1 Objectif

Mettre en œuvre les cartes de Kohonen avec Tanagra.

Les cartes de Kohonen sont des réseaux de neurones artificiels orientés, constitués de 2 couches. Dans la couche d'entrée, les neurones correspondent aux variables décrivant les observations. La couche de sortie, elle, est le plus souvent organisée sous forme de grille (de carte) de neurones à 2 dimensions. Chaque neurone représente un groupe d'observations similaires.

Le réseau de Kohonen est donc une technique de classification automatique (clustering, apprentissage non supervisé). L'objectif est de produire un regroupement de manière à ce que les individus situés dans la même case soient semblables, les individus situés dans des cases différentes soient différents. En y regardant de plus près, on se rend compte d'ailleurs que l'algorithme d'apprentissage est une version sophistiquée de la méthode des K-Means (on parle de « nuées dynamiques » en français, bien que cette dernière intègre elle aussi d'autres types d'améliorations par rapport aux K-Means de Forgy [1965]).

Les cartes de Kohonen constituent également une technique de visualisation. En effet, les neurones de la couche de sortie sont organisés de manière à ce que deux cellules adjacentes dans la grille correspondent à des groupes d'observations proches dans l'espace de représentation initial. On parle de cartes auto organisatrices (SOM : Self Organisation Map). De ce point de vue, le réseau de Kohonen se positionne par rapport aux techniques factorielles de réduction de dimensionnalité. A la différence que la projection est non linéaire.

Dans ce didacticiel, nous montrons comment mettre en œuvre l'algorithme de Kohonen dans Tanagra. Nous visualiserons graphiquement les résultats. L'idée est de vérifier cette fameuse proximité entre les cellules de la grille dans l'espace de représentation. Puis, nous comparons les groupes obtenus avec ceux de la méthode des K-Means, très largement répandue au sein de la communauté de l'apprentissage automatique. Enfin, nous montrons comment nous pouvons affiner les résultats en lançant une classification ascendante hiérarchique (CAH) à partir des cellules de la carte. Cette stratégie est une approche alternative de la classification mixte (K-MEANS + CAH ; Lebart et al., 2000[1]). Elle est particulièrement recommandée pour les fichiers comportant un grand nombre d'observations.

2 Données

Nous utilisons les données WAVEFORM de Breiman et al. (1984). Il s'agit de données artificielles comportant 21 variables. Nous disposons de 5000 observations (WAVEFORM_UNSUPERVISED.XLS[2]). Dans le fichier originel[3], les observations sont classées, *a priori*, dans 3 catégories différentes. Nous n'utiliserons pas cette information dans ce didacticiel.

[1] L. Lebart, A. Morineau, M. Piron, « Statistique Exploratoire Multidimensionnelle », Dunod, 2000 ; section 2.3, pages 177 à 180.

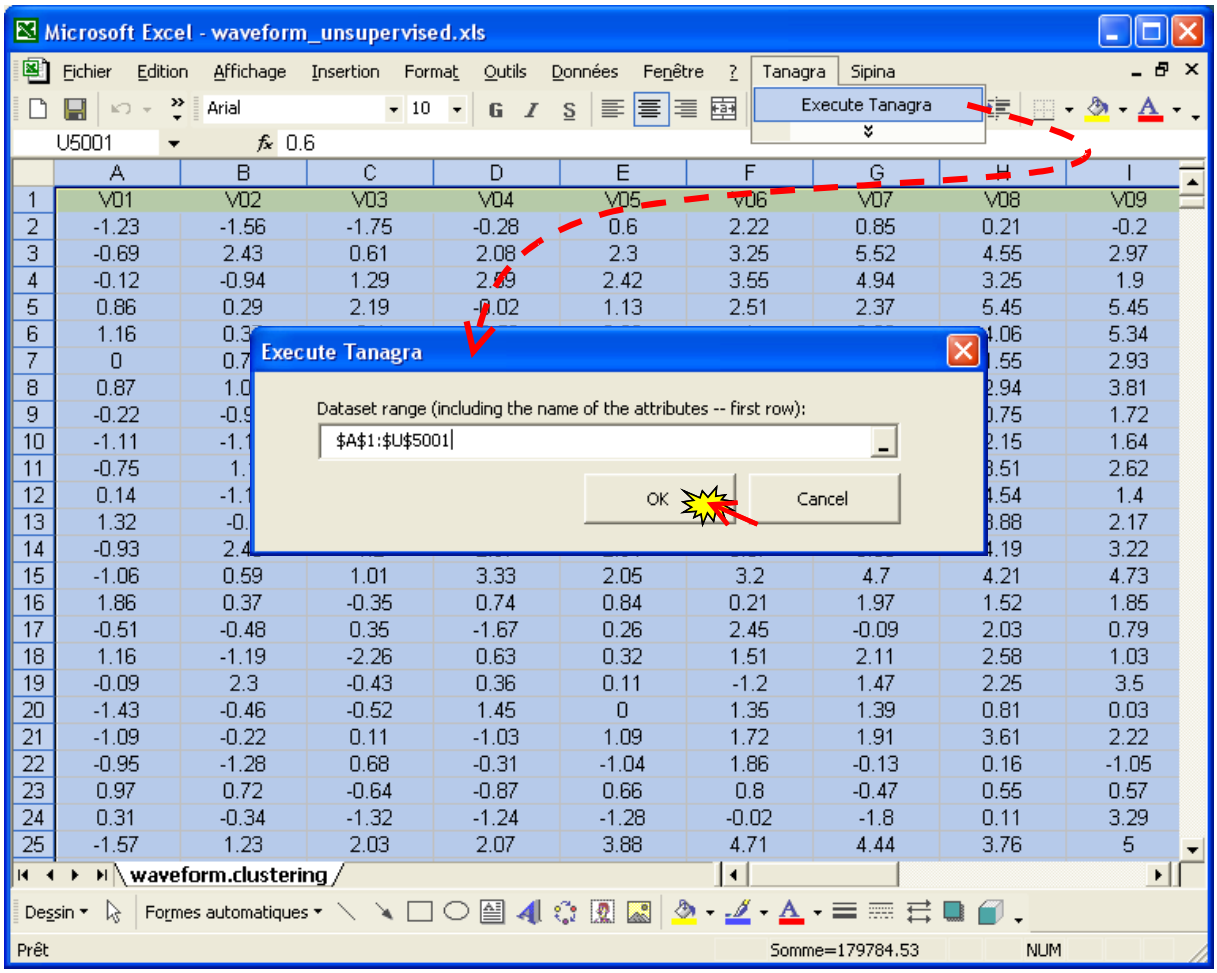
[2] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/waveform_unsupervised.xls

[3] http://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+(Version+1)

3 Les cartes de Kohonen avec Tanagra

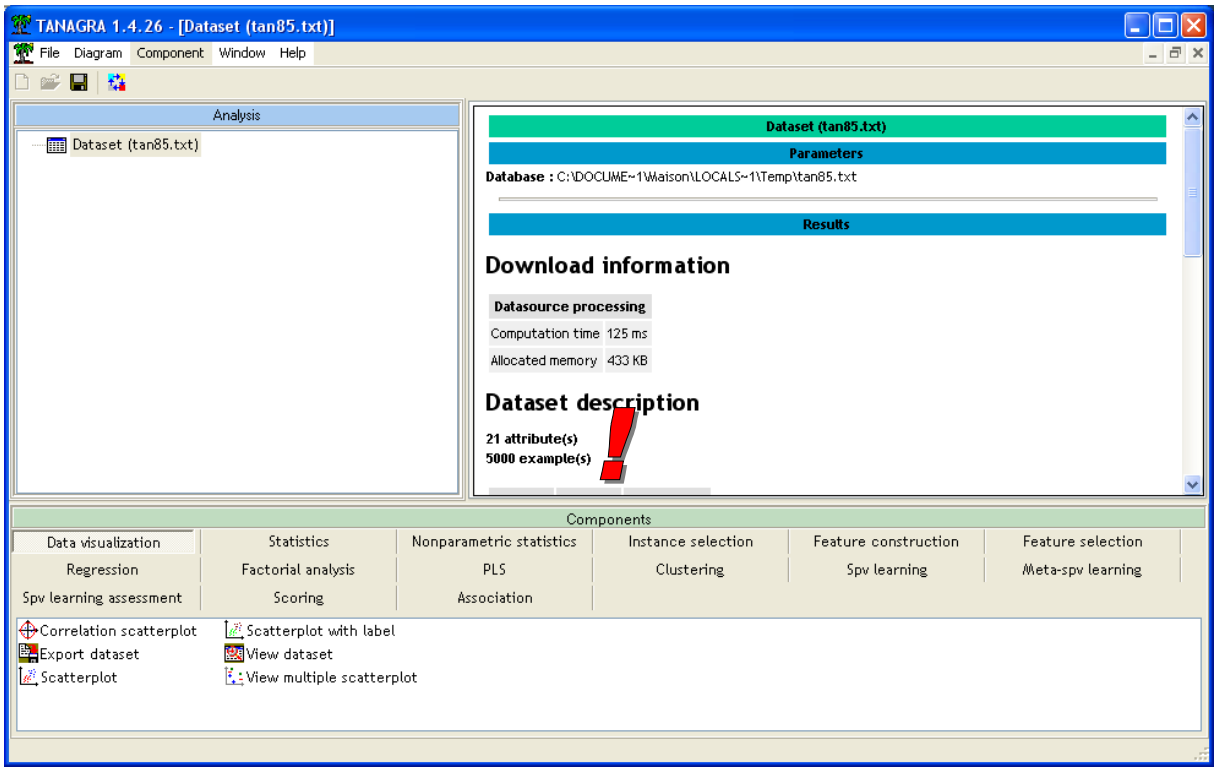
3.1 Importation des données

Le plus simple pour lancer Tanagra et charger les données est d'ouvrir le fichier XLS dans le tableur EXCEL. Nous sélectionnons la plage de données. La première ligne doit correspondre au nom des variables. Puis nous activons le menu TANAGRA / EXECUTE TANAGRA qui a été installé avec la macro complémentaire TANAGRA.XLA[4]. Une boîte de dialogue apparaît. Nous vérifions la sélection. Si tout est en règle, nous validons en cliquant sur le bouton OK.



