CLASSIFICATION DE VARIABLES – APPLICATION À LA BASE PERMANENTE DES ÉQUIPEMENTS

Brigitte GELEIN (*), Olivier SAUTORY (**)

(*) Ensai
(**) Cepe

Introduction

Une classification d'un ensemble de variables, i.e. la construction d'une partition formée de classes regroupant des variables "qui se ressemblent", peut répondre à différents objectifs.

Description d'un ensemble de variables

Dans une analyse en composantes principales, l'étude de la représentation des variables dans le 1er plan principal (le "cercle des corrélations"), fait souvent apparaître des groupes de variables bien corrélées. Cette analyse peut être enrichie par une approche de type classification, qui permet de regrouper ces variables par un algorithme automatique, et non par un procédé "visuel".

Réduction du nombre de variables

Lorsque l'on dispose d'un nombre de variables trop important, une classification de ces variables permet d'obtenir un nombre réduit de classes de variables corrélées. Chaque classe peut alors être représentée par une nouvelle variable synthétisant les variables de la classe, ou bien par celle des variables analysées qui représente le mieux les classes (optique sélection de variables).

L'analyse factorielle ("*factor analysis*") avec des techniques de rotation orthogonale ou oblique des facteurs permet également de mettre en évidence des variables synthétiques bien corrélées aux variables de départ, mais elle ne conduit pas directement à des partitions de l'ensemble des variables.

Pour réaliser une classification de variables, on peut utiliser une méthode de classification ascendante hiérarchique (CAH), comme pour une classification d'individus. Il convient pour cela de définir une mesure de dissimilarité adaptée aux variables, la stratégie d'agrégation pouvant être par exemple la méthode de Ward, la plus usitée en CAH.

On trouve dans la littérature plusieurs mesures de dissimilarité : voir par exemple Nakache et Confais [6], Qannari et alii [7].

Certaines de ces mesures de dissimilarité sont fondées sur le coefficient de corrélation linéaire r, par exemple $1 - r$, $1 - |r|$, ou $1 - r^2$.

La méthode de classification de variables mise en œuvre par la procédure **VARCLUS** de **SAS** est une méthode descendante, fondée sur un critère de division d'un groupe de variables en deux classes.

Chavent et alii [1] proposent une version simplifiée de **VARCLUS**, fondée également sur une méthode divisive.

Qannari et alii [7] proposent une approche de classification de variables autour de composantes latentes (nommée CLV), qui est une méthode ascendante présentant des similarités avec la "méthode" **VARCLUS**.

L'objectif modeste de ce papier est de décrire dans une première partie la méthode de classification mise en œuvre par la procédure **VARCLUS,** en détaillant les sorties de la procédure, et dans une seconde partie de présenter les résultats d'une application de cette méthode à des données issues de la Base Permanente des Équipements de l'Insee.

1. La procédure VARCLUS de SAS

1.1. Aspects théoriques

1.1.1. Caractéristiques de la méthode

La procédure **VARCLUS** de **SAS** permet de réaliser une classification de variables numériques selon une méthode descendante. Elle peut également être mise en œuvre sur des variables binaires.
L'objectif est d'obtenir une partition des variables en classes homogènes, i.e. regroupant des variables fortement corrélées entre elles (en valeur absolue), et telles que deux variables de classes différentes soient faiblement corrélées (en valeur absolue).
Chaque classe est représentée par une combinaison linéaire des variables de la classe, appelée *composante*, et la méthode vise à maximiser la somme, sur l'ensemble des classes de la partition, des variances de ces composantes.

Partant de la classe constituée par l'ensemble des variables, on divise cet ensemble en deux classes homogènes. Si l'une (au moins) des classes n'est pas suffisamment homogène, au sens d'un certain critère, cette classe est à son tour divisée en deux classes homogènes.
À chaque étape de l'algorithme, on sélectionne la classe la moins homogène, et on procède à sa division si elle n'est pas suffisamment homogène au sens du critère choisi.
L'algorithme s'arrête lorsqu'aucune classe ne peut être divisée. On obtient ainsi *in fine* une partition de l'ensemble des variables en un nombre K de classes disjointes.

L'utilisateur peut intervenir sur K en lui imposant une valeur minimale ou une valeur maximale.

La méthode utilisée produit *de facto* des partitions en 2, 3,…, K−2, K−1 classes. Avec l'option **HIERARCHY** de l'instruction **PROC VARCLUS** (qui n'est pas l'option par défaut), on peut imposer que ces partitions soient emboîtées : la partition en k+1 classes se déduit de la partition en k classes par division en 2 d'une des classes de cette partition, les autres classes restant inchangées. Les classes constituent alors une *hiérarchie*, permettant en particulier de représenter ces partitions sous la forme d'un arbre de classification.

Par défaut, les variables analysées sont réduites (divisées par leur écart-type). Avec l'option **COVARIANCE**, les variables ne sont pas réduites.
Comme on le verra par la suite, la méthode met en œuvre des analyses en composantes principales (ACP) sur des groupes de variables. Lors d'une ACP, la matrice diagonalisée sera donc :

- par défaut, la matrice des corrélations : toutes les variables ont le même poids dans l'analyse ;
- avec l'option **COVARIANCE,** la matrice des covariances : les variables jouent un rôle d'autant plus grand dans l'analyse que leurs variances sont plus élevées.

1.1.2. Caractéristiques d'une classe : composante, variance, valeurs propres

Une classe CL composée des variables Y_j est représentée par une *composante* notée C, qui est une combinaison linéaire des variables : par défaut, C est la 1ère composante principale des variables Y_j ; si on spécifie l'option **CENTROID**, la *composante* C est la moyenne arithmétique des Y_j.

La *variance de la classe* CL est égale à la somme des variances V_j des variables Y_j :

$$V_{CL} = \sum_{Y_j \in CL} V_j$$

Si les variables sont réduites, elle est égale au nombre de variables de la classe.

Cette quantité ne doit pas être confondue avec la variance de la *composante*, notée V_C.

On notera dans suite λ_1, λ_2 ($\lambda_1 \geq \lambda_2$) les variances des deux premières composantes principales de la classe CL ; ce sont aussi les deux premières valeurs propres de la matrice (des corrélations ou des

covariances) diagonalisée en ACP lors du calcul de ces composantes principales. On les appellera les *valeurs propres de la classe*.
Lorsque la classe est représentée par la 1ère composante principale des variables Y_j , alors $V_C = \lambda_1$.

1.1.3. Critère de division d'une classe.

Le principe est le suivant : une classe est divisée en deux si elle n'est pas suffisamment homogène, i.e. si la *composante* représentant la classe ne "résume" pas à elle seule l'ensemble des variables de la classe, au sens d'un certain critère. Une deuxième *composante* est nécessaire pour représenter les variables de la classe, qui doit donc être divisée en deux classes.

La procédure **VARCLUS** propose deux critères pour décider de diviser ou non une classe.

1.1.3.1. la deuxième valeur propre

On décide de diviser la classe CL en deux classes si sa *2ème valeur propre* λ_2 est supérieure à un certain seuil λ : cela signifie que la 2ème composante principale est un résumé suffisamment "informatif" des variables de la classe CL, car elle a une variance supérieure au seuil λ.

Ce seuil peut être choisi en renseignant le paramètre **MAXEIGEN** = λ. La valeur par défaut du seuil est :

- cas de variables réduites : $\lambda = 1$; on retrouve le classique critère de Kaiser utilisé en ACP normée pour sélectionner axes principaux et composantes principales : les composantes principales doivent avoir une variance supérieure à la variance, égale à 1, de chaque variable initiale réduite ;
- cas de variables non réduites : λ = moyenne des variances des variables de la classe.

Si $\lambda_2 \leq \lambda$, la classe CL n'est pas divisée.

Ce critère sera appelé *critère* λ.

1.1.3.2. la part de variance expliquée par la composante de la classe

On décide de diviser la classe CL en deux classes si la part de variance expliquée par la *composante* C de la classe (i.e. soit la 1ère composante principale, soit la moyenne des variables, voir § 1.1.2.) est inférieure à un certain seuil *p*. Cela signifie que deux composantes sont nécessaires pour représenter la classe, qui dont donc être divisée.

Ce seuil peut être choisi en renseignant le paramètre **PROPORTION** = *p*. La valeur par défaut du seuil est $p = 0.75$.

Si $V_C / V_{CL} > p$, la classe C n'est pas divisée.

Ce critère sera appelé *critère p*.

Remarques

En l'absence de l'option **CENTROID**, le *critère* λ est choisi par défaut par la procédure.

En présence de l'option **CENTROID**, le *critère* λ ne peut pas être utilisé.

1.1.4. Construction des classes.

Supposons que l'on soit à la $k^{ème}$ étape de l'algorithme descendant : la partition P_k répartit les variables de départ en k classes. Les différentes étapes conduisant à la partition P_{k+1} en k+1 classes sont les suivantes.

1.1.4.1. Choix de la classe à diviser.

Le choix de la classe à diviser dépend du critère de division utilisé :

- avec le *critère* λ, on sélectionne la classe ayant la plus forte deuxième valeur propre λ_2. Si $\lambda_2 \leq \lambda$, l'algorithme s'arrête, sinon, la classe est divisée ;

- avec le *critère p* : on sélectionne la classe CL ayant la plus petite part de variance expliquée par sa *composante*. Si $V_C / V_{CL} > p$ l'algorithme s'arrête, sinon, la classe est divisée.

1.1.4.2. Initialisation du processus de division de la classe

On réalise une analyse en composantes principales (ACP) sur les variables de la classe à diviser.
Plutôt que de conserver telles quelles les deux premières composantes principales de l'analyse, on préfère effectuer sur ces composantes principales une *rotation orthoblique suivant la méthode quartimax*.

L'objectif d'une rotation dans le 1^{er} plan factoriel des variables est d'obtenir deux nouvelles composantes plus facilement interprétables en fonction des variables initiales, car mieux corrélées (en valeur absolue) avec certaines de ces variables, et donc moins bien corrélées avec les autres variables.
Cette rotation peut être *orthogonale*, ce qui signifie que les nouvelles composantes sont non corrélées, comme celles issues de l'ACP "classique", ou bien *oblique*, auquel cas les composantes peuvent être corrélées.
Il existe plusieurs variantes pour chacun des deux types de rotation, dont la méthode *quartimax* utilisée ici. Pour plus de précisions, voir par exemple Harman [4].

1.1.4.3. Affectation des variables dans les classes

On note $CL_1 \ldots CL_i \ldots CL_{k-1}, CL_k$ les k classes de la partition P_k, où CL_k désigne la classe qui va être divisée en deux classes CL_{k1} et CL_{k2}.
Les classes $CL_1 \ldots CL_i \ldots CL_{k-1}$ sont représentées par leurs *composantes*, les classes CL_{k1} et CL_{k2} sont représentées par les deux premières composantes principales de l'ACP oblique réalisée sur CL_k.

L'affectation des variables dans les classes $CL_1 \ldots CL_i \ldots CL_{k-1}$, CL_{k1}, CL_{k2}, s'effectue selon une procédure en deux phases.

$1^{ère}$ phase : *NCS* (*nearest component sorting*)

L'algorithme mis en œuvre est similaire à la méthode usuelle des centres mobiles.

1. Chaque variable est affectée à la classe dont la *composante* est la plus corrélée avec la variable (au sens du carré du coefficient de corrélation linéaire r^2).
2. On calcule la *composante* de chacune des nouvelles classes ainsi constituées, et on réaffecte chaque variable à la classe dont la *composante* est la plus corrélée avec la variable (au sens du r^2).
3. Le processus est itéré jusqu'à ce que la composition des classes ne varie plus.

