Objectif

Mettre en œuvre la classification de variables dans TANAGRA.

Classification de variables. Dans la majorité des ouvrages, la classification de variables est décrite très sommairement. Les auteurs se contentent le plus souvent de la présenter comme un cas particulier de la typologie où le coefficient de corrélation *r* est utilisé pour mesurer la proximité entre les variables, (1- *r*) étant alors un indice de dissimilarité naturel.

Pourtant, la classification de variables peut être très utile dans la recherche des structures sous-jacentes dans les données. Elle permet de repérer les groupes de variables redondantes, emmenant le même type d'information ; de distinguer les groupes de variables orthogonales, rapportant des informations complémentaires. Nous disposons ainsi de précieuses indications sur l'architecture des données.

Cette méthode peut également être mise en œuvre dans une stratégie de réduction/sélection des variables. Pour chaque groupe de variables, une variable « moyenne » unique pourra être produite et utilisée dans les analyses ultérieures, en tant que variable prédictive présentée aux méthodes supervisées par exemple, réduisant considérablement la dimensionnalité.

Avec la version 1.4.16, nous introduisons dans TANAGRA plusieurs techniques de classification de variables inspirées par la lecture de l'ouvrage de Nakache et Confais (2005)[1]. Nous nous sommes plus particulièrement penchés sur les techniques de classification autour de composantes latentes basées sur les travaux de Vigneau et Qannari (2003)[2], à l'origine notamment de la fameuse procédure VARCLUS (*Variable Clustering*) implémentée dans le logiciel SAS.

Classification de variables autour de composantes latentes. Cette approche repose sur l'idée suivante pour représenter un groupe de variables, la variable « moyenne » en quelque sorte, nous utilisons une variable latente, c.-à-d. le premier facteur de l'analyse en composantes principales. Intellectuellement, c'est très satisfaisant. J'ai personnellement toujours eu des réticences à implémenter la typologie de variables parce que j'avais du mal à intégrer ce que représentait une variable synthétique qui serait une moyenne non pondérée des variables composant le groupe. Au moins, nous savons très bien interpréter un axe factoriel. La lecture des résultats est autrement plus intuitive. Dans TANAGRA, seules les techniques fondées sur le carré de la corrélation (r^2) sont disponibles, nous retrouvons ainsi l'interprétation classique des composantes principales en termes de concordance ou d'opposition de variables.

Partant de ce principe, un groupe de variables est représenté par le premier axe de l'ACP, nous pouvons décliner les différentes stratégies de classification de variables. La première, VARKMEANS, implémente les nuées dynamiques. L'utilisateur fixe un nombre de classes, chaque variable est affectée itérativement au groupe qui leur est le plus proche, au sens du carré de la corrélation. Après le passage de toutes les variables, les composantes principales sont remises à jour. Le processus se poursuit jusqu'à ce qu'il n'y ait plus d'amélioration du critère d'évaluation de la partition, c.-à-d. la somme des valeurs propres du premier axe factoriel associé à chaque groupe.

[1] J.P. Nakache et J. Confais, « Approche Pragmatique de la Classification », TECHNIP, 2005, chapitre 9, pages 219 à 239. Il nous servira de référence lors de la description des résultats de TANAGRA. Une autre référence est disponible en ligne, il s'agit de la documentation de la procédure VARCLUS du logiciel SAS version 8.0, chapitre 68 : http://www2.stat.unibo.it/ManualiSas/stat/chap68.pdf

[2] E. Vigneau et E. Qannari, « Clustering of variables around latent components », Simulation and Computation, 32(4), 1131-1150.

La seconde approche, VARHCA, implémente une approche hiérarchique. La démarche de construction est celle de la classification ascendante hiérarchique. Seul le critère d'agrégation change ici. A chaque étape, nous fusionnons les deux groupes engendrant la plus petite perte de variabilité expliquée, quantifiée par l'écart entre la somme de leurs premières valeurs propres et la première valeur propre du groupe constitué. La difficulté est, tout comme pour la typologie des individus, la détection du nombre de classes optimales. Il y a bien entendu la solution visuelle. L'utilisateur, à la lecture du dendrogramme, spécifie le bon nombre de classes. TANAGRA intègre également une seconde approche, originale, qui consiste à détecter le coude de la courbe des variations expliquées en fonction du nombre de classes. Nous détaillerons cette approche plus loin.

Enfin, et c'est un aspect très intéressant, nous disposons également d'une approche descendante. La méthode VARCLUS implémentée dans SAS en est une illustration. La méthode implémentée dans TANAGRA en est directement inspirée. Pour rappel, il s'agit à chaque étape, pour subdiviser un groupe de variables, de calculer les deux premiers facteurs de l'ACP, de les faire pivoter de manière à ce que l'alignement des variables sur les axes soit plus tranché. Si la valeur propre associée au second axe est supérieure à 1 (*paramètre modifiable*), les variables sont subdivisées en 2 sous-ensembles selon l'axe qui leur est le plus proche (au sens du r^2). Par rapport à la technique originelle, j'ai introduit deux variantes : (1) TANAGRA utilise une rotation VARIMAX qui semble plus efficace ; (2) je n'ai pas intégré les procédures de ré-affectation à chaque étape visant à maintenir la structure hiérarchique. Le dessin de l'arbre de partitionnement correspond donc à une description des séquences de découpage plutôt qu'à un dendrogramme reflétant les variations expliquées. Ces modifications améliorent significativement le temps de calcul sans dégrader les performances. La méthode propose un partitionnement au moins aussi efficace que les autres. Pour détecter le bon nombre de classes, nous utilisons une technique similaire à celle de la VARHCA.

Dans ce tutoriel, nous montrons la mise en œuvre des outils de classification de variables dans TANAGRA. Nous verrons comment lire les résultats, les comparer, modifier les paramétrages. Nous verrons également qu'il est possible de produire les variables synthétiques représentatives de chaque axe pour les utiliser dans les analyses subséquentes.

Fichier de données

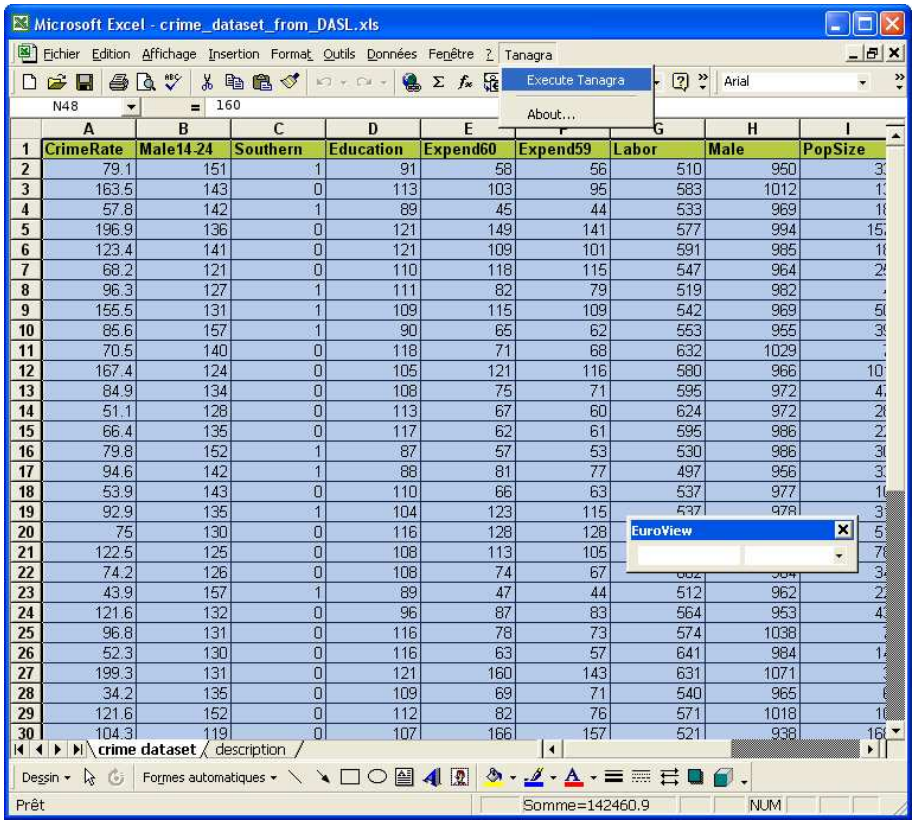
Nous utilisons le fichier CRIME_DATASET_FROM_DASL.XLS que l'on peut récupérer sur le portail DASL (The Data Story and Library -- http://lib.stat.cmu.edu/DASL/). Il recense des données relatives à la criminalité dans différents états des USA au début des années 1960. Le fichier contient 47 observations et 14 variables. Notre idée est de synthétiser aussi simplement que possible les principales informations que nous pouvons tirer de ce fichier.

