# INDICE

## Introduzione

Nell'elaborato viene focalizzata l'attenzione sull'analisi di raggruppamento o denominata anche: "*CLUSTER ANALYSIS*" .

Operativamente, in ambito economico-aziendale la cluster analysis viene utilizzata per individuare possibili raggruppamenti fra gli individui della popolazione, sostituendo le singole unità con opportuni "tipi".

Questi gruppi sono costruiti in modo da essere massimamente omogenei al loro interno e massimamente eterogenei fra loro. Dal contrasto tra questi due criteri, si origina l'articolazione finale dei gruppi.

Le variabili considerate sono i consumi medi/giornalieri dei principali prodotti alimentari, in sedici Paesi Europei.

I dati verranno elaborati mediante il software statistico SPSS.

Per poter effettuare una cluster analysis occorre un data set da cui partire, quello che ho utilizzato è stato costruito considerando le statistiche di consumo cronico degli alimenti, pubblicate dall'Autorità Europea per la sicurezza alimentare (EFSA).

Dopo aver opportunamente sintetizzato i dati, è stato possibile effettuare l'analisi e valutare i risultati.

Scopo di questo elaborato è quello di mostrare in senso pratico la suddivisione della popolazione in gruppi considerando vari metodi statistici, e confrontarli .

La parte teorica sarà solo brevemente accennata.

# Capitolo I

***La cluster Analysis : cenni teorici.***

La cluster analysis consiste in un insieme di tecniche statistiche atte ad individuare gruppi di unità tra loro simili rispetto ad un insieme di variabili prese in considerazione e secondo uno specifico criterio. L'obiettivo che si pone è sostanzialmente quello di riunire unità tra loro eterogenee in più sottoinsiemi tendenzialmente omogenei e mutuamente esaustivi.

La cluster analysis consente allora di pervenire ai seguenti risultati [1]:

- ridurre i dati in forma grafica semplice (immediatamente percepibile) e parsimoniosa;
- generare ipotesi di ricerca;
- identificare tipi: la ricerca tipologica mira all'individuazione di gruppi di unità con caratteristiche distintive che, nell'insieme, facciano percepire la fisionomia del sistema osservato;
- costruire sintesi di classificazione automatica;
- stratificare popolazioni da sottoporre a campionamento;
- attribuire alle entità valori noti con accuratezza solo per la classe.
- Questa proprietà può portare a trovare dati validi con cui: sostituire dati mancanti, trovare una modalità con cui confrontare una risposta elusiva, infine, stimare la probabilità che si verifichi un certo evento in campioni di numerosità esigua.

[1] Fabbris L. (1997), **8**: 303-304

# Capitolo II

***Costruzione del data set***

## 2.1 Identificazione delle variabili di classificazione.

Il data set dell'analisi contiene informazioni sul consumo medio giornaliero dei principali alimenti in 16 Paesi Europei.

Ho ottenuto questi dati applicando la proprietà associativa della media aritmetica[2] alle varie classi di età

in cui il consumo di ogni singolo prodotto era suddiviso.

Dopo un'analisi esplorativa iniziale le variabili considerate sono: cereali e prodotti a base di cereali, verdure e prodotti vegetali, radici e tuberi, frutta e prodotti ortofrutticoli, carne e derivati, pesce e frutti di mare, latte, uova e ovo prodotti, zucchero e pasticceria, grassi animali,vegetali e oli, infine i prodotti surgelati.

I dati sono stati rilevati da varie agenzie di sondaggi nei seguenti paesi: Belgio, Bulgaria, Cipro, Danimarca, Finlandia, Francia, Germania, Irlanda, Italia, Lettonia, Paesi Bassi, UK, Repubblica Ceca, Spagna, Svezia e Ungheria.