La composition des classes $CL_1 \ldots CL_i \ldots CL_{k-1}$ peut changer au cours de cet algorithme.
Avec l'option **HIERARCHY**, l'algorithme précédent ne concerne que les variables de la classe CL_k, qui sont affectées soit à la classe CL_{k1}, soit à la classe CL_{k2}. La composition des classes $CL_1 \ldots CL_i \ldots CL_{k-1}$

n'est donc pas remise en cause, ce qui permet d'obtenir une structure hiérarchique des classes successivement constituées.

Une fois cet algorithme achevé, on peut alors calculer la *variance expliquée par la partition*, définie comme la somme des variances des *composantes* de chacune des k+1 classes.

2ème phase : *Search*

À l'issue de la 1ère phase, on teste chaque variable pour voir si l'affectation de cette variable à une autre classe augmente la *variance expliquée par la partition*.

Si c'est le cas, on change donc la variable de classe, la *composante* de chacune des deux classes concernées par le transfert est recalculée avant le test de la variable suivante.

Avec l'option **HIERARCHY**, cette phase de recherche ne concerne que les variables des nouvelles classes CL_{k1}, et CL_{k2}. La composition des classes $CL_1 \ldots CL_j \ldots CL_{k-1}$ n'est donc pas remise en cause, ce qui permet d'obtenir une structure hiérarchique des classes successivement constituées.

À l'issue de ces deux phases, on obtient donc la partition P_{k+1} en k+1 classes. Si l'une de ses classes peut être divisée (voir § 1.1.4.1.), la partition en k+2 classes est construite selon le processus qui vient d'être décrit. Sinon, l'algorithme descendant s'achève, et la partition P_k est la partition finale.

Remarque

La documentation SAS (voir [8]) indique que :

- la phase *NCS* converge rapidement, avec un risque de convergence vers un optimum local ; la phase *Search* peut être longue lorsqu'elle opère sur un grand nombre de variables ;
- sans l'option **CENTROID**, la phase *Search* apporte rarement une amélioration substantielle sur les résultats de la phase *NCS* ;
- avec l'option **CENTROID**, la phase *NCS* peut ne pas augmenter la *variance expliquée par la partition*, elle est donc limitée à une itération.

1.1.4.4. Autres critères d'arrêt de l'algorithme

On peut aussi agir directement sur le nombre de classes de la partition finale avec deux paramètres.

Avec l'option **MAXCLUSTERS=**m, la partition finale a au plus m classes. Par défaut, m = nombre de variables analysées.

Avec l'option **MINCLUSTERS=**n, la partition finale a au moins n classes. Par défaut, n = 2.

1.2. Exemple de mise en œuvre de la procédure

Les données analysées sont issues d'une enquête d'opinion réalisée par l'association Agoramétrie en 1987 auprès de 698 personnes. On a demandé à chacune des personnes interrogées de se prononcer sur 27 "thèmes d'actualité" (il s'agit de l'actualité de 1987…) à travers une échelle d'accord à 5 positions :

1 : Pas du tout d'accord ; 2 : Pas tellement d'accord ; 3 : Peut-être d'accord ;
4 : Bien d'accord ; 5 : Entièrement d'accord

Les libellés détaillés des thèmes figurent en annexe 1.

On a utilisé les options par défaut de la procédure **VARCLUS** :

- les variables sont réduites ;
- la *composante* d'une classe est sa 1ère composante principale ;
- le critère de division est le *critère* λ, avec le seuil $\lambda = 1$;
- on n'impose pas de structure hiérarchique aux classes.

1.2.1. Étape 1 de l'algorithme descendant

1.2.1.1. La première partition

Tableau 1 : partition en 1 classe

Cluster Summary for 1 Cluster

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	27	27	4.58125	0.1697	2.6110

Total variation explained = 4.58125 Proportion = 0.1697

Cluster 1 will be split because it has the largest second eigenvalue, 2.610956, which is greater than the MAXEIGEN=1 value.

La première partition est constituée d'une seule classe, contenant les 27 variables (*Members*).

La variance de cette classe (*Cluster Variation*) est égale à 27, puisque les variables sont réduites.

La variance expliquée (*Variation Explained*) par cette classe, égale à la variance de sa *composante*, donc ici sa 1ère composante principale, vaut λ_1=4.58125. Ceci représente une part de variance expliquée (*Proportion Explained*) égale à 16.97 %.

La seconde valeur propre (*Second Eigenvalue*), i.e. la variance de la 2ème composante principale, vaut λ_2=2.6110. Cette valeur étant supérieure à 1, la classe va être divisée.

1.2.1.2. Construction de la partition en deux classes

Tableau 2 : affectation des variables aux 2 classes

Oblique Principal Component Cluster Analysis

Phase	Iteration	Variance Accounted For	famille_ cellule_ de_base_ societe A	avoir_confiance_ en_la_justice B	force_de_ frappe_ indispensable C	haschich_ en_vente_ libre D	enseignants_ consciencieux E
Split	0	4.581250	1	1	1	1	1
NCS	1	6.846434	1	2	2	1	1
NCS	2	6.846434	1	2	2	1	1
Search	1	6.881277	1	2	1	1	1
Search	2	6.905070	1	1	1	1	2
Search	3	6.905070	1	1	1	1	2

Ce tableau indique les différentes étapes de l'algorithme de construction des deux classes, à partir de l'ACP oblique et des phases de réaffectation. Pour l'éditer, il faut spécifier l'option **TRACE**, qui permet d'obtenir un tableau semblable pour chacune des divisions de l'algorithme descendant.

Dans le tableau ci-dessus, seules 5 variables sur les 27 sont présentées, renommées par commodité A, B, C, D et E, et illustrant différents "destins".

Sur la 1ère ligne (*Split*), on trouve la composition de la partition avant la division : ici, toutes les variables appartiennent bien sûr à la même classe, notée 1, avec une variance expliquée, ou "prise en compte" (*Variance Accounted For*) égale 4.581250.

Les deux lignes suivantes concernent la phase d'affectation *NCS*. Sur la 1ère ligne (*Iteration 1*), on lit l'affectation des variables aux classes 1 (variables A, D et E) et 2 (variables B et C). La variance expliquée par cette partition en 2 classes vaut 6.846434. Sur la 2ème ligne (*Iteration 2*), on observe que les variables n'ont pas changé de classe, et que la variance expliquée par la partition est inchangée : il n'y a pas eu de réaffectation de variables suite au calcul des nouvelles *composantes* de chaque classe après l'itération 1, l'algorithme a convergé.

Les trois lignes suivantes concernent la phase d'affectation *Search*. Sur la 1ère ligne (*Iteration 1*), on observe un premier changement de classe : la variable C passe de la classe 2 à la classe 1, entraînant une augmentation de la variance expliquée, qui passe à 6.881277. À l'étape suivante (*Iteration 2*), la variable E passe de la classe 1 à la classe 2, la variable B passe de la classe 2 à la classe 1, et la variance expliquée par la nouvelle partition vaut 6.905070. Sur la 3ème ligne (*Iteration 3*), on observe que les variables n'ont pas changé de classe, et que la variance expliquée par la partition est inchangée : il n'y a pas eu de réaffectation de variables, l'algorithme a convergé.

Remarque : SAS n'édite pas les premières composantes résultant de l'ACP oblique.

1.2.1.3. Résumé de la partition en 2 classes

Tableau 3 : partition en 2 classes

Cluster Summary for 2 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	17	17	4.31369	0.2537	1.8186
2	10	10	2.59138	0.2591	1.2240

Total variation explained = 6.90507 Proportion = 0.2557

Cluster 1 will be split because it has the largest second eigenvalue, 1.818646, which is greater than the MAXEIGEN=1 value.

- La classe 1 contient 17 variables, sa variance vaut 17. La variance de sa *composante* vaut 4.31369, qui explique 25.37% de la variance de la classe. La seconde valeur propre vaut 1.8186.
- La classe 2 contient 10 variables, sa variance vaut 10. La variance de sa *composante* vaut 2.59138, qui explique 25.91% de la variance de la classe. La seconde valeur propre vaut 1.2240.

- La variance expliquée par la partition (*Total variation explained*), égale à la somme des variances des *composantes* de chaque classe, vaut 6.90507, ce qui représente une part (*Proportion*) de la variance totale (qui vaut 27) égale à 25.57%.

La classe 1 est celle qui a la plus grande seconde valeur propre λ_2=1.8186, qui est supérieure à 1 : cette classe va donc être divisée à l'étape suivante.

La procédure édite d'autres tableaux donnant des informations détaillées sur la composition des classes, sur la structure des corrélations dans les classes, et entre les classes. Seuls les tableaux illustrant la partition finale seront analysés, au § 1.2.3.

1.2.2. Étapes 2 à 6 de l'algorithme descendant

On trouvera ci-dessous les tableaux résumés correspondant aux 5 étapes suivantes de l'algorithme descendant.

Tableau 4 : partition en 3 classes

Cluster Summary for 3 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	10	10	2.940764	0.2941	1.4544
2	9	9	2.576148	0.2862	1.2233
3	8	8	2.994585	0.3743	1.0278

Total variation explained = 8.511497 Proportion = 0.3152

Cluster 1 will be split because it has the largest second eigenvalue, 1.454432, which is greater than the MAXEIGEN=1 value.

La classe 1 de la partition en 2 classes a été divisée en deux classes, notées 1 et 3. Le jeu des réaffectations entre classes, en l'absence de l'option **HIERARCHY**, a fait que la classe 2 de la partition en 2 classes a perdu une variable.

Cette nouvelle partition explique 31.52% de la variance totale.

Tableau 5 : partition en 4 classes

Cluster Summary for 4 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	8	8	2.687041	0.3359	1.1175
2	9	9	2.576148	0.2862	1.2233
3	7	7	2.705006	0.3864	1.0192
4	3	3	1.874096	0.6247	0.6599

Total variation explained = 9.842292 Proportion = 0.3645

Cluster 2 will be split because it has the largest second eigenvalue, 1.223275, which is greater than the MAXEIGEN=1 value.

La classe 1 de la partition en 3 classes a été divisée en deux classes, notées 1 et 4. Le jeu des réaffectations entre classes a fait que la classe 3 de la partition en 3 classes a perdu une variable, la classe 2 restant inchangée.

Cette nouvelle partition explique 36.45% de la variance totale.

Tableau 6 : partition en 5 classes

Cluster Summary for 5 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	8	8	2.687041	0.3359	1.1175
2	5	5	2.201973	0.4404	0.9219

3	7	7	2.705006	0.3864	1.0192
4	3	3	1.874096	0.6247	0.6599
5	4	4	1.5408	0.3852	0.8833

Total variation explained = 11.00892 Proportion = 0.4077

Cluster 1 will be split because it has the largest second eigenvalue, 1.117519, which is greater than the MAXEIGEN=1 value.

La classe 2 de la partition en 4 classes a été divisée en deux classes, notées 2 et 5. Les autres classes sont inchangées.

Cette nouvelle partition explique 40.77% de la variance totale.

Tableau 7 : partition en 6 classes

Cluster Summary for 6 Clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	1.906793	0.4767	0.8581
2	5	5	2.201973	0.4404	0.9219
3	7	7	2.960164	0.4229	0.8852
4	3	3	1.874096	0.6247	0.6599
5	4	4	1.5408	0.3852	0.8833
6	4	4	1.734857	0.4337	0.9261

Total variation explained = 12.21868 Proportion = 0.4525

No cluster meets the criterion for splitting.

La classe 1 de la partition en 5 classes a été divisée en deux classes, notées 1 et 6. Les classes 2, 4 et 5 sont inchangées. La classe 3, même si elle a le même nombre de variables, a vu sa composition changer, ce qui l'a rendue plus homogène : la variance de sa *composante* a augmenté, et sa 2ème valeur propre a diminué, passant en dessous du seuil de 1. Il n'y a plus de classe à diviser.

