Cluster Analysis

Esempio 1

Stiamo studiando le abitudini alimentari nei Paesi europei. Sulla base dei dati a disposizione, ci chiediamo se si possano individuare sotto-aree con abitudini alimentari simili.

Dati: 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 1.** Possiamo raggruppare i paesi Europei in sotto-aree con comportamenti alimentari simili?
- **Domanda 2** Possiamo dare un'interpretazione ai gruppi (*cluster*) ottenuti? Cosa hanno in comune i Paesi che appartengono allo stesso gruppo?
- **Domanda 3** Quali variabili hanno maggiormente influenzato la determinazione dei gruppi?

Analisi

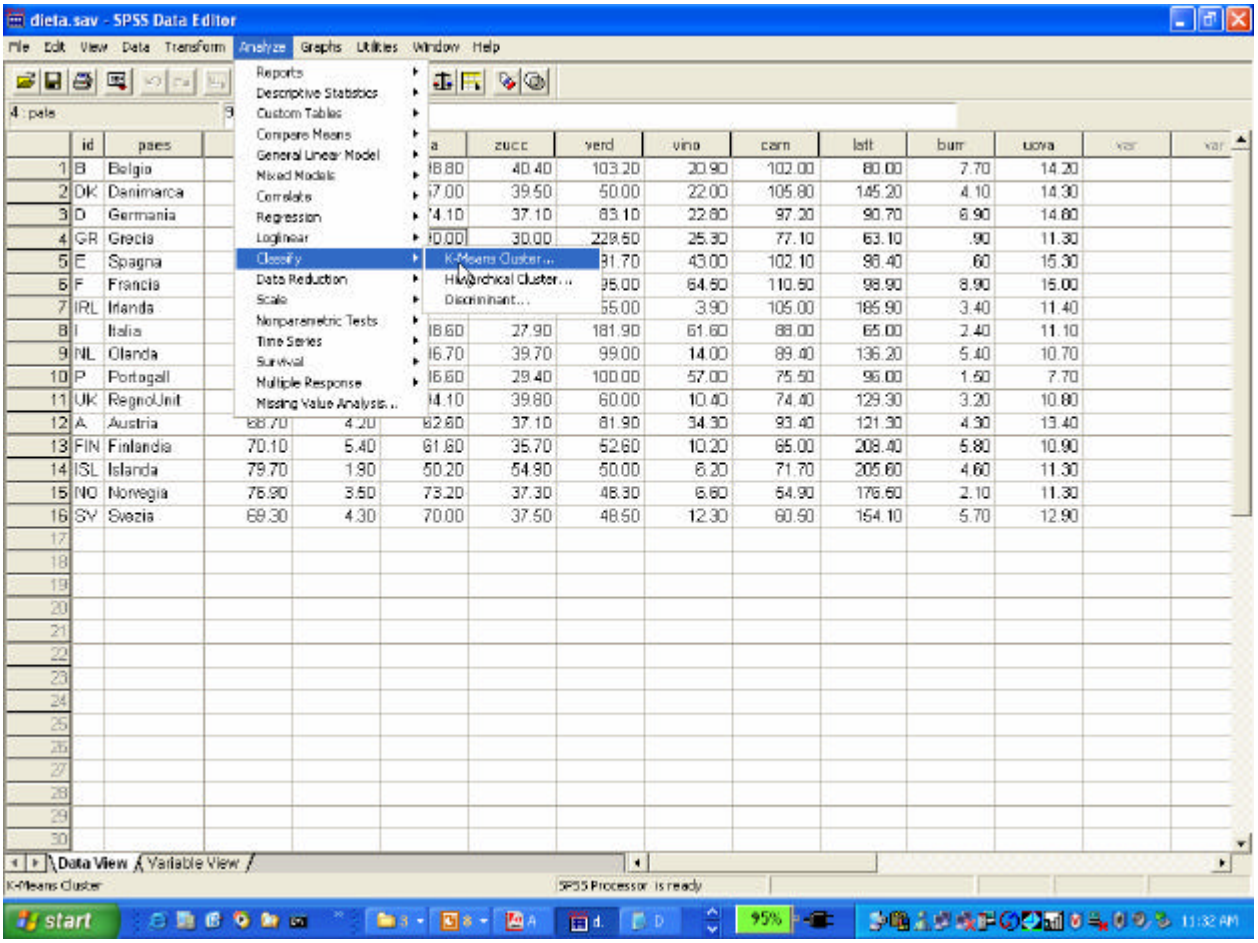
A tale scopo individuiamo la presenza di possibili gruppi mediante SPSS.