[2] In base alla quale: suddividendo in due o più gruppi i valori della variabile, la media aritmetica della variabile è uguale alla media aritmetica delle medie parziali dei diversi gruppi ponderate con il numero di elementi di ciascuno.

$$\bar{x} = \frac{\bar{x}_A \cdot n_A + \bar{x}_B \cdot n_B}{n_A + n_B}$$

Il data set ottenuto è il seguente:

| Id | paese | cereali | Verdure | Tuberi | frutta | carne | pesce | latte | uova | zucchero | grassi | Prodotti Surgel. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BE | Belgio | 206.27 | 105.89 | 99.88 | 124.11 | 102.59 | 21.00 | 214.13 | 7.42 | 23.23 | 28.61 | 113.17 |
| BU | Bulgaria | 87.50 | 75.84 | 34.45 | 54.58 | 46.28 | 3.23 | 191.43 | 8.03 | 11.93 | 13.34 | 13.06 |
| CI | Cipro | 161.00 | 103.70 | 68.00 | 95.50 | 72.50 | 18.50 | 228.70 | 2.30 | 12.60 | 6.00 | 139.20 |
| DA | Dani-marca | 212.32 | 152.87 | 105.44 | 146.84 | 127.78 | 17.27 | 406.17 | 15.67 | 35.84 | 31.03 | .0 |
| FI | Finlan-dia | 123.38 | 103.58 | 83.57 | 133.16 | 102.34 | 20.21 | 476.96 | 12.50 | 26.60 | 28.63 | 5.89 |
| FR | Francia | 224.02 | 131.72 | 66.26 | 121.21 | 126.49 | 27.27 | 228.62 | 14.50 | 29.46 | 24.44 | .50 |
| GE | -Germa nia | 228.16 | 109.37 | 61.96 | 167.44 | 107.66 | 16.69 | 185.03 | 6.82 | 21.28 | 23.26 | 73.56 |
| IR | Irlanda | 218.30 | 150.70 | 252.50 | 90.80 | 169.10 | 21.30 | 302.50 | 11.90 | 31.60 | 34.70 | 29.70 |
| IT | Italia | 248.20 | 222.38 | 50.21 | 190.96 | 109.55 | 44.64 | 197.32 | 20.81 | 19.55 | 38.50 | 12.04 |
| LE | Lettonia | 218.76 | 76.24 | 123.31 | 109.97 | 120.83 | 13.79 | 144.38 | 8.16 | 22.84 | 13.70 | 256.96 |
| PB | Paesi Bassi | 171.11 | 63.64 | 70.51 | 104.73 | 75.47 | 6.32 | 393.97 | 5.17 | 36.02 | 20.90 | 56.79 |
| UK | Regno Unito | 213.50 | 128.00 | 111.80 | 94.90 | 107.30 | 26.90 | 259.60 | 17.80 | 23.80 | 16.60 | 3.30 |
| RC | Repubblica Ceca | 263.53 | 118.43 | 94.06 | 139.01 | 167.34 | 15.77 | 202.32 | 18.98 | 29.31 | 39.74 | 16.88 |
| SP | Spagna | 207.81 | 141.16 | 61.34 | 144.92 | 157.07 | 57.42 | 425.09 | 13.65 | 14.14 | 30.80 | 39.66 |
| SV | Svezia | 217.80 | 43.17 | 101.82 | 111.60 | 74.25 | 19.67 | 425.94 | 6.25 | 22.33 | 12.96 | 171.96 |
| UN | Unghe-ria | 238.87 | 156.03 | 108.17 | 186.18 | 178.68 | 8.09 | 262.43 | 25.49 | 29.83 | 44.90 | 4.88 |

**2.2 Selezione della tecnica di raggruppamento delle entità.**

Le tecniche (o criteri o algoritmi) di analisi dei gruppi sono numerose.

Ho scelto due tecniche 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 dei cluster e massimizzare la distanza tra cluster.

**Metodo K-means cluster:** gli oggetti sono divisi in sottoinsiemi disgiunti, tali che ciascun oggetto appartiene ad uno ed un solo cluster.

Ogni cluster è associato con un centroide[3]; ogni oggetto viene associato al cluster il cui centroide risulta più vicino.

**Hierarchical Cluster**: consiste in un insieme di cluster gerarchici organizzati tramite un "albero gerarchico" (dendrogramma).

L'algoritmo si basa su una matrice di distanze tra gli oggetti.

### 2.3 Identificazione del numero di gruppi entro i quali ripartire le entità.

Nell'analisi non gerarchica con il metodo delle k-medie il numero dei cluster deve essere specificato inizialmente.

Ho considerato i seguenti casi: 3 e 4.

L'analisi gerarchica non necessita di specificare a priori il numero di cluster; il numero di cluster può essere ottenuto spezzando il dendrogramma a diverse altezze.

[3] Si dice centroide il punto d'incontro delle medie di una distribuzione multivariata.