La partition finale explique 45.25% de la variance totale.

1.2.3. Description de la partition finale en 6 classes

1.2.3.1. Composition des classes

Tableau 8 : composition des classes

6 Clusters		R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio
Cluster 1	famille_cellule_de_base_societe	0.5863	0.0576	0.4390
	respecter_les_convenances	0.4696	0.0881	0.5816
	ne_plus_se_marier	0.3837	0.0592	0.6551
	haschich_en_vente_libre	0.4671	0.0992	0.5915
Cluster 2	adopter_semaine_des_35h	0.5272	0.0782	0.5129
	abaisser_age_retraite	0.4869	0.0438	0.5366
	retablir_IGF	0.4173	0.0643	0.6227
	embaucher_dans_services_publics	0.4094	0.0268	0.6069
	reduire_ecarts_entre_revenus	0.3612	0.0553	0.6762
Cluster 3	on_ne_se_sent_plus_en_securite	0.3896	0.0689	0.6556
	isoler_les_malades_du_sida	0.4231	0.0365	0.5987
	etudiants_parasites_societe	0.3081	0.0521	0.7299
	retablir_la_peine_de_mort	0.5372	0.0908	0.5090
	trop_de_travailleurs_immigres	0.5267	0.0655	0.5064
	homosexuels_comme_les_autres	0.3820	0.1486	0.7259
	controle_identite_indispensable	0.3933	0.1707	0.7316
Cluster 4	lutter_contre_la_pornographie	0.7081	0.1347	0.3373
	censurer_certains_livres	0.5575	0.1562	0.5245
	bien_de_voir__femmes_nues_tele	0.6085	0.0339	0.4052
Cluster 5	soutenir_mouvements_ecologiques	0.5230	0.0927	0.5258
	developper_energie_solaire	0.3730	0.0225	0.6414
	adherer_assoc_def_consommateur	0.3502	0.0451	0.6805
	pollution_preoccupante	0.2947	0.0181	0.7183
Cluster 6	enseignants_consciencieux	0.2927	0.0231	0.7240
	avoir_confiance_en_la_justice	0.5129	0.0452	0.5101
	police_remplit_bien_sa_mission	0.5725	0.0869	0.4681
	force_de_frappe_indispensable	0.3567	0.0764	0.6966

Pour chaque variable, on lit :

- *R-squared with Own Cluster* : carré du coefficient de corrélation linéaire entre la variable Y_j et la *composante* $C_{k\,own}$ de sa classe $CL_{k\,own}$: $r^2(Y_j, C_{k\,own})$
- *R-squared with Next Closest* : carré du coefficient de corrélation linéaire entre la variable Y_j et la *composante* $C_{k\,next}$ de la classe "la plus proche" au sens de la corrélation au carré, autre que sa propre classe : $r^2(Y_j, C_{k\,next}) = \max_{k' \neq k\,own} r^2(Y_j, C_{k'})$

  N.B. Pour savoir de quelle classe il s'agit, il faut se reporter au tableau 10 ci-dessous, *Cluster Structure*.
- *1−R**2 Ratio* : $\left(1 - r^2(Y_j, C_{k\,own})\right) / \left(1 - r^2(Y_j, C_{k\,next})\right)$ [1]

Les classes sont d'autant plus homogènes, et bien séparées, que les $r^2(Y_j, C_{k\,own})$ sont élevés, que les $r^2(Y_j, C_{k\,next})$ sont faibles, et que les *1−R**2 Ratio* sont faibles.

Les variables ont été réordonnées en entrée de la procédure **VARCLUS**, dans l'instruction **VAR**, de telle façon qu'elles apparaissent dans les tableaux suivants regroupées par classe.

Quelques commentaires

Les variables sont regroupées en grands thèmes, que l'on pourrait caractériser très (très) schématiquement ainsi : famille-convenances, travail-revenus, la peur de l'autre, censure, écologie, les 4 piliers de la société.

On lit par exemple que la variable **famille_cellule_de_base_societe** est bien représentative de la classe 1, puisqu'elle est très corrélée (r^2=0.5863) avec la composante de cette classe, et faiblement corrélée (r^2=0.0576) à la classe dont elle est le plus proche.

[1] Ce coefficient peut être supérieur à 1 si on utilise l'option **HIERARCHY.**

C'est la classe 4 qui semble la plus homogène, et la mieux séparée des autres classes, car les coefficients *1−R**2 Ratio* sont plus faibles que dans les autres classes. Même si on soupçonne que les 3 variables de cette classe ne sont pas toutes corrélées positivement, on ne lit pas cette information dans ce tableau.

1.2.3.2. Coefficients des composantes des classes

Tableau 9 : coefficients des composantes des classes

Standardized Scoring Coefficients

Cluster	1	2	3	4	5	6
famille_cellule_de_base_societe	0.401574	0.000000	0.000000	0.000000	0.000000	0.000000
respecter_les_convenances	0.359389	0.000000	0.000000	0.000000	0.000000	0.000000
ne_plus_se_marier	-.324865	0.000000	0.000000	0.000000	0.000000	0.000000
haschich_en_vente_libre	-.358444	0.000000	0.000000	0.000000	0.000000	0.000000
adopter_semaine_des_35h	0.000000	0.329753	0.000000	0.000000	0.000000	0.000000
abaisser_age_retraite	0.000000	0.316882	0.000000	0.000000	0.000000	0.000000
retablir_IGF	0.000000	0.293368	0.000000	0.000000	0.000000	0.000000
embaucher_dans_services_publics	0.000000	0.290574	0.000000	0.000000	0.000000	0.000000
reduire_ecarts_entre_revenus	0.000000	0.272926	0.000000	0.000000	0.000000	0.000000
on_ne_se_sent_plus_en_securite	0.000000	0.000000	0.210866	0.000000	0.000000	0.000000
isoler_les_malades_du_sida	0.000000	0.000000	0.219750	0.000000	0.000000	0.000000
etudiants_parasites_societe	0.000000	0.000000	0.187526	0.000000	0.000000	0.000000
retablir_la_peine_de_mort	0.000000	0.000000	0.247602	0.000000	0.000000	0.000000
trop_de_travailleurs_immigres	0.000000	0.000000	0.245181	0.000000	0.000000	0.000000
homosexuels_comme_les_autres	0.000000	0.000000	-.208796	0.000000	0.000000	0.000000
controle_identite_indispensable	0.000000	0.000000	0.211856	0.000000	0.000000	0.000000
lutter_contre_la_pornographie	0.000000	0.000000	0.000000	0.449014	0.000000	0.000000
censurer_certains_livres	0.000000	0.000000	0.000000	0.398396	0.000000	0.000000
bien_de_voir__femmes_nues_tele	0.000000	0.000000	0.000000	-.416243	0.000000	0.000000
soutenir_mouvements_ecologiques	0.000000	0.000000	0.000000	0.000000	0.469338	0.000000
developper_energie_solaire	0.000000	0.000000	0.000000	0.000000	0.396363	0.000000
adherer_assoc_def_consommateur	0.000000	0.000000	0.000000	0.000000	0.384067	0.000000
pollution_preoccupante	0.000000	0.000000	0.000000	0.000000	0.352312	0.000000
enseignants_consciencieux	0.000000	0.000000	0.000000	0.000000	0.000000	0.311870
avoir_confiance_en_la_justice	0.000000	0.000000	0.000000	0.000000	0.000000	0.412818
police_remplit_bien_sa_mission	0.000000	0.000000	0.000000	0.000000	0.000000	0.436150
force_de_frappe_indispensable	0.000000	0.000000	0.000000	0.000000	0.000000	0.344250

Pour une classe CL_k donnée, sa *composante* C_k est une combinaison linéaire des variables de sa classe, de variance V_{C_k} . On peut écrire :

$$C_k = \sqrt{V_{C_k}} \sum_{Y_j \in CL_k} \alpha_j Y_j$$

On lit dans la colonne k (k=1…6) du tableau les α_j, coefficients de la *composante* "standardisée" (i.e. réduite), désignés par *Standardized Scoring Coefficients*.

Remarque : les *Standardized Scoring Coefficients* sont reliés aux coefficients de corrélation des variables avec les composantes de leur classe par la relation :

$$r(Y_j, C_{k\,own}) = V_{C_{k\,own}}\, \alpha_j$$

Quelques commentaires

Ce tableau permet de calculer la valeur de la *composante* de chaque classe. On obtient par exemple pour la classe CL_4 :

$C_4 / \sqrt{1.874} = 0.449 \times$ (contre pornographie) $+ 0.398 \times$ (censurer livres) $- 0.416 \times$ (femmes nues bien)

On constate que, en général, pour une classe donnée, les variables ont des poids similaires dans la *composante* de la classe.

1.2.3.3. Structure des corrélations

Tableau 10 : corrélations variables-composantes

Cluster Structure

Cluster	1	2	3	4	5	6
famille_cellule_de_base_societe	0.765718	-.134183	0.240062	0.233802	0.081852	0.231711
respecter_les_convenances	0.685279	-.065123	0.296854	0.260318	0.068007	0.215960
ne_plus_se_marier	-.619451	0.235739	-.160830	-.243402	0.104275	-.221958
haschich_en_vente_libre	-.683478	0.119424	-.314929	-.292810	0.130651	-.204968
adopter_semaine_des_35h	-.217113	0.726108	-.164954	-.073385	0.279612	-.154919
abaisser_age_retraite	-.178200	0.697766	0.020603	-.052389	0.167287	-.209339
retablir_IGF	-.105322	0.645989	-.053144	0.009449	0.253545	-.163579
embaucher_dans_services_publics	-.079852	0.639835	-.069374	0.006791	0.163849	-.076723
reduire_ecarts_entre_revenus	-.053186	0.600976	-.020539	0.082941	0.235227	-.054734
on_ne_se_sent_plus_en_securite	0.262561	0.072907	0.624199	0.221123	0.017520	-.042079
isoler_les_malades_du_sida	0.114866	-.015396	0.650496	0.191000	-.188511	0.063616
etudiants_parasites_societe	0.037740	0.000803	0.555108	0.131251	-.228208	0.086385
retablir_la_peine_de_mort	0.301255	-.105076	0.732941	0.188947	-.155669	0.098204
trop_de_travailleurs_immigres	0.256000	-.072721	0.725775	0.168386	-.112379	0.117643
homosexuels_comme_les_autres	-.266577	0.113209	-.618069	-.385523	0.167521	-.143204
controle_identite_indispensable	0.413106	-.159906	0.627128	0.254639	-.080083	0.268186
lutter_contre_la_pornographie	0.366960	-.020240	0.252545	0.841495	0.018064	0.168944
censurer_certains_livres	0.326540	0.015500	0.395206	0.746632	0.029158	0.131449
bien_de_voir__femmes_nues_tele	-.184053	0.025239	-.155868	-.780079	0.038884	-.051316
soutenir_mouvements_ecologiques	-.113712	0.304524	-.253601	-.051026	0.723156	-.078862
developper_energie_solaire	0.018921	0.149938	-.089235	-.007844	0.610717	0.013587
adherer_assoc_def_consommateur	0.016779	0.212474	-.056936	0.058665	0.591770	-.016775
pollution_preoccupante	0.045500	0.134671	-.063311	0.022900	0.542843	-.083164
enseignants_consciencieux	0.015184	0.086084	-.151979	-.011501	0.060869	0.541050
avoir_confiance_en_la_justice	0.212621	-.139040	0.049125	0.098424	-.061352	0.716181
police_remplit_bien_sa_mission	0.294738	-.226156	0.232722	0.197996	-.057351	0.756657
force_de_frappe_indispensable	0.276484	-.213371	0.252140	0.076067	-.104611	0.597225

Ce tableau donne les coefficients de corrélation linéaire entre chaque variable Y_j et la *composante* C_k de chaque classe CL_k : $r(Y_j, C_k)$.