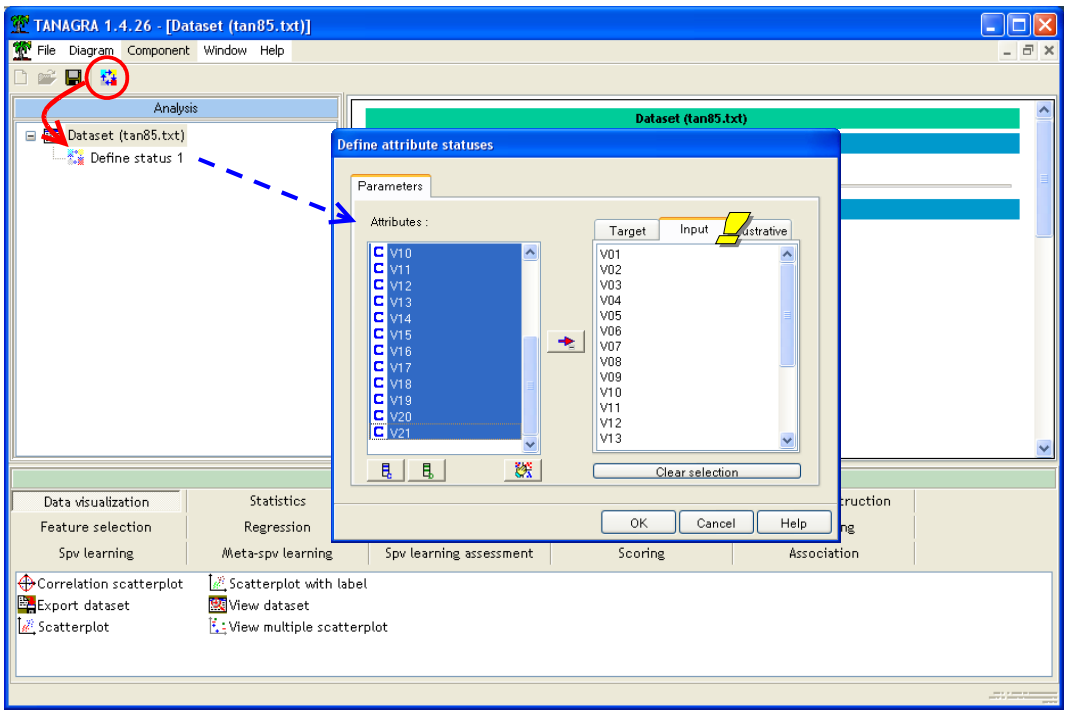
Tanagra est automatiquement lancé, un nouveau diagramme est créé, les données sont chargées. Nous vérifions que nous avons bien 5000 observations et 21 variables.

[4] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.

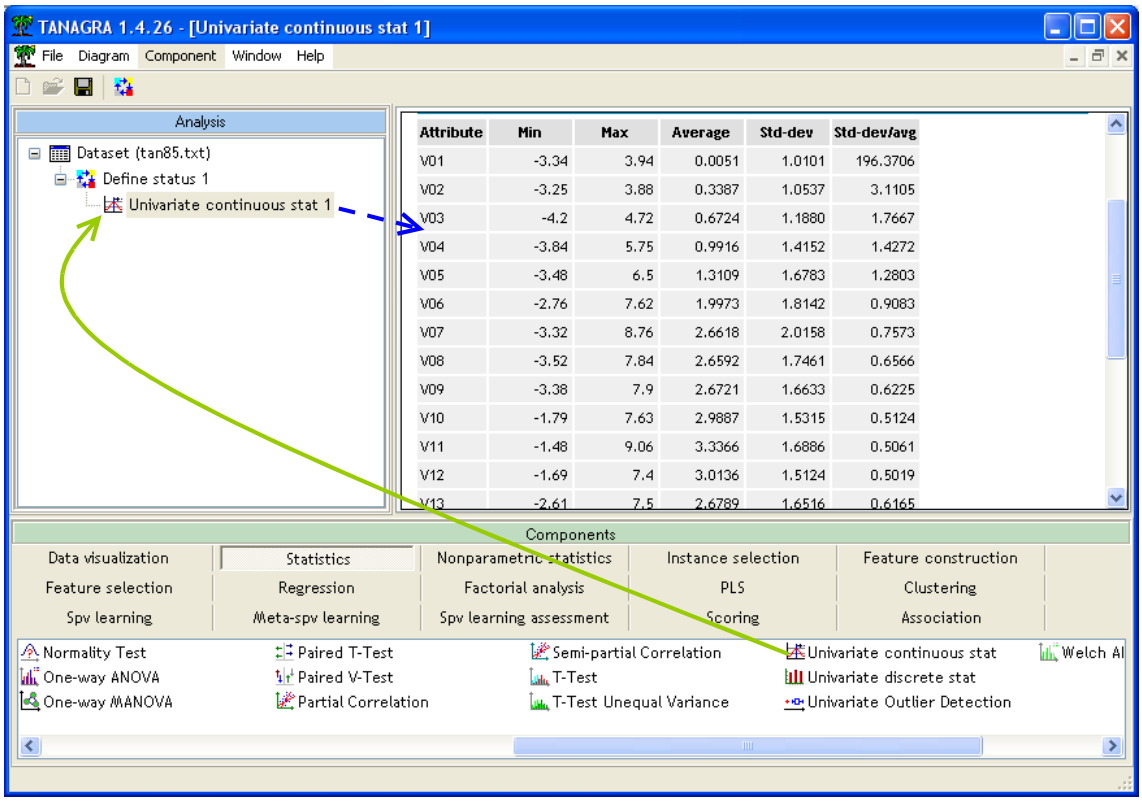


3.2 Statistique descriptives et détection des points atypiques

Première étape toujours, vérifions l'intégrité des données en calculant quelques indicateurs statistiques. Nous insérons DEFINE STATUS dans le diagramme, via le raccourci de la barre d'outils, nous plaçons en INPUT toutes les variables disponibles. **Attention, elles doivent être continues (C) si nous souhaitons leur appliquer l'algorithme de Kohonen**.



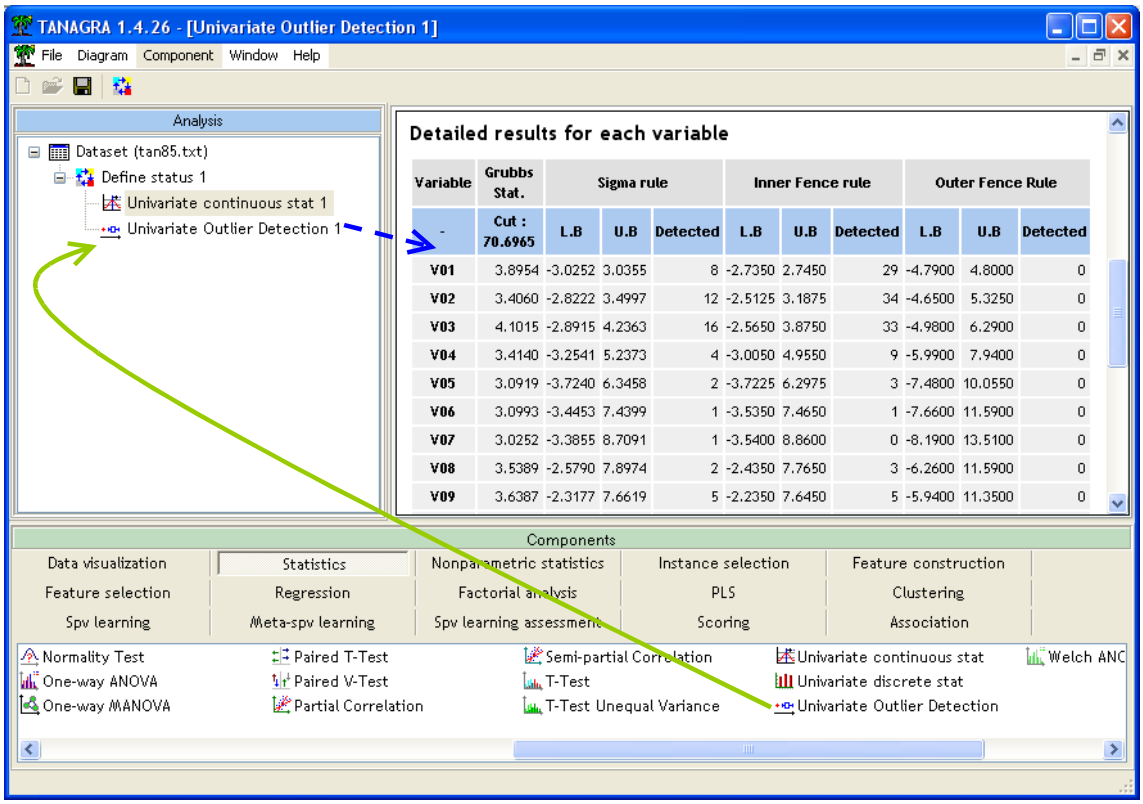
Puis nous insérons le composant UNIVARIATE CONTINUOUS STAT (onglet STATISTICS). Nous activons le menu VIEW.