# Capitolo III

***La cluster Analysis con l'utilizzo di SPSS***

## *3.1) Metodo delle K-Means*

Dopo aver costruito il data set, ho selezionato le variabili da considerare nell'analisi e scelto come "label cases by" la variabile nominale id (o paese).

Numero di cluster: 3

*Analisi dell'output*.

Comincio dalla tabella 1 che presenta il riassunto dell'analisi.

In particolare, ci sono 3 cluster, a cui appartengono rispettivamente 6, 9 oggetti e viene individuato un cluster con un solo oggetto.

**Tabella 1:**

**Numero di casi in ogni cluster**

| | | |
|---|---|---|
| Cluster | 1 | 6.000 |
| | 2 | 9.000 |
| | 3 | 1.000 |
| Validi | | 16.000 |
| Mancanti | | .000 |

La tabella 2: Cluster di appartenenza ci dice a quale cluster appartiene ciascun oggetto.

**Tabella 2: Cluster di appartenenza**

| Numero di caso | ID | Cluster | Distanza |
|---|---|---|---|
| 1 | BE | 2 | 79.431 |
| 2 | BU | 2 | 177.295 |
| 3 | CI | 2 | 126.294 |
| 4 | DA | 1 | 76.273 |
| 5 | FI | 1 | 114.445 |
| 6 | FR | 2 | 50.655 |
| 7 | GE | 2 | 67.852 |
| 8 | IR | 1 | 191.555 |
| 9 | IT | 2 | 130.806 |
| 10 | LE | 3 | .000 |
| 11 | PB | 1 | 83.322 |
| 12 | RU | 2 | 76.342 |
| 13 | RC | 2 | 87.510 |
| 14 | SP | 1 | 88.579 |
| 15 | SV | 1 | 150.132 |
| 16 | UN | 2 | 118.630 |

Al primo cluster appartengono: Danimarca, Finlandia, Irlanda, Paesi Bassi, Spagna e Svezia.

Al secondo appartengono: Belgio, Bulgaria, Cipro, Francia,. Germania, Italia, Regno Unito, Repubblica Ceca e Ungheria.

Infine, al terzo cluster troviamo la Lettonia.

L'ultima colonna rappresenta la distanza dal punto al centroide del cluster di riferimento, dove la metrica utilizzata da SPSS è la metrica euclidea[4].

[4]In formula: $${}_2 d_{hk} = \left\{ \sum_{v=1}^{p} w_v \left( x_{hv} - x_{kv} \right)^2 \right\}^{1/2}$$

Ora ci domandiamo se possiamo dare un'interpretazione ai gruppi ottenuti.

Cosa hanno in comune i Paesi che appartengono allo stesso gruppo?

Comincio col vedere quali siano i centroidi finali.

**Tabella 3: Centri dei cluster finali**

| | Cluster | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| grano | 191.79 | 207.89 | 218.76 |
| verdure | 109.19 | 127.93 | 76.24 |
| tuberi | 112.53 | 77.20 | 123.31 |
| frutta | 122.01 | 130.43 | 109.97 |
| carne | 117.67 | 113.15 | 120.83 |
| pesce | 23.70 | 20.23 | 13.79 |
| latte | 405.11 | 218.84 | 144.38 |
| uova | 10.86 | 13.57 | 8.16 |
| zucchero | 27.75 | 22.33 | 22.84 |
| grassi | 26.50 | 26.15 | 13.70 |
| surgelati | 50.67 | 41.84 | 256.96 |

Sapendo che 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 pesce, latte, zucchero, grassi e un consumo medio alto di tuberi.

Un basso consumo di grano. Paesi con una dieta tendenzialmente calorica .

Al gruppo 2 appartengono i paesi con un alto consumo di verdure, frutta, uova

Un medio alto consumo di grano, carne, pesce, grassi. un basso consumo di tuberi, prodotti surgelati e zucchero.

I paesi con un regime alimentare variato e completo.

Infine, al gruppo 3 appartiene il paese con un alto consumo di grano, tuberi, carne, e surgelati.

Un basso consumo di verdure, frutta, pesce, uova,  grassi.

Quindi una dieta che a differenza del primo gruppo è sempre calorica ma, più proteica.