Quelques commentaires

Ce tableau complète le tableau 8, puisqu'on lit les corrélations (signées) de chaque variable avec la *composante* de sa classe, et qu'on mesure la proximité de chaque variable avec chaque classe.

On retrouve par exemple que la variable **famille_cellule_de_base_societe** est bien corrélée avec la *composante* de sa classe (r=0.77), et faiblement corrélée avec les autres *composantes* ($r \leq 0.24$).

À l'inverse, la variable **controle_identite_indispensable** a une corrélation de 0.63 avec la *composante* de sa classe CL_3, mais aussi une corrélation de 0.41 avec la *composante* de la classe CL_1, ce qui explique la forte valeur de son *1−R**2 Ratio* dans le tableau 8.

Tableau 11 : corrélations entre classes

Inter-Cluster Correlations

Cluster	1	2	3	4	5	6
1	1.00000	-0.19668	0.36822	0.37147	-0.02340	0.31624
2	-0.19668	1.00000	-0.08922	-0.01342	0.33141	-0.20264
3	0.36822	-0.08922	1.00000	0.33572	-0.19857	0.16118
4	0.37147	-0.01342	0.33572	1.00000	0.00354	0.14959
5	-0.02340	0.33141	-0.19857	0.00354	1.00000	-0.06737
6	0.31624	-0.20264	0.16118	0.14959	-0.06737	1.00000

Ce tableau donne les coefficients de corrélation linéaire entre les *composantes* de chaque classe :

$$r(C_k, C_{k'})$$

Quelques commentaires

On lit par exemple une corrélation de 0.33 entre les *composantes* des classes CL_2 et CL_5, qui sont les classes qui ont été créées à la dernière étape de l'algorithme descendant, et une corrélation de 0.37 entre les composantes des classes CL_1 et CL_4, qui sont les classes qui ont été créées à l'étape précédente.

1.2.4. Récapitulatif de la classification

Tableau 11 : caractéristiques des partitions successives obtenues

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.581250	0.1697	0.1697	2.610956	0.0052	
2	6.905070	0.2557	0.2537	1.818646	0.0247	0.9768
3	8.511497	0.3152	0.2862	1.454432	0.0510	0.9580
4	9.842292	0.3645	0.2862	1.223275	0.0688	0.9435
5	11.008917	0.4077	0.3359	1.117519	0.0688	0.9435
6	12.218684	0.4525	0.3852	0.926063	0.2927	0.7316

Number of Clusters : nombre K de classes de la partition P_K.

Total Variation Explained by Clusters : variance expliquée par la partition P_K

$V(P_K) = \sum_{k=1}^{K} V_{C_k}$, où C_k désigne la composante de la classe CL_k.

Proportion of Variation Explained by Clusters : part de la variance totale expliquée par la partition

$$V(P_K) / \sum_{j=1}^{J} V(Y_j)$$

Minimum Proportion Explained by a Cluster : valeur minimale, sur l'ensemble des K classes, de la part de variance de la classe expliquée par sa *composante*

$$\underset{k=1..K}{\text{Min}}\ V_{C_k} / V_{CL_k} \quad \text{où } V_{CL_k} = \sum_{Y_j \in CL_k} V(Y_j)$$

Maximum Second Eigenvalue in a Cluster : valeur maximale, sur l'ensemble des K classes, de la seconde valeur propre de chaque classe (i.e. de la variance de la seconde composante principale de chaque classe).

Minimum R-squared for a Variable : valeur minimale, sur l'ensemble des variables, des coefficients *R-squared with Own Cluster*.

*Maximum 1−R**2 Ratio for a Variable :* valeur maximale, sur l'ensemble des variables, des coefficients *1−R**2 Ratio.*

Quelques commentaires

Ce tableau montre bien que les partitions contiennent des classes de plus en plus homogènes :

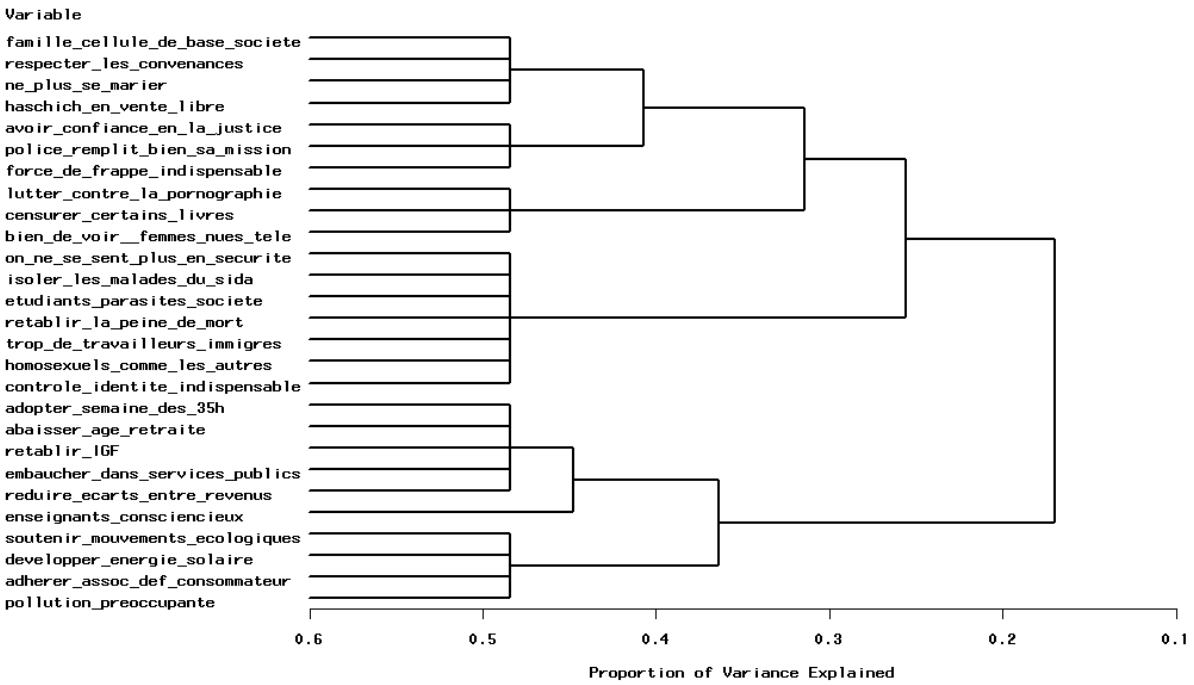
- la part de variance expliquée par la partition augmente ;
- les *composantes* des classes expliquent une part croissante des variances des classes ;
- les variances des deuxièmes composantes principales des classes diminuent ;
- les variables sont de plus en plus corrélées avec les *composantes* de leur propre classe, et de moins en moins corrélées avec les *composantes* des autres classes.

1.2.5. Classification avec structure hiérarchique

Les résultats obtenus avec l'option **HIERARCHY** diffèrent très légèrement des résultats précédents : la partition à 6 classes est identique à la partition décrite ci-dessus, excepté la variable **enseignants_consciencieux**, qui appartient à la classe 2 et non à la classe 6. La variance expliquée par la partition vaut 12.085, valeur légèrement inférieure à la variance obtenue plus haut.

La 2ème valeur propre de la classe 2 valant 1.025, cette classe est divisée, et la variable **enseignants_consciencieux** constitue la 7ème classe à elle toute seule.

On peut alors obtenir, grâce à la procédure **TREE**, un arbre de classification présentant la hiérarchie des classes :



2. Gammes d'équipements

Mise à jour annuellement par l'Insee à partir de sources administratives diverses, la base permanente des équipements (BPE) propose une information finement localisée. Le champ actuel recouvre les équipements dans les domaines des services, marchands ou non, des commerces, de la santé et de l'action sociale, de l'enseignement et du tourisme.

L'accès aux équipements et aux services constitue une problématique importante pour les acteurs locaux. Cet accès contribue en effet à l'attractivité d'un territoire.
On peut se demander par exemple :

- si une commune dispose d'équipements de telle ou telle catégorie,
- quel est le temps nécessaire pour accéder aux équipements s'ils ne sont pas sur sa commune.

Dans le premier cas, on peut décrire les communes avec des **variables binaires** de présence-absence des différents équipements.
Dans le second cas, on peut décrire les communes avec des **variables quantitatives** représentant le temps routier nécessaire pour accéder aux différents équipements les plus proches aux heures de pointe.

Cependant, dans les deux cas, les communes sont décrites par un très grand nombre de variables. On peut souhaiter recourir à la classification de variables pour réduire le nombre de descripteurs de nos communes. On peut aussi simplement souhaiter repérer des grands groupes d'équipements. C'est davantage dans cette dernière optique qu'on se placera : description des variables en vue d'une identification de gammes d'équipements en fonction de leur présence ou absence sur les communes.

2.1. Présence – absence des équipements sur les communes

La table utilisée construite à l'aide de la BPE de 2007 comporte :

- en ligne les communes (36 000 environ) traitées comme unités statistiques,
- en colonne les équipements (plus de 120 catégories d'équipements) traités comme des **variables qualitatives binaires** (présence de l'équipement : modalité = 1 ; absence de l'équipement : modalité = 0).

Une gamme d'équipements a déjà été élaborée par le PSAR (Pôle de Services de l'Action Régionale) Synthèses Locales de l'Insee mais avec une méthode différente de celle présentée ci-dessous [5].

La procédure **VARCLUS** est utilisée ici avec toutes les options par défaut mais on impose à l'algorithme de maintenir une structure hiérarchique pour les classes. C'est en effet ce que l'on obtient par le recours à l'option **OUTTREE** qui implique l'option **HIERARCHY**. Avec l'option **OUTTREE** on crée une table contenant l'information nécessaire à la construction de l'arbre – édité ensuite grâce la procédure **TREE**.

L'algorithme de classification descendante hiérarchique s'arrête de lui-même à 13 classes d'équipements. Il n'y a en effet plus, à ce niveau-là, de groupe de variables présentant une seconde dimension significative (pas de seconde valeur propre supérieure à 1). Dans une logique de réduction de dimensions, on retiendrait ces 13 classes, chaque classe étant bien représentée par une seule variable synthétique (la première composante). **Dans l'optique de description**, on souhaite un résumé un peu plus synthétique de la réalité avec un niveau de division moindre. **On accepte donc que les classes de variables ne soient pas forcément unidimensionnelles.**