Classification de variables dans TANAGRA

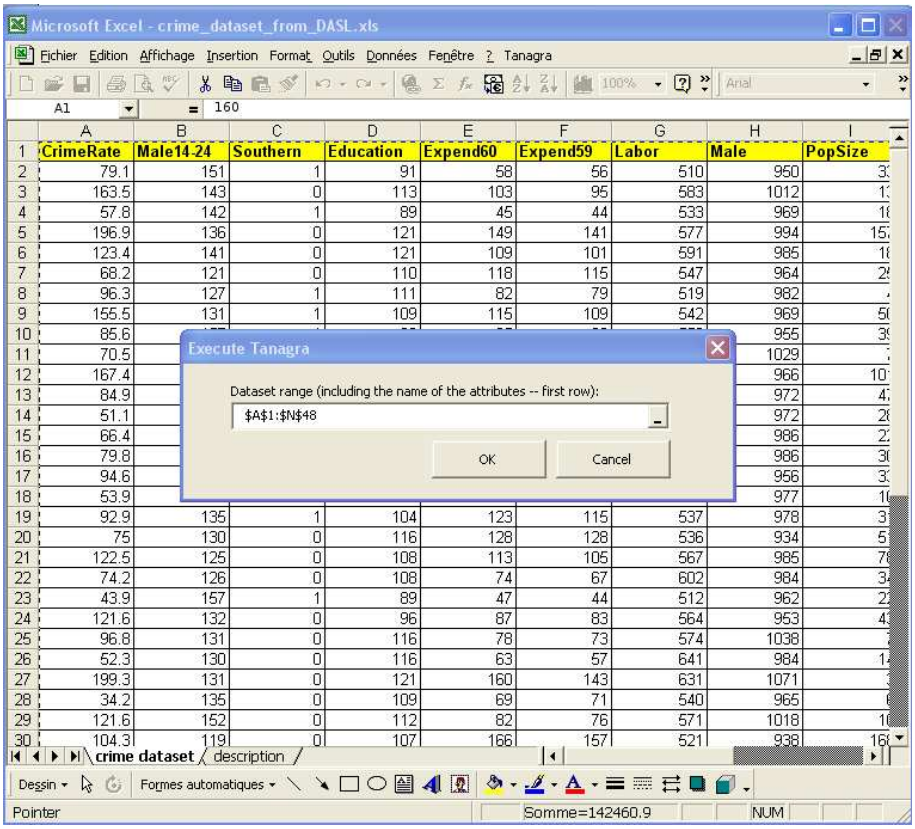
Création d'un diagramme

Le plus simple est d'ouvrir le fichier dans le tableur EXCEL, de sélectionner la plage de cellules puis d'activer le menu TANAGRA/EXECUTE TANAGRA installé avec la macro complémentaire TANAGRA.XLA[3].

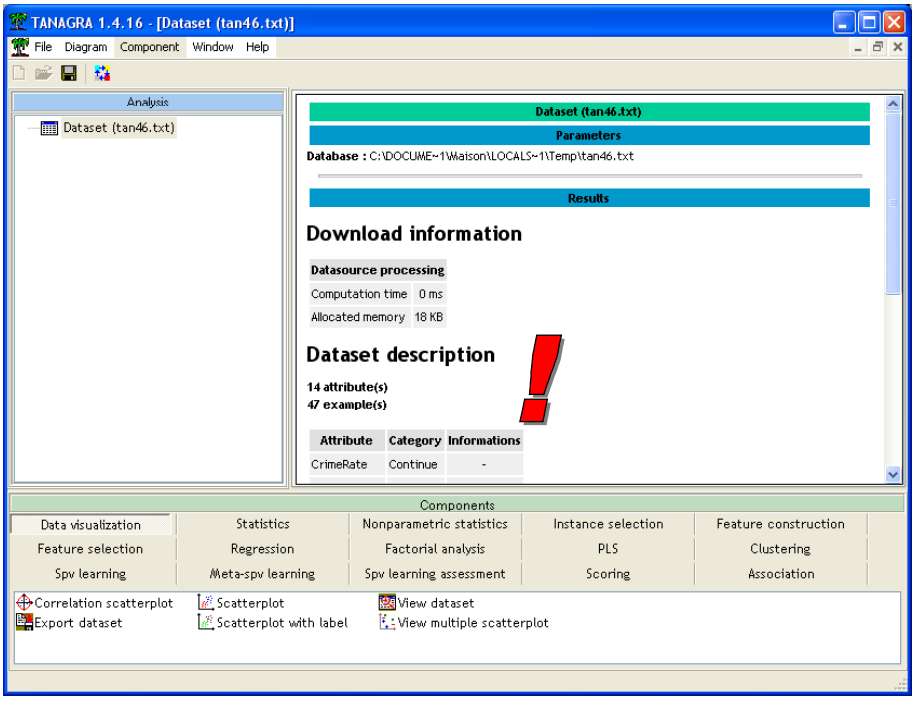
[3] Voir les didacticiels relatifs à l'installation de la macro complémentaire sur le site web si elle n'a pas encore été installée et activée. Elle est disponible depuis la version 1.4.11 de TANAGRA.



Une boîte de dialogue apparaît, demandant confirmation des coordonnées de la plage de cellules. Nous confirmons si la sélection est correcte.



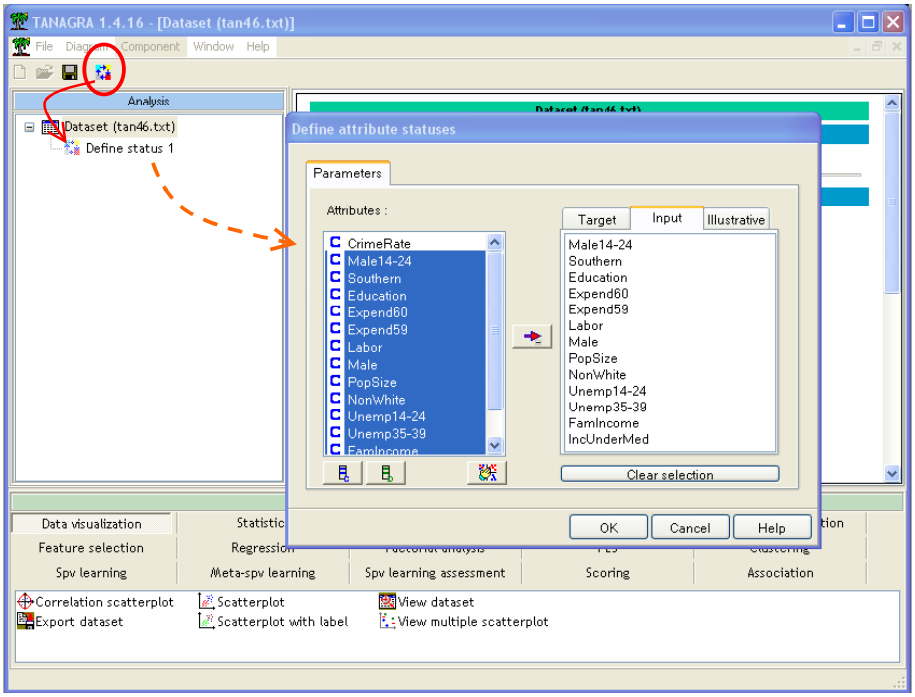
TANAGRA est alors automatique lancé, les données chargées. Nous vérifions que nous disposons bien de 14 variables et 47 observations.



