Cluster Analysis (2 parte)

Esempio 2

Data set:

Nel data set **Dieta (**Dieta.txt, Dieta.sav) sono contenute informazioni sul consumo medio dei principali alimenti in 16 paesi Europei.

Paese	Nome del paese
Cereali (Ce)	Consumo medio annuale in Kg
Riso (R)	Consumo medio annuale in Kg
Patate (P)	Consumo medio annuale in Kg
Zucchero (Z)	Consumo medio annuale in Kg
Verdure (Ver)	Consumo medio annuale in Kg
Vino (Vi)	Consumo medio annuale in litri
Carne (Ca)	Consumo medio annuale in Kg
Latte (L)	Consumo medio annuale in litri
Burro (B)	Consumo medio annuale in Kg
Uova (U)	Consumo medio annuale in Kg

Domanda: Possiamo raggruppare i paesi Europei in sotto-aree con comportamenti alimentari simili?

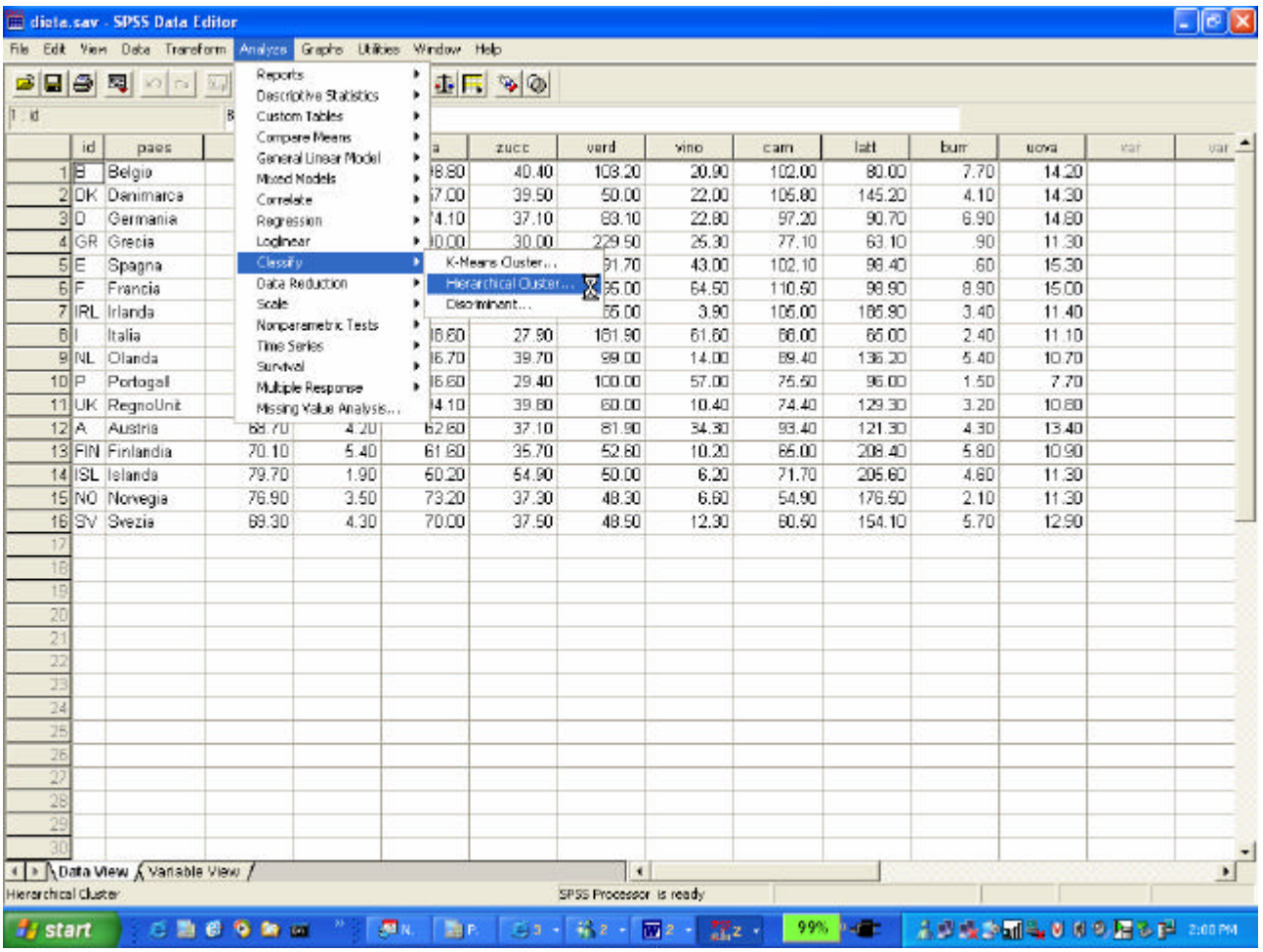
Analisi

A tale scopo individuiamo la presenza di possibili cluster mediante SPSS.

Dal menu **Analyse**, selezioniamo **Classify**, SPSS permette di scegliere due diversi approcci per la cluster analysy: **K-Means Cluster** e **Hierarchical Cluster**

Hierarchical Cluster:

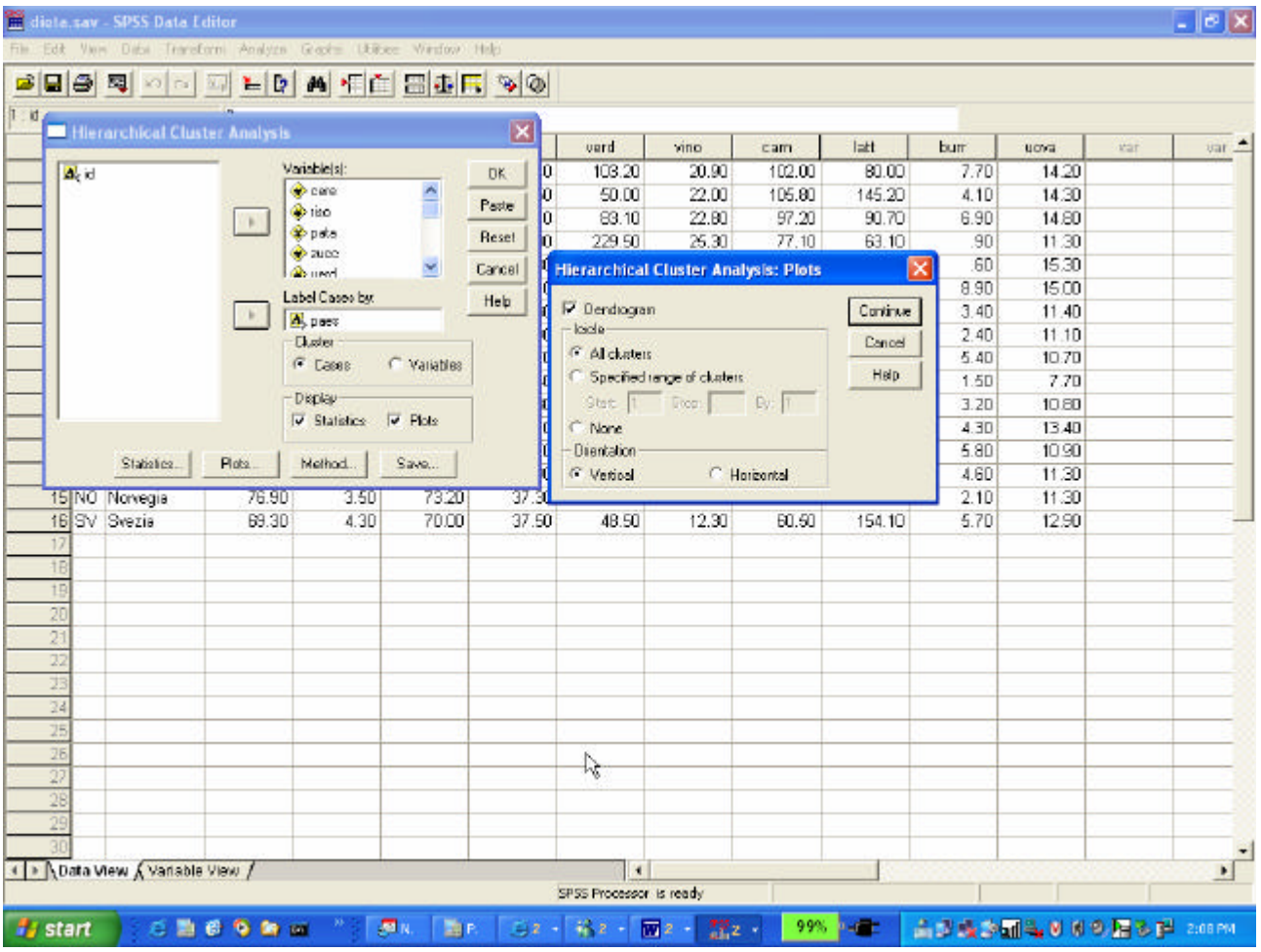
dal menu **Analyse**, selezioniamo **Classify**, e poi **Hierarchical Cluster.**



