I modelli di analisi statistica multidimensionale dei dati:

La Cluster Analysis Gerarchica

Obiettivi dell'unità didattica

- Comprendere l'insieme delle procedure che si prefiggono di raggruppare individui in classi tali che gli individui all'interno di una classe siano molto simili e che ogni classe sia relativamente distinta dalle altre.
- Acquisire le informazioni fondamentali sui principali metodi di classificazione gerarchica
- Imparare ad utilizzare Tanagra per effettuare una CA gerachica

4.6 La Cluster Analysis Gerarchica

- **Criteri di classificazione**
- **Procedure per l'individuazione dei gruppi di elementi**
- **Il Dendrogramma**
- **Esempio: il dataset Automobili**

Ambiti di applicazione della CA

L'obiettivo della Cluster Analysis (CA) è la ricerca della partizione dell'insieme E in K (con K < n e intero) sottoinsiemi (gruppi o cluster) tali che le unità appartenenti ad uno stesso sottoinsieme siano il più possibile omogenei tra loro.

Gli ambiti di applicazione sono molteplici, come, per esempio:

- Segmentazione di mercato (riferita ad esempio a tipi di prodotto, a tipi di consumatori, etc.)
- Classificazione dei comuni di una regione in gruppi omogenei in base a pluralità di indicatori demografici, economici, sociali

I criteri di raggruppamento

Gerarchici

Prevedono l'individuazione di n partizioni ciascuna caratterizzata da un diverso numero di k di gruppi (k = 1,....,n); le partizioni individuate costituiscono una struttura gerarchica di raggruppamento.

Consentono di ottenere una famiglia di partizioni con numero di gruppi da *n* a 1, partendo da quella banale con numero di gruppi = *n* per giungere a quella, anch'essa banale, in cui tutti gli elementi sono riuniti in un unico gruppo.

Non Gerarchici

Prevedono l'individuazione di una sola partizione delle unità in k gruppi, dove k puó essere fissato a-priori o derivare dal processo di raggruppamento stesso.

Forniscono un'unica partizione delle *n* unità in *g* gruppi, con *g* fissato a priori.

La Cluster Gerarchica

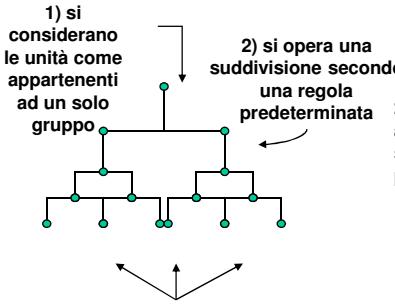
In un metodo di partizione gerarchica:

- Si considerano tutti i livelli di distanza
- I gruppi che si ottengono ad ogni livello di distanza comprendono i gruppi ottenuti ai livelli inferiori
- Quando due o più unità si uniscono (o si dividono) non possono più essere divise (o unite) nei passi successivi

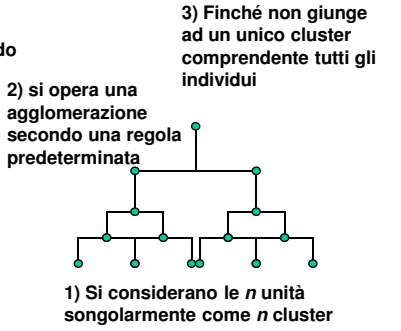
I criteri di classificazione

I criteri di classificazione possono essere:

SCISSORI



AGGLOMERATIVI



Tipologia di algoritmi di classificazione

Ricapitolando:

- Direttamente alle partizioni ⇒ **Classificazione non gerarchica**
- Alla costruzione di classi per aggregazioni successive di coppie di oggetti ⇒ **Classificazione gerarchica ascendente**
- Alla costruzione di classi per dicotomizzazioni successive dell'insieme degli oggetti ⇒ **Classificazione gerarchica discendente**