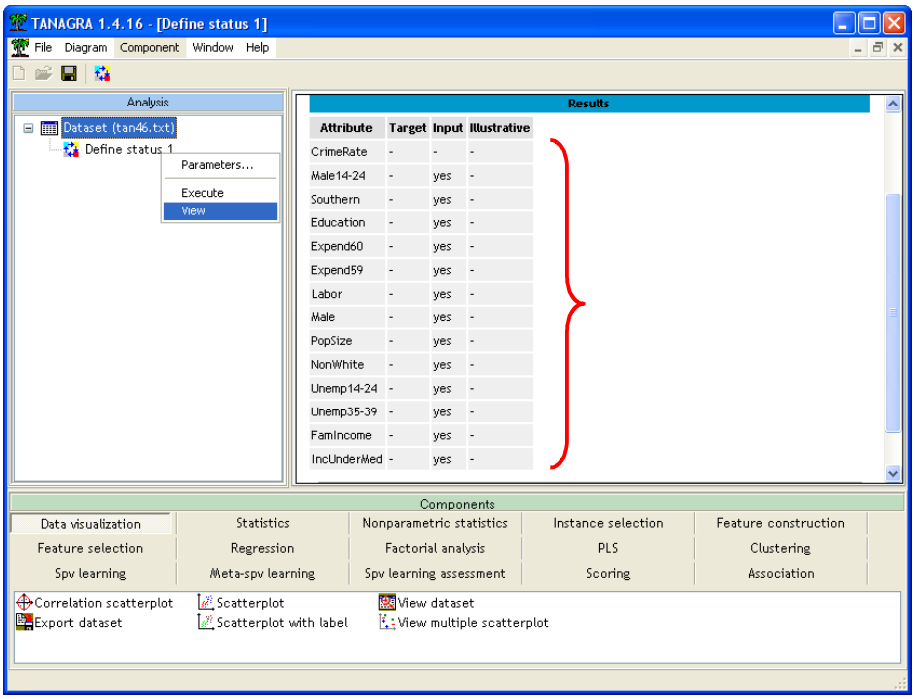
VARHCA

Définir le traitement

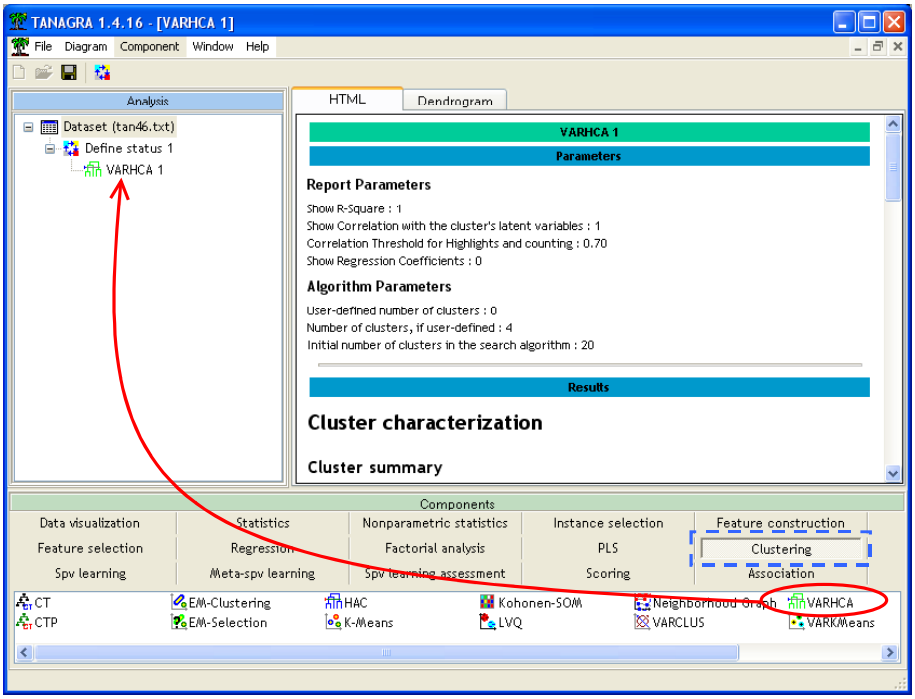
Bien entendu, dans un premier temps nous devons définir les variables de l'analyse. Nous insérons le composant DEFINE STATUS dans le diagramme en utilisant le raccourci de la barre d'outil.



Nous plaçons en INPUT toutes les variables, excepté la variable CRIME RATE qui a un rôle à part, nous l'utiliserons plus tard.



Il nous reste alors à placer le composant **VARHCA** (onglet CLUSTERING) dans le diagramme, par glisser-déposer. Pour visualiser les résultats, nous activons le menu contextuel VIEW.



Lecture des résultats

Pour comprendre au mieux les résultats, nous nous référons à l'ouvrage de Nakache et Confais (2005). Nous détaillons les différents tableaux.

Résumé de la partition. CLUSTER SUMMARY recense les classes qui ont été construites. Nous observons le nombre de variables dans chaque groupe ; la valeur propre du premier facteur de l'ACP dans chaque groupe ; la proportion de variabilité expliquée à l'intérieur du groupe – la valeur 1 indique que toute la variabilité est résumée par le premier facteur de l'ACP dans ce groupe. Au

final, la somme des valeurs propres indique la qualité globale du partitionnement, la proportion de variabilité expliquée.

Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	2	1.7459	0.8730
2	3	2.3843	0.7948
3	6	4.4051	0.7342
4	2	1.5136	0.7568
Total		10.0489	0.7730

Liste des variables par cluster. CLUSTER MEMBERS AND R-SQUARE VALUES recense les variables dans chaque groupe. Plusieurs indicateurs permettent d'apprécier la qualité de l'affectation : OWN CLUSTER indique le R^2 de la variable avec son groupe c.-à-d. le carré de la corrélation de la variable avec le représentant de la classe, le premier axe de l'ACP sur les variables composant le groupe ; NEXT CLOSEST indique le R^2 de la variable avec le groupe le plus proche, si cette valeur est plus grande que la première, il y a matière à s'inquiéter.

L'indicateur (1-R^2 ratio) indique justement le rapport entre (1-R^2 own cluster) et (1-R^2 next closest). Plus petite est sa valeur, meilleure est l'affectation de la variable au groupe. Si elle est supérieure à 1, cela voudrait dire que la variable est plus corrélée avec un autre cluster qu'avec son propre groupe d'appartenance.

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R^2 ratio
1	Unemp14-24	0.8730	0.0050	0.1277
	Unemp35-39	0.8730	0.0638	0.1357
2	Expend60	0.9334	0.3436	0.1015
	Expend59	0.9260	0.3569	0.1150
	PopSize	0.5249	0.0159	0.4827
3	Southern	0.7441	0.1011	0.2847
	NonWhite	0.6944	0.0213	0.3123
	Male14-24	0.5988	0.2473	0.5331
	Education	0.7396	0.1537	0.3076
	FamIncome	0.7798	0.5376	0.4762
	IncUnderMed	0.8485	0.3085	0.2191
4	Labor	0.7568	0.1738	0.2944
	Male	0.7568	0.0811	0.2647

Dans notre exemple, VARHCA a proposé une typologie en 4 classes, 77,3% de la variabilité totale est restituée par ce partitionnement. Les variables semblent bien assorties à leurs classes respectives. Dans le pire des cas, 1-R^2 ratio est égal à 0.533 pour la variable MALE14-24 dans le $3^{ème}$ cluster.

Interprétation des classes. Le tableau des corrélations des variables avec les clusters (CLUSTER CORRELATIONS – STRUCTURE) permet d'interpréter les groupes de variables. Il faut le lire en parallèle avec le tableau précédent.

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Male14-24	1	-0.2511	-0.4973	0.7738	-0.1090
Southern	1	-0.0539	-0.3180	0.8626	-0.4714
Education	1	-0.1057	0.3920	-0.8600	0.5737
Expend60	1	0.0757	0.9661	-0.5862	0.0892
Expend59	1	0.0629	0.9623	-0.5974	0.0743
Labor	1	-0.3479	0.0546	-0.4169	0.8699
Male	1	0.1783	-0.1019	-0.2848	0.8699
PopSize	1	0.1243	0.7245	-0.1259	-0.3071
NonWhite	1	-0.0404	-0.1460	0.8333	-0.3842
Unemp14-24	1	0.9343	-0.0502	-0.1286	0.0704
Unemp35-39	1	0.9343	0.2255	0.0133	-0.2526
FamIncome	2	0.0733	0.7332	-0.8830	0.2726
IncUnderMed	1	-0.0258	-0.5554	0.9211	-0.2512

Nous disposons des corrélations de chaque variable avec l'ensemble des classes. Lorsque que la corrélation est supérieure à 0.7 (ou inférieure à –0.7), ce paramètre est modifiable, elle est mise en surbrillance et elle est recensée dans la colonne MEMBERS. Dans l'idéal, chaque variable ne devrait être significativement corrélée qu'avec une et une seule classe.