Dal menu **Analyse**, selezioniamo **Classify**. SPSS permette di scegliere due diversi approcci per la *cluster analysis*: **K-Means Cluster** e **Hierarchical Cluster**: i due metodi cercano entrambi gruppi di oggetti tali che all'interno dello stesso gruppo (*cluster*) gli oggetti siano "simili" tra loro, e oggetti appartenenti a gruppi diversi siano "differenti" tra loro: lo scopo è minimizzare la distanza all'interno del *cluster* e massimizzare la distanza tra *cluster*:

K-Means Cluster: gli oggetti sono divisi in sottoinsiemi disgiunti, tale che ciascun oggetto appartiene ad uno ed un solo cluster. Ogni *cluster* è associato con un centroide; ogni oggetto viene assegnato al *cluster* il cui centroide risulta più vicino. Il numero di *cluster* deve essere specificato inizialmente!

Hierarchical cluster: consiste in un insieme di *cluster* gerarchici organizzati tramite un "albero gerarchico" (dendogramma). Non necessita di specificare a priori del numero di *cluster* ; il numero di *cluster* può essere ottenuto spezzando il dendogramma a diverse altezze. L'algoritmo si basa su una matrice di distanze tra gli oggetti (con la metrica desiderata).

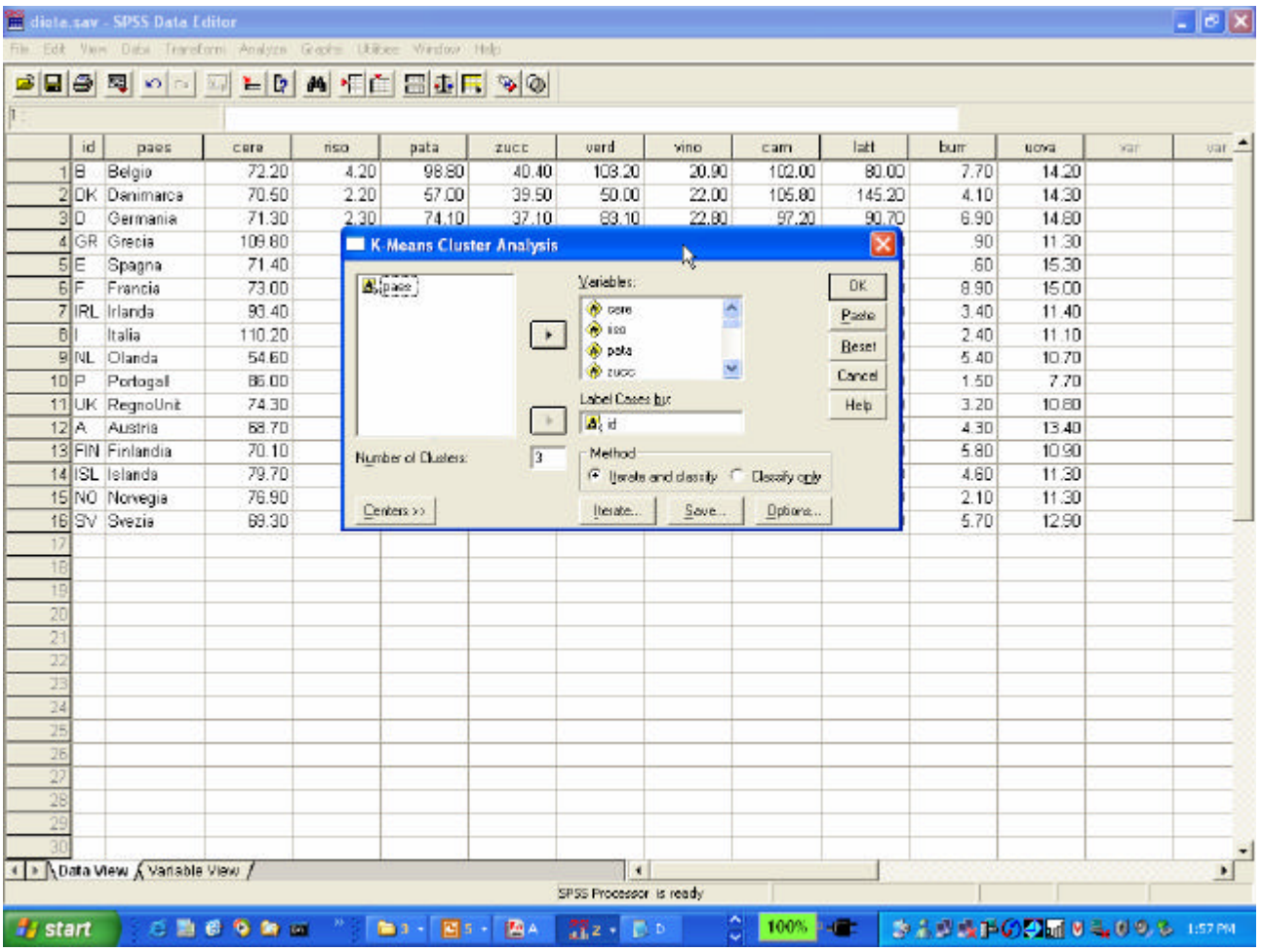
In questo esempio utilizziamo il metodo **K-Means Clustering.** Il metodo gerarchico sarà illustrato nell'esempio2.
Dal menu **Analyse**, selezioniamo **Classify**, e poi **K-Means Cluster**



Selezioniamo le variabili da considerare nell'analisi (in questo caso possiamo selezionarle tutte, alternativa potrebbe essere selezionare solo alcune variabili tramite un'analisi esplorativa iniziale); la variabile nominale *id* (o *paese*) sarà selezionata come **Label Cases by**.

Dobbiamo decidere il numero di *cluster*. Consideriamo i seguenti casi: 3 e 4.

Number of clusters: 3



Analisi dell'output

Cominciamo dall'ultima tabella che presenta il riassunto dell'analisi; in particolare, ci sono 3 cluster, a cui appartengono rispettivamente 3, 6 e 7 oggetti.

Number of Cases in each Cluster

Cluster	1	3.000
	2	6.000
	3	7.000
Valid		16.000
Missing		.000

La tabella Cluster Membership ci dice a quale *cluster* appartiene ciascun oggetto; è un risultato opzionale e va ottenuto selezionando dal riquadro **Options** la casella **Cluster Information for each case**.

Cluster Membership

Case Number	ID	Cluster	Distance
1	B	3	37.110
2	DK	2	50.765
3	D	3	24.082
4	GR	1	41.498
5	E	1	47.905
6	F	3	40.083
7	IRL	2	81.845
8	I	1	51.019
9	NL	3	39.395
10	P	3	43.374
11	UK	3	46.586
12	A	3	28.269
13	FIN	2	36.225
14	ISL	2	41.307
15	NO	2	23.462
16	SV	2	32.202

Al primo *cluster* appartengono Grecia, Spagna e Italia, al secondo Danimarca, Irlanda, Finlandia, Islanda, Norvegia, Svezia, infine al terzo *cluster* appartengono Belgio, Germania, Francia, Olanda, Portogallo, UK e Austria. L'ultima colonna rappresenta la distanza dal punto al centroide del *cluster* di riferimento, dove la metrica utilizzata da SPSS è la metrica euclidea.

Domanda 2. Possiamo dare un'interpretazione ai gruppi ottenuti? Cosa hanno in comune i Paesi che appartengono allo stesso gruppo?

Cominciamo col vedere quali siano i centroidi finali.

Final Cluster Centers

	Cluster		
	1	2	3
CERE	97.13	76.65	71.44
RISO	5.33	3.42	4.31
PATA	78.80	77.25	85.87
ZUCC	28.23	39.95	36.80
VERD	201.03	50.73	88.89
VINO	43.30	10.20	31.99
CARN	89.07	77.15	91.77
LATT	75.50	179.28	107.49
BURR	1.30	4.28	5.41
UOVA	12.57	12.02	12.37

I "final cluster centers" di un gruppo sono costituiti dalle medie di ogni variabile all'interno del gruppo, e ci aiutano a capire le caratteristiche degli oggetti appartenenti a ciascun gruppo.

Quali sono i paesi appartenenti al cluster 1?
Al gruppo 1 appartengono i paesi con un alto consumo di cereali e riso, basso consumo di zucchero, alto consumo di verdure e di vino e basso consumo di latte e burro: infatti Grecia, Spagna e Italia sono tre paesi caratterizzati da una dieta mediterranea.
Al gruppo 2 appartengono paesi con basso consumo di riso, alto consumo di zucchero e latte, basso consumo di verdure e medio alto consumo di burro: cioè paesi con una dieta molto calorica, i paesi Scandinavi: Danimarca, Irlanda, Finlandia, Islanda, Norvegia, Svezia.
Infine al gruppo 3 appartengono i paesi con alto consumo di carne, patate e burro, e medio alto consumo di uova, una dieta sempre calorica ma più proteica.

Selezionando l'optione **Cluster Information for each case** si ha a disposizione anche la seguente tabella:

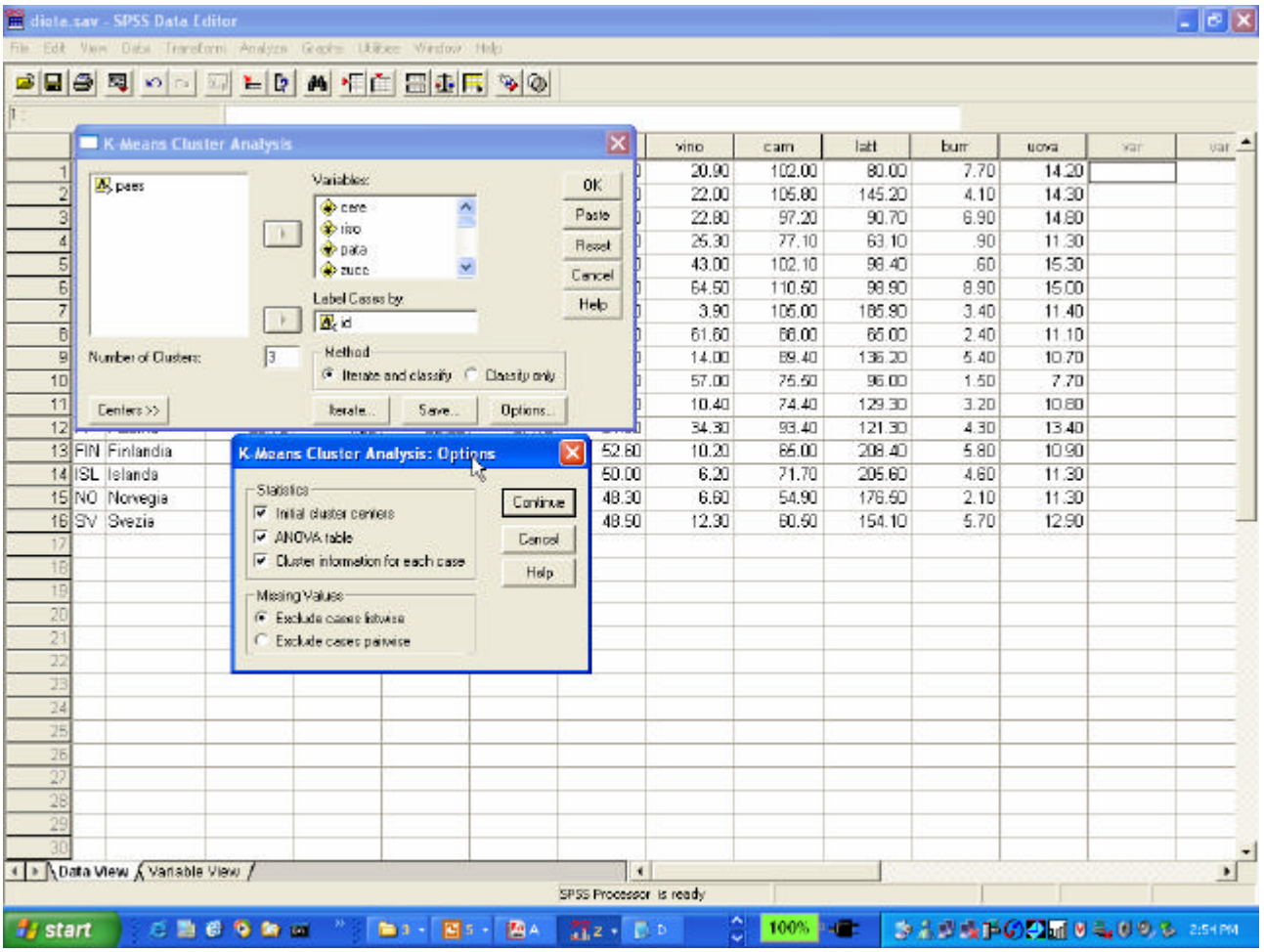
Distances between Final Cluster Centers

Cluster	1	2	3
1		187.539	120.569
2	187.539		86.096
3	120.569	86.096	

La precedente tabella mostra la distanza euclidea tra i centroidi dei gruppi finali: chiaramente maggiore è tale distanza, maggiore sarà la dissomiglianza tra i tre gruppi. I tre gruppi sembrano distanti tra loro; la distanza maggiore si osserva tra il primo e il secondo, mentre il secondo e il terzo sembrano molto vicini (intuitivamente si poteva già arrivare a tale risultato).

Domanda 3: Quali variabili hanno maggiormente influenzato la determinazione dei cluster?

Selezionando dal riquadro **Options** la casella **Anova Table**.



Si ottiene la seguente tabella:

ANOVA

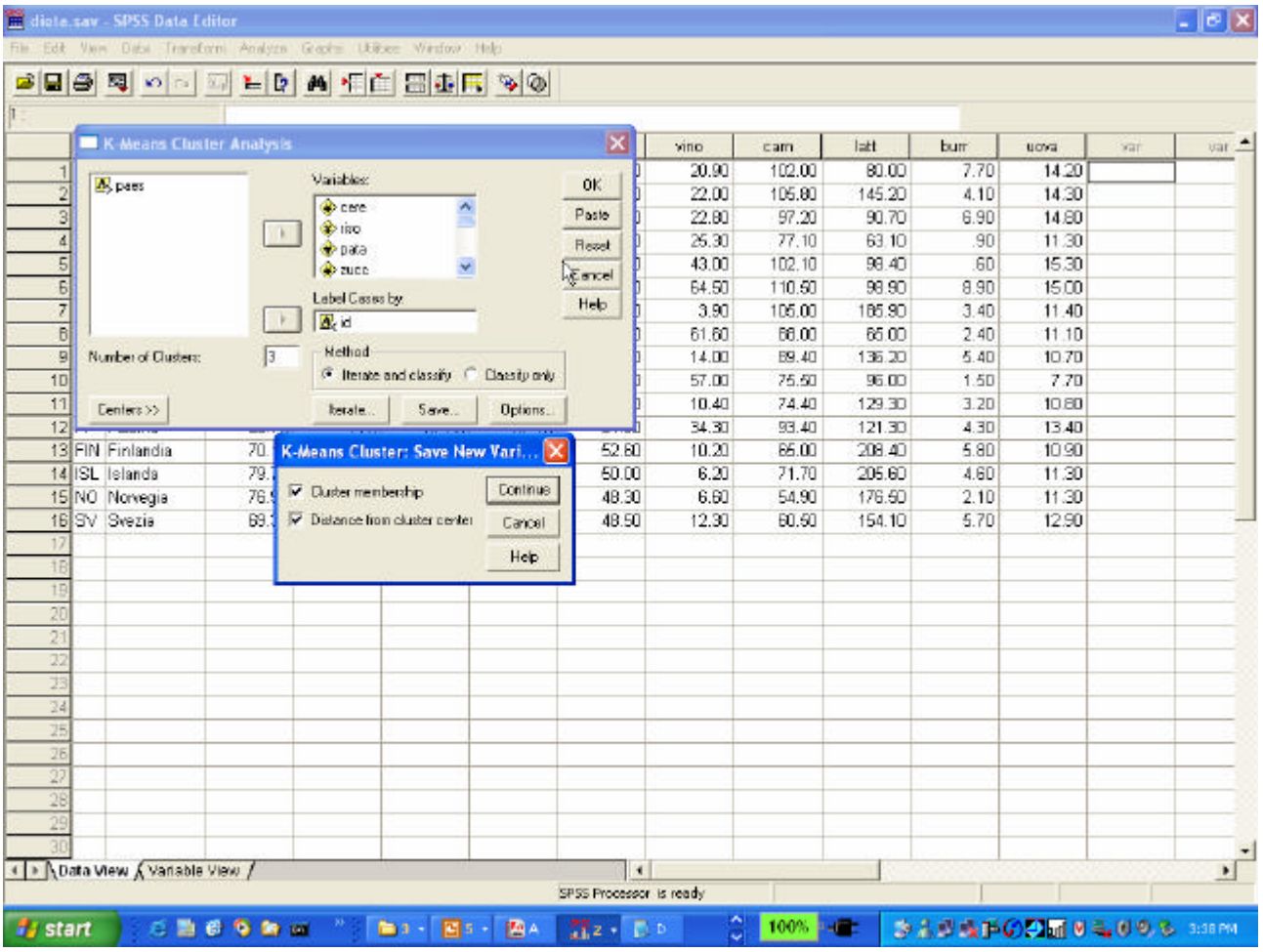
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
CERE	704.719	2	148.643	13	4.741	.028
RISO	3.805	2	1.197	13	3.179	.075
PATA	131.724	2	842.722	13	.156	.857
ZUCC	138.404	2	29.169	13	4.745	.028
VERD	22871.120	2	206.307	13	110.860	.000
VINO	1323.224	2	273.556	13	4.837	.027
CARN	365.274	2	301.532	13	1.211	.329
LATT	13495.880	2	531.557	13	25.389	.000
BURR	17.794	2	4.069	13	4.373	.035
UOVA	.360	2	4.936	13	.073	.930

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

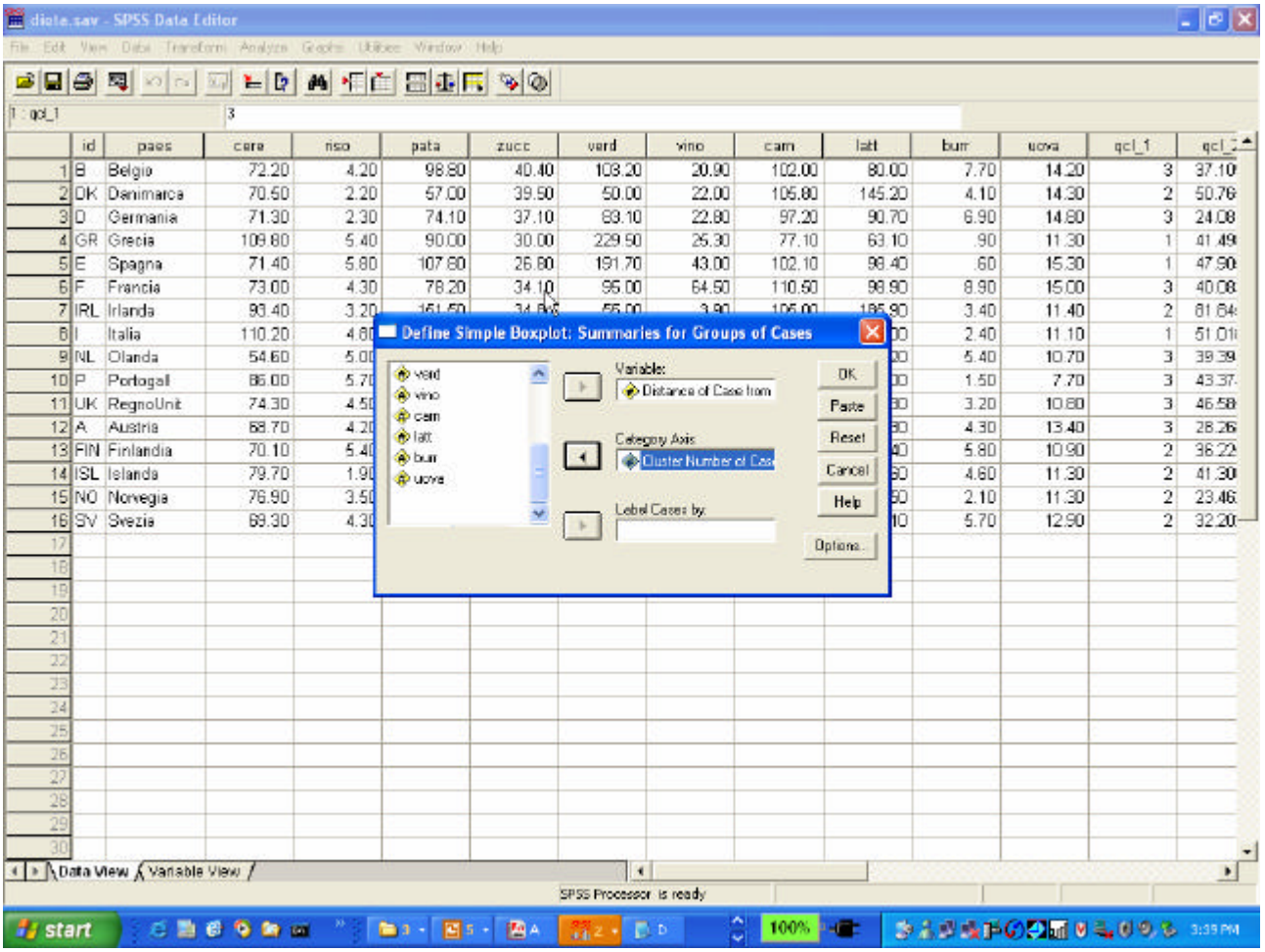
La tabella ANOVA indica quali variabili hanno maggiormente contribuito all'individuazione dei cluster. Latte e Verdura risultano le due variabili significativamente associate ai cluster individuati, a seguire Cereali, Zucchero e Vino. Uova e Patate risultano invece le meno influenti nella divisione in gruppi così ottenuta. (Ricordiamo che la procedura ANOVA di SPSS richiede i gruppi bilanciati e in questo caso non lo sono, quindi i risultati ottenuti dalla precedente tabella hanno un'interpretazione solo descrittiva).

Le tabelle Initial Cluster Centers and Iteration History riassumono i passi necessari all'algoritmo per trovare tali cluster.