Le possibili partizioni

Problema ⇨ **Numero delle partizioni possibili**

Es.: 4 elementi (A,B,C,D) e 2 gruppi

(A) (B,C,D) (B) (A,C,D) (C) (A,B,D) (D) (A,B,C) (A,B) (C,D) (A,C) (B,D) (A,D) (B,C)

Numero delle partizioni (P) ⇨ $2^{n-1} - 1$

n=4 ⇨ P = 7

n=10 ⇨ P = 511

n=100 ⇨ P = 1,000,000,000,000,000,000,000,000,000,00 - 1

$= 10^{29}-1$

Procedura per l'individuazione dei gruppi di elementi

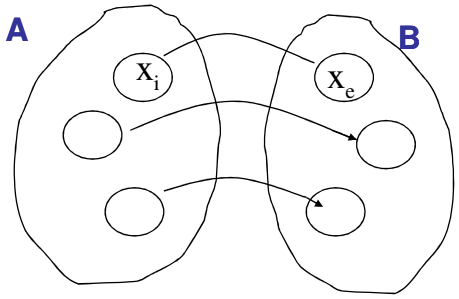
Con impiego iniziale di una matrice delle distanze (o di una matrice di indici di dissimilarità) D:

1. Si individuano in D le due unità con minor distanza (più simili) e si riuniscono a formare il primo gruppo
2. Si ricalcola la distanza del gruppo ottenuto dagli altri gruppi ricavando una nuova matrice delle distanze D1 con dimensioni diminuite di uno
3. Si individua in D1 la coppia di unità (o gruppi) con minore distanza e la si riunisce
4. Si ripetono le fasi 2 e 3 fino a che tutte le unità si riuniscono in un unico gruppo

Metodo del legame singolo

Metodo del legame completo (*complete linkage*) o del vicino più lontano (*furthest neighbour*)

$$d(A,B) = \min[d(X_i, X_e)] \quad x_i \in A, x_e \in B$$

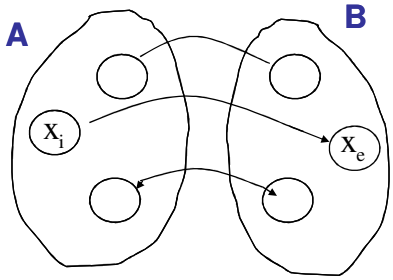


La distanza tra due gruppi è definita come il minimo delle distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo

Metodo del legame completo

Metodo del legame singolo (*single linkage*) o del vicino più prossimo (*nearest neighbour*)

$$d(A,B) = \max[d(X_i, X_e)]$$



$$x_i \in A, x_e \in B$$

La distanza tra due gruppi è definita come il massimo delle distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo

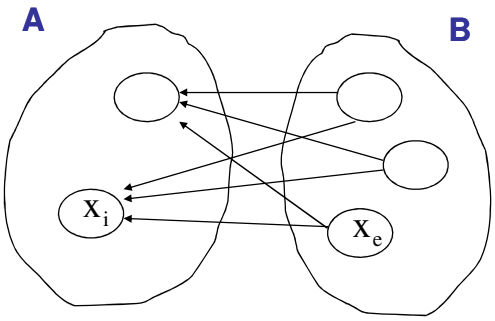
Metodo del legame medio

Metodo del legame medio (*average linkage*) tra i gruppi

$$d(A,B) = (1/n_A n_B)\Sigma d(x_i, x_e)$$

$x_i \in A$

$x_e \in B$



La distanza tra due gruppi è definita come la media aritmetica delle distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo

Esempio: Metodo del legame medio

Metodo del legame medio (*average linkage*) tra i gruppi: esempio

Primo gruppo (A): (1,2,3)
Secondo gruppo (B): (4,5)

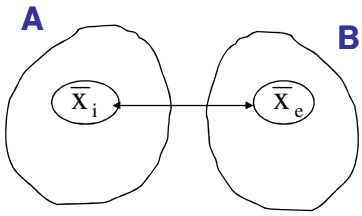
d(A,B) = media delle distanze tra le 6 coppie di unità:
(1,4); (2,4); (3,4); (1,5); (2,5); (3,5).