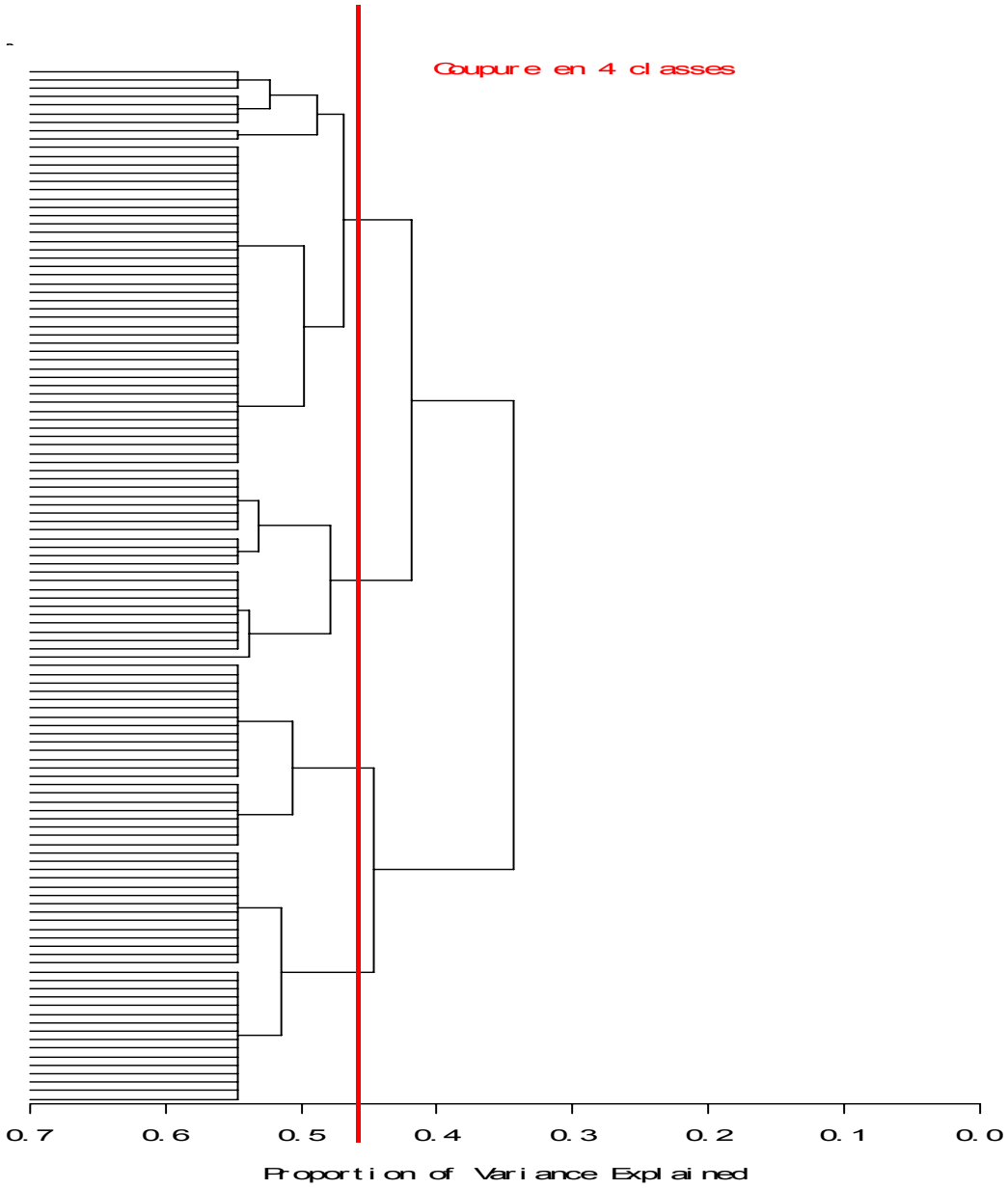
Si l'on examine l'évolution de la proportion de variance expliquée au cours des 13 itérations, on peut constater des ruptures dans le rythme de croissance de cette proportion (Tableau 12). Passer de 3 à 4 classes induit un gain de variance expliquée plus important (plus de 2%) que passer de 4 à 5 classes (1% seulement). On retiendra dans notre exemple une partition des variables et donc des équipements en 4 classes (voir coupure de l'arbre Figure 2).

Tableau 12 : résumé des étapes de l'algorithme descendant hiérarchique sur variables binaires

Nombre de classes d'équipements	Variance expliquée totale	Proportion de variance expliquée	Différence de proportion expliquée entre 2 niveaux de division successifs	Proportion expliquée minimale	Seconde valeur propre maximale	R^2 minimal pour une variable	$1-R^2$ maximal pour une variable
1	41.986577	0.3442		0.3442	11.143935	0.0070	
2	51.112163	0.4190	0,075	0.3856	3.997611	0.0090	0.9951
3	54.485196	0.4466	0,028	0.3582	3.398508	0.0090	1.0672
4	57.225466	0.4691	**0,023**	0.3582	1.512640	0.0096	0.9963
5	58.413912	0.4788	0,010	0.3582	1.297307	0.0096	0.9966
6	59.521686	0.4879	0,009	0.3415	1.273584	0.0096	0.9966
7	60.779223	0.4982	**0,010**	0.3415	1.187751	0.0096	0.9973
8	61.758491	0.5062	0,008	0.3415	1.181420	0.0096	0.9973
9	62.773974	0.5145	0,008	0.3415	1.113266	0.0096	0.9973
10	63.722046	0.5223	0,008	0.3415	1.062020	0.0107	0.9972
11	64.768045	0.5309	0,009	0.3415	1.049775	0.0107	0.9986
12	65.739647	0.5388	0,008	0.3415	1.034948	0.0107	0.9986
13	66.723542	0.5469	0,008	0.3741	0.996438	0.0107	0.9986

Le choix d'une structure hiérarchique a été retenu car en termes de variance expliquée totale les performances de l'algorithme non hiérarchique étaient très proches de celles de l'algorithme hiérarchique. De plus, la création de gammes d'équipements s'apparente à celle d'une nouvelle nomenclature. Or, dans le domaine des nomenclatures, on peut souhaiter disposer de plusieurs niveaux de détails. Les différents niveaux sont alors imbriqués les uns dans les autres de façon hiérarchique.

Figure 2 : Arbre de classification



Si on demande l'exécution de la procédure **VARCLUS** en utilisant l'option **MAXCLUSTERS** = 4, on obtient alors les ensembles de variables ayant les caractéristiques suivantes (la composition des classes est donnée en annexe). En fonction des objectifs de l'étude, on pourra souhaiter un plus ou moins grand niveau de division (voir le détail pour une classification en 7 gammes d'équipements en annexe).

Se reporter à la partie 1, pour une description détaillée du contenu des tableaux suivants.

Tableau 13 : Description générale des 4 classes

Cluster Summary for 4 Clusters					
Cluster	**Members**	**Cluster Variation**	**Variation Explained**	**Proportion Explained**	**Second Eigenvalue**
1	47	47	22.12561	0.4708	1.5126
2	30	30	16.02305	0.5341	1.1133
3	23	23	8.238803	0.3582	1.2973
4	22	22	10.838	0.4926	1.1814

Total variation explained = 57.22547 Proportion = 0.4691

Tableau 14 : Corrélation entre les 4 composantes des classes

Inter-Cluster Correlations				
Cluster	**1**	**2**	**3**	**4**
1	1.00000	0.74310	0.69958	0.45724
2	0.74310	1.00000	0.40638	0.78203
3	0.69958	0.40638	1.00000	0.23341
4	0.45724	0.78203	0.23341	1.00000

La corrélation élevée entre les composantes des classes tient à la taille des classes et à la structure des données (variables binaires globalement liées entre elles).

Les variables les plus caractéristiques (les plus liées à la composante de la classe) sont :

- **Pour la classe 1 :**
  - D202 : Spécialiste en cardiologie
  - D203 : Spécialiste en dermatologie et vénérologie
  - D206 : Gastro-entérologie
  - D208 : Spécialiste Ophtalmologie
  - D209 : Spécialiste en Oto-rhino-laryngologie
  - D212 : Radio diagnostic Imagerie médicale
  - D210 : Spécialiste pédiatrie
  - C_R3 : Lycée d'enseignement général et/ou technologique
  - C_R4 : Lycée d'enseignement professionnel
- **Pour la classe 2 :**
  - A304 : École de conduite
  - A502 : Vétérinaire
  - A506 : Blanchisserie teinturerie
  - B102 : Supermarché
  - B304 : Magasin de chaussures

- C201 : Collège
- D235 : Orthophoniste
- D237 : Pédicure-podologue
- D302 : Laboratoire d'analyses médicales

- **Pour la classe 3 :**
  - C403 : Formation commerce
  - C501 : UFR
  - C502 : Institut universitaire
  - C509 : Autre enseignement supérieur
  - C701 : Résidence universitaire
  - C702 : Restaurant universitaire

- **Pour la classe 4 :**
  - B203 : Boulangerie
  - A201 : Bureau de poste
  - A203 : Banque caisse d'épargne
  - A501 : Coiffure
  - B204 : Boucherie charcuterie
  - D201 : Médecin omnipraticien
  - D221 : Chirurgien dentiste
  - D233: Masseur kinésithérapeute
  - D301 : Pharmacie

2.2. Caractérisation des gammes d'équipements

La logique de constitution de ces gammes d'équipements n'est pas sectorielle (voir le détail de composition des classes en annexe 2) mais correspond plutôt à des densités de population différentes ou encore à des répartitions entre les rôles des communes. Ainsi, par exemple, un équipement tel qu'un institut universitaire ne se situera pas dans de petites villes contrairement à un collège par exemple.

La classe 3 relèverait plutôt d'une gamme d'équipements que l'on pourrait qualifier de "métropolitaine", la classe 1 relèvera d'une gamme d'équipements "supérieure", la classe 2 réunit des équipements de type "intermédiaire", et la classe 4 rassemble des équipements de "proximité".

Caractériser les composantes des classes par d'autres variables n'ayant par participé à l'analyse nécessite le calcul des valeurs prises par chaque commune sur ces composantes. L'option **OUTSTAT** de la procédure **VARCLUS** permet de créer une table contenant les coefficients à appliquer aux variables de base (ici la présence ou non des équipements) pour calculer les composantes des classes. C'est en appliquant la procédure **SCORE** à cette nouvelle table que l'on obtient les valeurs des composantes des classes pour chaque commune. On peut ensuite réaliser différentes analyses sur les liens entre les composantes et d'autres variables comme la population totale, ou encore le type de commune (chef lieu de canton, d'arrondissement, de département, de région ou pas chef lieu).

On constate ainsi que les composantes des classes 1 et 3 sont fortement corrélées positivement avec la population totale des communes de 2006.

Pearson Correlation Coefficients,				
	Clus1	**Clus2**	**Clus3**	**Clus4**
Population totale 2006	**0.72**	**0.51**	**0.80**	**0.35**

Les composantes de classes sont également liées avec le type de commune tel que défini ci-dessus.