Attribute	Min	Max	Average	Std-dev	Std-dev/avg
V01	-3.34	3.94	0.0051	1.0101	196.3706
V02	-3.25	3.88	0.3387	1.0537	3.1105
V03	-4.2	4.72	0.6724	1.1880	1.7667
V04	-3.84	5.75	0.9916	1.4152	1.4272
V05	-3.48	6.5	1.3109	1.6783	1.2803
V06	-2.76	7.62	1.9973	1.8142	0.9083
V07	-3.32	8.76	2.6618	2.0158	0.7573
V08	-3.52	7.84	2.6592	1.7461	0.6566
V09	-3.38	7.9	2.6721	1.6633	0.6225
V10	-1.79	7.63	2.9887	1.5315	0.5124
V11	-1.48	9.06	3.3366	1.6886	0.5061
V12	-1.69	7.4	3.0136	1.5124	0.5019
V13	-2.61	7.5	2.6789	1.6516	0.6165

Nous constatons qu'il n'y a pas de constante dans le fichier c.-à-d. des variables avec un écart type (Std-Dev) nul. Les variables sont définies sur des plages de valeurs à peu près identiques.

Nous pouvons compléter cette première inspection par la détection univariée des points atypiques, toujours pour avoir une première expertise sur les données. Nous introduisons le composant UNIVARIATE OUTLIER DETECTION (onglet STATISTICS).



Detailed results for each variable

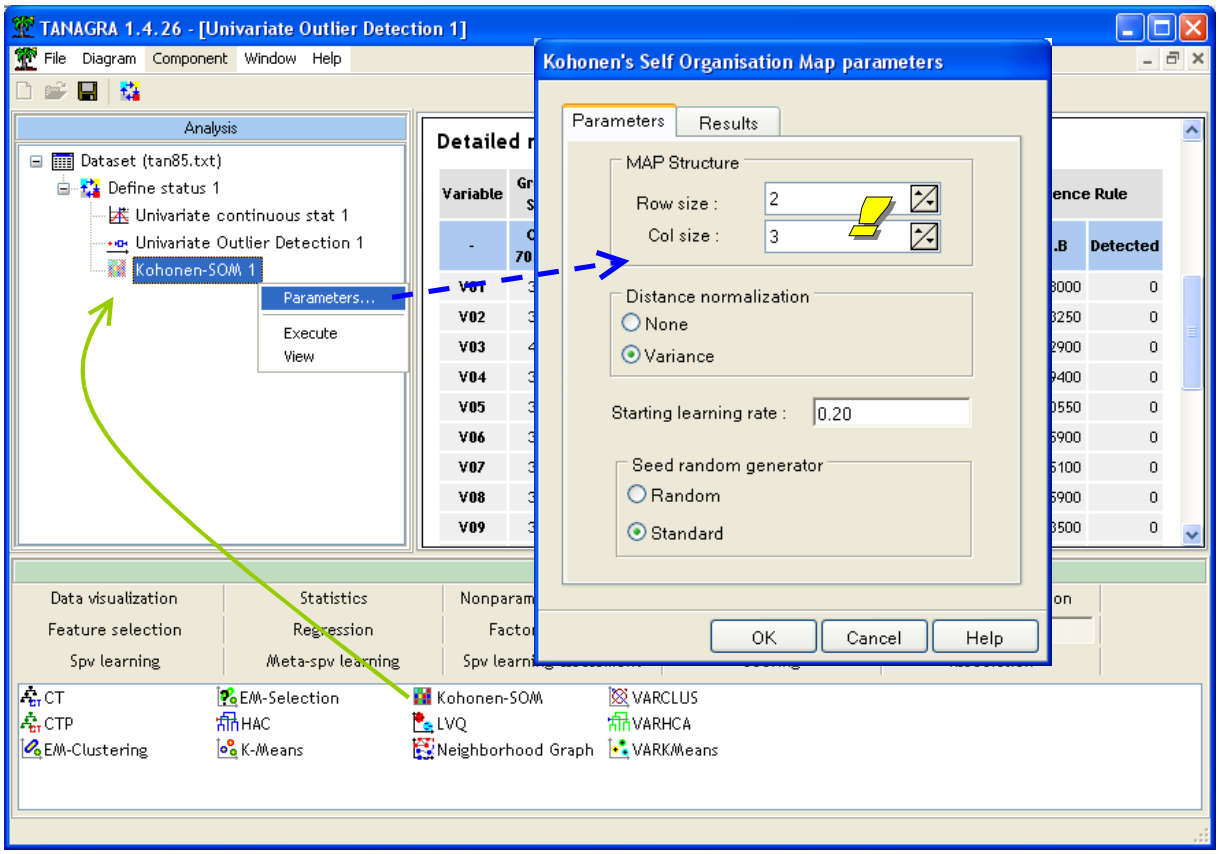
Variable	Grubbs Stat.	Sigma rule			Inner Fence rule			Outer Fence Rule		
-	Cut : 70.6965	L.B	U.B	Detected	L.B	U.B	Detected	L.B	U.B	Detected
V01	3.8954	-3.0252	3.0355	8	-2.7350	2.7450	29	-4.7900	4.8000	0
V02	3.4060	-2.8222	3.4997	12	-2.5125	3.1875	34	-4.6500	5.3250	0
V03	4.1015	-2.8915	4.2363	16	-2.5650	3.8750	33	-4.9800	6.2900	0
V04	3.4140	-3.2541	5.2373	4	-3.0050	4.9550	9	-5.9900	7.9400	0
V05	3.0919	-3.7240	6.3458	2	-3.7225	6.2975	3	-7.4800	10.0550	0
V06	3.0993	-3.4453	7.4399	1	-3.5350	7.4650	1	-7.6600	11.5900	0
V07	3.0252	-3.3855	8.7091	1	-3.5400	8.8600	0	-8.1900	13.5100	0
V08	3.5389	-2.5790	7.8974	2	-2.4350	7.7650	3	-6.2600	11.5900	0
V09	3.6387	-2.3177	7.6619	5	-2.2350	7.6450	5	-5.9400	11.3500	0

Quelle que soit la variable analysée, il n'y a pas de points atypiques[5]. Ce n'est guère étonnant, les données ont été générées artificiellement.

3.3 Le composant KOHONEN-SOM

Il est temps de lancer l'analyse, nous souhaitons créer une grille 2 X 3 (2 lignes et 3 colonnes), soit une classification en 2 x 3 = 6 clusters.

Nous insérons le composant KOHONEN-SOM (onglet CLUSTERING) dans le diagramme. Nous activons le menu contextuel PARAMETERS.