Metodo del centroide

Metodo del centroide

Utilizza anche la matrice dei dati, oltre a quella delle distanze

$$d(A,B) = d(\bar{x}_i, \bar{x}_e) \quad \begin{cases} \bar{X}_i = \text{centroide di } \mathbf{A} \\ \bar{X}_e = \text{centroide di } \mathbf{B} \end{cases}$$



La distanza tra due gruppi è definita come la distanza tra i rispettivi centroidi

Nel metodo del legame medio si considera la media delle distanze tra le unità dei due gruppi, nel metodo del centroide si individua prima un centro di ogni gruppo e poi si misura la distanza tra essi

Metodo di Ward

Metodo di Ward, o della devianza minima

Considera la decomposizione della Devianza Totale delle p variabili in Devianza Between e Devianza Within:

$$T = B + W$$

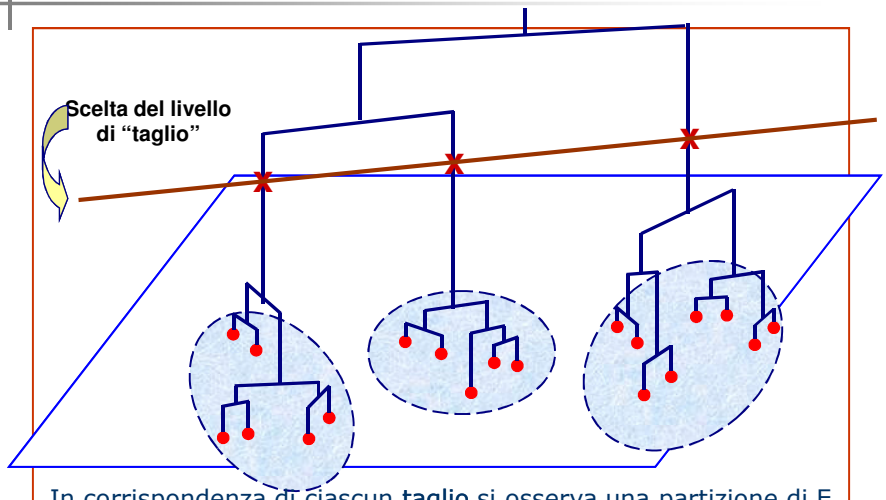
Ad ogni passo della procedura gerarchica si aggregano tra di loro i gruppi che comportano il minor incremento della Devianza Within, cioè che assicurano la maggior coesione interna possibile

Il grafico della CA

Gli algoritmi forniscono una gerarchia di partizioni che si presentano sotto forma di albero detto anche *Dendrogramma* e che contiene n-1 partizioni.

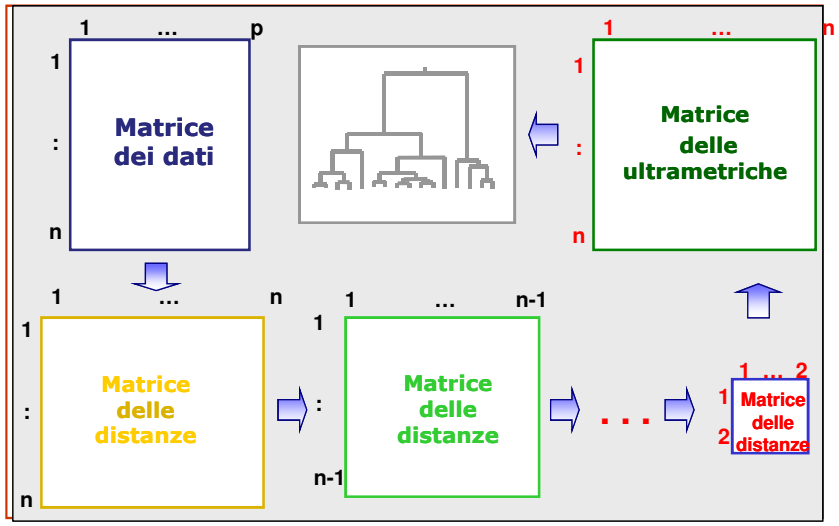
L'importanza della lettura del dendrogramma è nella possibilità di suggerire il numero di classi effettivamente presenti nell'insieme osservato.

