Subject

Using variable clustering components in TANAGRA.

Variable clustering can be viewed like a clustering of the individuals where we would have transposed the dataset. But, instead of the utilization of the euclidean distance in order to compute the similarities between examples, we use the correlation coefficient (or the squared correlation coefficient).

Variable clustering may be useful in several situations. It can be used in order to detect the main dimensionality in the dataset; it may be used also in a feature selection process, in order to select the most relevant attributes for the subsequent analysis. The synthesized variable which represents a group, the main factor of PCA, may be used also.

Variable clustering around latent components. In the 1.4.16 TANAGRA version, we add some methods about variable clustering. They rely on the same principle: the group is depicted by the first factor of PCA (Principal Components Analysis). It is a weighted average of the variables that explains as much variance as possible. The famous SAS procedure is one of the illustrations of this approach[1]. Our implementation is based on the Vigneau and Qannari's works (2003)[2].

Only methods based on the squared correlation coefficients are available. Thus, for the groups' interpretation, we obtain the usual interpretation of the latent factors i.e. in a group, we can have positive or negative correlation on the factor.

Such as in the cases clustering, we implement several techniques. A bottom-up approach, named VARHCA, which is a hierarchical agglomerative approach. VARKMEANS which rely on the K-MEANS algorithm. The users specify the right number of clusters, the algorithm detects the configuration of greatest possible distinction.

A top down method is also available. It is suggested by the famous VARCLUS procedure. But, the iterative reassignment phase is ignored. When a variable is assigned to a group during the splitting process, we do not check if it is more correlated to other groups. Thus, the hierarchical structure of the tree is not maintained, the graphical representation, i.e. the tree, displays only the succession of the operations.

[1] J.P. Nakache et J. Confais, « Approche Pragmatique de la Classification », TECHNIP, 2005, chapter 9, pages 219 to 239, is our main reference. It is in French. Other reference is the SAS version 8,0 on line documentation (chapter 68). It is available here: 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.

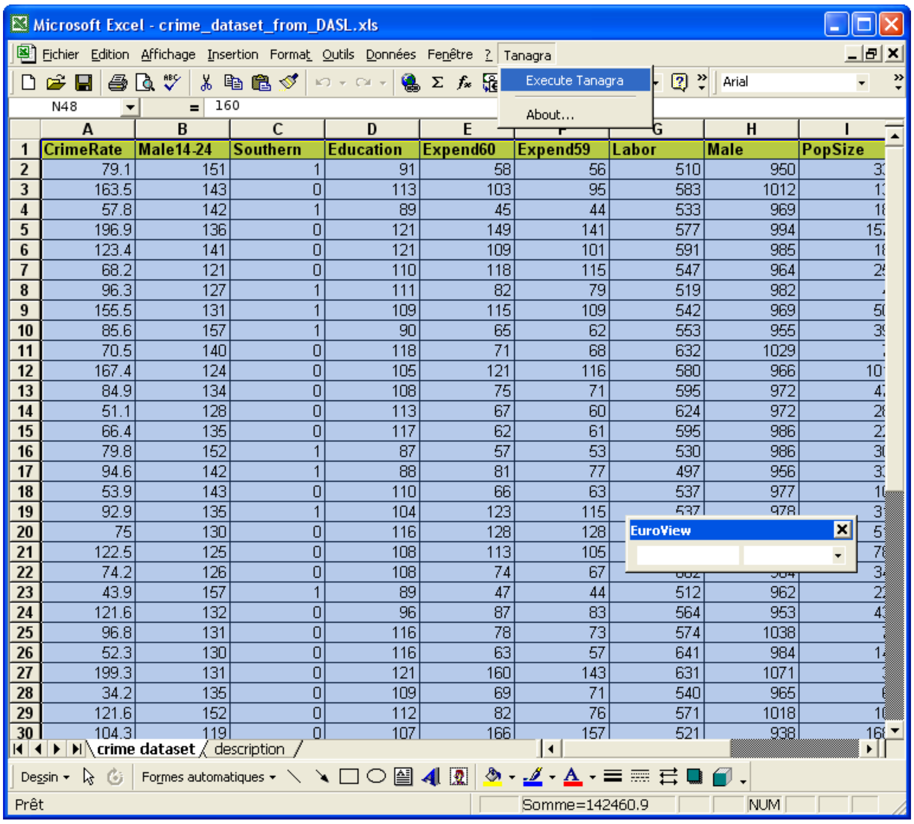
Dataset

In this tutorial, we use the CRIME_DATASET_FROM_DASL.XLS dataset from the DASL library (The Data Story and Library -- http://lib.stat.cmu.edu/DASL/). It depicts the criminality in US states in the 1960's. There are 47 examples and 14 variables in the file.

Clustering Variables with TANAGRA

Creating a new diagram

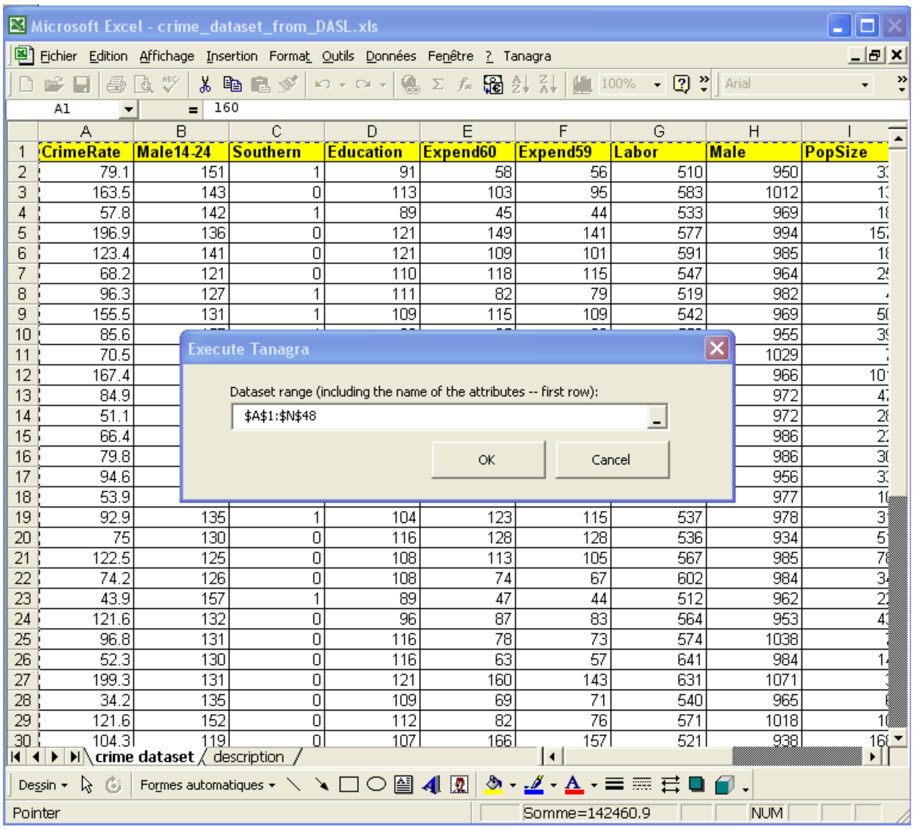
The simplest way to make an analysis with TANAGRA is to open the dataset in a spreadsheet software such as EXCEL. With the add-in TANAGRA.XLA[3], a new menu appears and we can start the process by selecting the whole dataset and clicking on the menu TANAGRA / EXECUTE TANAGRA.