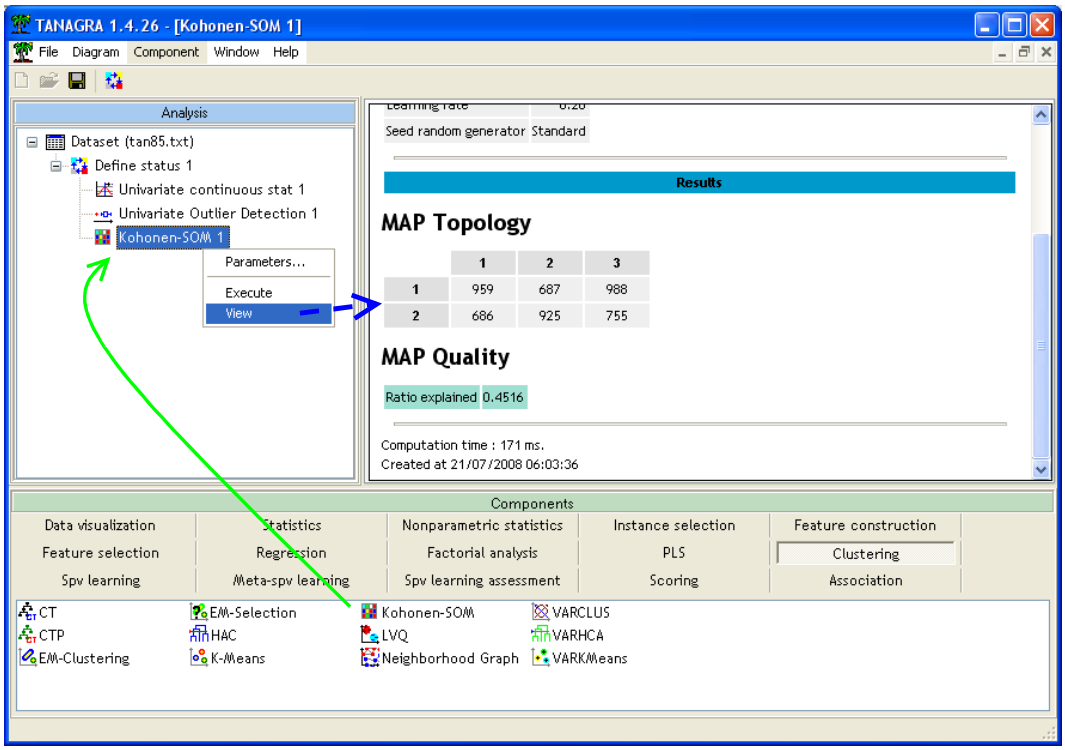


Le nombre de lignes dans la grille est 2 (Row Size), le nombre de colonnes 3 (Col Size). Les données sont normalisées c.-à-d. divisées par leur écart type de manière à les ramener à la même échelle. Cela peut être utile lorsque les variables sont définies dans des échelles très différentes. Dans notre cas, à la lumière des statistiques descriptives ci-dessus, cette opération est un peu superflue. Nous conservons les paramètres par défaut néanmoins. Nous ne modifions pas non plus les autres paramètres[6].

Nous validons ces choix en cliquant sur OK, puis nous activons le menu contextuel VIEW afin d'obtenir les résultats.

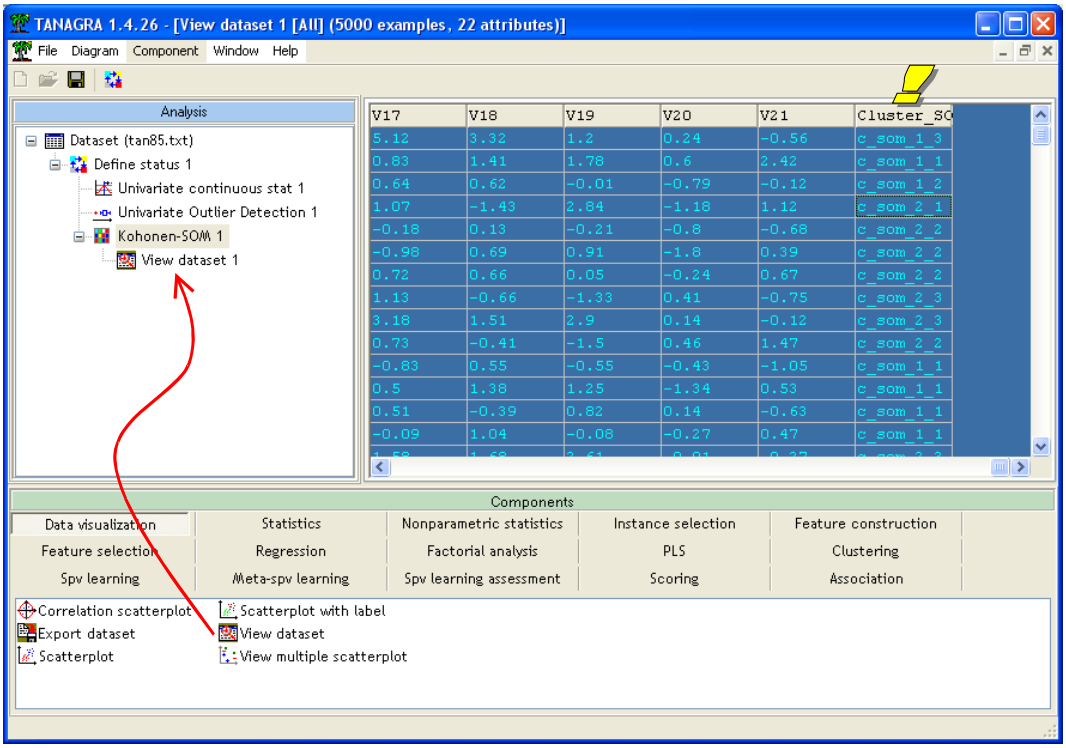
[5] Pour plus de détails sur les techniques et la lecture approfondie des résultats dans Tanagra, voir le didacticiel http://tutoriels-data-mining.blogspot.com/2008/05/dtection-univarie-des-points-aberrants.html

[6] Pour plus de détails sur la technique et son paramétrage, voir http://en.wikipedia.org/wiki/Self-organizing_map



La grille est affichée. Nous distinguons les effectifs dans chaque cellule. Tanagra nous indique que **45.16%** de l'inertie est expliquée par le partitionnement en 6 groupes. Nous pourrons comparer ce résultat avec celui des autres méthodes.

3.4 Affectation des individus aux groupes (cellules)



Le composant KOHONEN-SOM ajoute automatiquement une nouvelle colonne à l'ensemble de données, disponible pour les éventuels traitements en aval. Nous pouvons la visualiser en insérant le composant VIEW DATASET (onglet DATA VISUALIZATION). La nouvelle variable est rajoutée en

dernière position. Ainsi la première observation est affectée à la cellule n° (1 ; 3) c.-à-d. 1ère ligne et 3ème colonne, etc.

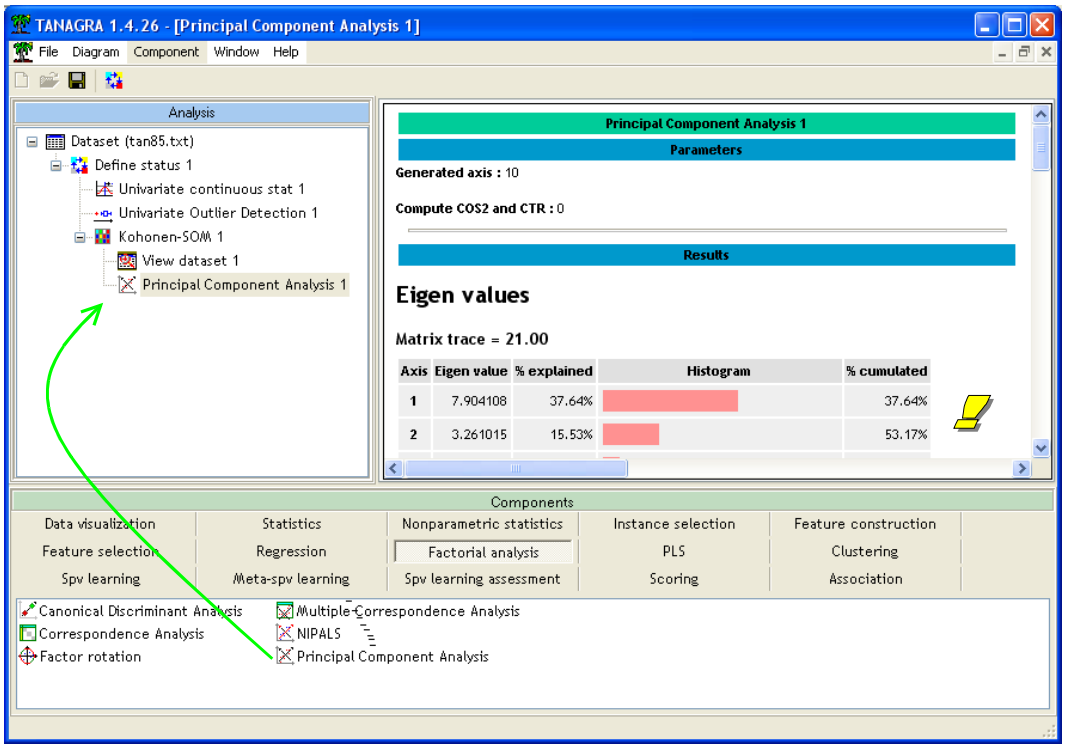
A l'aide de ce dispositif, nous pouvons classer des individus supplémentaires, n'ayant pas participé aux calculs. Cette fonctionnalité est très importante. Le **déploiement des modèles** est une des principales finalités du data mining[7].

4 Proximité des groupes dans la carte

La proximité des cellules dans la grille équivaut à une proximité dans l'espace initial disions nous plus haut. Vérifions cela sur la carte 2 x 3 fournie par Tanagra pour les données WAVEFORM.

Il est impossible de représenter le nuage de points directement dans l'espace à 21 dimensions. Nous utilisons une analyse en composantes principales pour réduire la dimensionnalité. Si le premier plan factoriel est de qualité suffisante, nous essaierons de visualiser les positions relatives des groupes c.-à-d. des individus associés aux cellules de la grille.