Selezioniamo le variabili da considerare nell'analisi; la variabile nominale *id* (o *paese*) sarà selezionata come **Label Cases by**.

Analisi dell'output

Per poter meglio interpretare l'output selezioniamo nel menù **Plots…**: **Dendogram**; in questo modo il dendogramma ci permetterà di meglio interpretare il risultato ottenuto.



I metodi di classificazione gerarchica si ripartiscono in metodi divisivi e metodi agglomerativi (o ascendenti). I metodi divisivi definiscono partizioni sempre più fini dell'insieme iniziale, i metodi agglomerativi realizzano una successione di partizioni in n classi, n-1 classi, n-2 classi connesse le une alle altre in modo tale che la partizione in k classi è ottenuta raggruppando due delle classi delle partizioni in k+1 classi. SPSS procede con il secondo metodo.

Il risultato dell'analisi è dato dal **Dendogram** e dalla tabella **Agglomeration Schedule**. Il dendogramma rappresenta una sintesi grafica del risultato ottenuto dall'analisi del *cluster* gerarchico, mentre la seguente tabella rappresenta una sintesi numerica.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	15	16	649.850	0	0	7
2	13	14	680.100	0	0	7
3	1	3	1171.550	0	0	8
4	9	12	1738.030	0	0	5
5	9	11	2384.655	4	0	9
6	6	10	2422.320	0	0	8
7	13	15	2509.665	2	1	13
8	1	6	2782.430	3	6	10
9	2	9	2931.363	0	5	10
10	1	2	4892.781	8	9	14
11	4	5	5431.110	0	0	12
12	4	8	7214.295	11	0	15
13	7	13	10644.298	0	7	14
14	1	7	11477.340	10	13	15
15	1	4	23101.344	14	12	0

Osservando la precedente tabella vediamo che al primo livello i casi 15 e 16 (corrispondenti a Norvegia e Svezia) sono raggruppati, in quanto hanno la distanza minima. Il nuovo *cluster* così creato riappare allo stage 7. Se passiamo subito alla riga corrispondente allo stage 7, vediamo che i *cluster* 13 e 15 vengono raggruppati; ma chi sono i *cluster* 13 e 15? Dalla quinta e sesta colonna leggiamo a che livello sono stati creati tali *cluster*, (livello 0 indica che il *cluster* è composto da un singolo oggetto). Dalla settima riga, leggiamo che il *cluster* 13 è stato creato allo stage 2, è l'unione degli oggetti 13 e 14 (corrispondenti a Finlandia e Islanda), mentre il cluster 15 è il gruppo creato al primo livello di agglomerazione. Quando le osservazioni sono tante, la lettura della precedente tabella può essere complicata e si preferisce la rappresentazione grafica data dal dendogramma**.** E' importante non sottovalutare l'informazione data nella quarta colonna che fornisce la distanza tra gli oggetti. L'osservazione della quarta colonna ci aiuta nel cercare il più grande gap nei valori di distanza tra oggetti; vediamo che gli stadi 13 e 15 rappresentano i maggiori passi nell'agglomerazione.