Il Dendrogramma



In corrispondenza di ciascun **taglio** si osserva una partizione di E ed un determinato numero di k.

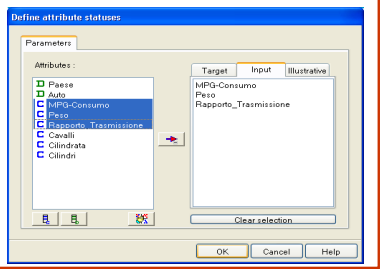
I passi di una classificazione gerarchica



Esempio: il dataset "Automobili"

Si apra il dataset automobili.xls e si importino i dati in Tanagra come mostrato nelle precedenti unit.

Una volta aperto il software, si clicchi sul simbolo di Define Status e, poi, con il tasto destro del mouse si selezionino i parametri come mostrato nella figura successiva.

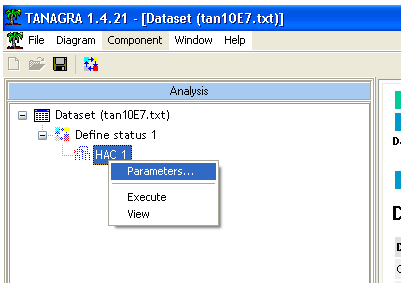


In pratica come descrittori (Input) del dataset automobili, indichiamo il Consumo, il Peso e il Rapporto di trasmissione.

Esempio: il comando HAC

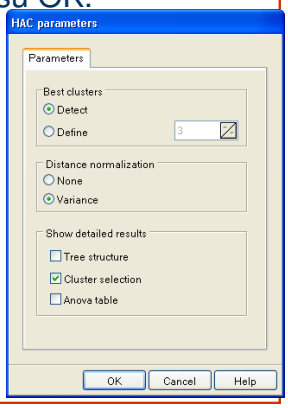
Per iniziare la CA gerarchica bisogna trascinare il comando *HAC (*che si trova in Clustering nel menù Components) nell'icona di Define Status.

Si clicchi, poi, con il tasto destro alla voce Parameters come mostrato nella figura successiva.



Esempio: HAC parameters

Nella finestra che si aprirà successivamente si lascino i comandi che appaiono di default (come mostrato nell'immagine), cliccando su OK.



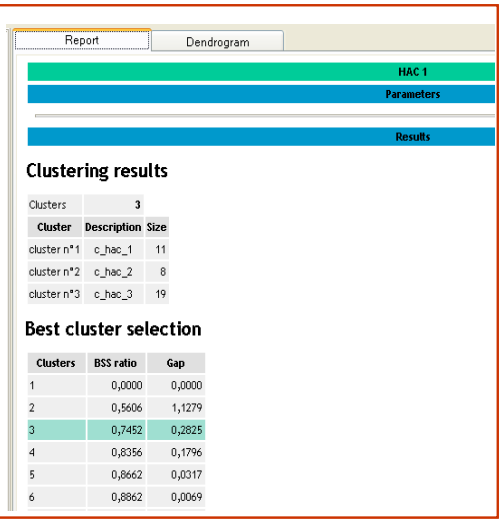
In particolare, non definiamo il numero di cluster a priori (in Best clusters) e chiediamo come report la sezione dei cluster (in Show detailed results).

A questo punto clicchiamo con il tasto destro su HAC e selezioniamo View per avere il report e il dendrogramma.