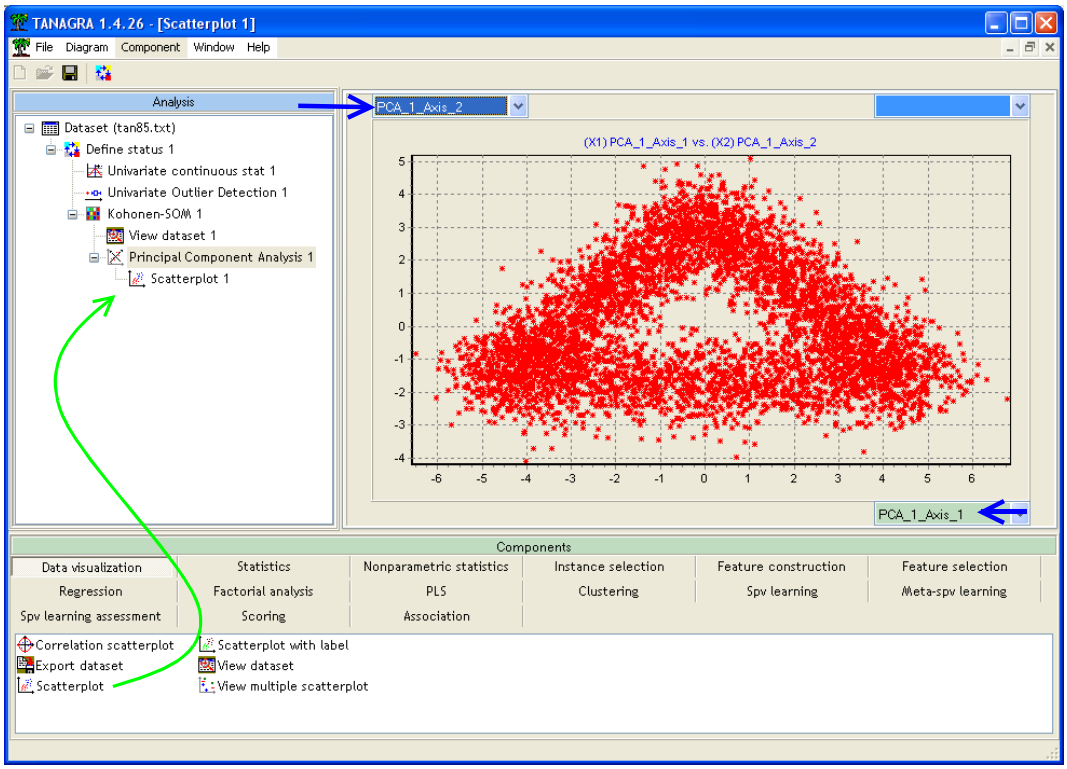
Nous insérons le composant PRINCIPAL COMPONENT ANALYSIS (onglet FACTORIAL ANALYSIS) à la suite de la méthode Kohonen. Nous activons le menu VIEW.



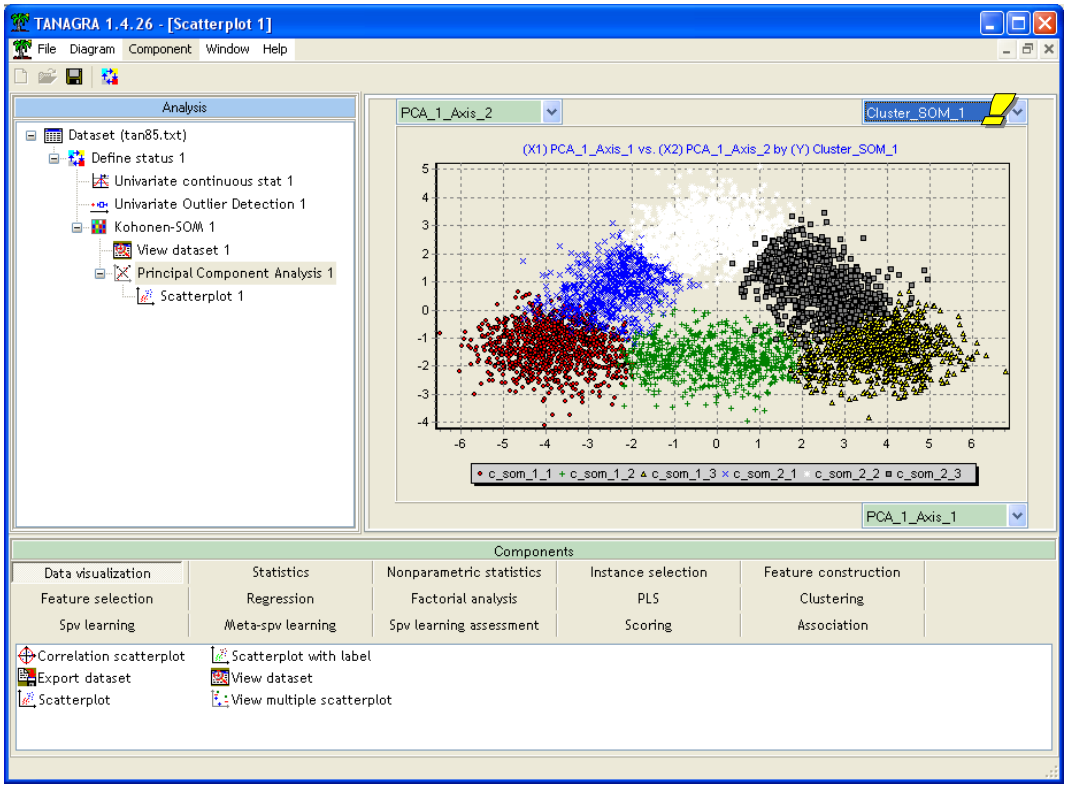
Le premier axe factoriel explique 53.17% de la variabilité totale. La valeur semble faible. Mais la pratique assidue de ce fichier nous a enseigné que ce résultat est amplement suffisant pour positionner correctement les points.

Nous insérons le composant SCATTERPLOT (onglet DATA VISUALIZATION) pour projeter les points dans le premier plan factoriel : en abscisse le premier axe, en ordonnée le second.

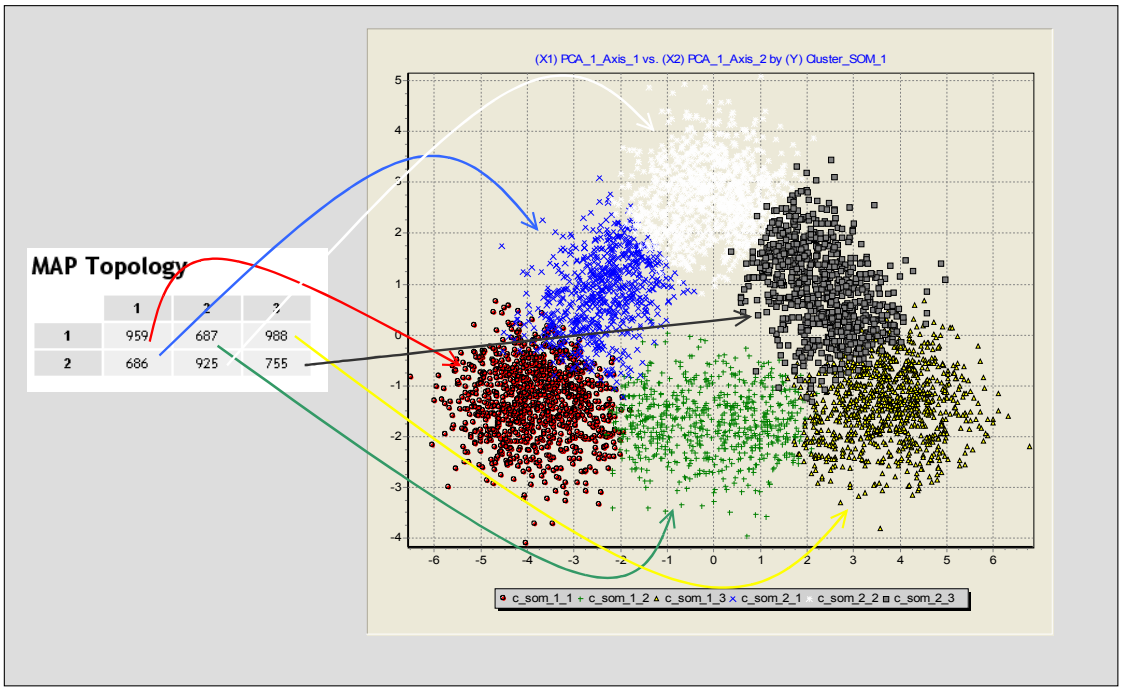
[7] Voir http://tutoriels-data-mining.blogspot.com/2008/03/dploiement-de-modles-avec-tanagra.html concernant le déploiement des modèles en apprentissage supervisé avec Tanagra. Nous sommes dans une situation un peu différente puisque nous souhaitons appliquer un modèle non supervisé. Mais la trame globale est bien la même.



Etape décisive, nous illustrons les points selon leur groupe d'appartenance issu de la méthode Kohonen. Nous utilisons pour ce faire la variable générée automatiquement CLUSTER_SOM_1.