**tabella 4:**

**Distanze tra i centri dei cluster finali**

| Cluster | 1 | 2 | 3 |
|---|---|---|---|
| 1 | | 191.758 | 336.024 |
| 2 | 191.758 | | 239.666 |
| 3 | 336.024 | 239.666 | |

Analizzando la tabella 4 notiamo la distanza euclidea tra i centroidi dei gruppi finali e sappiamo che maggiore è la distanza, maggiore sarà la dissomiglianza tra i tre gruppi.

I tre gruppi sembrano distanti tra loro; in particolare la distanza maggiore si osserva tra il primo e il terzo, mentre il primo e il secondo sembrano più vicini.

Quali variabili hanno influenzato la determinazione dei cluster?

Per rispondere osserviamo la tabella 5[5] che indica quali variabili hanno maggiormente contribuito all'individuazione dei cluster.

[5]Ricordiamo che la procedura ANOVA di SPSS richiede che i gruppi siano bilanciati e in questo casi non lo sono, quindi i risultati ottenuti dalla tabella hanno un'interpretazione solo descrittiva.

**Tabella 5: ANOVA**

| | Cluster | | Errore | | | |
|---|---|---|---|---|---|---|
| | Media dei quadrati | df | Media dei quadrati | df | F | Sig. |
| grano | 607.462 | 2 | 2326.388 | 13 | .261 | .774 |
| verdure | 1547.725 | 2 | 1927.980 | 13 | .803 | .469 |
| tuberi | 2726.278 | 2 | 2379.146 | 13 | 1.146 | .348 |
| frutta | 264.681 | 2 | 1478.670 | 13 | .179 | .838 |
| carne | 52.827 | 2 | 1686.332 | 13 | .031 | .969 |
| pesce | 50.354 | 2 | 205.917 | 13 | .245 | .787 |
| latte | 72851.036 | 2 | 1773.578 | 13 | 41.076 | .000 |
| uova | 22.046 | 2 | 43.995 | 13 | .501 | .617 |
| zucchero | 54.225 | 2 | 56.541 | 13 | .959 | .409 |
| grassi | 74.567 | 2 | 130.223 | 13 | .573 | .578 |
| surgelati | 21125.695 | 2 | 3258.090 | 13 | 6.484 | .011 |

Latte e Sergelati risultano le due variabili significativamente associate ai cluster individuati, a seguire Tuberi e Verdure.

Uova, Pesce, Carne e Zucchero risultano invece le meno influenti nella divisione in gruppi così ottenuta.

Numero di cluster:4

Inizio col valutare la tabella 6 che indica 4 cluster a cui appartengono rispettivamente 1, 4, 5 e 6 oggetti.

**Tabella 6:**

**Numero di casi in ogni cluster**

| | | |
|---|---|---|
| Cluster | 1 | 1.000 |
| | 2 | 4.000 |
| | 3 | 5.000 |
| | 4 | 6.000 |
| Validi | | 16.000 |
| Mancanti | | .000 |

Anche in questo caso osserviamo la tabella 7 “cluster di appartenenza” per vedere a quale cluster appartiene ciascun oggetto.

**Tabella 7: Cluster di appartenenza**

| Numero di caso | ID | Cluster | Distanza |
|---|---|---|---|
| 1 | BE | 2 | 64.382 |
| 2 | BU | 2 | 135.524 |
| 3 | CI | 2 | 64.538 |
| 4 | DA | 3 | 90.256 |
| 5 | FI | 3 | 95.428 |
| 6 | FR | 4 | 60.694 |
| 7 | GE | 2 | 88.845 |
| 8 | IR | 4 | 162.391 |
| 9 | IT | 4 | 125.896 |
| 10 | LE | 1 | .000 |
| 11 | PB | 3 | 69.598 |
| 12 | RU | 4 | 68.624 |
| 13 | RC | 4 | 68.109 |
| 14 | SP | 3 | 83.111 |
| 15 | SV | 3 | 141.070 |
| 16 | UN | 4 | 68.166 |

Al primo cluster appartiene la Lettonia.

Al secondo appartengono: Belgio, Bulgaria, Cipro e Germania.

Al terzo troviamo: Danimarca, Finlandia, Paesi Bassi, Spagna e Svezia.

Infine, al quarto appartengono: Francia, Irlanda, Italia, Regno Unito, Repubblica Ceca e Ungheria.