La première classe associe les variables de chômage : UNEMP14-24, UNEMP35-39. Ces variables semblent caractéristiques d'un phénomène fort, les états à taux de chômage élevé. Les autres variables sont très peu corrélées avec cette classe.

Avec la seconde classe, il semble que lorsque les dépenses liées à la sécurité sont élevées (EXPEND59 et EXPEND60), la population est importante (POPSIZE), et le niveau de revenu est élevé (FAMINCOME). On pourrait penser que les riches, quand ils sont nombreux, deviennent paranoïaques. Il faut y voir plutôt la caractérisation des états avec un plus grand nombre de grands centres urbains où le niveau de revenu est élevé par rapport aux campagnes. Tout à fait logiquement, la variable INCUNDERMED (la proportion des personnes ayant un revenu en dessous de la médiane nationale), caractéristique des zones à faible revenu, est corrélée négativement avec cette classe.

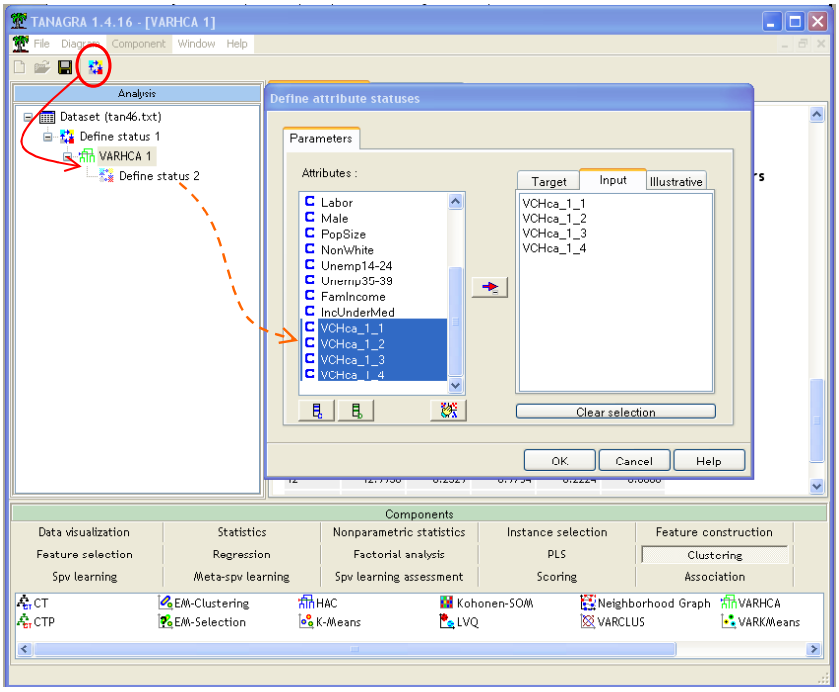
La troisième classe est assez intéressante car elle traduit un contraste : les états du sud (SOUTHERN), à proportion élevée de population de couleur (NONWHITE), à faible revenu (INCUNDERMED), et avec proportionnellement beaucoup de jeunes (MALE14-24), sont opposées aux états avec un niveau de revenu élevé (FAMINCOME) avec un niveau d'éducation moyen élevé (EDUCATION). Nous retrouvons des éléments de lecture que nous observons habituellement dans l'analyse en composante principale. Notons également que la variable FAMINCOME intervient de nouveau ici. C'est un des intérêts de cette approche : une variable peut être affectée mécaniquement à un groupe, mais en étudiant les corrélations, nous pouvons voir quelle est son influence sur l'ensemble des groupes. Nous pouvons relativiser le rôle discriminant d'une variable si elle est corrélée (colonne MEMBERSHIP) à tous les groupes ou au contraire à aucun des groupes.

Enfin, la quatrième classe traduit curieusement l'idée que la proportion d'hommes (MALE) est liée avec la proportion de personnes actives (LABOR). Gardons-nous bien de nous lancer dans des interprétations hasardeuses.

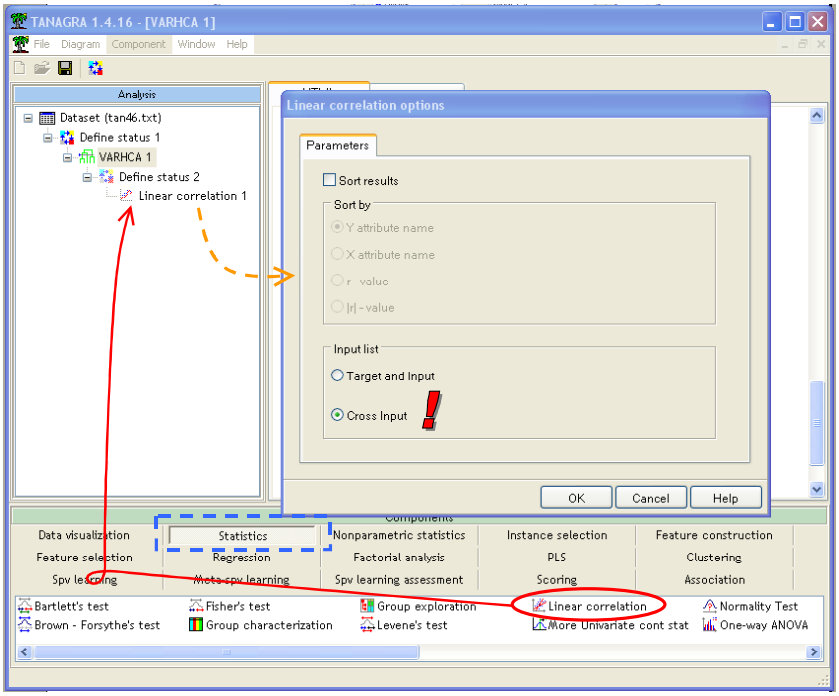
Corrélation entre les classes. Une autre manière d'évaluer la partition est de calculer la corrélation entre les classes, plus précisément la corrélation entre les variables qui représentent les

classes. Ces variables sont produites automatiquement par TANAGRA, elles correspondent au premier facteur de l'ACP dans chaque groupe.

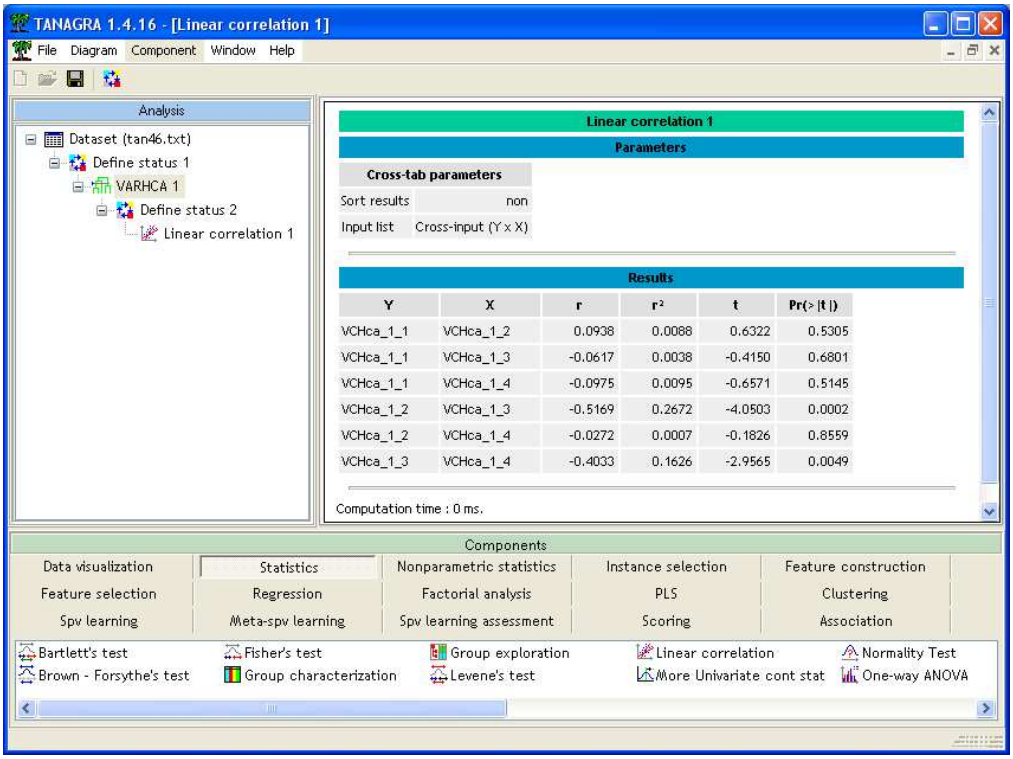
Nous plaçons un nouveau composant DEFINE STATUS dans le diagramme. Nous plaçons en INPUT les 4 nouvelles variables représentatives des 4 classes.