L'osservazione del **dendogramma** risulta più facile ed immediata. Nell'asse verticale di sinistra leggiamo gli oggetti presenti nell'analisi, l'asse orizzontale mostra la distanza tra i cluster quando sono uniti. L'albero fornisce vari livelli di aggregazione: la scelta del livello a cui "tagliare" l'albero deve rappresentare un giusto compromesso tra numero di gruppi e omogeneità degli stessi. Generalmente il taglio va fatto prima delle aggregazioni corrispondenti a salti molto grossi tra i valori dell'indice. Se tagliamo lo schema gerarchico prima del valore 10 a cui corrisponde il gap maggiore nel coefficiente riscalato, troviamo lo stesso risultato che avevamo trovato con il metodo delle k-medie, cioè un gruppo di paesi mediterranei (Grecia, Spagna e Italia), un gruppo più mitteleuropeo/centroeuropeo (Belgio, Germania, Francia, Portogallo, Olanda, Austria, UK e Danimarca), un gruppo scandinavo (Norvegia, Svezia, Finlandia e Islanda) ed infine un gruppo con un unico paese: l'Irlanda. Interessante osservare che si ottengono gli stessi risultati con il metodo di raggruppamento tramite K-medie.

```
* * * * * * H I E R A R C H I C A L  C L U S T E R  A N A L Y S I S * * * * * *

 Dendrogram using Average Linkage (Between Groups)

                          Rescaled Distance Cluster Combine

    C A S E          0         5        10        15        20        25
  Label      Num     +---------+---------+---------+---------+---------+

  Norvegia    15     ─┐
  Svezia      16     ─┘─┐
  Finlandia   13     ─┐ │
  Islanda     14     ─┘ │
  Irlanda      7      ──┼──────────────────┐
  Belgio       1     ─┐ │                  │
  Germania     3     ─┘─┼─┐                │
  Francia      6      ──┘ │                │
  Portogall   10     ─┐   │                │
  Olanda       9     ─┘─┐ │                │
  Austria     12     ─┐ │ │                │
  RegnoUnit   11     ─┘─┘ │                │
  Danimarca    2      ────┘                │
  Grecia       4     ─┐                    │
  Spagna       5     ─┘─┐                  │
  Italia       8      ──┘                  │
```