**Tabella 8: Centri dei cluster finali**

| | Cluster | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| grano | 218.76 | 170.73 | 186.48 | 234.40 |
| verdure | 76.24 | 98.70 | 100.88 | 151.21 |
| tuberi | 123.31 | 66.07 | 84.54 | 113.83 |
| frutta | 109.97 | 110.41 | 128.25 | 137.18 |
| carne | 120.83 | 82.26 | 107.38 | 143.08 |
| pesce | 13.79 | 14.86 | 24.18 | 24.00 |
| latte | 144.38 | 204.82 | 425.63 | 242.13 |
| uova | 8.16 | 6.14 | 10.65 | 18.25 |
| zucchero | 22.84 | 17.26 | 26.99 | 27.26 |
| grassi | 13.70 | 17.80 | 24.86 | 33.15 |
| surgelati | 256.96 | 84.75 | 54.86 | 11.22 |

Al gruppo 1 appartengono i paesi con un basso consumo di verdure, pesce, grassi e latte.

Un alto consumo di tuberi e prodotti surgelati. Un medio alto consumo di grano e carne.

Al gruppo 2 appartengono i paesi con basso consumo di grano, tuberi, carne, e uova.

Al gruppo 3 appartengono i paesi con alto consumo di pesce, latte.

Medio alto consumo di grano, zucchero e grassi.

Infine al gruppo 4 appartengono paesi con alto consumo in quasi tutte le categorie alimentari, ad esclusione dei prodotti surgelati.

Quindi un’alimentazione varia.

**Tabella 9: Distanze tra i centri dei cluster finali**

| Cluster | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | | 202.334 | 351.948 | 278.817 |
| 2 | 202.334 | | 226.809 | 144.472 |
| 3 | 351.948 | 226.809 | | 206.742 |
| 4 | 278.817 | 144.472 | 206.742 | |

Analizzando la tabella 9 notiamo che il secondo e il quarto gruppo presentano una distanza euclidea tra i centroidi minore, mentre il primo e il terzo gruppo presentano la distanza maggiore.

**Tabella 10: ANOVA**

| | Cluster | | Errore | | | |
|---|---|---|---|---|---|---|
| | Media dei quadrati | df | Media dei quadrati | df | F | Sig. |
| grano | 3896.855 | 3 | 1647.284 | 12 | 2.366 | .122 |
| verdure | 3771.442 | 3 | 1403.739 | 12 | 2.687 | .094 |
| tuberi | 2259.893 | 3 | 2466.815 | 12 | .916 | .462 |
| frutta | 668.095 | 3 | 1478.982 | 12 | .452 | .721 |
| carne | 3113.510 | 3 | 1057.286 | 12 | 2.945 | .076 |
| pesce | 102.360 | 3 | 205.879 | 12 | .497 | .691 |
| latte | 51792.464 | 3 | 1115.099 | 12 | 46.447 | .000 |
| uova | 131.503 | 3 | 18.460 | 12 | 7.124 | .005 |
| zucchero | 96.270 | 3 | 46.223 | 12 | 2.083 | .156 |
| grassi | 243.032 | 3 | 92.745 | 12 | 2.620 | .099 |
| surgelati | 18540.825 | 3 | 2415.340 | 12 | 7.676 | .004 |

Anche in questo caso visualizzando la tabella 10 notiamo che Latte e Surgelati sono le due variabili più significative per i cluster individuati; mentre lo Zucchero risulta il meno influente.

### 3.2) *Metodo Gerarchico*

Utilizzando lo stesso data set ho individuato la presenza di possibili cluster mediante SPSS ma, scegliendo l'approccio Hierarchical Cluster[6].

Le tecniche di analisi gerarchica si distinguono in:

- agglomerative, se procedono a una successione di fusioni delle n unità, a partire dalla situazione di base nella quale ognuna costituisce un gruppo a sé stante e fino allo stadio n-1 nel quale si forma un gruppo che le comprende tutte;
- divisive, o scissorie, quando l'insieme delle n unità, in n-1 passi, si ripartisce in gruppi che sono, a ogni passo dell'analisi, sottoinsiemi di un gruppo formato allo stadio di analisi precedente, e che termina con la situazione in cui ogni gruppo è composto da un'unità.

[6] Ho selezionato le variabili considerate nell'analisi e ho assegnato la variabile nominale Id(o paese) come "Label Cases by" .

Il software SPSS procede con il primo metodo.

Il risultato dell'analisi è dato dal dendrogramma e dalla tabella "programma di agglomerazione".