Microsoft Excel - crime_dataset_from_DASL.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Tanagra
Execute Tanagra
About...
Arial
N48
= 160
EuroView
crime dataset
description
Dessin
Formes automatiques
Prêt
Somme=142460.9
NUM

	A	B	C	D	E	F	G	H	I
1	CrimeRate	Male14-24	Southern	Education	Expend60	Expend59	Labor	Male	PopSize
2	79.1	151	1	91	58	56	510	950	3
3	163.5	143	0	113	103	95	583	1012	1
4	57.8	142	1	89	45	44	533	969	1
5	196.9	136	0	121	149	141	577	994	15
6	123.4	141	0	121	109	101	591	985	1
7	68.2	121	0	110	118	115	547	964	2
8	96.3	127	1	111	82	79	519	982	4
9	155.5	131	1	109	115	109	542	969	5
10	85.6	157	1	90	65	62	553	955	3
11	70.5	140	0	118	71	68	632	1029	7
12	167.4	124	0	105	121	116	580	966	10
13	84.9	134	0	108	75	71	595	972	4
14	51.1	128	0	113	67	60	624	972	2
15	66.4	135	0	117	62	61	595	986	2
16	79.8	152	1	87	57	53	530	986	3
17	94.6	142	1	88	81	77	497	956	3
18	53.9	143	0	110	66	63	537	977	1
19	92.9	135	1	104	123	115	537	978	3
20	75	130	0	116	128	128	[illegible]	[illegible]	5
21	122.5	125	0	108	113	105	[illegible]	[illegible]	7
22	74.2	126	0	108	74	67	[illegible]	[illegible]	3
23	43.9	157	1	89	47	44	512	962	2
24	121.6	132	0	96	87	83	564	953	4
25	96.8	131	0	116	78	73	574	1038	7
26	52.3	130	0	116	63	57	641	984	1
27	199.3	131	0	121	160	143	631	1071	3
28	34.2	135	0	109	69	71	540	965	6
29	121.6	152	0	112	82	76	571	1018	1
30	104.3	119	0	107	166	157	521	938	16

In the following dialog box, we check if the selection is right and we validate the process.

[3] See the tutorials about the installation of add-in on the website if it are not installed and activated. The add-in is available since version 1.4.11 of TANAGRA.



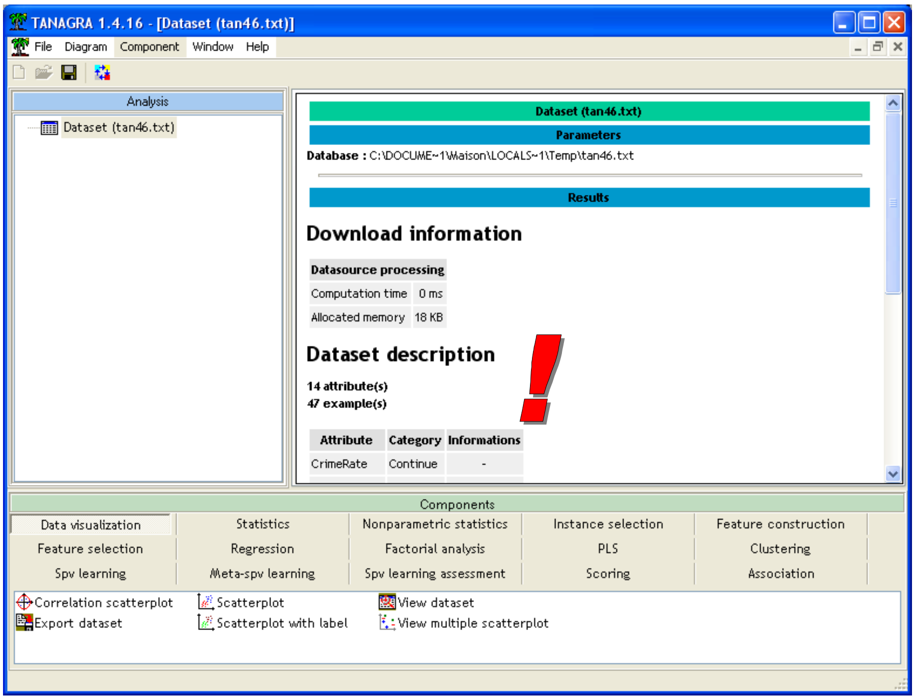
Microsoft Excel - crime_dataset_from_DASL.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Tanagra
A1 = 160

	A	B	C	D	E	F	G	H	I
1	CrimeRate	Male14-24	Southern	Education	Expend60	Expend59	Labor	Male	PopSize
2	79.1	151	1	91	58	56	510	950	33
3	163.5	143	0	113	103	95	583	1012	13
4	57.8	142	1	89	45	44	533	969	18
5	196.9	136	0	121	149	141	577	994	157
6	123.4	141	0	121	109	101	591	985	18
7	68.2	121	0	110	118	115	547	964	25
8	96.3	127	1	111	82	79	519	982	4
9	155.5	131	1	109	115	109	542	969	50
10	85.6							955	39
11	70.5							1029	7
12	167.4							966	101
13	84.9							972	47
14	51.1							972	28
15	66.4							986	22
16	79.8							986	30
17	94.6							956	33
18	53.9							977	10
19	92.9	135	1	104	123	115	537	978	31
20	75	130	0	116	128	128	536	934	51
21	122.5	125	0	108	113	105	567	985	78
22	74.2	126	0	108	74	67	602	984	34
23	43.9	157	1	89	47	44	512	962	22
24	121.6	132	0	96	87	83	564	953	43
25	96.8	131	0	116	78	73	574	1038	7
26	52.3	130	0	116	63	57	641	984	14
27	199.3	131	0	121	160	143	631	1071	3
28	34.2	135	0	109	69	71	540	965	6
29	121.6	152	0	112	82	76	571	1018	10
30	104.3	119	0	107	166	157	521	938	168

Execute Tanagra

Dataset range (including the name of the attributes -- first row): $A$1:$N$48

TANAGRA is automatically opened, the dataset is downloaded. We check again if we have really 14 variables and 47 examples.



TANAGRA 1.4.16 - [Dataset (tan46.txt)]
File Diagram Component Window Help
Analysis
Dataset (tan46.txt)
!

Dataset (tan46.txt)

Parameters

Database : C:\DOCUME~1\Maison\LOCALS~1\Temp\tan46.txt

Results

Download information

Datasource processing	
Computation time	0 ms
Allocated memory	18 KB

Dataset description

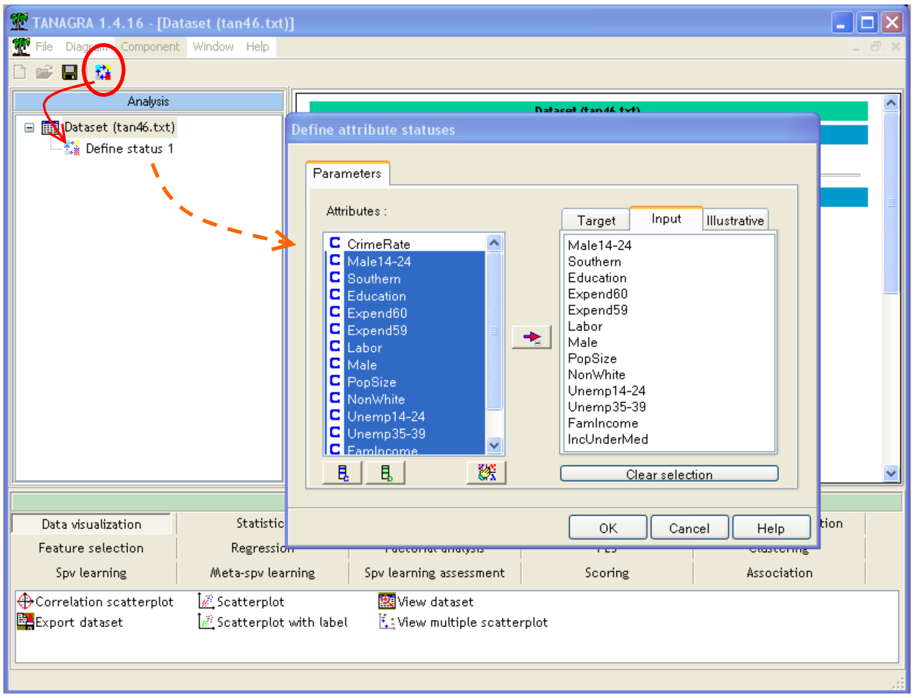
14 attribute(s)
47 example(s)