Esempio: il primo report

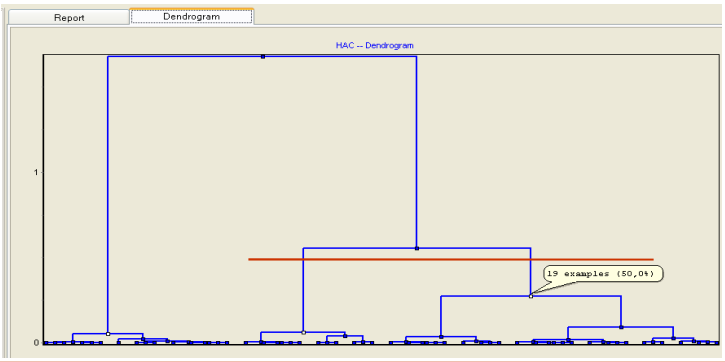
Tanagra individua automaticamente il "salto" maggiore all'interno del dendrogramma (nell'esempio riportato i cluster individuati sono 3).

Cliccando sulla voce Dendrogram appare il grafico.



Esempio: il dendrogramma

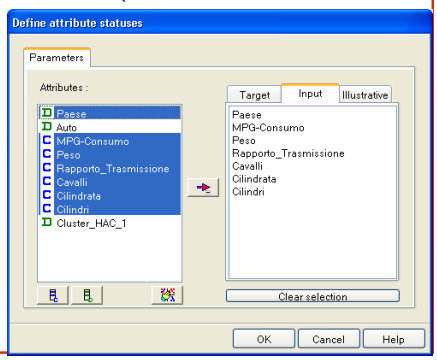
Cliccando nel punto in cui si incrociano i gruppi e si avrà la consistenza del nuovo gruppo in termini assoluti e relativi. La linea rossa indica il punto in cui sono stati individuati i tre cluster.



Esempio: Analisi dei cluster

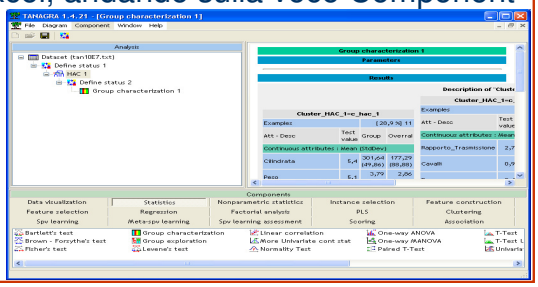
Per comprendere in che modo sono stati individuati i cluster, si clicchi sul simbolo di Define Status (che apparirà sotto il comando di HAC). Si indichino come Input tutte le variabili (eccetto Auto che rappresenta ovviamente le osservazioni).

Come Target si indichi il Cluster che si è venuto a formare precedentemente (Cluster_HAC_1).



Esempio: Group Characterization

Una volta definiti i parametri si aggiunga il comando *Group Characterization* che si trova sotto la voce Statistics in Components, quindi, tasto destro del mouse e cliccare su View. Per una migliore comprensione dei risultati, si suggerisce di copiare i risultati in un file Excel, andando sulla voce Component del menù in alto e, poi, si clicchi su Copy Results ed, infine, si incollino i risultati su un file Excel.



Esempio: il primo cluster

Il primo cluster (11 auto) è formato dai veicoli di grandi dimensioni. Essi sono pesanti, potenti, consumano molto e sono, soprattutto, auto statunitensi.