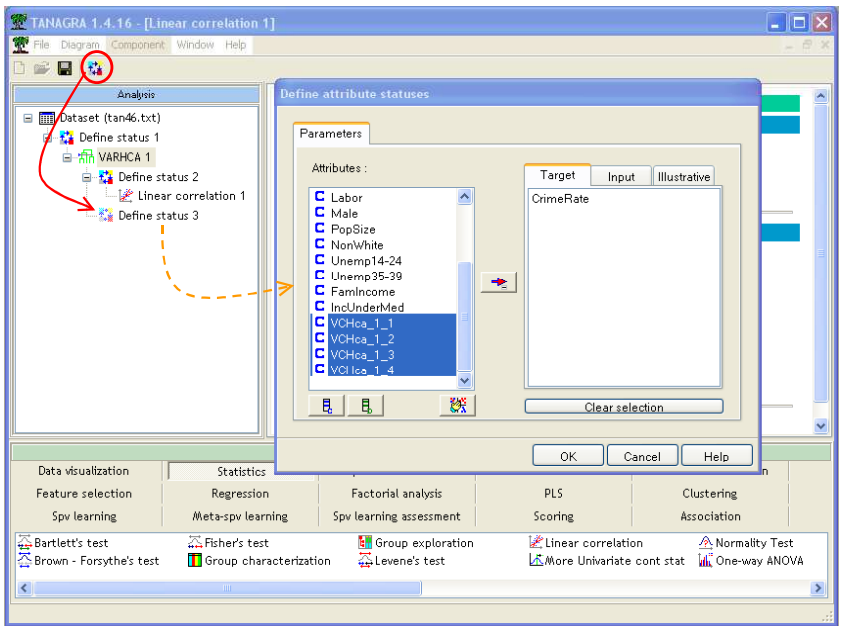
Puis nous insérons le composant LINEAR CORRELATION (onglet STATISTICS), nous la paramétrons pour que les calculs soient réalisés sur le croisement des variables en INPUT.



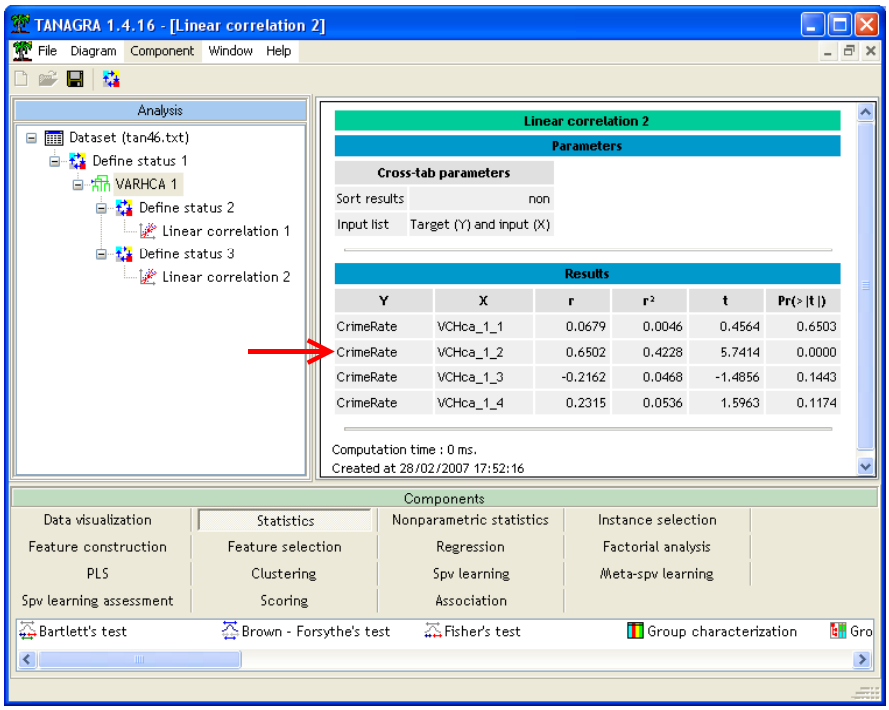
Nous constatons que la typologie est assez satisfaisante. Les classes sont faiblement corrélées, excepté le cluster 2 et le cluster 3 qui sont corrélés négativement : les états à forte population avec des dépenses liées à la sécurité élevées sont opposés aux états du sud relativement pauvres. Il faut surtout y voir une opposition entre les états « ruraux » et « urbains » vraisemblablement.



Utilisation des variables supplémentaires. Un autre analyse intéressante est de mettre à contribution notre variable supplémentaire, qui est en réalité notre principale variable d'intérêt ici (CRIME RATE). Quel est le groupe de variables qui expliquerait le mieux le haut niveau de criminalité ? Pour ce faire, nous insérons de nouveau le composant DEFINE STATUS. Nous plaçons en TARGET la variable CRIME RATE, en INPUT les variables issues de la typologie.

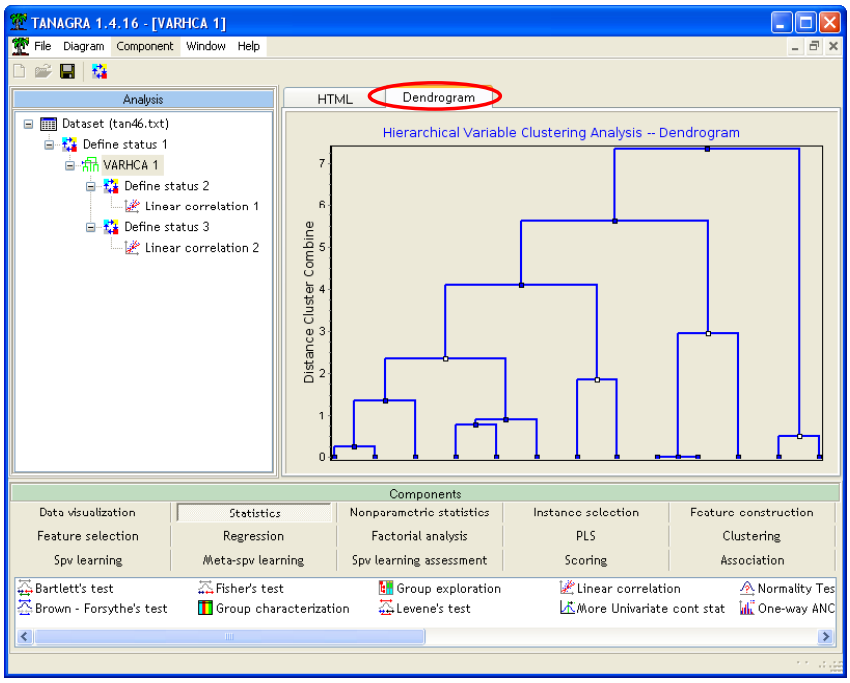


Puis nous insérons de nouveau le composant LINEAR CORRELATION en conservant les paramètres par défaut. Le résultat est assez édifiant : les états qui souffrent le plus de la criminalité sont celles à forte population engageant des dépenses élevées en matière de sécurité (corrélation positive de 0.65). Nous avons vu par ailleurs que ces états se distinguent par une forte population et un haut niveau de revenu, ce qui laisse à penser qu'elles ont une proportion plus élevée de grandes zones urbaines, variable absente dans notre fichier.