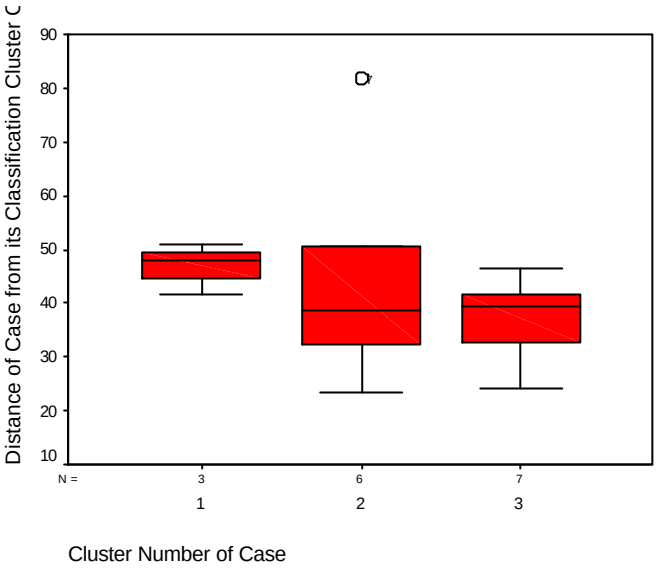
Dal menù **Save** selezioniamo **Cluster Membership** e **Distance from cluster center**; in questo modo nel file Dieta.sav, SPSS Data Editor, compariranno due colonne con le informazioni selezionate.



Ora dal Menu **Graph**, selezioniamo **Box-Plot** e poi Distance *of Cases from Cluster* è **Variable**, e poi *Cluster Number of Cases* come **Category Axis**. E' possibile in questo modo



Il risultato è un grafico diagnostico che permette di osservare le differenze tra i gruppi e la presenza di eventuali *outlier*. Il gruppo 2 presenta un *outlier*, Irlanda, come è osservabile dalla tabella Cluster Membership, che mostra come l'Irlanda sia il paese più distante dal centroide.



Proviamo a ricercare con la stessa tecnica **4 cluster**. Otteniamo le seguenti tabelle:

Cluster Membership

Case Number	ID	Cluster	Distance
1	B	3	37.110
2	DK	2	48.999
3	D	3	24.082
4	GR	4	41.498
5	E	4	47.905
6	F	3	40.083
7	IRL	1	.000
8	I	4	51.019
9	NL	3	39.395
10	P	3	43.374
11	UK	3	46.586
12	A	3	28.269
13	FIN	2	32.011
14	ISL	2	34.328
15	NO	2	21.352
16	SV	2	27.989

Viene individuato un cluster con un solo oggetto, l'Irlanda (ricordiamo che risultava *outlier* nell'analisi precedente). Al secondo gruppo appartengono Danimarca, Finlandia, Islanda, Norvegia e Svezia. Al terzo gruppo appartengono Belgio, Germania, Francia, Olanda, Portogallo, UK e Austria. Infine, al quarto gruppo appartengono Grecia, Spagna e Italia. Osservando l'analisi dell'Anova, risulta interessante notare che la variabile "consumo di patate" assume importanza nel discriminare i gruppi; infatti è proprio l'alto consumo di patate (visibile anche dalla Tabella Final Cluster Center) a determinare un gruppo a cui appartiene come unico paese l'Irlanda.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
CERE	582.038	3	132.974	12	4.377	.027
RISO	2.556	3	1.292	12	1.978	.171
PATA	2293.041	3	361.643	12	6.341	.008
ZUCC	102.878	3	28.948	12	3.554	.048
VERD	15254.695	3	221.679	12	68.814	.000
VINO	898.026	3	292.383	12	3.071	.069
CARN	553.765	3	249.097	12	2.223	.138
LATT	9014.766	3	571.475	12	15.775	.000
BURR	12.175	3	4.330	12	2.812	.085
UOVA	.392	3	5.309	12	.074	.973

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Final Cluster Centers

	Cluster			
	1	2	3	4
CERE	93.40	73.30	71.44	97.13
RISO	3.20	3.46	4.31	5.33
PATA	151.50	62.40	85.87	78.80
ZUCC	34.80	40.98	36.80	28.23
VERD	55.00	49.88	88.89	201.03
VINO	3.90	11.46	31.99	43.30
CARN	105.00	71.58	91.77	89.07
LATT	185.90	177.96	107.49	75.50
BURR	3.40	4.46	5.41	1.30
UOVA	11.40	12.14	12.37	12.57