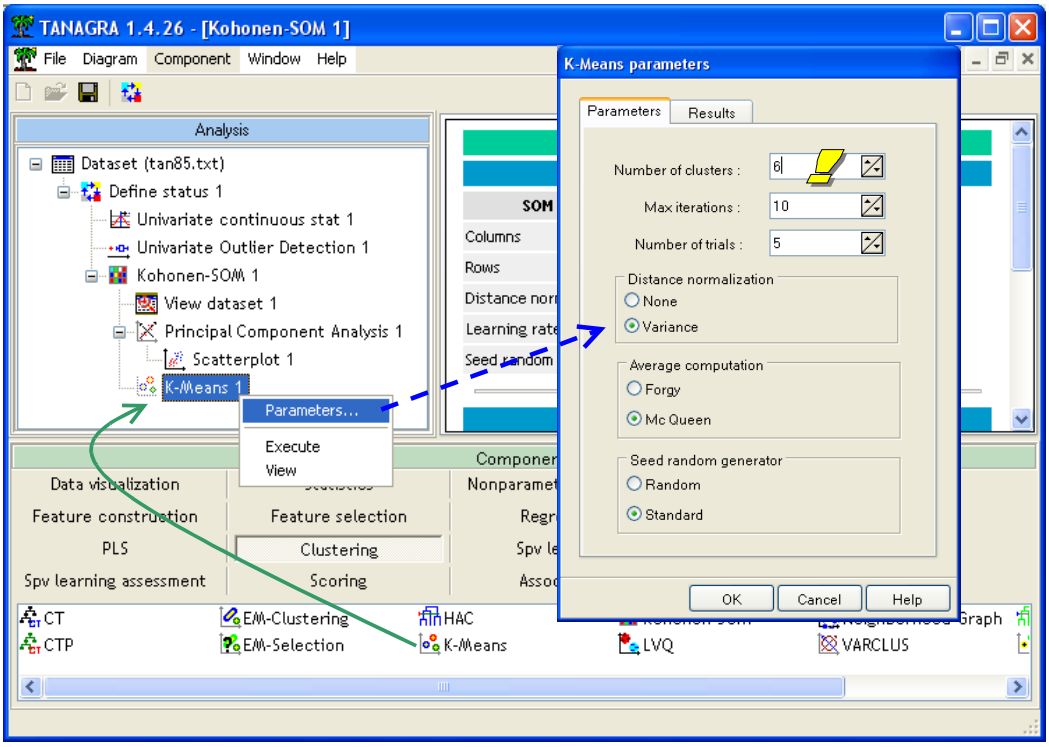
Nous retrouvons les correspondances en faisant le parallèle entre les proximités dans la carte et les proximités dans le premier plan factoriel. Les observations situées dans des cellules adjacentes sont proches dans l'espace initial de représentation (via la représentation issue de l'analyse en composantes principales).



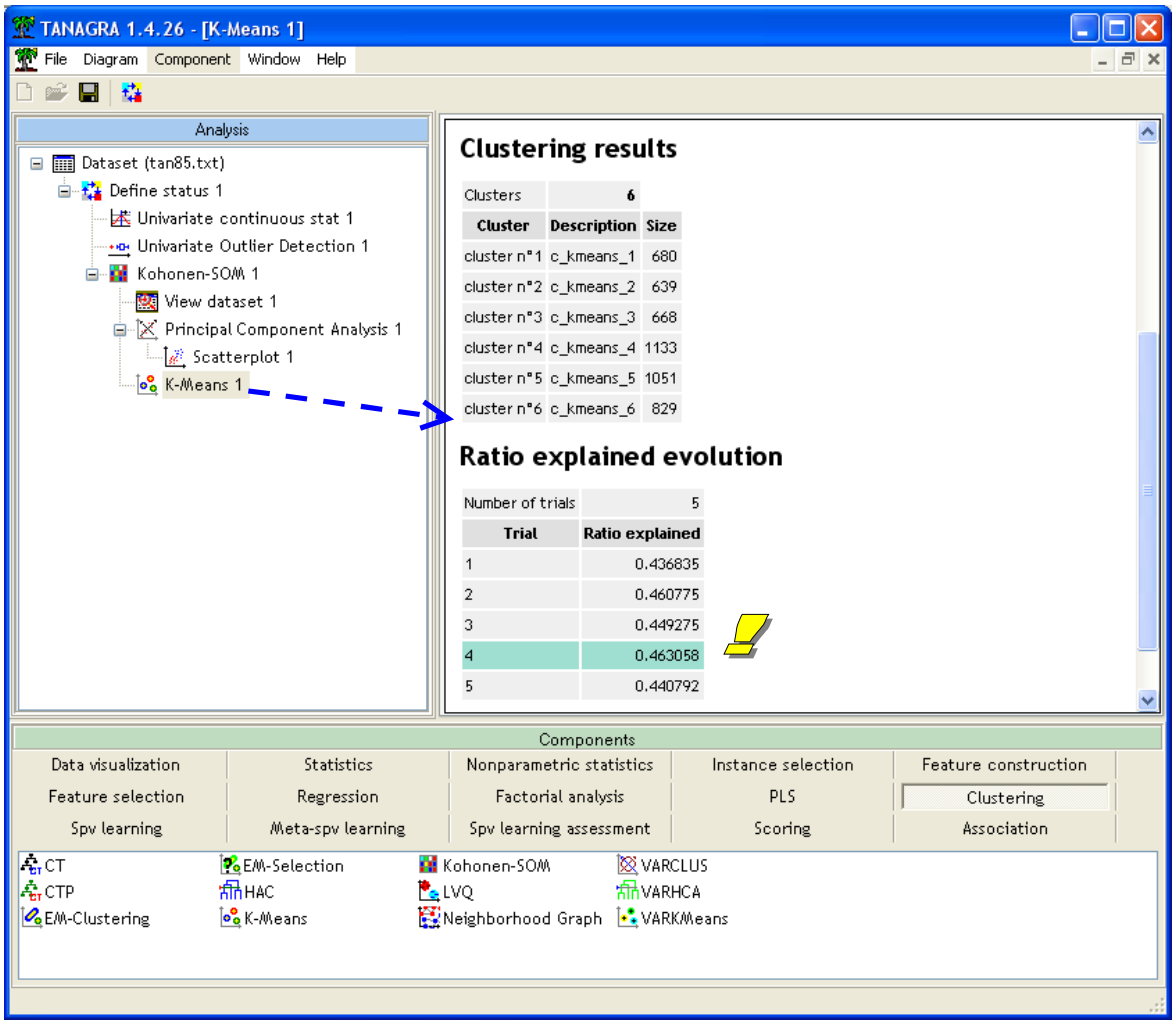
5 Comparaison avec la méthode des K-MEANS

5.1 Performances comparées

Nous souhaitons maintenant comparer les performances de SOM (Self Organization Map) de Kohonen avec la méthode de K-Means qui constitue une référence depuis plusieurs dizaines d'années. Nous insérons le composant K-MEANS (onglet CLUSTERING) à la suite de KOHONEN-SOM. Nous le paramétrons (menu PARAMETERS) de manière à produire 6 clusters. Il n'y aucune contrainte concernant le positionnement relatif des groupes dans cette méthode. Nous ne modifions pas les autres paramètres. Par défaut, les variables sont également normalisées.



Nous validons et nous cliquons sur le menu contextuel VIEW pour accéder aux résultats.

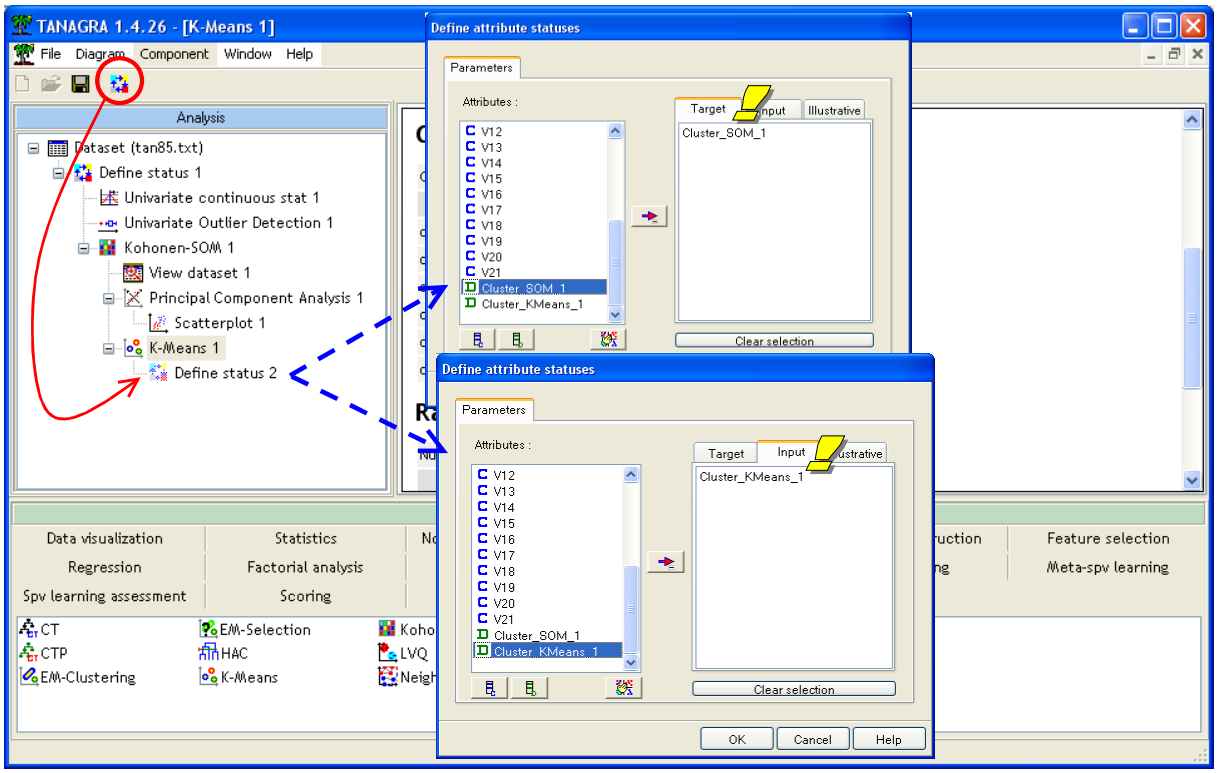


Premier résultat important, l'inertie expliquée par le regroupement est de **46.31%**, à rapprocher avec les 45.16% de SOM. Les performances sont comparables. Malgré l'absence de contrainte sur le positionnement des groupes, la méthode des K-Means ne fait guère mieux.