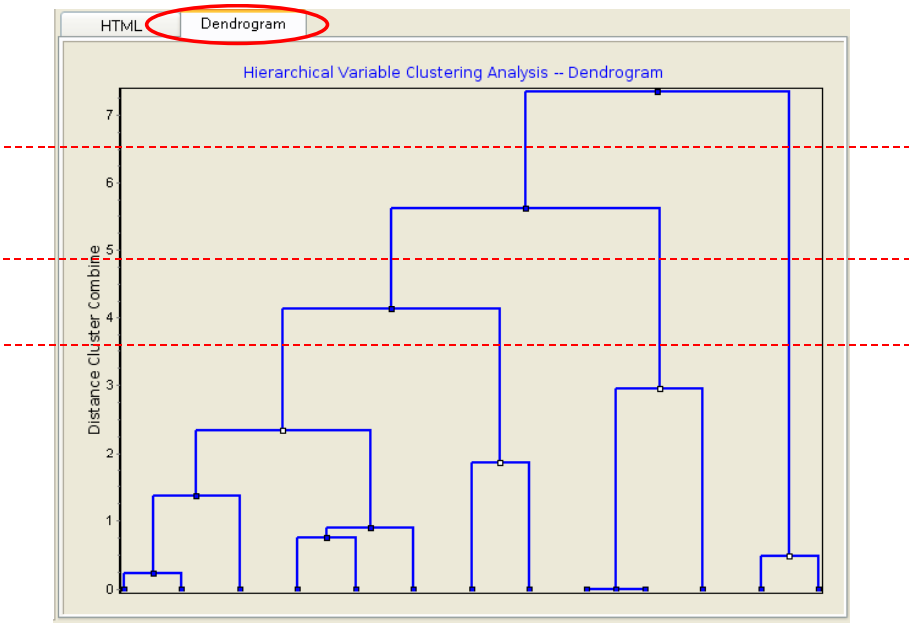
Les corrélations avec les autres groupes sont négligeables. L'intérêt de regrouper les variables en classes aussi orthogonales que possible est manifeste ici. En structurant au préalable les informations apportées par les données, nous discernons mieux les associations, et surtout les causalités que nous pouvons en déduire. Il ne s'agissait surtout pas d'affirmer sur la base des simples corrélations brutes que des dépenses liées à la sécurité élevées entraînaient une criminalité accrue.

Dendrogramme. Tout comme pour la typologie sur les individus (Composant HAC), nous disposons d'un dendrogramme qui retrace l'évolution des fusions et les indices d'agrégation.

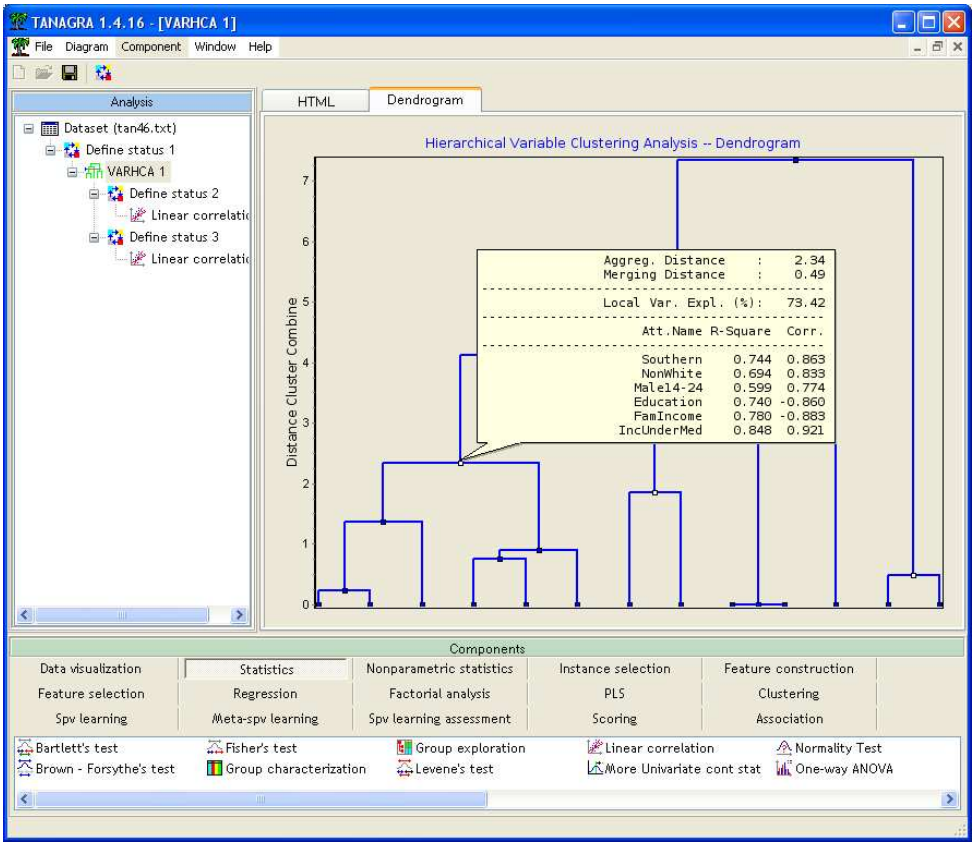


Nous observons les agrégations successives. Lorsque la distance d'agrégation -- l'écart en hauteur entre deux nœuds successifs -- est faible, cela indique que les classes fusionnées sont assez

semblables. En nous basant sur la hauteur des paliers, nous pouvons proposer les valeurs les plus appropriées du bon nombre de classes. Intuitivement, dans notre exemple, nous choisirons entre des partitions en 4, 3 et 2 groupes. Bien entendu, à l'instar de la classification des individus, il convient toujours de se méfier de la subdivision en 2 groupes. Comme il s'agit de la première partition, la variabilité expliquée varie fortement de manière mécanique.



Le dendrogramme de TANAGRA rapporte un autre type d'information. Les sommets correspondant à la typologie définie par le logiciel sont surlignés en blanc. Nous avons la possibilité d'observer la liste des variables qui sont attribués à chaque sommet, ce qui permet de suivre le détail du processus d'agrégation. Il suffit pour cela de cliquer sur les sommets de l'arbre hiérarchique.



Plusieurs informations sont disponibles : la hauteur du sommet dans l'arbre, la perte de variabilité expliquée consécutive à la fusion, la variabilité expliquée localement, et les corrélations des variables à la classe.

Détection automatique du nombre de classes. Nous avions vu dans le dendrogramme que les partitions en 3 ou 4 classes semblent les plus appropriées dans cet exemple. Or nous constatons que TANAGRA propose automatiquement une partition en 4 classes. Comment a-t-il procédé ?

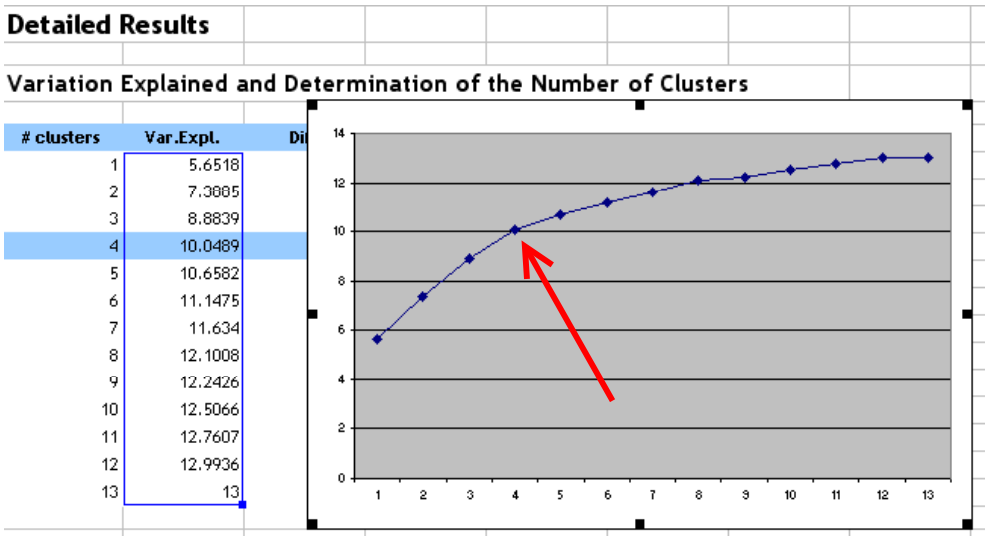
TANAGRA s'appuie sur la courbe d'évolution de la variabilité expliquée en fonction du nombre de classes. Ce tableau est disponible dans la section DETAILED RESULTS de la fenêtre de résultats.