Figure 3 : Distribution de la composante de la classe 1 selon le type de commune

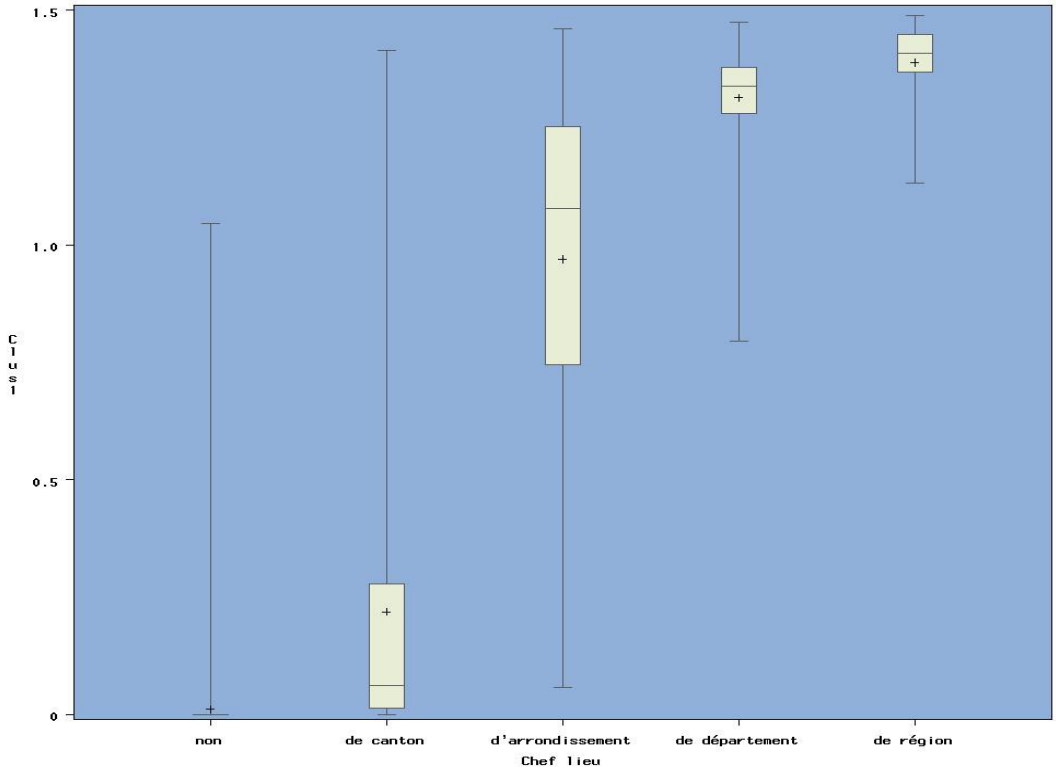


Figure 4 : Distribution de la composante de la classe 2 selon le type de commune

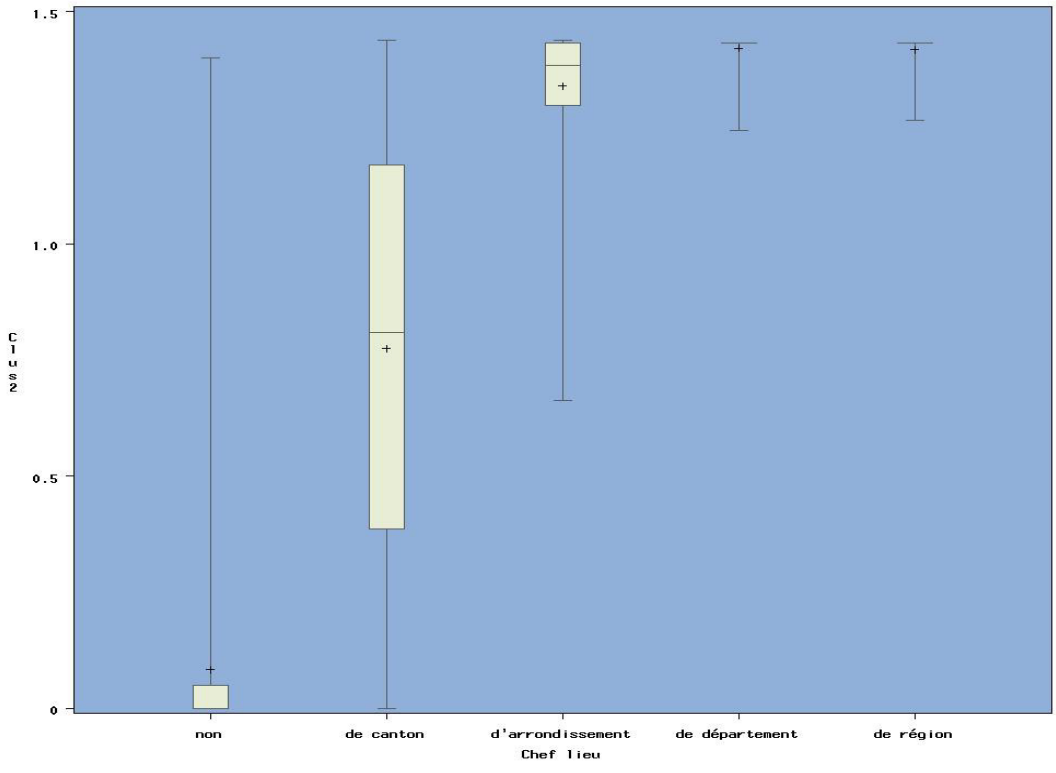


Figure 5 : Distribution de la composante de la classe 3 selon le type de commune

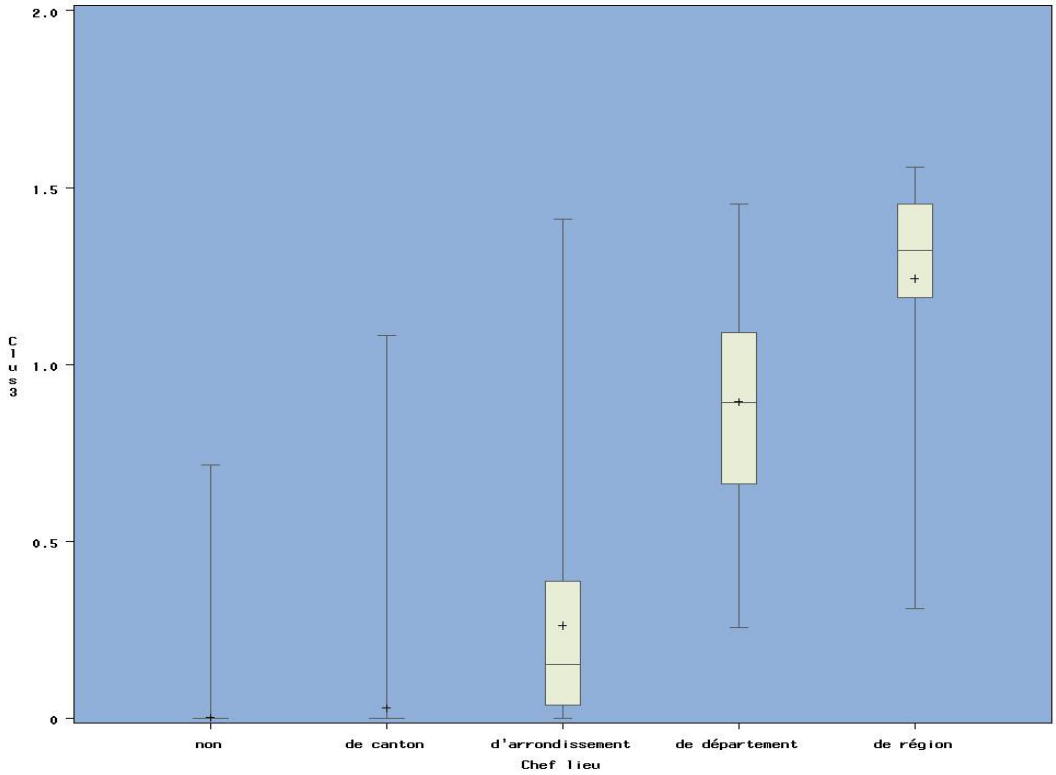
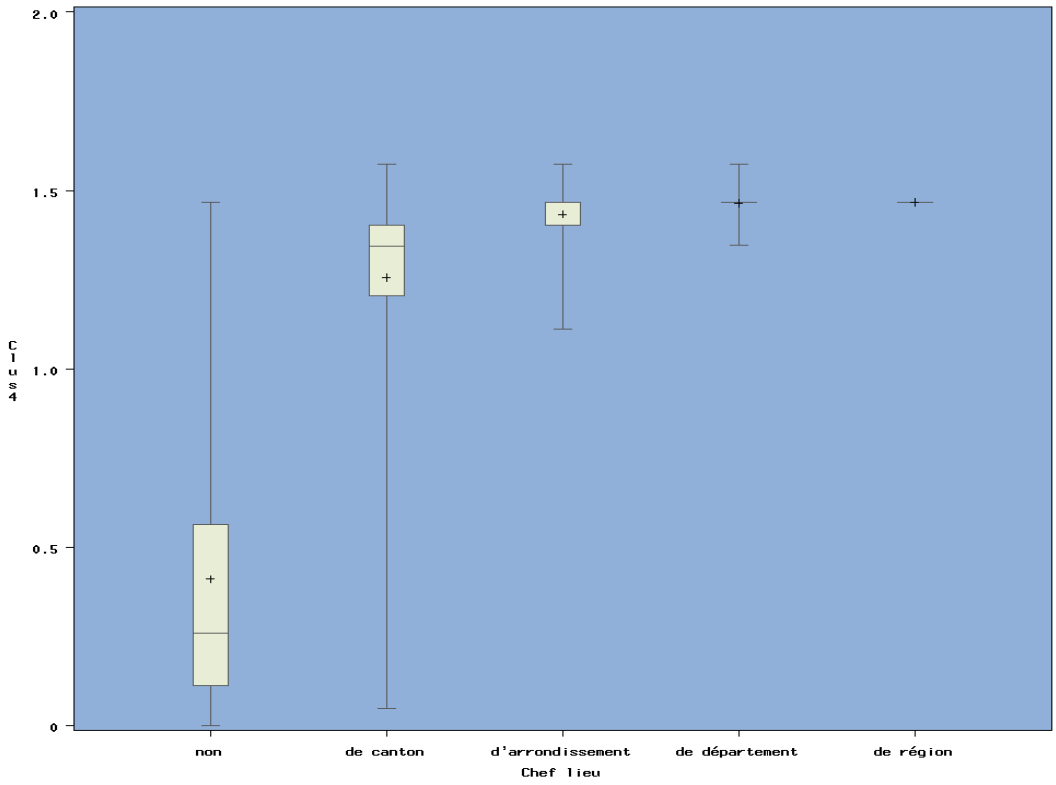


Figure 6 : Distribution de la composante de la classe 4 selon le type de commune



Les classes d'équipements n'étant pas unidimensionnelles, on a également utilisé le coefficient RV pour mesurer le lien les groupes d'équipements. Ce coefficient prend en compte l'ensemble des variables d'une classe et non plus seulement une variable composante représentative d'une classe. Un coefficient RV(*j,l*) est toujours compris entre 0 et 1. Il prend la valeur 1 si les nuages d'individus (ici les communes) induits par les deux groupes de variables j et k sont homothétiques (voir Escofier et Pagès [2]).

Ce coefficient n'est pas calculé par la procédure **VARCLUS**. On peut en obtenir le calcul par une macro SAS d'Analyse Factorielle Multiple [3].

Notons Lg(*j,l*) l'indice de liaison entre deux groupes de variables *j* et *l,* la somme pondérée par les poids des variables de l'AFM, des carrés des covariances (coefficients de corrélation pour les variables continues réduites) entre chaque colonne (variable continue ou indicatrice) du groupe *j* et chaque colonne du groupe *l*.

Cas particuliers:

- Les groupes *j* et *l* sont constitués chacun par une seule variable continue réduite : Lg est alors le carré du ***coefficient de corrélation*** entre les deux variables.
- Les groupes *j* et *l* sont constitués chacun par une seule variable nominale : Lg est alors le χ^2 *divisé par l'effectif total du tableau de contingence* croisant les deux variables.
- Les groupes *j* et *l* sont constitués l'un par une seule variable continue et l'autre par une seule variable nominale : Lg correspond alors au ***rapport de corrélation***.
  Lg(*j,l*) vaut 0 si chaque colonne du groupe *j* est non corrélée avec chaque colonne du groupe *l*. Il est d'autant plus élevé que les groupes de variables possèdent des directions d'inertie importante en commun.

On a alors : $RV(j,l) = Lg(j,l) / [Lg(j,j) \quad Lg(l,l)]^{1/2}$

Tableau 15 : Mesure du lien entre les 4 groupes d'équipements par le coefficient RV

RV Coefficients,				
Cluster	**1**	**2**	**3**	**4**
1	1.000	**0.562**	0.466	0.222
2	**0.562**	1.000	0.175	**0.596**
3	0.466	0.175	1.000	0.060
4	0.222	**0.596**	0.060	1.000

On a également utilisé le coefficient RV pour mesurer le lien entre chaque groupe d'équipements et la population totale des communes en 2006. Le lien le plus fort apparaît encore pour les classes 1 et 3.

Tableau 16 : Mesure du lien entre les 4 groupes d'équipements et la population totale des communes en 2006 par le coefficient RV

RV Coefficients,				
	Clus1	**Clus2**	**Clus3**	**Clus4**
Population totale 2006	**0.52**	**0.27**	**0.62**	**0.12**

Le niveau commune peut bien sûr faire l'objet d'une discussion. Il peut être intéressant de descendre à des niveaux plus fins dans le maillage du territoire. On pourrait également exclure de l'analyse certains équipements trop atypiques et les traiter en éléments supplémentaires : les établissements thermaux par exemple.