Attribute	Category	Informations
CrimeRate	Continue	-

VARHCA

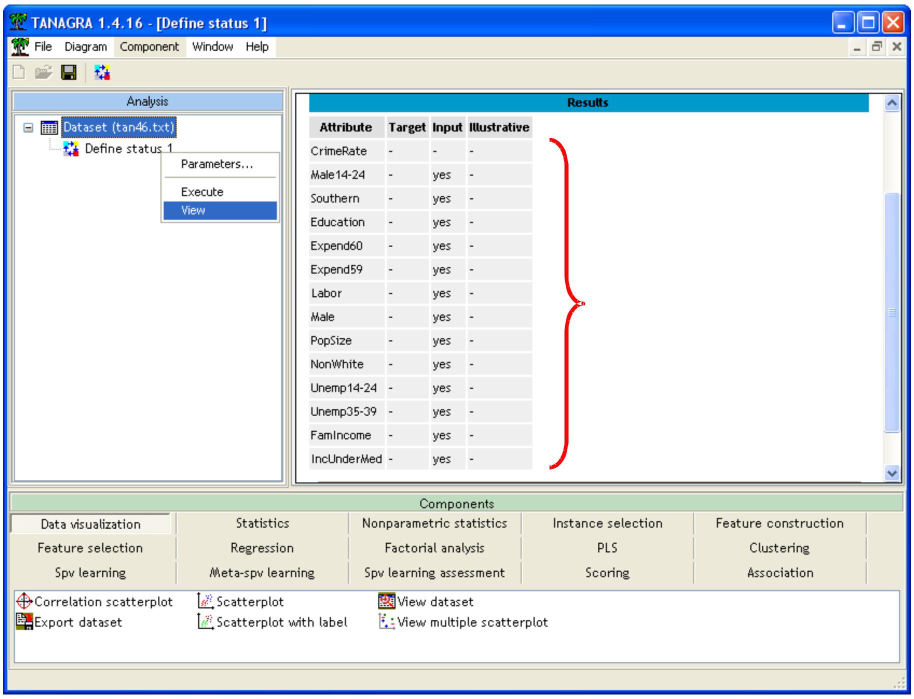
Defining the analysis

Of course, we must specify first the variables of the analysis. We insert the DEFINE STATUS component in the diagram by using the shortcut on the toolbar.



Input:
Male14-24
Southern
Education
Expend60
Expend59
Labor
Male
PopSize
NonWhite
Unemp14-24
Unemp35-39
FamIncome
IncUnderMed

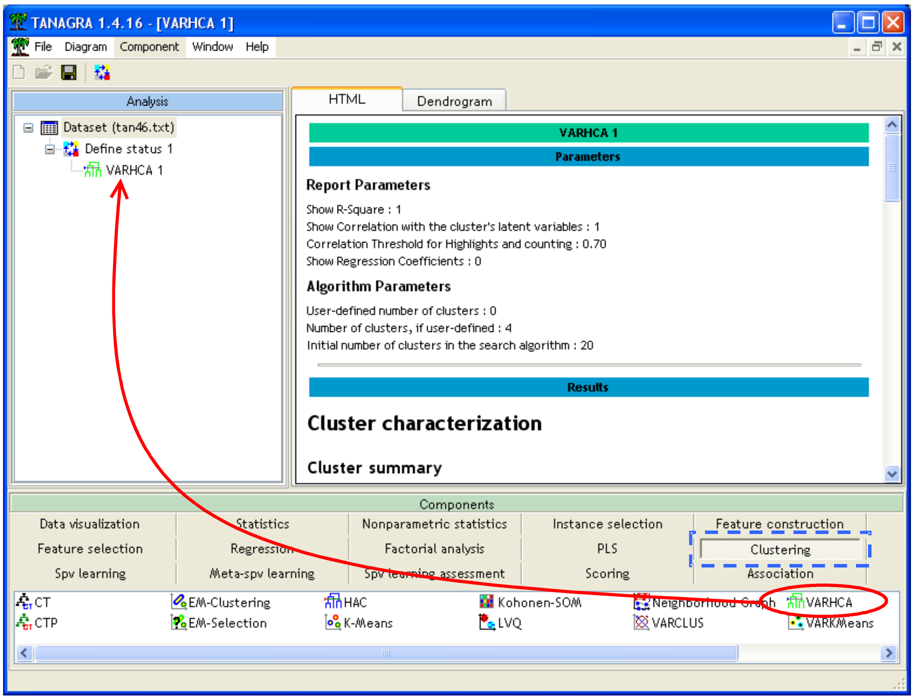
We set as INPUT all the variables, except the variable CRIME RATE which has a particular status in this dataset, we will use it later.



Results

Attribute	Target	Input	Illustrative
CrimeRate	-	-	-
Male14-24	-	yes	-
Southern	-	yes	-
Education	-	yes	-
Expend60	-	yes	-
Expend59	-	yes	-
Labor	-	yes	-
Male	-	yes	-
PopSize	-	yes	-
NonWhite	-	yes	-
Unemp14-24	-	yes	-
Unemp35-39	-	yes	-
FamIncome	-	yes	-
IncUnderMed	-	yes	-

Then, we add the VARHCA component (CLUSTERING tab) in the diagram, by drag-and-drop operation. To view the results, we activate the contextual menu VIEW.



VARHCA 1

Parameters

Report Parameters

Show R-Square : 1
Show Correlation with the cluster's latent variables : 1
Correlation Threshold for Highlights and counting : 0.70
Show Regression Coefficients : 0

Algorithm Parameters

User-defined number of clusters : 0
Number of clusters, if user-defined : 4
Initial number of clusters in the search algorithm : 20

Results

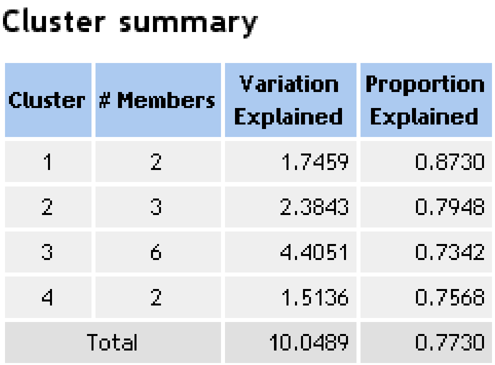
Cluster characterization

Cluster summary

Reading the results

For a good understanding of the results, we refer to the SAS (version 8.0) on-line documentation, see http://www2.stat.unibo.it/ManualiSas/stat/chap68.pdf

Cluster Summary. It displays the number of clusters, the number of variables in each cluster, the variation explained in clusters and the total variation explained (value and proportion).

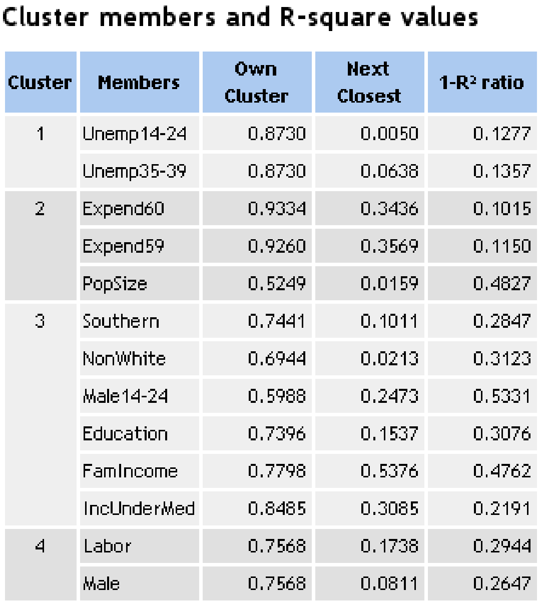


Cluster summary

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

Cluster members and R-square values.

It details the variables in each cluster. Three indicators are available: R² with own cluster, R² with the nearest cluster, and the 1-R² ratio. Small value of this ratio indicates good clustering. If this value is larger to 1, it means that the variable has a larger correlation with another cluster than its group.