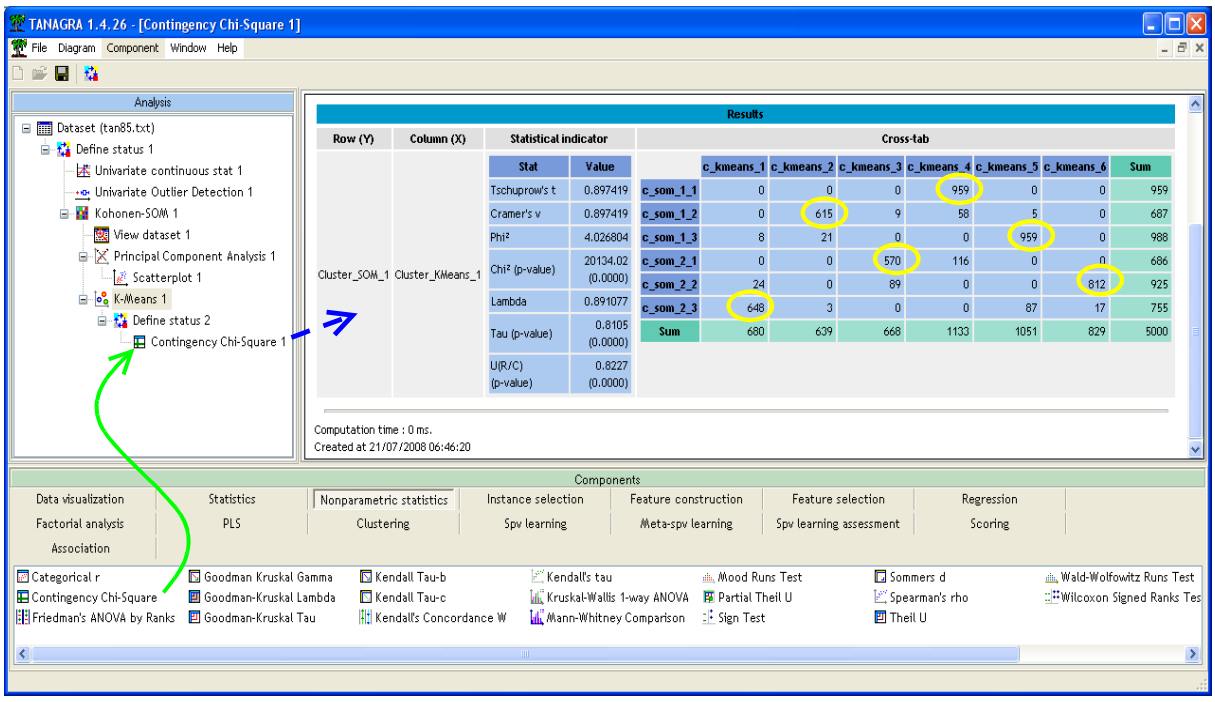
5.2 Correspondance entre les groupes

Les performances sont très proches. Mais qu'en est-il des groupes ? Est-ce qu'il y a une correspondance entre les clusters produits par les deux approches ?

Pour le savoir, nous allons croiser les deux variables, indicatrices des groupes d'appartenance, produites automatiquement par les deux composants. Nous introduisons un nouveau DEFINE STATUS dans le diagramme, à la suite de K-MEANS. Nous plaçons en TARGET la variable produite par SOM (CLUSTER_SOM_1), en INPUT celle fournie par les K-MEANS (CLUSTER_KMEANS_1).



Nous introduisons alors le composant tableau de contingence (CONTINGENCY CHI-SQUARE, onglet NONPARAMETRIC STATISTICS). Nous cliquons sur VIEW.



Nous constatons que les deux approches produisent finalement une classification très similaire. La relation est quasi univoque. Dans le tableau suivant nous montrons les correspondances :

Cluster	Cluster SOM	Cluster K-MEANS
A	(1 ; 1)	4
B	(1 ; 2)	2
C	(1 ; 3)	5
D	(2 ; 1)	3
E	(2 ; 2)	6
F	(2 ; 3)	1

6 Classification mixte

Le choix du nombre adéquat des groupes est un problème récurrent de la classification automatique. C'est peut être aussi un débat sans fin. Une approche raisonnable est d'y associer les connaissances du domaine et les objectifs de l'analyse.

Par rapport aux méthodes de réallocation telles que les K-MEANS ou SOM, la classification ascendante hiérarchique (CAH) propose, avec le dendrogramme, une aide à la détection de la « bonne » solution. En effet, le partitionnement est hiérarchisé. L'utilisateur a la possibilité d'élaborer des scénarios qu'il peut valider *a posteriori* en interprétant les groupes.

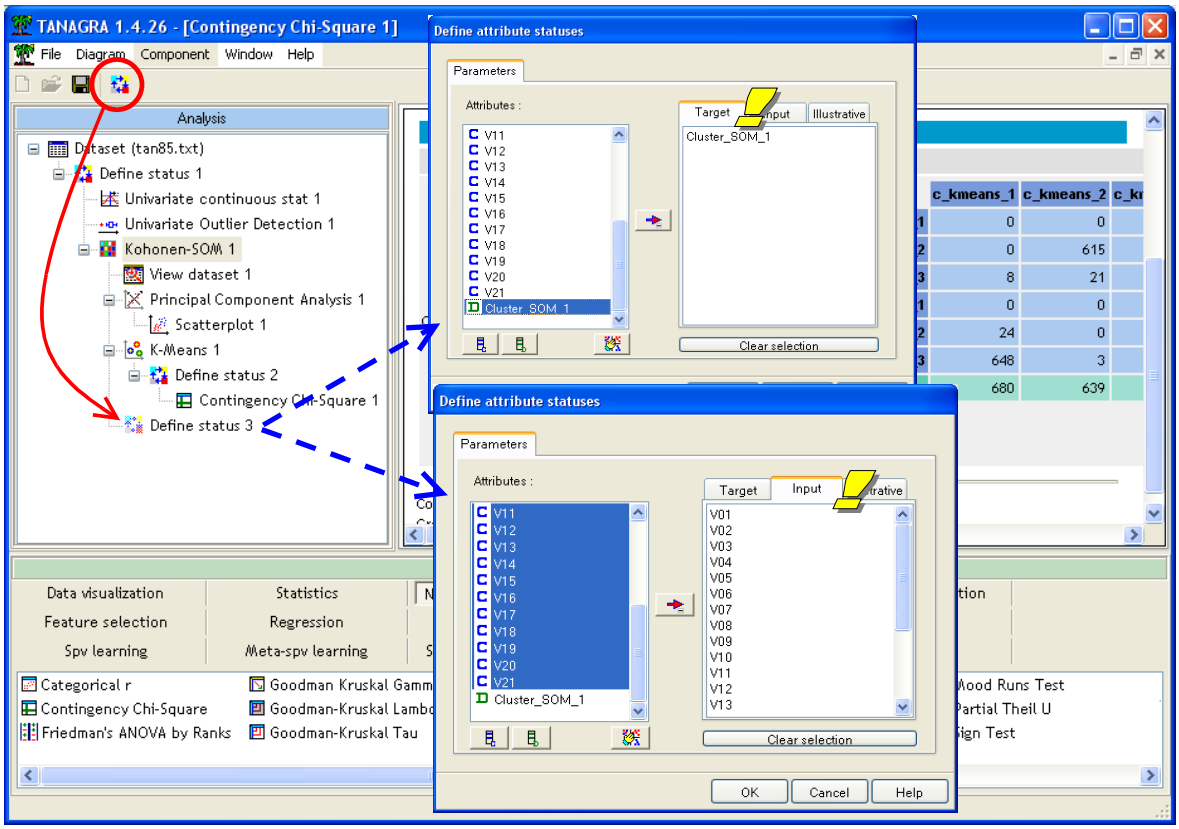
Cet avantage est contrebalancé par une limitation rédhibitoire. En effet, il est illusoire de vouloir lancer directement une CAH sur un grand nombre d'observations. Même si les capacités des machines augmentent constamment, le temps de calcul devient prohibitif dès que les effectifs dépassent le millier d'observations.