En conclusion, l'algorithme divisif de la procédure **VARCLUS** permet de s'affranchir des limites en capacités de calcul inhérentes aux algorithmes ascendants. De plus, il offre une aide à la décision quant au choix du nombre de classes.

Bibliographie

[1] Chavent M., Kuentz V., Saracco J., "Une approche divisive de classification hiérarchique de variables quantitatives", *14èmes rencontres de la Société Francophone de Classification (SFC)*, Paris, septembre 2007.

[2] Escofier B., Pagès J., *Analyses factorielles simples et multiples*, 4ème édition, 2008, Dunod.

[3] Gelein B., Sautory O., De nouvelles macros SAS d'analyse des données à l'Insee : comment réaliser une Analyse Factorielle Multiple, VIIIèmes Journées de Méthodologie Statistique, 2002

[4] Harman H., *Modern Factor Analysis*, 3rd edition, 1976, University of Chicago Press.

[5] Insee PSAR Synthèses Locales – *Guide de l'utilisateur du Kit Synthèses locales*, Insee, 2009.

[6] Nakache J.-P., Confais J., *Approche pragmatique de la classification*, 2005, TECHNIP.

[7] Qannari E.M., Vigneau E., Courcoux Ph., "Une nouvelle distance entre variables. Application en classification", *Revue de statistique appliquée*, 1988, vol 46, n°2, p. 21-32.

[8] SAS Institute, "The VARCLUS Procedure", *SAS/STAT User'Guide 9.1*, 2004, p. 4798-4827.

[9] Tufféry, *Data mining et statistique décisionnelle*, 2007, TECHNIP.

[10] Vigneau E., Qannari E.M., Sahmer K., Ladiray D., "Classification de variables autour de composantes latentes", *Revue de statistique appliquée*, 2006, vol 54, n°1, p. 27-45.

Annexe 1

Liste des variables de l'enquête Agoramétrie

Nom de variable	Libellé du thème
reduire_ecarts_entre_revenus	Il faut réduire au maximum les écarts entre les revenus
avoir_confiance_en_la_justice	On peut avoir confiance en la justice
force_de_frappe_indispensable	La force de frappe est indispensable
etudiants_parasites_societe	Les étudiants vivent en parasites de la société
trop_de_travailleurs_immigres	Il y a trop de travailleurs immigrés
lutter_contre_la_pornographie	On doit lutter énergiquement contre la pornographie
pollution_preoccupante	La pollution est terriblement préoccupante
famille_cellule_de_base_societe	La famille doit rester la cellule de base de la société
adherer_assoc_def_consommateur	Il faut adhérer aux associations de défense du consommateur
ne_plus_se_marier	On ne devrait plus se marier
censurer_certains_livres	Il est nécessaire de censurer certains livres
respecter_les_convenances	Il faut respecter les convenances
on_ne_se_sent_plus_en_securite	On ne se sent plus en sécurité
developper_energie_solaire	Il faut développer au maximum l'utilisation de l'énergie solaire
soutenir_mouvements_ecologiques	Il faut soutenir les mouvements écologiques
adopter_semaine_des_35h	Il faut adopter la semaine des 35 heures
embaucher_dans_services_publics	Il faut embaucher dans les services publics
abaisser_age_retraite	Il faut abaisser l'âge de la retraite
homosexuels_comme_les_autres	Les homosexuels sont des gens comme les autres
retablir_la_peine_de_mort	Il faut rétablir la peine de mort
haschich_en_vente_libre	Le haschisch devrait être en vente libre
controle_identite_indispensable	Les contrôles d'identité sont indispensables
bien_de_voir__femmes_nues_tele	C'est bien de voir des femmes nues à la télé
isoler_les_malades_du_sida	Il faut isoler les malades du SIDA
police_remplit_bien_sa_mission	La police remplit bien sa mission
enseignants_consciencieux	Les enseignants sont consciencieux
retablir_IGF	Il faut rétablir l'impôt sur les grandes fortunes

Annexe 2 : Composition des gammes d'équipements en fonction de leur présence – absence sur les communes – 4 classes d'équipements

4 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 1	A103	0.6620	0.3479	0.5183	ANPE
	A303	0.4010	0.2481	0.7967	Location auto-utilitaires légers
	A503	0.5059	0.3802	0.7972	Agce travail temporaire
	B205	0.3972	0.2292	0.7821	Produits surgelés
	B310	0.4827	0.4319	0.9106	Parfumerie
	C601	0.3643	0.2648	0.8647	Centre formation apprentis
	D101	0.5950	0.3526	0.6255	Etab santé court séjour
	D104	0.4819	0.2925	0.7323	Etab psychiatrique
	D106	0.5997	0.3062	0.5769	Urgence
	D107	0.5839	0.3652	0.6554	Maternité
	D109	0.4593	0.3564	0.8401	Structures psychiatriques en ambulatoire
	D111	0.4599	0.2772	0.7472	Dialyse
	D202	0.7345	0.3881	0.4339	Spéc Cardiologie
	D203	0.7117	0.3833	0.4675	Spéc Dermatologie Vénéréologie
	D204	0.5745	0.3008	0.6086	Spéc Gynécologie médicale
	D205	0.6481	0.3109	0.5107	Spéc Gynécologie obstetrique
	D206	0.6873	0.3514	0.4820	Spéc Gastro-entérologie
	D208	0.6919	0.4519	0.5622	Spéc Ophtalmologie
	D209	0.7054	0.3191	0.4326	Spéc Oto-rhino-laryngologie
	D211	0.5561	0.3687	0.7032	Spéc Pneumologie
	D212	0.6906	0.4606	0.5736	Spéc Radio diagnostic Imagerie médicale
	D213	0.5202	0.3657	0.7565	Spéc Stomatologie
	D231	0.4156	0.2782	0.8096	Sage-femme
	D236	0.6591	0.3551	0.5286	Orthoptiste

4 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	D238	0.4532	0.2774	0.7568	Audio prothésiste
	F301	0.4219	0.3188	0.8486	Cinéma
	D601	0.3117	0.2000	0.8604	Enfants handicapés hébergement
	B101	0.4043	0.2941	0.8438	Hypermarché
	B308	0.3556	0.2207	0.8269	Mag revêtements murs et sols
	D604	0.2635	0.1644	0.8815	Adultes handicapés services
	B206	0.3200	0.2968	0.9670	Poissonnerie
	D603	0.2568	0.2278	0.9625	Adultes handicapés hébergement
	D108	0.3643	0.2919	0.8978	Centre de santé
	D210	0.6652	0.3299	0.4997	Spéc Pédiatrie
	D605	0.2874	0.1983	0.8888	Travail protégé
	D102	0.4074	0.3074	0.8556	Etab santé moyen séjour
	D207	0.6134	0.3123	0.5622	Spéc Psychiatrie
	D701	0.2986	0.1952	0.8716	Aide sociale à l'enfance hébergement
	C303	0.0841	0.0675	0.9822	Lycée enseignement technologique / profe
	C402	0.5106	0.4006	0.8165	Formation santé
	D103	0.3383	0.2391	0.8696	Etab santé long séjour
	D240	0.3123	0.2699	0.9419	Psychomotricien
	D709	0.4112	0.3069	0.8494	Autre étab adultes et familles en diffic
	C603	0.0404	0.0268	0.9860	CFPPA
	D404	0.1041	0.0816	0.9756	Foyers restaurants
	C_R3	0.6445	0.3736	0.5675	Lycées d'enseignement général et/ou technologique
	C_R4	0.6690	0.3653	0.5216	Lycées d'enseignement professionnel
Cluster 2	A102	0.5324	0.3101	0.6778	Trésorerie
	A205	0.4664	0.2834	0.7446	Pompes funèbres

4 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	A302	0.5687	0.3202	0.6345	Contrôle technique automobile
	A304	0.6495	0.4633	0.6531	Ecoles de conduite
	A406	0.2632	0.2314	0.9586	Entr gale bâtiment
	A502	0.6174	0.3738	0.6110	Vétérinaire
	A506	0.6629	0.3693	0.5344	Blanchisserie teinturerie
	A507	0.5375	0.4332	0.8159	Soins de beauté
	B102	0.6509	0.4273	0.6096	Supermarché
	B301	0.6047	0.4914	0.7773	Librairie papeterie
	B302	0.6189	0.4066	0.6422	Mag vêtements
	B303	0.4657	0.3421	0.8122	Mag équipements du foyer
	B304	0.6120	0.4108	0.6585	Mag chaussures
	B305	0.5456	0.3162	0.6644	Mag électroménager
	B306	0.4505	0.3058	0.7916	Mag meubles
	B307	0.4568	0.2729	0.7471	Mag articles de sports loisirs
	B311	0.5922	0.4640	0.7608	Horlogerie Bijouterie
	B312	0.5988	0.5194	0.8348	Fleuriste
	C201	0.6864	0.4171	0.5380	Collège
	D234	0.6033	0.4631	0.7389	Opticien-lunetier
	D235	0.6273	0.4167	0.6390	Orthophoniste
	D237	0.6477	0.4060	0.5931	Pedicure-podologue
	D302	0.6445	0.5554	0.7996	Laboratoire d'analyses médicales
	D303	0.5132	0.3512	0.7502	Ambulances
	D401	0.5045	0.4407	0.8859	Pers agees hébergement
	D501	0.4792	0.3247	0.7713	Garde enfants d'âge préscolaire
	D403	0.2667	0.2226	0.9433	Pers âgées service d aide
	D305	0.0096	0.0059	0.9963	Etab thermal

4 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	A_R1	0.5521	0.3729	0.7141	Gendarmerie et Police
	B_R1	0.5945	0.3632	0.6369	Grande surface de bricolage et Droguerie quincallerie bricol
Cluster 3	C503	0.4186	0.1362	0.6731	Ecoles d'ingénieurs
	C602	0.4728	0.3245	0.7805	GRETA
	D703	0.4331	0.3753	0.9075	CHRS Centre hébergement et réadaptation
	D110	0.3910	0.2348	0.7959	Centre médecine préventive
	D306	0.0379	0.0181	0.9798	Etab lutte contre l'alcoolisme
	C403	0.5377	0.1896	0.5704	Formation Commerce
	C501	0.5917	0.2198	0.5233	UFR
	C502	0.6323	0.3196	0.5405	Institut universitaire
	D304	0.4892	0.2871	0.7165	Transfusion sanguine
	D702	0.4426	0.3493	0.8566	Aide sociale à l'enfance action éducativ
	C609	0.3154	0.2501	0.9129	Autre formation continue
	D705	0.2532	0.1694	0.8991	Centre accueil demandeur d'asile
	C509	0.5826	0.2256	0.5390	Autre enseignement supérieur
	D704	0.1167	0.0410	0.9210	Centre provisoire d'hébergement
	C701	0.5513	0.2022	0.5624	Résidence universitaire
	C401	0.3463	0.1484	0.7676	STS CPGE
	C409	0.4657	0.2328	0.6964	Autre formation post bac non universitai
	C604	0.0969	0.0383	0.9391	Formation métiers sport
	C702	0.5944	0.2231	0.5220	Restaurant universitaire
	D239	0.0943	0.0511	0.9544	Ergothérapeute
	D105	0.1736	0.0387	0.8597	Centre lutte contre le cancer
	C504	0.1843	0.0615	0.8692	Enseignement général supérieur privé
	C203	0.0171	0.0073	0.9901	SET Section Enseignement Technologique