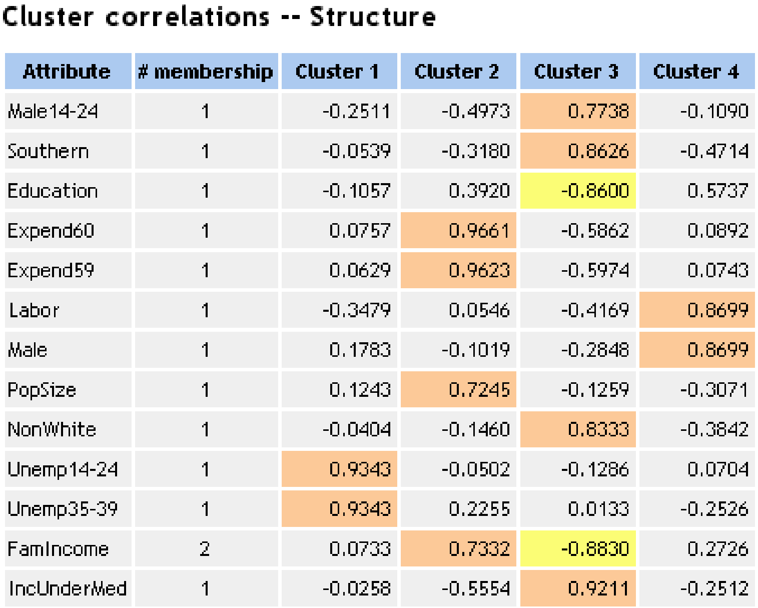


Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² 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

In our dataset, VARHCA detects 4 groups. The variables seems well assigned to their groups. The largest 1-R² ratio is 0.533 for MALE14-24 attribute in the third cluster.

Cluster correlations - Structure. This table displays the cluster structure i.e. the correlation of each variable to clusters. It enables to interpret the clusters. We underline the correlations higher than 0.7 or lower than -0.7 (this threshold can be modified) and we count these situations in the MEMBERS column. If the variables are well clustered, each variable must be associated to one and only one cluster.



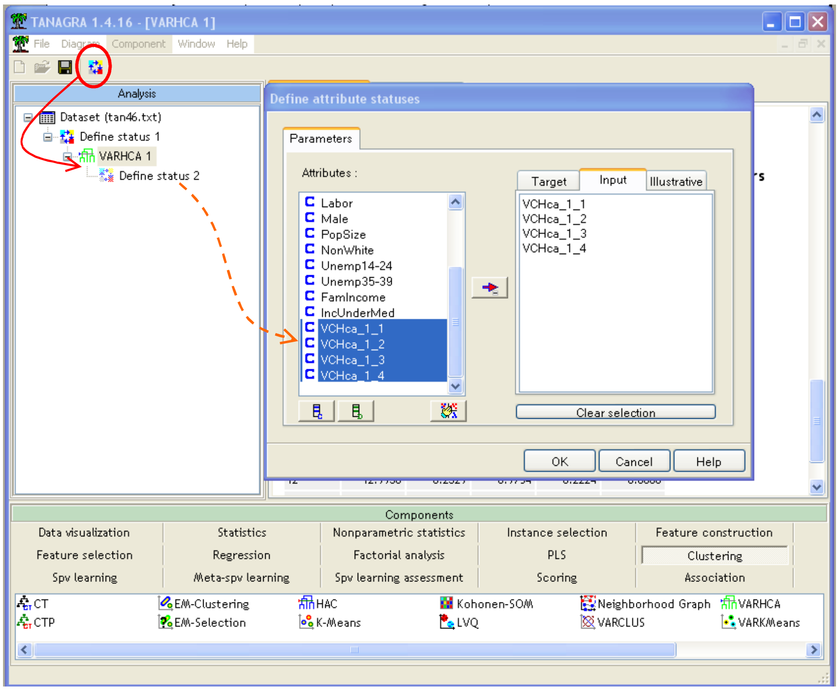
Cluster correlations -- Structure

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

In our dataset, we obtain a good association between variables and their cluster. Only FAMINCOME seems highly correlated to the second and the third cluster. We see in the next subsection that these clusters are themselves correlated.

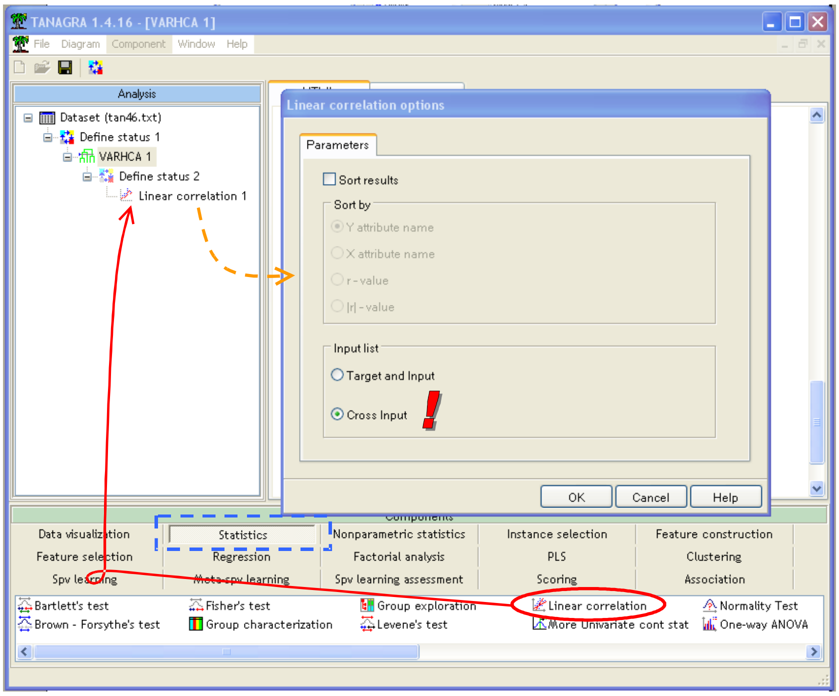
Inter-cluster correlations. An another way to evaluate the orthogonality between the clusters is to compute the correlation between the components. They represent the clusters, they correspond to the first factor of the PCA in each group. TANAGRA automatically computes these factors. They are available for the subsequent operators in the diagram.

We add again the DEFINE STATUS component into the diagram, using the shortcut. We set as INPUT these new variables.



Input:
VCHca_1_1
VCHca_1_2
VCHca_1_3
VCHca_1_4

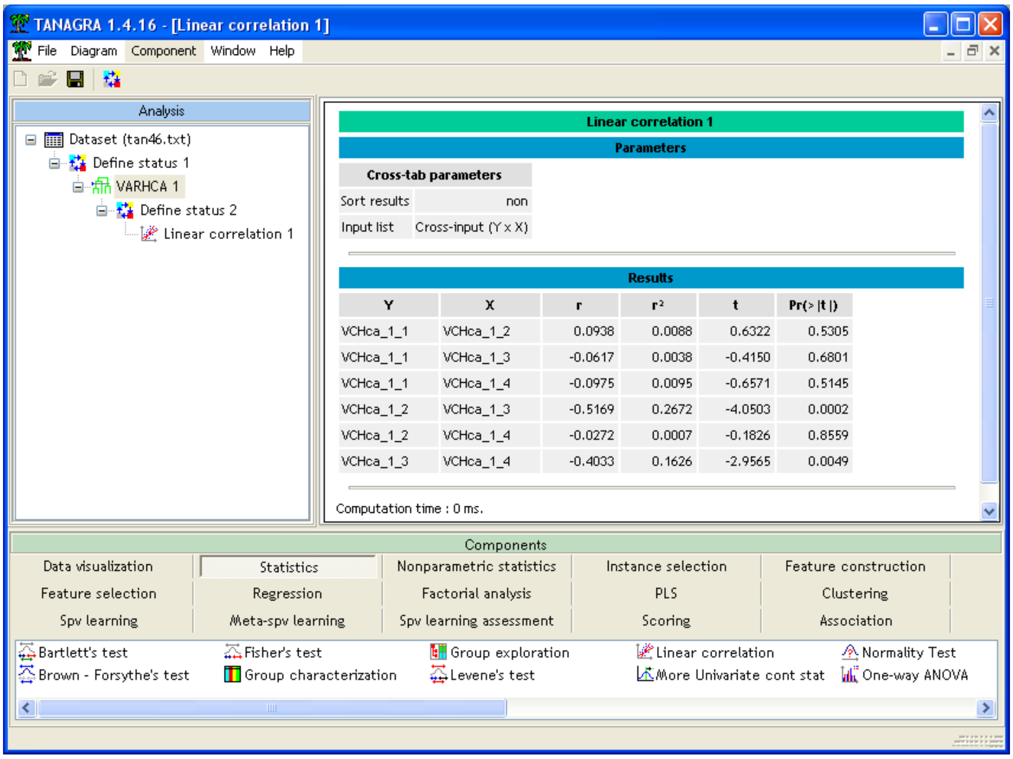
Then we add the LINEAR CORRELATION component (STATISTICS tab) into the diagram. We modify the default settings, we want to compute the correlations between the INPUT attributes.



□ Sort results

Input list: ● Cross Input

We note that, except the cluster 2 and 3, the components are poorly correlated.



Linear correlation 1

Parameters

Cross-tab parameters	
Sort results	non
Input list	Cross-input (Y x X)

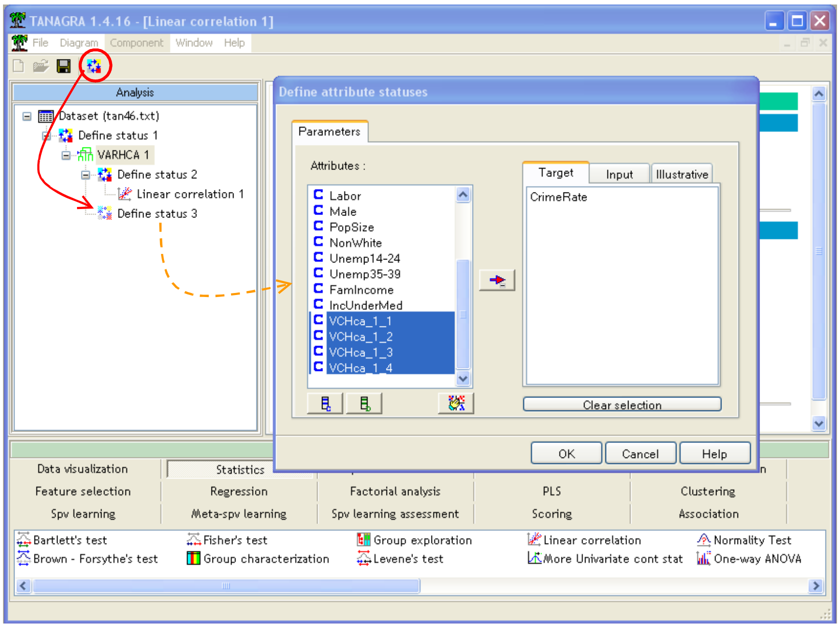
Results

Y	X	r	r²	t	Pr(>|t|)
VCHca_1_1	VCHca_1_2	0.0938	0.0088	0.6322	0.5305
VCHca_1_1	VCHca_1_3	-0.0617	0.0038	-0.4150	0.6801
VCHca_1_1	VCHca_1_4	-0.0975	0.0095	-0.6571	0.5145
VCHca_1_2	VCHca_1_3	-0.5169	0.2672	-4.0503	0.0002
VCHca_1_2	VCHca_1_4	-0.0272	0.0007	-0.1826	0.8559
VCHca_1_3	VCHca_1_4	-0.4033	0.1626	-2.9565	0.0049

Computation time : 0 ms.

Using supplementary variables. Now we want to insert the CRIME RATE in the analysis. We want to know which dimension gives the best explanation of this variable.

We add a new DEFINE STATUS component. We set as TARGET the CRIME RATE attribute, and as INPUT the factors of the clusters.

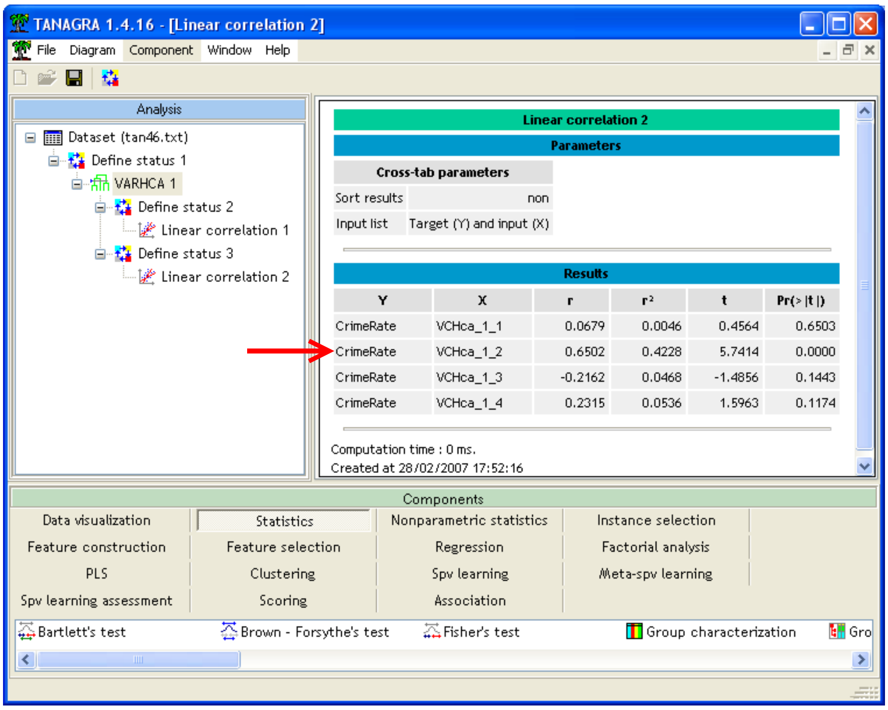


Attributes :
- C Labor
- C Male
- C PopSize
- C NonWhite
- C Unemp14-24
- C Unemp35-39
- C FamIncome
- C IncUnderMed
- C VCHca_1_1
- C VCHca_1_2
- C VCHca_1_3
- C VCHca_1_4

Target | Input | Illustrative

CrimeRate

Then we insert component LINEAR CORRELATION again. We do not modify the default settings. We see that the CRIME RATE variable is mainly correlated to the second cluster.



Linear correlation 2

Parameters

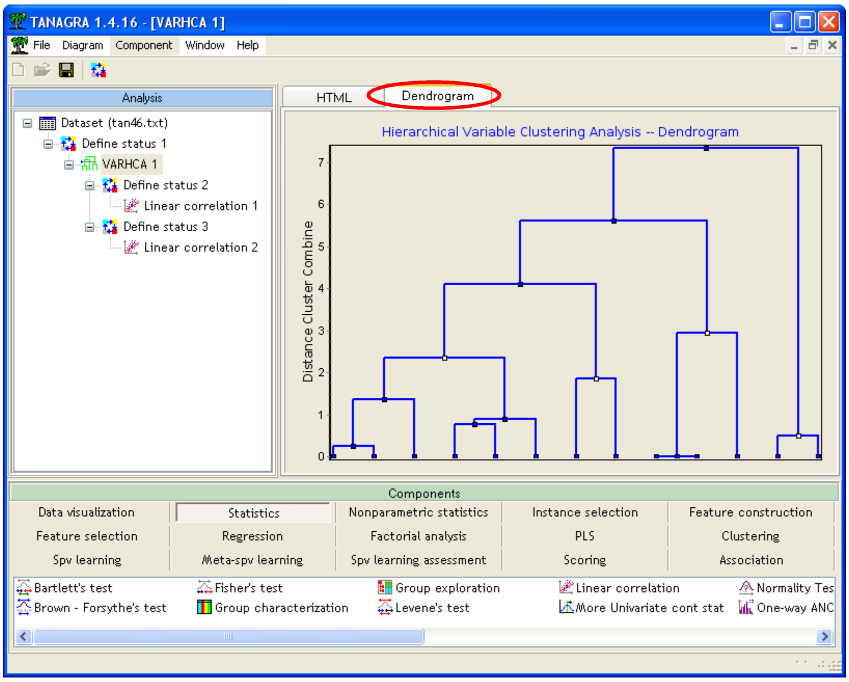
Cross-tab parameters	
Sort results	non
Input list	Target (Y) and input (X)

Results

Y	X	r	r²	t	Pr(>|t|)
CrimeRate	VCHca_1_1	0.0679	0.0046	0.4564	0.6503
CrimeRate	VCHca_1_2	0.6502	0.4228	5.7414	0.0000
CrimeRate	VCHca_1_3	-0.2162	0.0468	-1.4856	0.1443
CrimeRate	VCHca_1_4	0.2315	0.0536	1.5963	0.1174

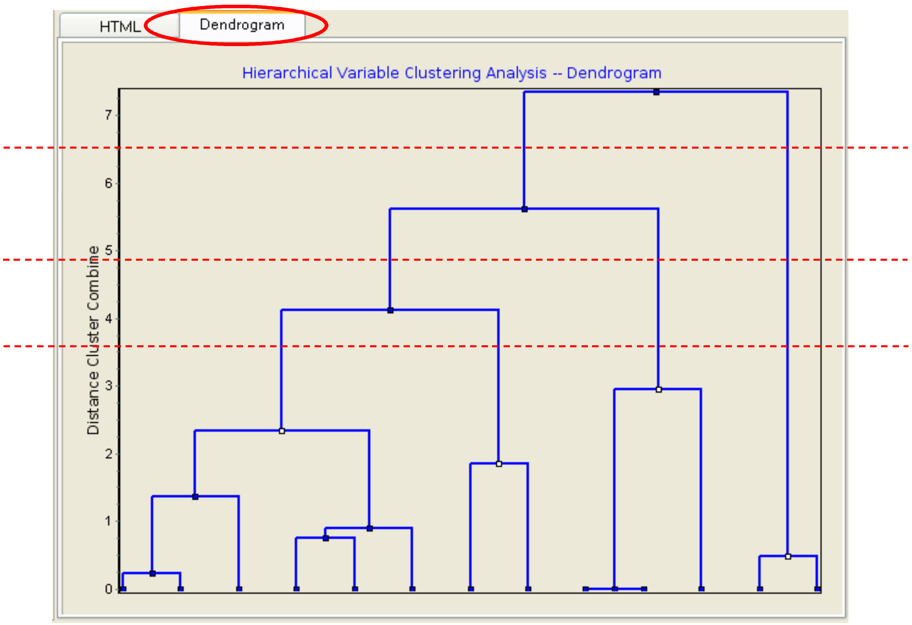
Computation time : 0 ms.
Created at 28/02/2007 17:52:16

Dendrogram. Such as in the cases clustering, we can produce a tree diagram of the cluster structure. We can see at each node the variables which are merged, the distance between the two groups, and the aggregating index.

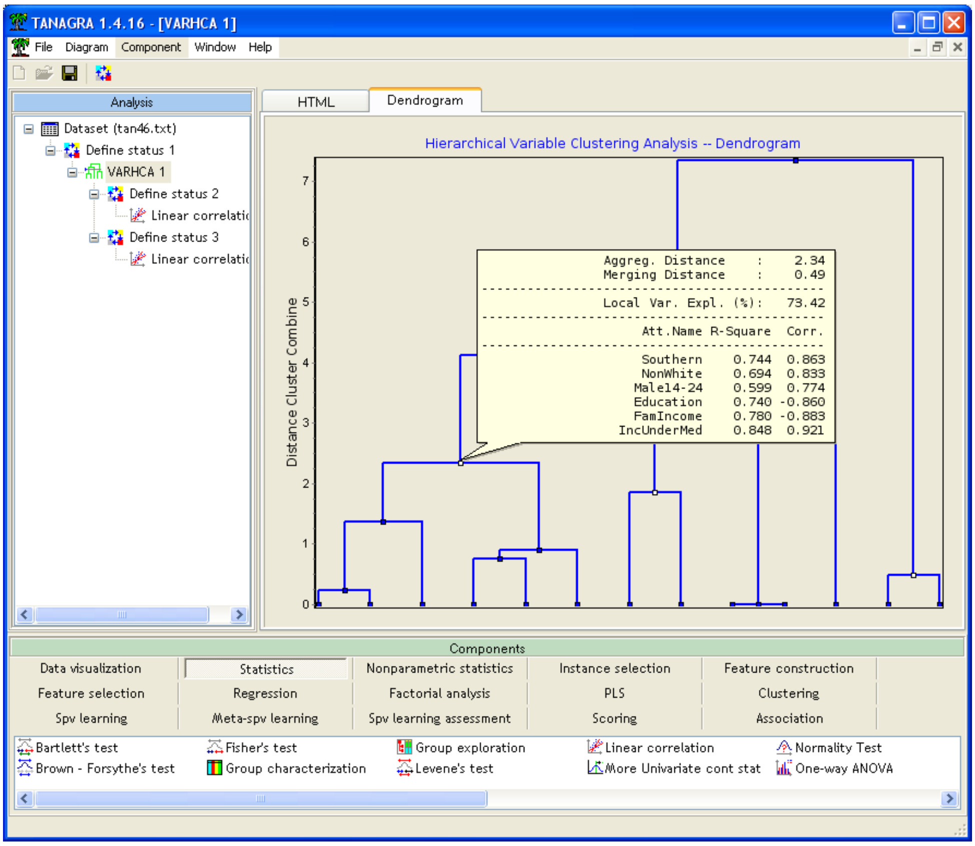


A low-level merging distance means that the merged groups are rather similar. Using the distance between each stage of the dendrogram, we intuitively choose 4, 3 or 2 groups here.

Note that the subdivision into two groups is often not relevant. The merging distance is mechanically large because it is the first partitioning of the variables.



By clicking on the nodes of the tree, we can visualize the list of variables. We can thus follow the aggregating process.



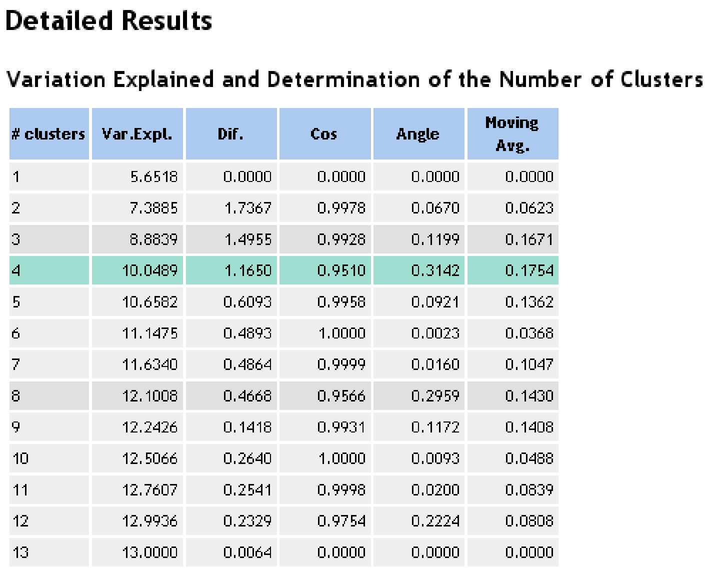
```
        Aggreg. Distance    :    2.34
        Merging Distance    :    0.49
  --------------------------------------
        Local Var. Expl. (%):   73.42
  --------------------------------------
           Att.Name R-Square  Corr.
  --------------------------------------
           Southern    0.744  0.863
           NonWhite    0.694  0.833
          Male14-24    0.599  0.774
          Education    0.740 -0.860
          FamIncome    0.780 -0.883
        IncUnderMed    0.848  0.921
```

The selected groups correspond to white colored nodes.

Selecting the optimal number of clusters. TANAGRA selects 4 clusters in the VARHCA approach. How it proceeds ?

The detection of the right number of clusters is an opened problem. I think the software must only propose the most probable right number of clusters. Searching the “knee” in the explained variation curve is a very popular approach. TANAGRA produces a table (DETAILED RESULTS

Section) with the explained variation for each number of clusters, the angle between the semi tangents at each point (in fact, the angle between the two interpolation lines). The right number of clusters would be associated to area with a large value of this angle. The best solution is in green, and the two following one are gray.