Cluster_HAC_1=c_hac_1			
Examples			[28,9 %] 11
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Cilindrata	5,4	301,64 (49,86)	177,29 (88,88)
Peso	5,1	3,79 (0,29)	2,86 (0,71)
Cilindri	5	7,45 (0,93)	5,39 (1,60)
Cavalli	4,3	130,82 (15,50)	101,74 (26,44)
MPG-Consumo	-4,1	17,88 (1,40)	24,76 (6,55)
Rapporto_Trasmissione	-4,4	2,50 (0,20)	3,09 (0,52)
Discrete attributes : [Recall] Accuracy			
Paese=U.S.A.	3,3	[50,0 %] 100,0 %	57,90%
Paese=Italia	-0,6	[0,0 %] 0,0 %	2,60%
Paese=Francia	-0,6	[0,0 %] 0,0 %	2,60%
Paese=Svezia	-0,9	[0,0 %] 0,0 %	5,30%
Paese=Germania	-1,5	[0,0 %] 0,0 %	13,20%
Paese=Giappone	-1,8	[0,0 %] 0,0 %	18,40%

Esempio: il secondo cluster

Il secondo cluster (8 auto) è formato dai veicoli di medie dimensioni. Essi sono mediamente potenti, sono caratterizzate da un'elevata manovrabilità e sono auto europee.

Cluster_HAC_1=c_hac_2			
Examples			[21,1 %] 8
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Rapporto_Trasmissione	2,7	3,53 (0,30)	3,09 (0,52)
Cavalli	0,9	109,63 (15,22)	101,74 (26,44)
Peso	0,4	2,95 (0,25)	2,86 (0,71)
Cilindri	0	5,38 (0,92)	5,39 (1,60)
Cilindrata	-0,9	152,00 (27,53)	177,29 (88,88)
MPG-Consumo	-2,2	20,16 (2,28)	24,76 (6,55)
Discrete attributes : [Recall] Accuracy			
Paese=Svezia	2,8	[100,0 %] 25,0 %	5,30%
Paese=Francia	1,9	[100,0 %] 12,5 %	2,60%
Paese=Germania	1,1	[40,0 %] 25,0 %	13,20%
Paese=Giappone	-0,5	[14,3 %] 12,5 %	18,40%
Paese=Italia	-0,5	[0,0 %] 0,0 %	2,60%
Paese=U.S.A.	-2,1	[9,1 %] 25,0 %	57,90%

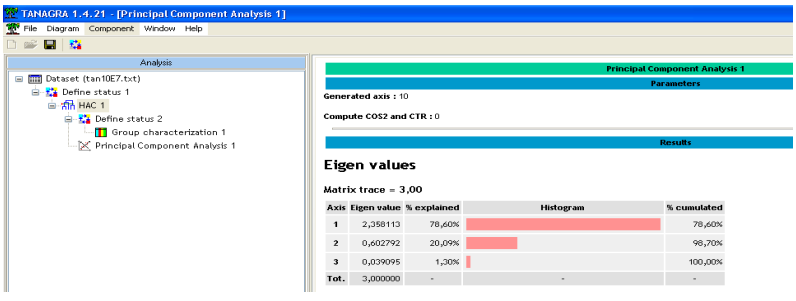
Esempio: il terzo cluster

Il terzo cluster (19 veicoli) è formato dai veicoli di piccole dimensioni. Essi sono caratterizzati da bassi consumi, potenza non alta e basso peso.

Cluster_HAC_1=c_hac_3			
Examples			[50,0 %] 19
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
MPG-Consumo	5,5	30,68 (3,12)	24,76 (6,55)
Rapporto_Trasmissione	1,8	3,25 (0,41)	3,09 (0,52)
Cilindrata	-4,2	115,95 (29,25)	177,29 (88,88)
Cilindri	-4,5	4,21 (0,63)	5,39 (1,60)
Cavalli	-4,6	81,58 (15,51)	101,74 (26,44)
Peso	-4,9	2,29 (0,28)	2,86 (0,71)
Discrete attributes : [Recall] Accuracy			
Paese=Giappone	2,1	[85,7 %] 31,6 %	18,40%
Paese=Italia	1	[100,0 %] 5,3 %	2,60%
Paese=Germania	0,5	[60,0 %] 15,8 %	13,20%
Paese=Francia	-1	[0,0 %] 0,0 %	2,60%
Paese=U.S.A.	-1,3	[40,9 %] 47,4 %	57,90%
Paese=Svezia	-1,4	[0,0 %] 0,0 %	5,30%