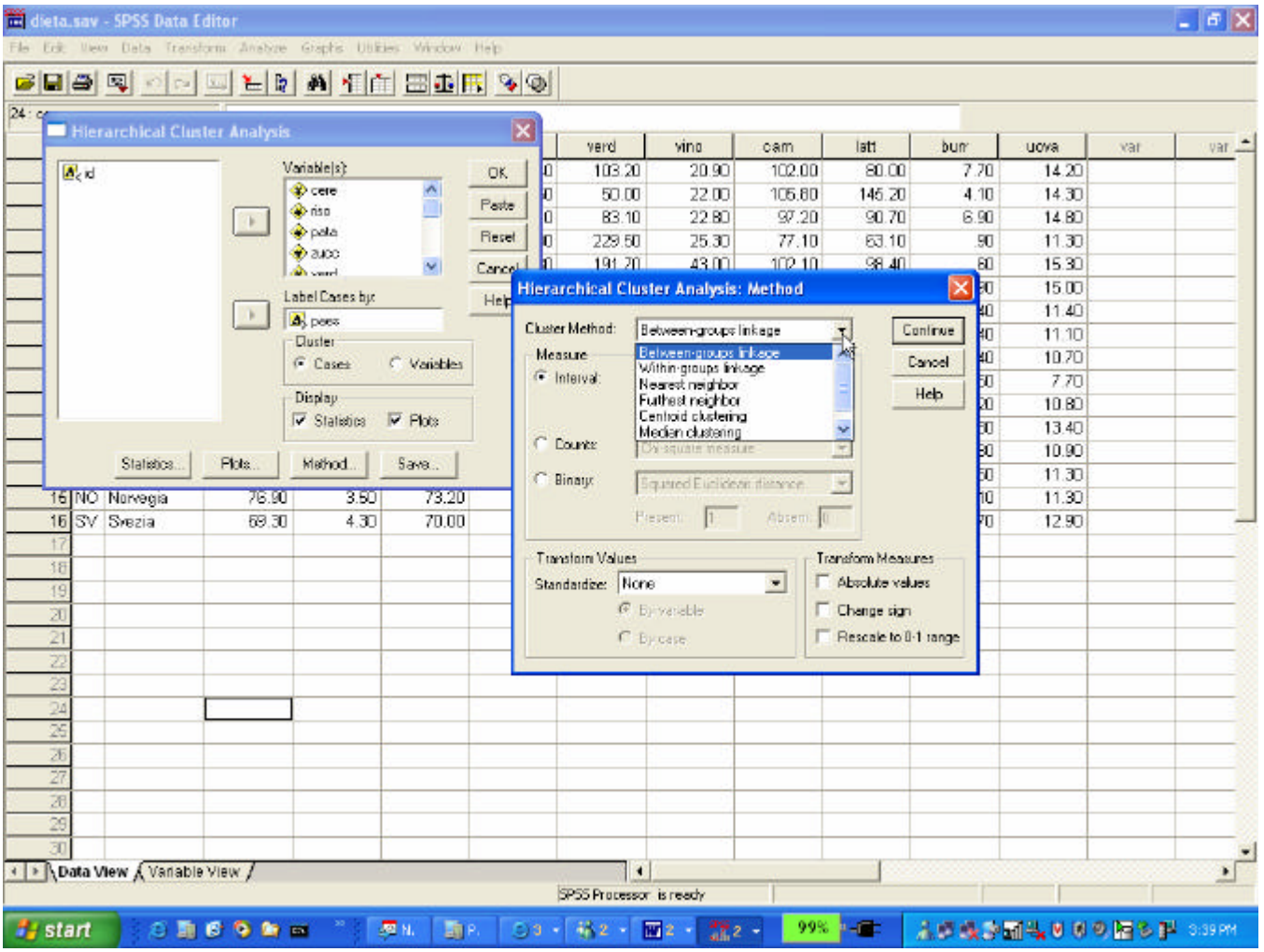
Il problema principale dei metodi di classificazione gerarchica consiste nel definire il criterio di raggruppamento di due oggetti (due unità, due gruppi, un'unità e un gruppo), cosa che equivale a definire una distanza tra oggetti. Tutti gli algoritmi di classificazione gerarchica si sviluppano nella maniera seguente: si ricercano ad ogni tappa le due classi più vicine, le si riunisce e si continua fino a che non si abbia una sola classe.

Quale distanza considerare? Quale metodo utilizzare per il cluster gerarchico?

Nel Menù Method possiamo scegliere il metodo agglomerativo e il tipo di distanza da utilizzare. I metodi disponibili sono:

Between-groups linkage (metodo utilizzato di default da SPSS) e *Within-groups linkage*: la distanza tra c_1 e c_2 è data dalla media delle distanze tra tutti i punti di c_1 da tutti i punti di c_2. Con *Furthest neighbor* e *Nearest neighbor* la distanza tra c_1 e c_2 è data rispettivamente dalla distanza massima e minima tra tutti i punti di c_1 da tutti i punti di c_2. Con *centroid clustering* la distanza tra c_1 e c_2 è data dalla distanza tra i baricentri di c_1 e c_2. Com *median clustering*, la distanza tra c_1 e c_2 è data rispettivamente dalla distanza mediana tra le distanze tra tutti i punti di c_1 e da tutti i punti di c_2. *Ward's method* è basato sulla minimizzazione della varianza all'interno dei gruppi

Le distanze utilizzate sono: Euclidean distance, Squared Euclidean distance, Cosin, Pearson correlation, Chebychev, Block, Minkowski, Customized.



Nel menù Method è anche possibile scegliere il tipo di standardizzazione delle variabili. Standardizzando le variabili si ottengono risultati differenti.

Osserviamo infatti il dendogramma ottenuto scegliendo la **standardizzazione** tramite la deviazione standard. Si ottengono risultati differenti; in particolare, nel *cluster* dei paesi mediterranei vediamo incluso anche il Portogallo, la Gran Bretagna è invece unita ai paesi Scandinavi e l'Islanda rimane in un cluster da sola e solo al 13-mo step viene agglomerata ad un gruppo contenente sia i paesi dell'Europa centrale che scandinava.

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine