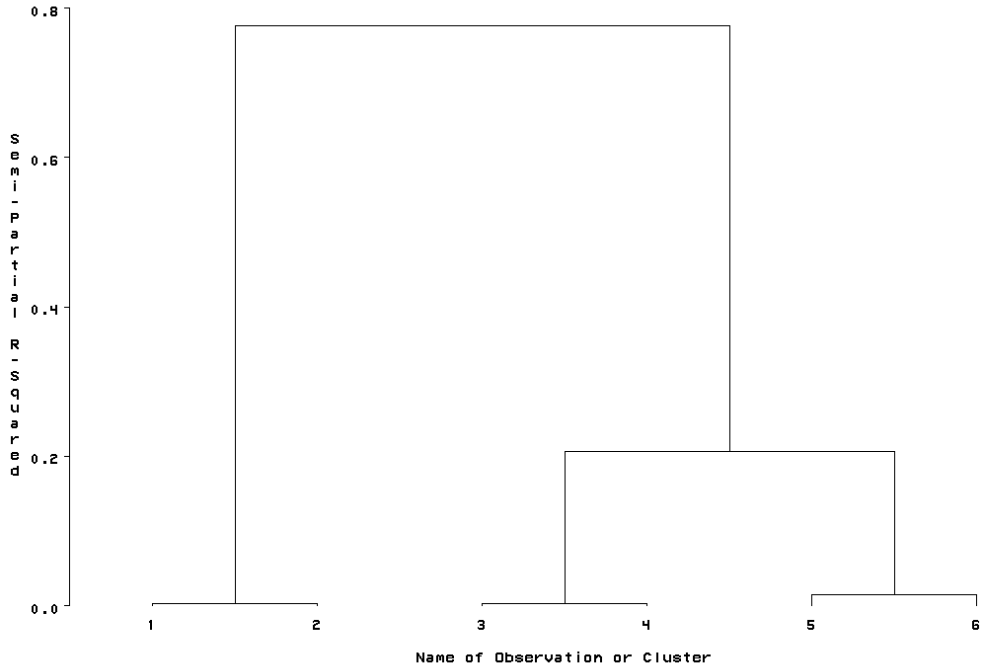


FIGURE 3 – Dendogramme

```
#Calcul des moyennes de chaque variable par groupe
pour faire la typologie
> rbind(mean(reveduc3[groupe2==1,]),
+ mean(reveduc3[groupe2==2,]),
+ mean(reveduc3[groupe2==3,]))
      rev educ groupe2
[1,]  5.5  5.5       1
[2,] 15.5 14.5       2
[3,] 27.5 19.5       3

#Graphique et calcul des R2 par variable : voir ci-dessus
```