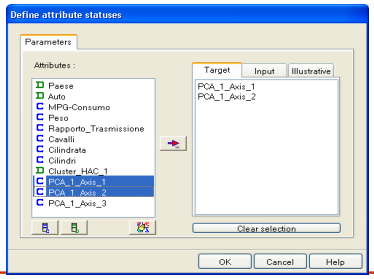
Esempio: ACP e CA

Per visualizzare i cluster, si può utilizzare l'analisi fattoriale. Si trascina il comando *Principal Components Analysis* su HAC e, con il tasto destro del mouse sul nuovo componente, si clicca su View. I primi due assi raggiungono il 98,7% dell'informazione disponibile.



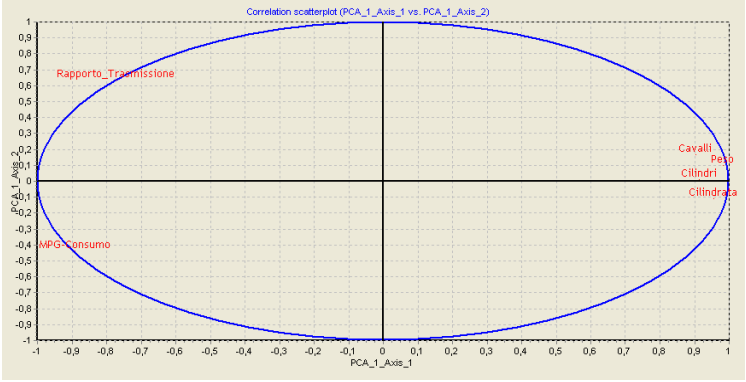
Esempio: rappresentazione dei cluster

Per ottenere il Cerchio delle Correlazioni si aggiunge ancora un Define Status e come Target si individua i primi due assi. Come Input tutti gli altri descrittori: quelli che hanno reso possibile la costruzione degli assi (MPG, Peso e Rapporto di trasmissione) e quelli che si usano per interpretare gli assi (Cavalli, Cilindrata e Cilindri).



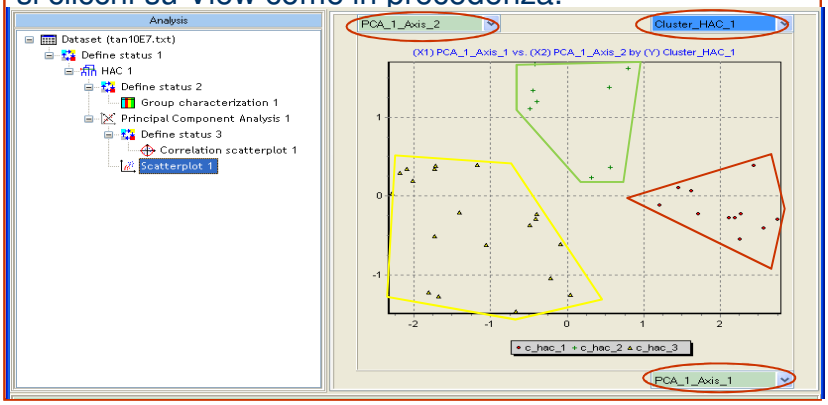
Esempio: Correlation Scatterplot

Si aggiunga il comando Correlation Scatterplot per visualizzare il Cerchio delle Correlazioni tra i due assi principali.



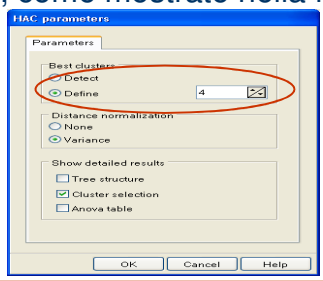
Esempio: i cluster sul piano fattoriale

A questo punto per una migliore visualizzazione dei cluster si aggiunga il componente *Scatterplot* all'ACP e si clicchi su View come in precedenza.



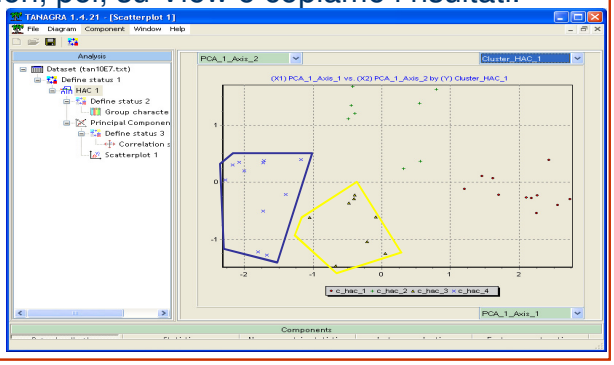
Esempio: modificare il numero di cluster

Nel piano fattoriale precedente era possibile individuare chiaramente i tre gruppi,ma se si volesse modificare il numero di cluster (perché per esempio crediamo che il terzo cluster ha troppi individui) è sufficiente cliccare con il tasto destro del mouse su HAC e ridefinire il numero di cluster, come mostrato nella figura in basso.



Esempio: il nuovo cluster

Come si può notare dal nuovo piano fattoriale il cluster 3 è stato diviso in due gruppi. Clicchiamo nuovamente con il tasto destro del mouse sul componente Group Characterization, poi, su View e copiamo i risultati.



Esempio: definizione dei due nuovi cluster

Il nuovo terzo cluster (8 auto) deve essere contrapposto al secondo: infatti, ora è composto dalle auto di medie dimensioni fabbricate in USA. Il quarto (11 automobili) è composto dalle citycar.

Cluster_HAC_1=c_hac_3			
Examples			[21,1 %] 8
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
MPG-Consumo	1,8	28,55 (2,24)	24,76 (6,55)
Cavalli	-1,1	92,63 (16,07)	101,74 (26,44)
Cilindrata	-1,2	142,63 (25,43)	177,29 (88,88)
Rapporto_Trasmissione	-1,3	2,88 (0,22)	3,09 (0,52)
Peso	-1,4	2,56 (0,18)	2,86 (0,71)
Cilindri	-1,8	4,50 (0,93)	5,39 (1,60)
Discrete attributes : [Recall] Accuracy			
Paese=U.S.A.	1,1	[27,3 %] 75,0 %	57,90%
Paese=Giappone	0,5	[28,6 %] 25,0 %	18,40%
Paese=Italia	-0,5	[0,0 %] 0,0 %	2,60%
Paese=Francia	-0,5	[0,0 %] 0,0 %	2,60%
Paese=Svezia	-0,7	[0,0 %] 0,0 %	5,30%
Paese=Germania	-1,2	[0,0 %] 0,0 %	13,20%

Cluster_HAC_1=c_hac_4			
Examples			[28,9 %] 11
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
MPG-Consumo	4,4	32,23 (2,78)	24,76 (6,55)
Rapporto_Trasmissione	3,2	3,52 (0,28)	3,09 (0,52)
Cilindri	-3,4	4,00 (0,00)	5,39 (1,60)
Cilindrata	-3,5	96,55 (10,20)	177,29 (88,88)
Cavalli	-4,1	73,55 (9,15)	101,74 (26,44)
Peso	-4,2	2,09 (0,13)	2,86 (0,71)
Discrete attributes : [Recall] Accuracy			
Paese=Giappone	1,8	[57,1 %] 36,4 %	18,40%
Paese=Germania	1,6	[60,0 %] 27,3 %	13,20%
Paese=Italia	1,6	%] 9,1 %	2,60%
Paese=Francia	-0,6	[0,0 %] 0,0 %	2,60%
Paese=Svezia	-0,9	[0,0 %] 0,0 %	5,30%
Paese=U.S.A.	-2,4	[13,6 %] 27,3 %	57,90%