Detailed Results

Variation Explained and Determination of the Number of Clusters

# clusters	Var.Expl.	Dif.	Cos	Angle	Moving Avg.
1	5.6518	0.0000	0.0000	0.0000	0.0000
2	7.3885	1.7367	0.9978	0.0670	0.0623
3	8.8839	1.4955	0.9928	0.1199	0.1671
4	10.0489	1.1650	0.9510	0.3142	0.1754
5	10.6582	0.6093	0.9958	0.0921	0.1362
6	11.1475	0.4893	1.0000	0.0023	0.0368
7	11.6340	0.4864	0.9999	0.0160	0.1047
8	12.1008	0.4668	0.9566	0.2959	0.1430
9	12.2426	0.1418	0.9931	0.1172	0.1408
10	12.5066	0.2640	1.0000	0.0093	0.0488
11	12.7607	0.2541	0.9998	0.0200	0.0839
12	12.9936	0.2329	0.9754	0.2224	0.0808
13	13.0000	0.0064	0.0000	0.0000	0.0000

En recopiant ces valeurs dans un tableur, puis en construisant la courbe, nous obtenons le graphique suivant. Parmi les innombrables règles de détection du nombre adéquat des classes, il y a la fameuse « loi du coude » : nous cherchons visuellement la zone où la variation de courbure est la plus importante, traduisant l'idée d'une modification significative des proximités des variables à l'intérieur des groupes.



Si l'appréciation visuelle est simple, il est plus difficile de trouver une solution informatique. Dans TANAGRA, nous introduisons une idée simple : pour chaque point, nous calculons la demi-tangente à gauche et à droite, puis nous évaluons l'angle que font ces demi-droites. Plus l'angle est fort, plus nous suspectons la présence d'un coude. Les calculs locaux étant généralement très instables, pour

lisser les estimations, nous calculons une moyenne mobile sur 3 points. De plus, afin de simplifier les calculs, et ne disposant pas de toute manière de l'expression analytique de la courbe, nous substituons au calcul des demi-tangentes, le calcul des droites d'interpolation à gauche et à droite.

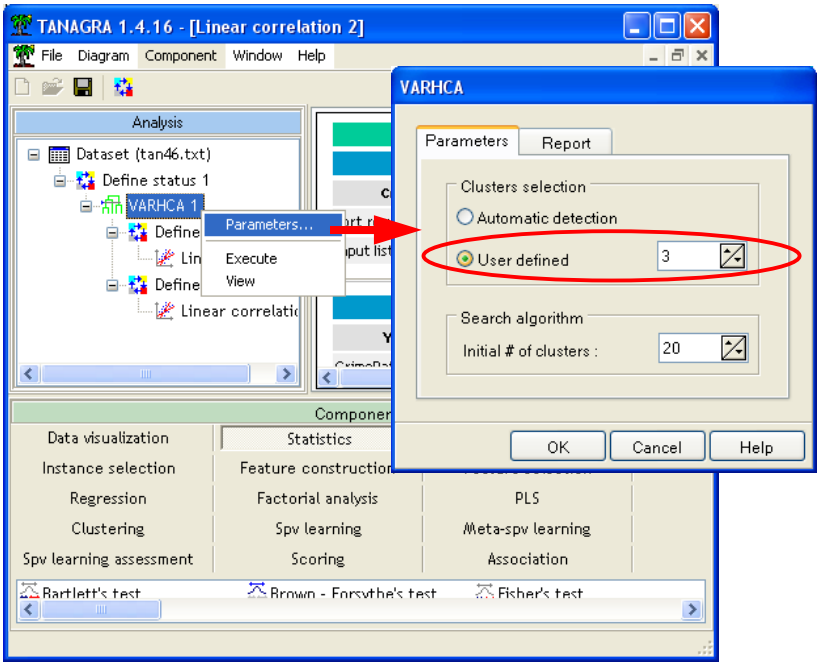
Ainsi, nous disposons du tableau détaillé des résultats. TANAGRA met en vert la solution correspondant au « coude » le plus significatif. Il met en gris foncé les deux autres zones de coupure de l'arbre possible. Nous constatons que la partition la plus intéressante semble être en 4 classes. La partition en 3 classes serait la solution suivante. Enfin, il semble également qu'une partition en 8 groupes serait à envisager.

Detailed Results

Variation Explained and Determination of the Number of Clusters

# clusters	Var.Expl.	Dif.	Cos	Angle	Moving Avg.
1	5.6518	0.0000	0.0000	0.0000	0.0000
2	7.3885	1.7367	0.9978	0.0670	0.0623
3	8.8839	1.4955	0.9928	0.1199	0.1671
4	10.0489	1.1650	0.9510	0.3142	0.1754
5	10.6582	0.6093	0.9958	0.0921	0.1362
6	11.1475	0.4893	1.0000	0.0023	0.0368
7	11.6340	0.4864	0.9999	0.0160	0.1047
8	12.1008	0.4668	0.9566	0.2959	0.1430
9	12.2426	0.1418	0.9931	0.1172	0.1408
10	12.5066	0.2640	1.0000	0.0093	0.0488
11	12.7607	0.2541	0.9998	0.0200	0.0839
12	12.9936	0.2329	0.9754	0.2224	0.0808
13	13.0000	0.0064	0.0000	0.0000	0.0000

Partition avec un nombre de classes prédéfini. Pour construire une partition avec un nombre de classes différent, en 3 classes par exemple, il faudrait activer le menu PARAMETERS, spécifier cette nouvelle option (CLUSTER SELECTION = USER DEFINED) et définir le nombre de classes (3).



Nous obtenons le résultat suivant.

Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	2	1.7459	0.8730
2	8	4.7537	0.5942
3	3	2.3843	0.7948
Total		8.8839	0.6834

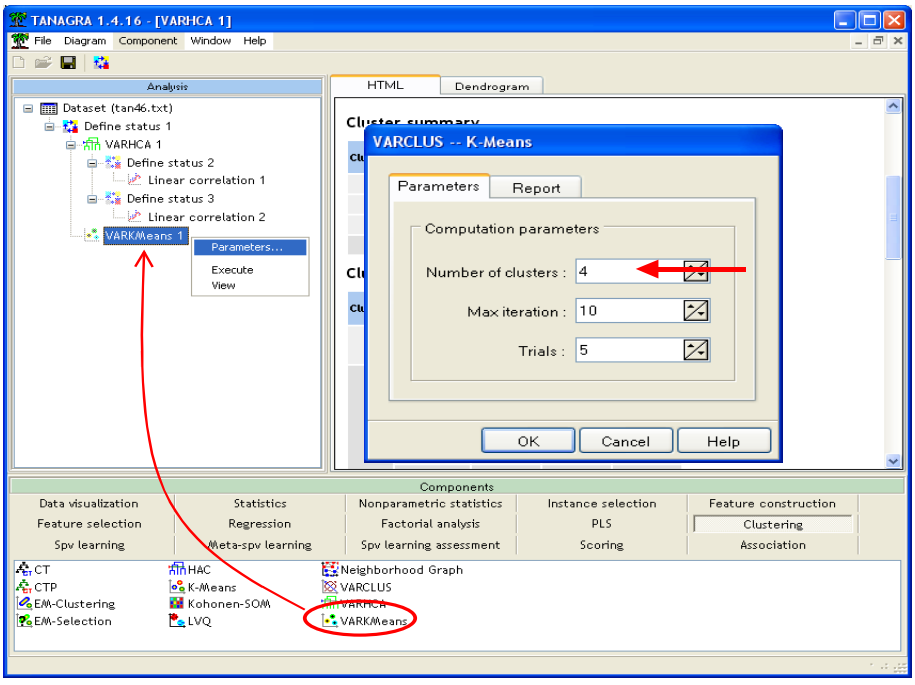
Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² ratio
1	Unemp14-24	0.8730	0.0025	0.1274
	Unemp35-39	0.8730	0.0508	0.1338
2	Southern	0.7671	0.1011	0.2591
	NonWhite	0.6898	0.0213	0.3169
	Male14-24	0.5232	0.2473	0.6334
	Education	0.7944	0.1537	0.2429
	FamIncome	0.7295	0.5376	0.5850
	IncUnderMed	0.7852	0.3085	0.3107
	Labor	0.2965	0.0030	0.7056
	Male	0.1680	0.0104	0.8407
3	Expend60	0.9334	0.3017	0.0954
	Expend59	0.9260	0.3102	0.1072
	PopSize	0.5249	0.0038	0.4769

VARKMEANS

VARKMEANS est une variante de la méthode de ré-allocation, adaptée aux variables. Elle repose toujours sur le principe des composantes latentes. L'utilisateur fixe le nombre de groupes qui sont formés au hasard dans un premier temps. Les variables sont itérativement affectées au groupe le plus proche, au sens du carré de la corrélation avec le premier facteur, jusqu'à ce qu'il y ait convergence. Le critère à optimiser est la variabilité totale expliquée. L'algorithme produit ses sorties conformes aux standards ci-dessus.