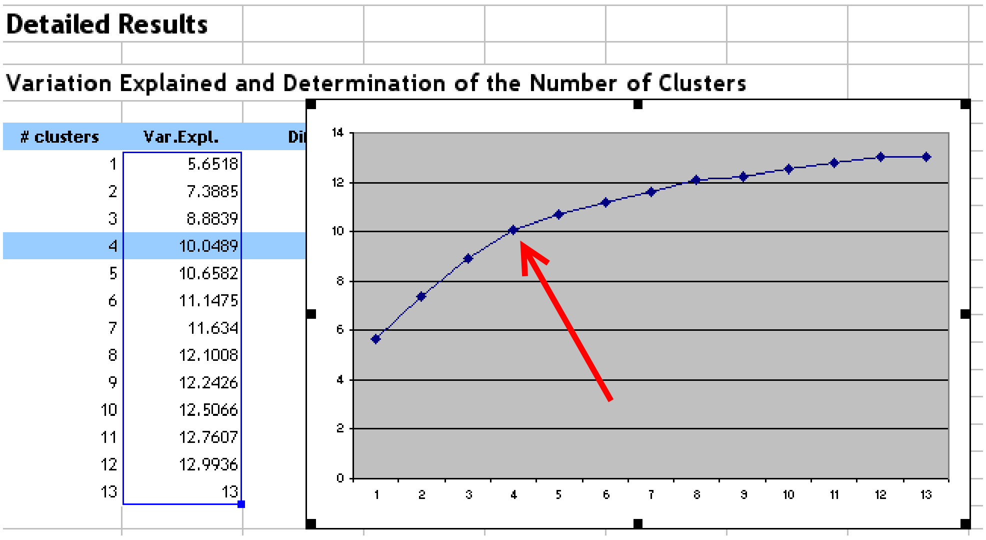


Detailed Results

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

If we copy and paste the results in a spreadsheet, we can draw the curve, it seems the proposed solutions are appropriate for this dataset. In order to avoid undesirable local variation, we compute the moving average of the values on 3 points. The criterion is better smoothed.

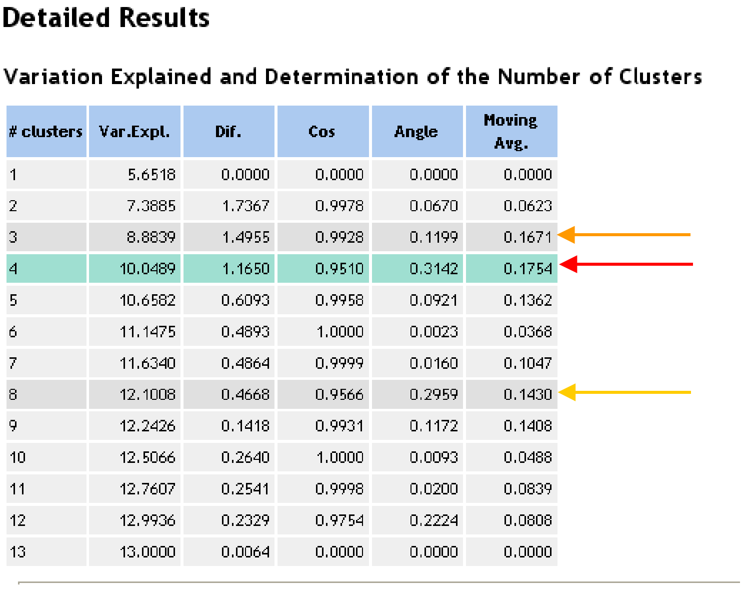


Detailed Results

Variation Explained and Determination of the Number of Clusters

# clusters	Var.Expl.	Di
1	5.6518	
2	7.3885	
3	8.8839	
4	10.0489	
5	10.6582	
6	11.1475	
7	11.634	
8	12.1008	
9	12.2426	
10	12.5066	
11	12.7607	
12	12.9936	
13	13	

According to TANAGRA, the proposed optimal number of clusters are 4, 3 and 8. The user can check these solutions with the domain knowledge.

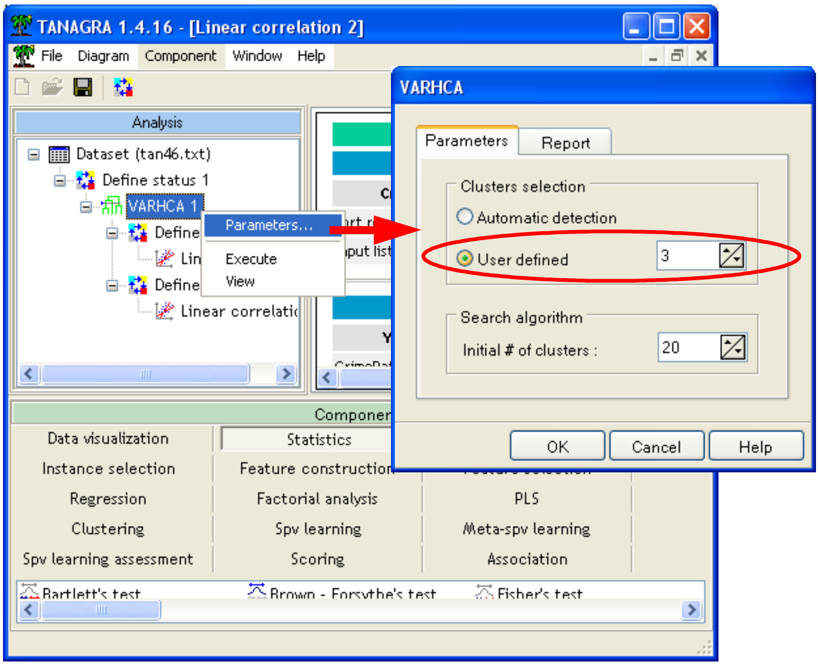


Detailed Results

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

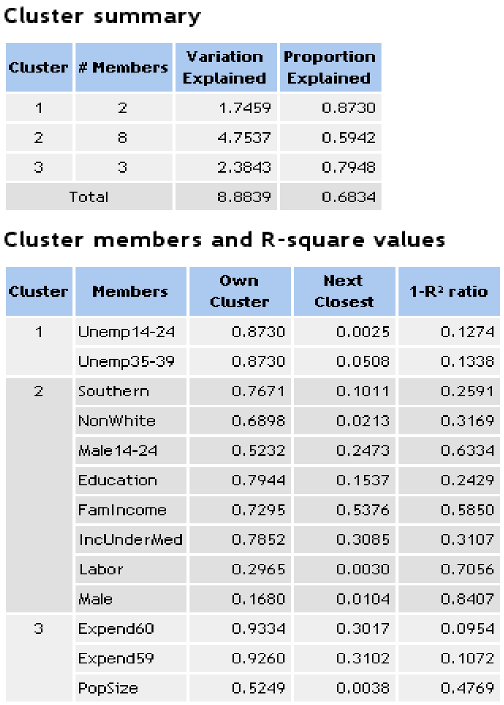
User-defined number of clusters. In some situations, the user want to set the desired number of cluster. To do that, we click on the PARAMETERS menu and modify the cluster selection option. We insert the right number of clusters.



Clusters selection
○ Automatic detection
● User defined: 3

Search algorithm
Initial # of clusters: 20

We obtain the following results.



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-R^2 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

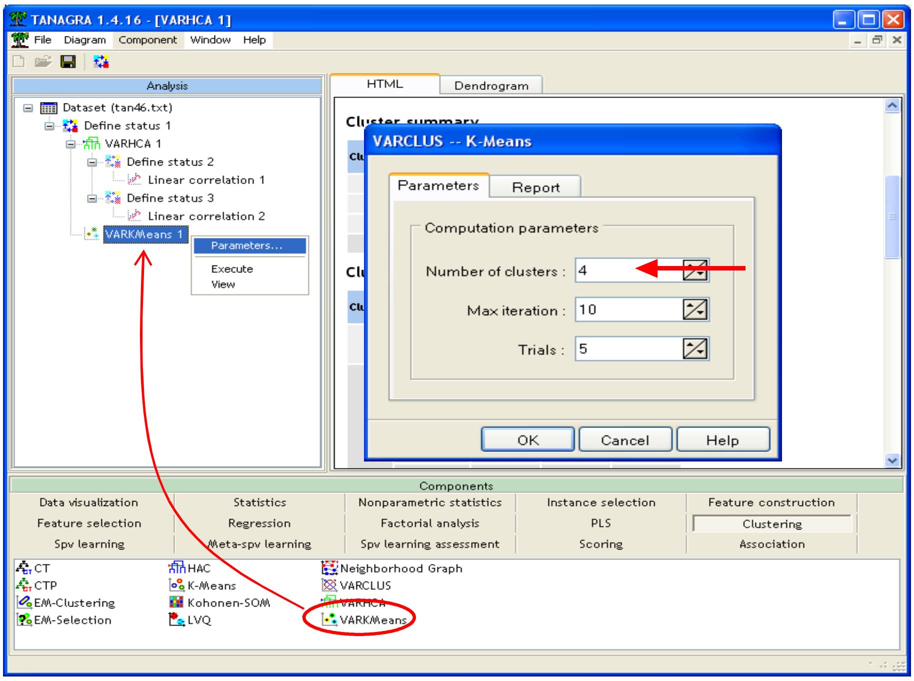
VARKMEANS

VARKMEANS is a variable clustering approach based on the K-MEANS framework. The number of clusters is fixed a priori. The "centroids" of the clusters is the first factor of PCA in our algorithm. In the first pass, the groups are randomly built. Then, we associate each variable to the nearest component. The "centroids" are refreshed and we check all the variables again. The algorithm stopped when there is no modifications of the clusters' structure.