Une solution simple est la classification mixte (Lebart et al., 2000). Grosso modo il s'agit de combiner les méthodes de réallocations et les méthodes hiérarchiques de la manière suivante[8] : les K-MEANS servent à produire un nombre de groupes relativement élevé, la CAH prend comme point de départ ces « pré clusters » pour construire le dendrogramme, propice à la détection de la solution la plus appropriée.

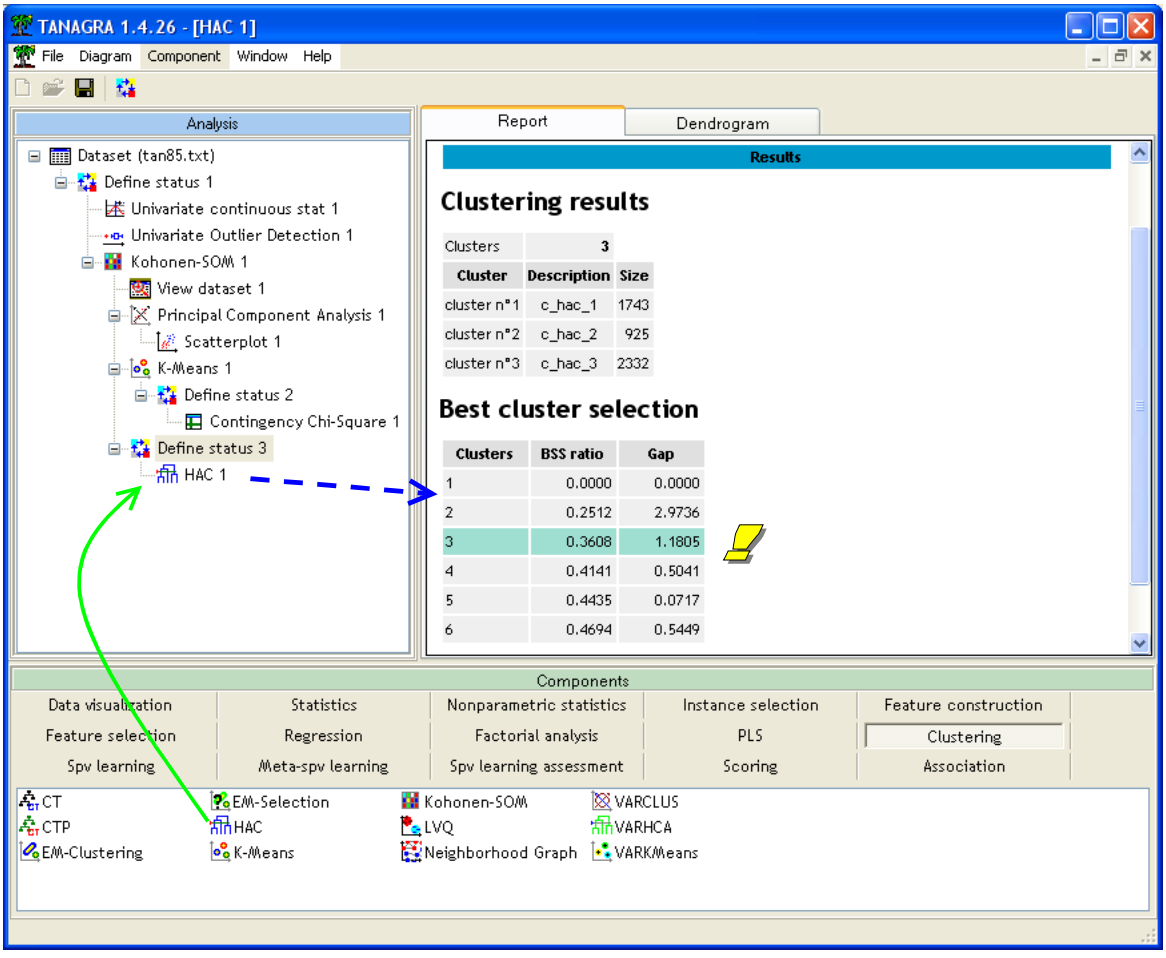
Les cartes de Kohonen présentent un atout dans ce contexte : la méthode permet de construire les « pré clusters », à l'instar des autres techniques, mais de plus, ces premiers regroupements correspondent à des zones adjacentes dans l'espace de représentation. L'interprétation du dendrogramme est renforcée.

Dans notre diagramme, nous insérons un nouveau DEFINE STATUS à la suite de KOHONEN-SOM. Nous plaçons en TARGET la variable générée automatiquement (CLUSTER_SOM_1), en INPUT les variables descriptives (V1 à V21).

[8] La méthode intègre d'autres subtilités que nous ne développons pas ici.

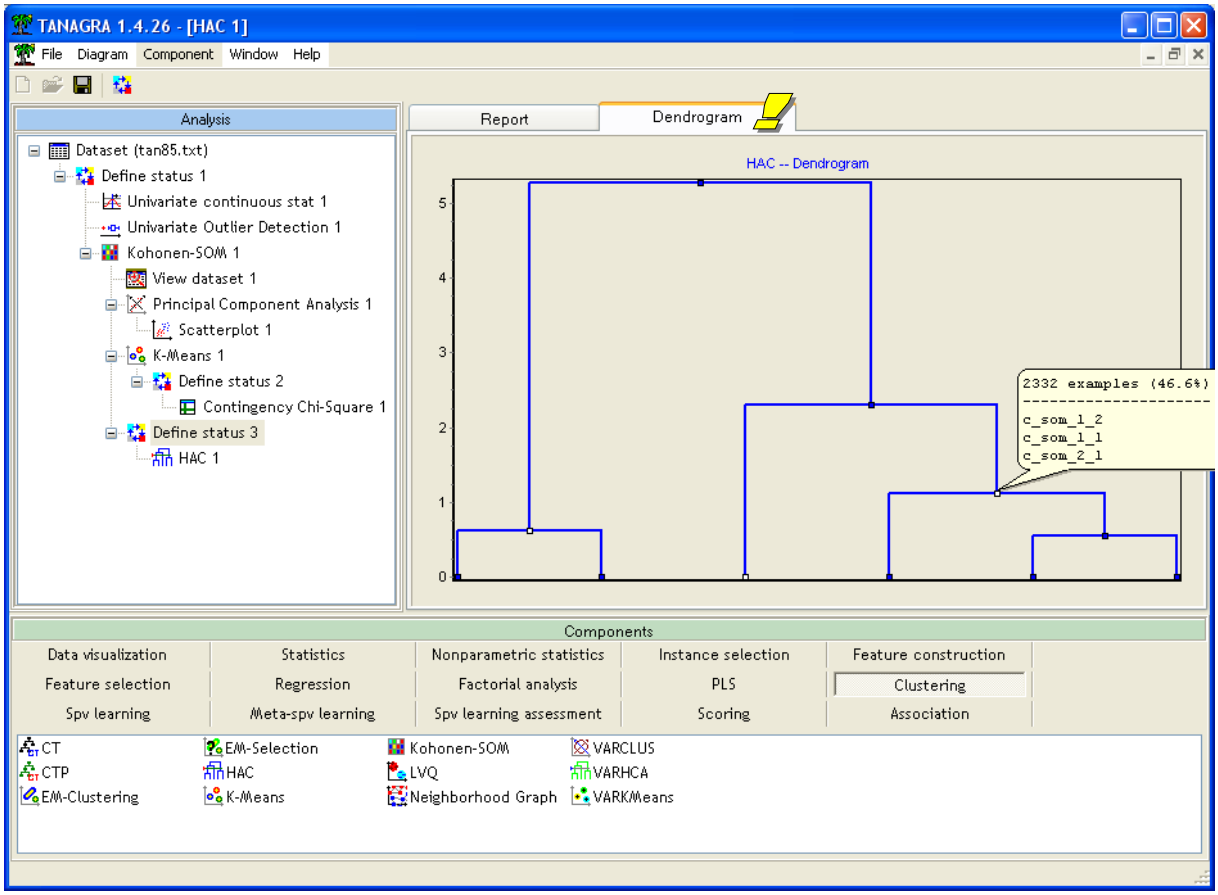


Nous introduisons alors le composant HAC (onglet CLUSTERING).



Le composant propose automatiquement une partition en 3 groupes. La détection est basée sur l'écart maximum entre deux paliers du dendrogramme. Il ne faut pas prendre pour argent comptant ce résultat.

Plus intéressant pour notre part est l'onglet DENDROGRAM de la fenêtre de visualisation. Nous visualisons le dendrogramme et, surtout, nous avons la possibilité en cliquant sur les nœuds de retrouver les « pré clusters » proposés par l'algorithme SOM.



Les sommets en blanc correspondent à la partition proposée. En cliquant sur le sommet blanc à droite, nous que les cases (1 ; 1), (1 ; 2) et (2 ; 1) ont été réunies pour former un cluster dans la partition en 3 classes. Dans le tableau ci-dessous nous montrons les correspondances entre le regroupement proposé par la CAH et les « pré clusters » initialement proposés par SOM.

Cluster CAH	Cluster SOM
1	(1 ; 3) + (2 ; 3)
2	(2 ; 2)
3	(1 ; 1) + (1 ; 2) + (2 ; 1)

Les résultats sont admirables de cohérence : la CAH a réuni en priorité les classes situées dans des cases adjacentes dans la carte de Kohonen.