Plaçons donc le composant VARKMEANS dans notre diagramme, à la suite de DEFINE STATUS 1. Nous le paramétrons de manière à produire 4 classes. Les autres paramètres ne sont pas modifiés. Nous cliquons sur le menu VIEW pour obtenir les résultats.



Nous constatons que les classes correspondent à peu près à la partition avec la VARHCA, avec une lecture identique. La variabilité expliquée est légèrement inférieure, elle est de 75.26% contre 77.30%. L'écart est négligeable.

Cluster characterization

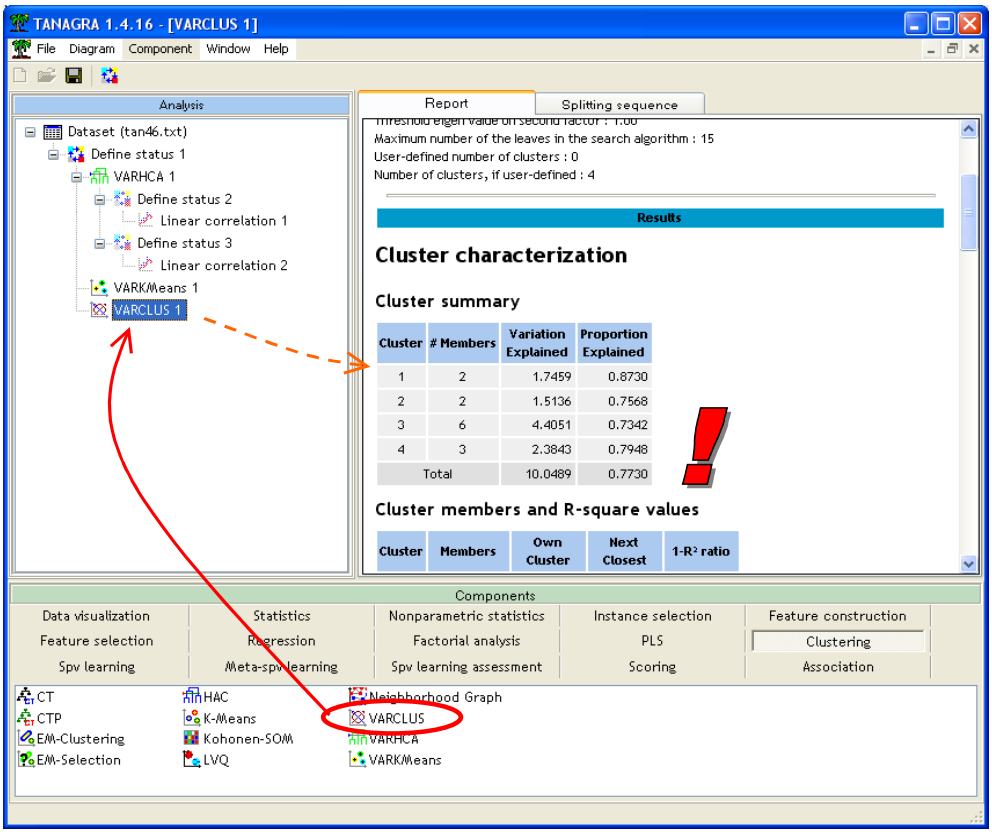
Cluster summary

Cluster	# Members	Variation Explained	Proportion Explained
1	4	3.1578	0.7895
2	4	3.1594	0.7899
3	2	1.7459	0.8730
4	3	1.7211	0.5737
Total		9.7843	0.7526

VARCLUS

VARCLUS est une approche descendante inspirée de la procédure du même nom du logiciel SAS (http://www2.stat.unibo.it/ManualiSas/stat/chap68.pdf). L'avantage, en temps de calcul, est manifeste lorsque le nombre de variables de départ est très élevé. En effet, l'exploration s'arrête dès qu'il n'y a plus de subdivisions pertinentes. A la différence de la procédure originelle, nous ne procédons pas aux ré-allocations à chaque étape de l'algorithme, cette opération étant très (trop) gourmande. La structure hiérarchique n'est donc pas préservée. Pour cette raison, l'arbre ne correspond pas à un dendrogramme, les hauteurs des sommets retranscrivent tout simplement la succession des opérations.

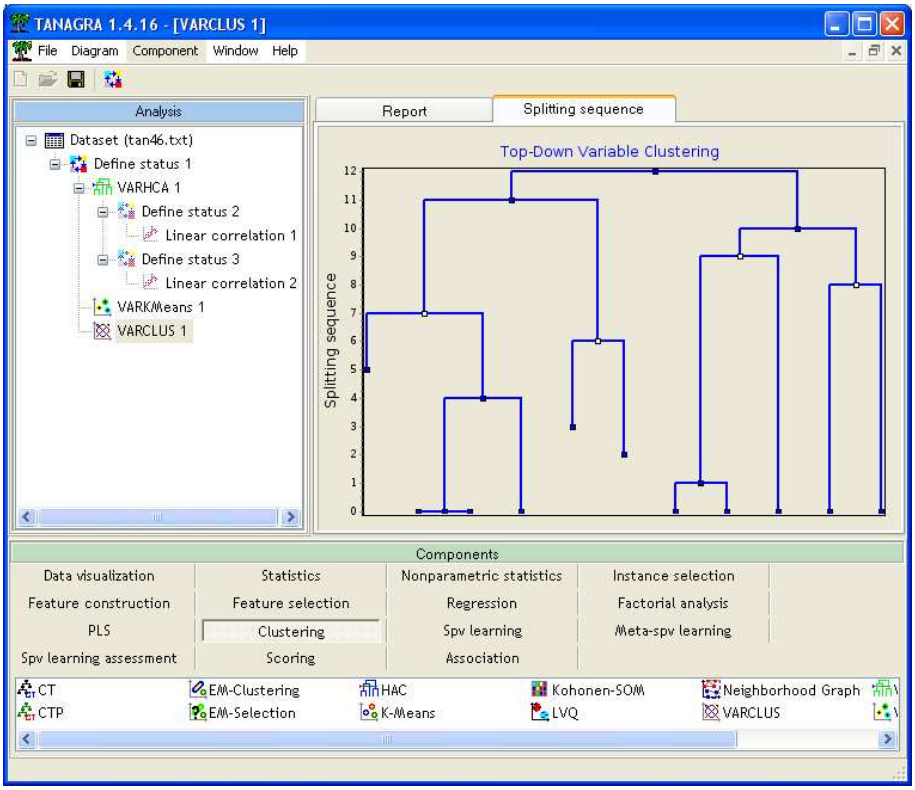
Nous plaçons le composant VARCLUS dans le diagramme, à la suite de DEFINE STATUS 1. Nous cliquons sur le menu VIEW pour accéder aux résultats. Le même formalisme est utilisé.



La procédure propose également un partitionnement en 4 classes, cohérente avec celle de VARHCA, avec une variabilité expliquée identique 77.30%.

Dans le second onglet de la fenêtre de résultat (SPLITTING SEQUENCE), le détail des opérations, la subdivision à chaque étape, est retranscrit graphiquement.

En cliquant sur les sommets, nous retrouvons les variables associées à chaque groupe. Les groupes produits par la typologie sont mis en évidence avec des sommets de couleur blanche.



N.B. : Attention, encore une fois, la hauteur des sommets retranscrit uniquement l'ordre des opérations ici.

Conclusion

La typologie des variables emmène souvent des informations intéressantes, complémentaires à la classification des individus.

En résumant les principales structures dans les données, elle propose de surcroît des pistes pour la sélection de variables. Nous avons également la possibilité d'utiliser directement les variables synthétiques associées à chaque groupe de variables.