Il dendrogramma rappresenta una sintesi grafica del risultato ottenuto dall'analisi del cluster gerarchico, mentre la tabella 11 rappresenta una sintesi numerica.

**Tabella 11: Programma di agglomerazione**

| Stadio | Cluster accorpati | | Coefficienti | Stadio di formazione del cluster | | Stadio successivo |
|---|---|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | | Cluster 1 | Cluster 2 | |
| 1 | 6 | 12 | 4331.001 | 0 | 0 | 5 |
| 2 | 1 | 7 | 6299.526 | 0 | 0 | 6 |
| 3 | 4 | 14 | 6981.792 | 0 | 0 | 8 |
| 4 | 13 | 16 | 8460.043 | 0 | 0 | 5 |
| 5 | 6 | 13 | 11542.919 | 1 | 4 | 9 |
| 6 | 1 | 3 | 11972.142 | 2 | 0 | 10 |
| 7 | 5 | 11 | 15447.482 | 0 | 0 | 8 |
| 8 | 4 | 5 | 18421.521 | 3 | 7 | 11 |
| 9 | 6 | 9 | 20576.891 | 5 | 0 | 10 |
| 10 | 1 | 6 | 25989.598 | 6 | 9 | 12 |
| 11 | 4 | 15 | 37102.277 | 8 | 0 | 14 |
| 12 | 1 | 2 | 48554.131 | 10 | 0 | 13 |
| 13 | 1 | 8 | 55092.652 | 12 | 0 | 14 |
| 14 | 1 | 4 | 65450.388 | 13 | 11 | 15 |
| 15 | 1 | 10 | 92859.917 | 14 | 0 | 0 |

Osservando la tabella 11 vediamo che al primo livello i casi 6 e 12 (corrispondenti a Francia e Regno Unito) sono raggruppati, in quanto hanno la distanza minima.

Il nuovo cluster così creato riappare al 5 stadio.

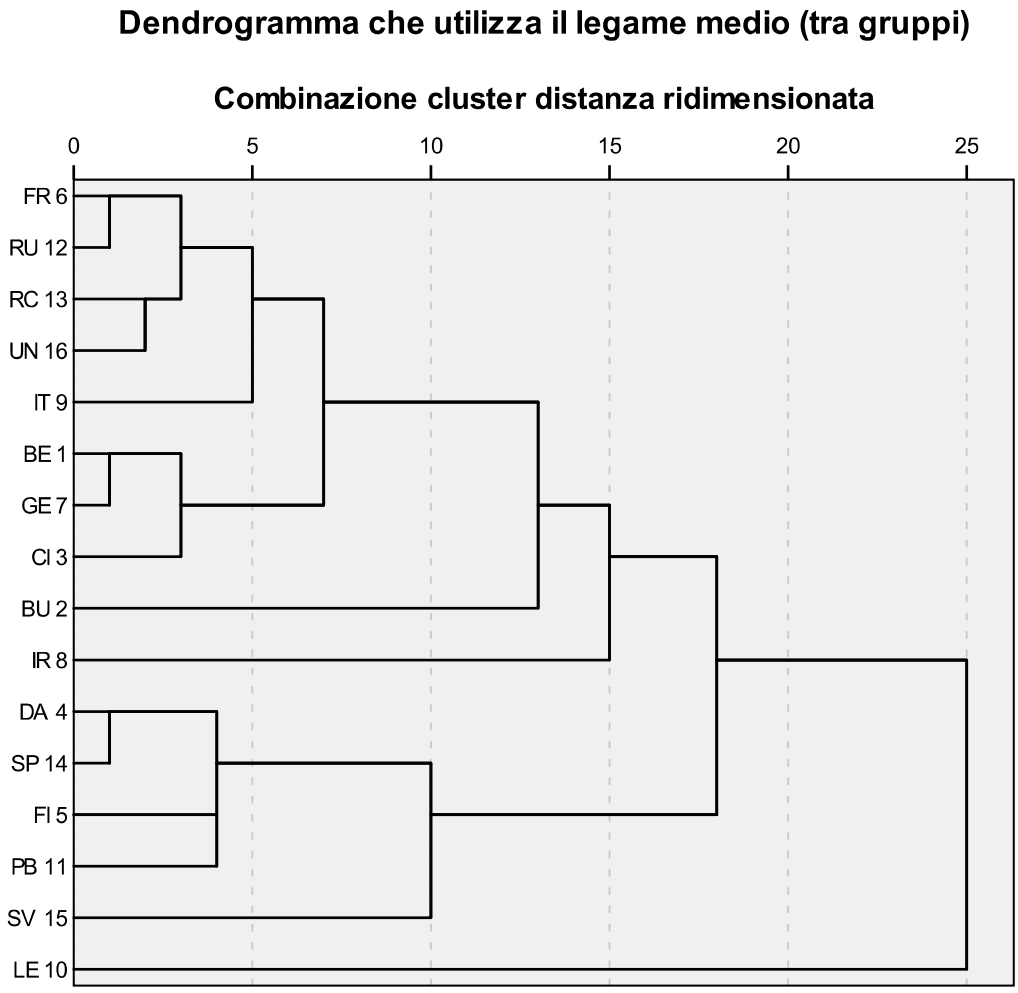
In corrispondenza della riga dello stadio 5, vediamo che i cluster 6 e 13 vengono raggruppati e notiamo anche che il cluster 13 è stato creato al quarto stadio ed è l'unione degli oggetti 13 e 16 (corrispondenti a Repubblica Ceca e Ungheria.), mentre il cluster 6 è il gruppo creato al primo livello di agglomerazione.