4 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 4	A402	0.3317	0.1715	0.8067	Platrier peintre
	A405	0.3396	0.1819	0.8073	Electricien
	A504	0.3173	0.1389	0.7929	Restaurant
	B203	0.5751	0.2515	0.5677	Boulangerie
	A201	0.6627	0.3661	0.5321	Bureau de poste
	A203	0.5895	0.5427	0.8977	Banque Caisse d'épargne
	A301	0.4365	0.2172	0.7198	Rép auto-mat agricole
	A401	0.2319	0.1096	0.8627	Maçon
	A403	0.2625	0.1267	0.8445	Menuis charpent serrurerie
	A404	0.3130	0.1479	0.8062	Plombier couvreur chauffagiste
	A501	0.6007	0.2791	0.5538	Coiffure
	A505	0.4260	0.3554	0.8905	Agence immobilière
	B204	0.5713	0.3889	0.7015	Boucherie charcuterie
	D201	0.7639	0.3982	0.3923	Médecin omnipraticien
	D221	0.7126	0.5814	0.6866	Chirurgien dentiste
	D232	0.6292	0.3710	0.5894	Infirmier
	D233	0.7180	0.5011	0.5654	Masseur kinésithérapeute
	D301	0.7936	0.4984	0.4114	Pharmacie
	E101	0.3754	0.2706	0.8564	Taxi
	B_R2	0.5045	0.3745	0.7921	Supérette - Epicerie
	C_R1	0.4095	0.3670	0.9328	Ecoles maternelles (y/c RPI)
	C_R2	0.2733	0.0903	0.7988	Ecoles élémentaires (y/c RPI)

Annexe 3 : Composition des gammes d'équipements en fonction de leur présence – absence sur les communes – 7 classes d'équipements

Cluster Summary for 7 Clusters					
Cluster	**Members**	**Cluster Variation**	**Variation Explained**	**Proportion Explained**	**Second Eigenvalue**
1	38	38	19.67846	0.5179	1.1878
2	30	30	16.02305	0.5341	1.1133
3	11	11	3.756703	0.3415	1.0349
4	22	22	10.838	0.4926	1.1814
5	7	7	3.492369	0.4989	1.0620
6	12	12	5.589874	0.4658	1.0498
7	2	2	1.400758	0.7004	0.5992

Total variation explained = 60.77922 Proportion = 0.4982

Inter-Cluster Correlations							
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1.00000	0.72952	0.74860	0.44254	0.78947	0.60591	0.27220
2	0.72952	1.00000	0.44754	0.78203	0.69102	0.33956	0.25300
3	0.74860	0.44754	1.00000	0.25823	0.58676	0.75038	0.22967
4	0.44254	0.78203	0.25823	1.00000	0.46713	0.19439	0.17138
5	0.78947	0.69102	0.58676	0.46713	1.00000	0.45186	0.25079
6	0.60591	0.33956	0.75038	0.19439	0.45186	1.00000	0.14425
7	0.27220	0.25300	0.22967	0.17138	0.25079	0.14425	1.00000

7 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 1	A103	0.6666	0.4193	0.5741	ANPE
	A303	0.4085	0.2481	0.7867	Location auto-utilitaires légers
	A503	0.5019	0.3802	0.8037	Agce travail temporaire
	B205	0.4060	0.2399	0.7815	Produits surgelés
	B310	0.4828	0.4319	0.9103	Parfumerie
	C601	0.3648	0.2615	0.8600	Centre formation apprentis
	D104	0.4716	0.3723	0.8419	Etab psychiatrique
	D106	0.5888	0.4513	0.7493	Urgence
	D107	0.5805	0.4283	0.7339	Maternité
	D109	0.4539	0.3564	0.8485	Structures psychiatriques en ambulatoire
	D111	0.4598	0.3183	0.7925	Dialyse
	D202	0.7437	0.4430	0.4601	Spéc Cardiologie
	D203	0.7245	0.4139	0.4700	Spéc Dermatologie Vénéréologie
	D204	0.5886	0.3252	0.6096	Spéc Gynécologie médicale
	D205	0.6596	0.3784	0.5476	Spéc Gynécologie obstetrique
	D206	0.6981	0.4025	0.5053	Spéc Gastro-entérologie
	D208	0.6977	0.4519	0.5516	Spéc Ophtalmologie
	D209	0.7170	0.4156	0.4841	Spéc Oto-rhino-laryngologie
	D211	0.5666	0.4131	0.7385	Spéc Pneumologie
	D212	0.6945	0.4606	0.5665	Spéc Radio diagnostic Imagerie médicale
	D213	0.5330	0.3852	0.7596	Spéc Stomatologie
	D231	0.4170	0.2782	0.8078	Sage-femme
	D236	0.6687	0.3915	0.5444	Orthoptiste
	D238	0.4556	0.3082	0.7870	Audio prothésiste
	F301	0.4176	0.3188	0.8549	Cinéma

7 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	B101	0.4033	0.2941	0.8453	Hypermarché
	B308	0.3593	0.2367	0.8394	Mag revêtements murs et sols
	B206	0.3199	0.2968	0.9672	Poissonnerie
	D108	0.3623	0.2919	0.9007	Centre de santé
	D210	0.6817	0.3682	0.5038	Spéc Pédiatrie
	D207	0.6266	0.3447	0.5698	Spéc Psychiatrie
	D701	0.2972	0.2158	0.8962	Aide sociale à l'enfance hébergement
	C402	0.5075	0.4380	0.8763	Formation santé
	D240	0.3218	0.2716	0.9311	Psychomotricien
	D709	0.4171	0.3503	0.8972	Autre étab adultes et familles en diffic
	D404	0.1040	0.0809	0.9749	Foyers restaurants
	C_R3	0.6436	0.4250	0.6199	Lycées d'enseignement général et/ou technologique
	C_R4	0.6667	0.4518	0.6080	Lycées d'enseignement professionnel
Cluster 2	A102	0.5324	0.3107	0.6785	Trésorerie
	A205	0.4664	0.2834	0.7446	Pompes funèbres
	A302	0.5687	0.3088	0.6241	Contrôle technique automobile
	A304	0.6495	0.4633	0.6531	Ecoles de conduite
	A406	0.2632	0.2314	0.9586	Entr gale bâtiment
	A502	0.6174	0.3738	0.6110	Vétérinaire
	A506	0.6629	0.3590	0.5259	Blanchisserie teinturerie
	A507	0.5375	0.4332	0.8159	Soins de beauté
	B102	0.6509	0.4273	0.6096	Supermarché
	B301	0.6047	0.4914	0.7773	Librairie papeterie
	B302	0.6189	0.4066	0.6422	Mag vêtements
	B303	0.4657	0.3348	0.8033	Mag équipements du foyer

7 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	B304	0.6120	0.3984	0.6450	Mag chaussures
	B305	0.5456	0.3162	0.6644	Mag électroménager
	B306	0.4505	0.2999	0.7850	Mag meubles
	B307	0.4568	0.2672	0.7412	Mag articles de sports loisirs
	B311	0.5922	0.4528	0.7452	Horlogerie Bijouterie
	B312	0.5988	0.5194	0.8348	Fleuriste
	C201	0.6864	0.4171	0.5380	Collège
	D234	0.6033	0.4550	0.7280	Opticien-lunetier
	D235	0.6273	0.4167	0.6390	Orthophoniste
	D237	0.6477	0.4060	0.5931	Pedicure-podologue
	D302	0.6445	0.5464	0.7838	Laboratoire d'analyses médicales
	D303	0.5132	0.3512	0.7502	Ambulances
	D401	0.5045	0.4407	0.8859	Pers agees hébergement
	D501	0.4792	0.3180	0.7637	Garde enfants d'âge préscolaire
	D403	0.2667	0.2226	0.9433	Pers âgées service d aide
	D305	0.0096	0.0069	0.9973	Etab thermal
	A_R1	0.5521	0.3562	0.6957	Gendarmerie et Police
	B_R1	0.5945	0.3632	0.6369	Grande surface de bricolage et Droguerie quincallerie bricol
Cluster 3	C602	0.5727	0.3292	0.6369	GRETA
	D703	0.5472	0.3808	0.7313	CHRS Centre hébergement et réadaptation
	D110	0.4778	0.2707	0.7161	Centre médecine préventive
	D306	0.0574	0.0229	0.9646	Etab lutte contre l'alcoolisme
	D304	0.5654	0.3520	0.6707	Transfusion sanguine
	D702	0.5494	0.3553	0.6989	Aide sociale à l'enfance action éducativ
	C609	0.3701	0.2506	0.8405	Autre formation continue

7 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	D705	0.3426	0.1702	0.7922	Centre accueil demandeur d'asile
	D704	0.1398	0.0825	0.9376	Centre provisoire d'hébergement
	D239	0.1124	0.0666	0.9509	Ergothérapeute
	C203	0.0218	0.0115	0.9895	SET Section Enseignement Technologique
Cluster 4	A402	0.3317	0.1715	0.8067	Platrier peintre
	A405	0.3396	0.1819	0.8073	Electricien
	A504	0.3173	0.1389	0.7929	Restaurant
	B203	0.5751	0.2515	0.5677	Boulangerie
	A201	0.6627	0.3661	0.5321	Bureau de poste
	A203	0.5895	0.5427	0.8977	Banque Caisse d'épargne
	A301	0.4365	0.2172	0.7198	Rép auto-mat agricole
	A401	0.2319	0.1096	0.8627	Maçon
	A403	0.2625	0.1267	0.8445	Menuis charpent serrurerie
	A404	0.3130	0.1479	0.8062	Plombier couvreur chauffagiste
	A501	0.6007	0.2791	0.5538	Coiffure
	A505	0.4260	0.3554	0.8905	Agence immobilière
	B204	0.5713	0.3889	0.7015	Boucherie charcuterie
	D201	0.7639	0.3982	0.3923	Médecin omnipraticien
	D221	0.7126	0.5814	0.6866	Chirurgien dentiste
	D232	0.6292	0.3710	0.5894	Infirmier
	D233	0.7180	0.5011	0.5654	Masseur kinésithérapeute
	D301	0.7936	0.4984	0.4114	Pharmacie
	E101	0.3754	0.2706	0.8564	Taxi
	B_R2	0.5045	0.3745	0.7921	Supérette - Epicerie
	C_R1	0.4095	0.3670	0.9328	Ecoles maternelles (y/c RPI)

7 Clusters		R-squared with		1-R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
	C_R2	0.2733	0.0903	0.7988	Ecoles élémentaires (y/c RPI)
Cluster 5	D101	0.6620	0.5500	0.7511	Etab santé court séjour
	D601	0.3943	0.2830	0.8448	Enfants handicapés hébergement
	D604	0.3938	0.2336	0.7910	Adultes handicapés services
	D603	0.4527	0.2278	0.7088	Adultes handicapés hébergement
	D605	0.4554	0.2523	0.7284	Travail protégé
	D102	0.6038	0.3590	0.6182	Etab santé moyen séjour
	D103	0.5305	0.2950	0.6660	Etab santé long séjour
Cluster 6	C503	0.4967	0.2113	0.6381	Ecoles d'ingénieurs
	C403	0.5632	0.3477	0.6697	Formation Commerce
	C501	0.6294	0.3689	0.5872	UFR
	C502	0.6124	0.4669	0.7270	Institut universitaire
	C509	0.6056	0.3808	0.6369	Autre enseignement supérieur
	C701	0.6047	0.3251	0.5857	Résidence universitaire
	C401	0.3689	0.2177	0.8067	STS CPGE
	C409	0.4938	0.2941	0.7171	Autre formation post bac non universitai
	C604	0.1021	0.0622	0.9575	Formation métiers sport
	C702	0.6640	0.3376	0.5072	Restaurant universitaire
	D105	0.2131	0.0828	0.8580	Centre lutte contre le cancer
	C504	0.2361	0.0800	0.8304	Enseignement général supérieur privé
Cluster 7	C303	0.7004	0.0725	0.3230	Lycée enseignement technologique / profe
	C603	0.7004	0.0347	0.3104	CFPPA