VARKMEANS produce the same report as the other variable clustering methods.

We add the VARKMEANS (CLUSTERING tab) after the DEFINE STATUS 1 component into the diagram. We activate the PARAMETERS menu in order to specify the number of clusters. Other parameters about the optimization process, the maximum number of iterations and the number of trials can be also modified.

We click on the VIEW menu, the clustering process is started and the results are available in a new window.



VARCLUS -- K-Means

Parameters | Report

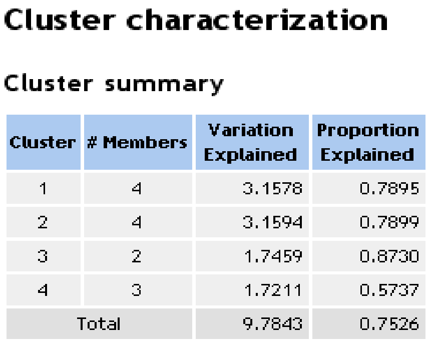
Computation parameters

Number of clusters : 4

Max iteration : 10

Trials : 5

We note that we obtain similar results to VARHCA approach. The variation explained is a little bit small but the difference is not really significant.



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

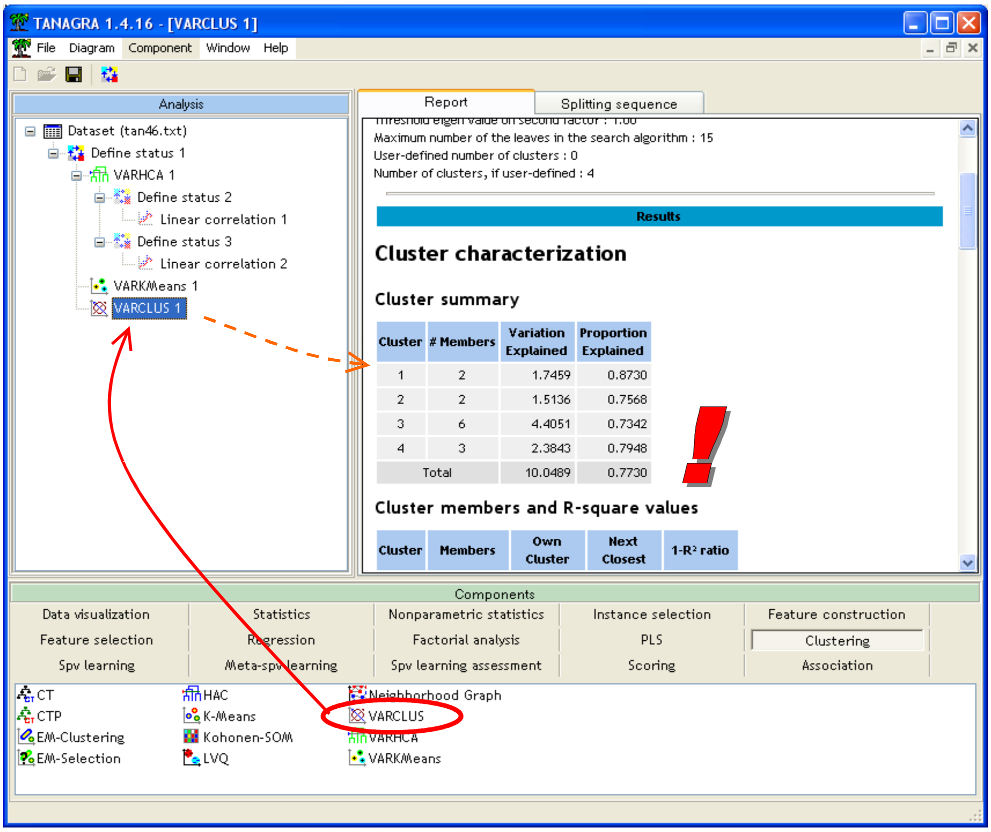
VARCLUS

VARCLUS is a top down approach which is similar to the SAS procedure. A detailed description of the original approach is available here (http://www2.stat.unibo.it/ManualiSas/stat/chap68.pdf).

In comparison to the previous approaches (VARHCA and VARKMEANS), this method is much more fast when the number of variables is large. The process stops when there is no relevant subdivision of the groups.

Unlike the original procedure, we do not proceed to reassignment at each stage of the algorithm, this operation being very (too much) time consuming. The hierarchical structure is thus not preserved. For this reason, *the tree does not correspond to a dendrogram, it details only the succession of the splitting operations*.

We add the VARCLUS component after DEFINE STATUS 1 into the diagram. We click on the VIEW menu. The results are available with the standard presentation mode.



Threshold eigen value on second factor : 1.00
Maximum number of the leaves in the search algorithm : 15
User-defined number of clusters : 0
Number of clusters, if user-defined : 4

Results

Cluster characterization

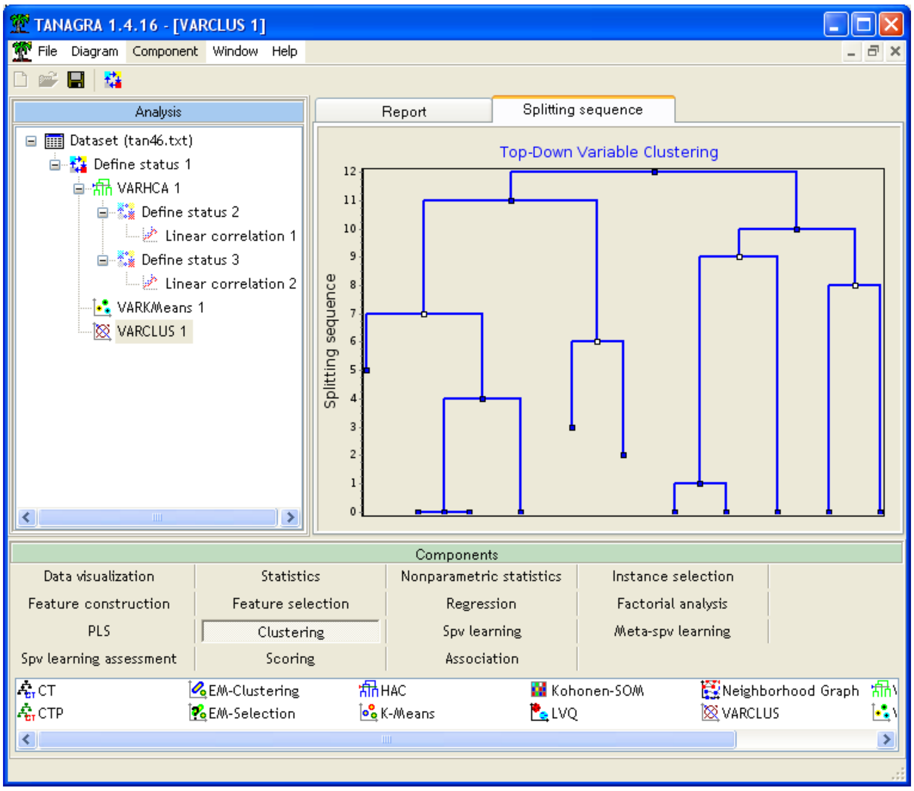
Cluster summary

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

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² ratio

We obtain the same groups as VARHCA. In the SLIPPTING SEQUENCE tab, we can follow the successive subdivisions of the variables. By clicking on the nodes, we can observe the local list of the variables and their correlation with the latent factor.



N.B. : Be careful again, the level of the nodes does not correspond to aggregation distance here.

Conclusion

Variable clustering is useful in various situations. It enables to summarize quickly the main dimensions in a large dataset. It can be used also as a variable-reduction technique. The interpretation of the groups is as easy as the interpretation of the factors of Principal Components Analysis.