Quando le osservazioni sono tante, la lettura della precedente tabella può essere complicata e si preferisce la rappresentazione grafica data dal dendrogramma.

È importante non sottovalutare l'informazione data nella quarta colonna che fornisce la distanza tra gli oggetti.

Per poter considerare tutte le osservazione in un modo più facile e immediato considero il dendrogramma della figura 1.

**figura 1.**



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.
Se tagliamo lo schema gerarchico prima del valore 20, troviamo tre gruppi. In uno appartiene la Lettonia; in un secondo gruppo appartengono: Danimarca, Spagna, Finlandia, Paesi Bassi, Svezia.
Nel terzo gruppo troviamo: Francia, UK, Repubblica Ceca, Ungheria, Italia, Belgio, Germania, Cipro, Bulgaria e Irlanda.
Se decidessimo di tagliare prima del valore 15, notiamo che si forma un ulteriore gruppo formato dall'Irlanda.

Il problema principale dei metodi di classificazione gerarchica consiste nel definire il criterio di raggruppamento di due oggetti, cosa che equivale a definire una distanza tra oggetti.
Tutti gli algoritmi di classificazione gerarchica si sviluppano in modo tale da ricercare ad ogni tappa le due classi più vicine, le si riunisce e si continua fino a che non si abbi una sola classe.
Con il programma SPSS possiamo scegliere il metodo agglomerativo e il tipo di distanza da utilizzare.
Il metodo utilizzato di default da SPSS è "between-groups linkage[7]".
La distanza che ho utilizzato per il precedente dendrogramma è "Squared Euclidean distance[8]"
È possibile anche scegliere il tipo di standardizzazione delle variabili.

Il dendrogramma riportato nella figura.2 è stato ottenuto scegliendo la standardizzazione tramite la devianza standard.
Si ottengono risultati differenti, in particolare notiamo che se replichiamo i tagli precedenti.
Nel primo caso abbiamo quattro gruppi. Uno formato da Irlanda, Ungheria e Repubblica Ceca.

---

$$D(X,Y) = \frac{1}{N_X \times N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d(x_i, y_j);$$

[7] in formula: $x_i \in X,\ y_i \in Y,$

[8] in formula: $d(p,q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_i - q_i)^2 + \ldots + (p_n - q_n)^2.$

Un secondo gruppo formato da Italia e Spagna.

Un terzo gruppo formato da Bulgaria e Cipro.

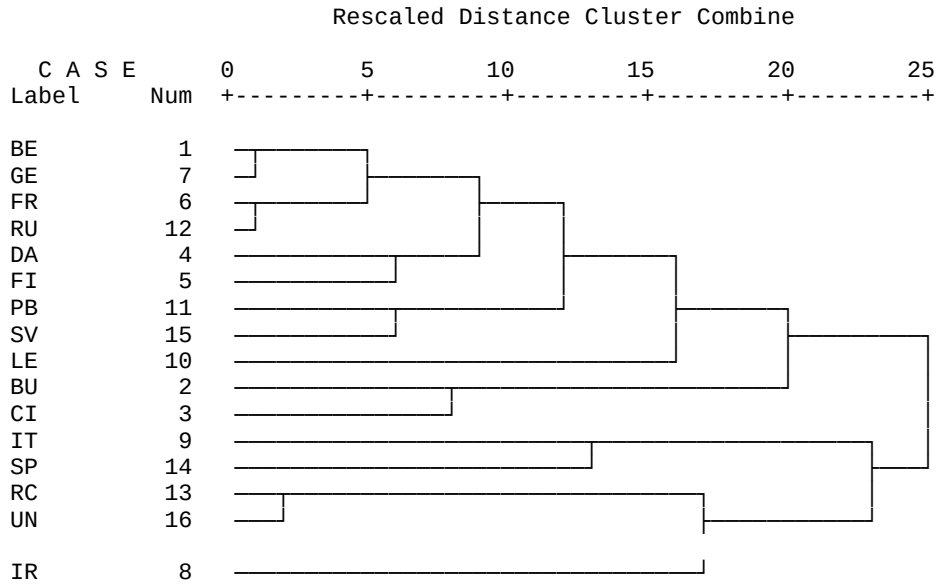
Infine, un gruppo formato da Belgio, Germania, Francia, UK, Danimarca, Finlandia, Paesi Bassi, Svezia e Lettonia.

Se invece consideriamo il taglio prima del valore 15 abbiamo ben 6 gruppi.

Il primo formato da: Belgio, Germania, Francia, UK, Danimarca, Finlandia, Paesi Bassi e Svezia.

Il secondo formato dalla Lettonia; il terzo da Bulgaria e Cipro; il quarto da Italia e Spagna; il quinto da Repubblica Ceca e Ungheria. Infine, l'ultimo contenente solo l'Irlanda.

**Figura 2**: Dendrogram using Average Linkage (Between Groups)



## Conclusioni

Dopo aver applicato i due metodi nei vari casi è possibile trarre delle conclusioni.

Si evince come con il metodo delle k-medie in entrambi i casi sia quando si è scelto di fissare il numero di cluster a 3, e sia nel caso in cui si è scelto di fissarlo a 4, ho riscontrato un gruppo formato solo dallo stato della Lettonia.

Inoltre si nota anche come il primo gruppo del primo caso coincida con il terzo gruppo del secondo caso, ad esclusione dell'Irlanda.

Quindi Danimarca, Finlandia, Paesi Bassi, Spagna e Svezia, li ritroviamo accorpati in un unico cluster, possiamo dedurre che presentano un consumo alimentare simile.

Può risultare anomalo come il consumo alimentare degli spagnoli sia associato a quello dei paesi Scandinavi ma, ovviamente questo risultato è frutto di determinati dati e quindi può variare se viene effettuata una diversa indagine.

Con il metodo gerarchico in entrambi i casi troviamo un gruppo formato solo dalla Lettonia e uno solo dall'Irlanda.

Quindi pur considerando le variabili in maniera differente questi due gruppi restano costanti.

Per decidere sulla strategia da adottare, si deve tener conto dell'obiettivo analitico.

In generale le tecniche non gerarchiche sono più informative di quelle gerarchiche perché danno risultati anche intermedi e vari indici per la misura della qualità del risultato.

I metodi gerarchici risentono di più della presenza di errori di misura e di altre fonti di variabilità spuria presenti nelle misure di prossimità, e sono scombinati dalla presenza di dati anomali[9].

I metodi gerarchici risentono di più della presenza di errori di misura e di altre fonti di variabilità spuria presenti nelle misure di prossimità, e sono scombinati daala presenza di dati anomali.

Per grandi linee, si può dire che, se si cercano gruppi di unità statistiche caratterizzate da alta omogeneità interne (nel senso di strettezza dei legami tra entità appartenenti a un gruppo), le tecniche gerarchiche sono meno efficaci delle tecniche non gerarchiche.

[9] Fabbris L. (1997, 8.3.1)

## Bibliografia:

FABBRIS L. (1997) Statistica multivariata: analisi esplorativa dei dati, McGraw-Hill, Milano.

## Sitografia:

http://www.spss.it

Software SPSS.

http://www.efsa.europa.eu/it/datexfoodcdb/datexfooddb.htm

Statistiche di consumo cronico degli alimenti (per Paese, indagine e classe di età) reported in grams/day.