# Metodologie di Analisi per dati multivariati

L’analisi statistica *multivariata* studia le proprietà di un insieme di *p variabili* rilevate su un insieme di elementi $I = \{I_1, I_2, \ldots\ldots, I_n\}$ *(prodotti, marchi, aziende, individui, …….)*

# *Matrice di dati multivariati*

I dati consistono in una matrice in cui *p* variabili vengono rilevate su *n* di soggetti, oggetti o altre entità di interesse. Tali dati possono essere rappresentati da una matrice *X* ovvero la *matrice dei dati multivariati*

- X = $\begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & & & & \cdot \\ x_{i1} & \cdot & \cdot & \cdot & x_{ip} \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}$

# La segmentazione del database

- La segmentazione del database ha come obiettivo la suddivisione  di un database in segmenti che contengono dei records che hanno in comune un determinato numero di proprietà, e che per questo motivo vengono considerati fra di loro omogenei.

# La segmentazione del database

- *Segmentazione* descrive l'operazione di data mining,
  - *segmenti* o *clustering* descrivono i gruppi risultanti dai records contenuti nel database.
  - I segmenti devono essere caratterizzati da un'elevata omogeneità interna (internamente al segmento) e da un'elevata eterogeneità esterna (fra i segmenti).
  - Per omogeneità interna intendiamo che i records sono il più possibile "vicini" l'uno all'altro.

# Fasi del processo di analisi dei clusters

- La *cluster analysis* tradizionale consiste in quattro semplici fasi strettamente collegate tra loro, la procedura può richiedere una serie di tentativi e di ripetizioni dei vari passaggi che vengono di seguito sintetizzati:
- scelta delle variabili;
- scelta dell'algoritmo di *clustering*;
- validazione dei *clusters*;
- interpretazione dei risultati.

# Fasi del processo di analisi dei clusters

1. Scelta delle unità di osservazione;
2. Scelta delle variabili;
3. Omogeneizzazione scale di misura;
4. Scelta della **misura di similarità** o diversità tra unità statistiche;
5. numero di gruppi;
6. Scelta del **criterio di raggruppamento**;
7. Scelta dell'**algoritmo di classificazione** ;
8. Interpretazione dei risultati ottenuti.

# Fasi del processo di analisi dei clusters in DM

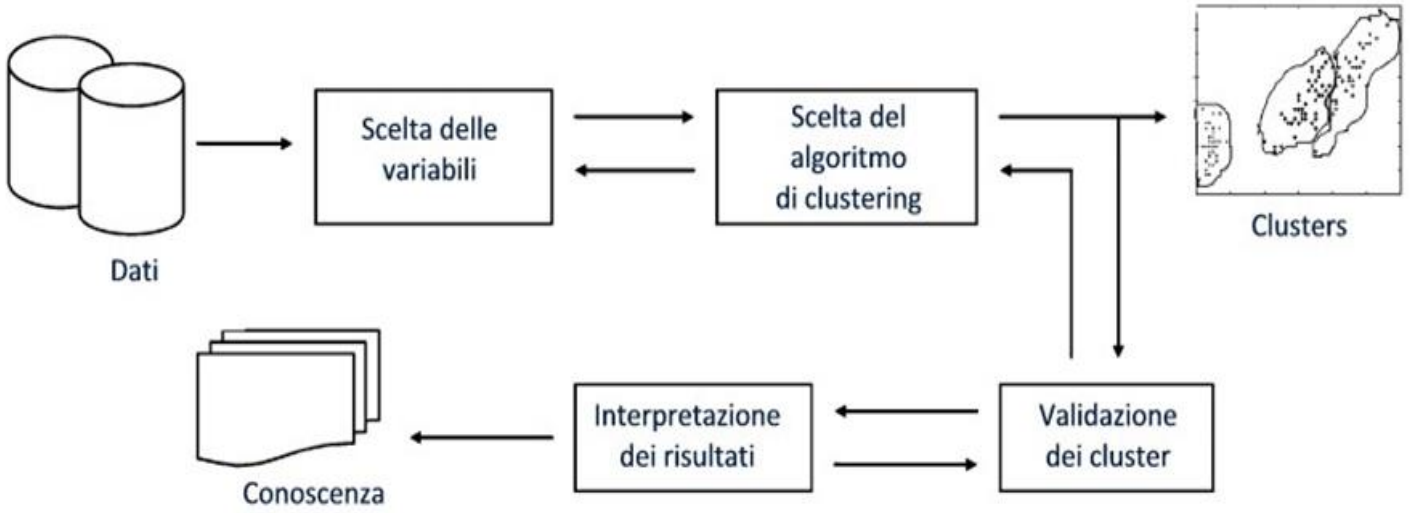



Figura 1.1- Le fasi di una procedura di clustering
Fonte: Xu, Wunsch (2005)

# Gli algoritmi di cluster analysis:

- Esistono due grandi famiglie di classificazione.
- *Gli algoritmi di classificazione gerarchica,* caratterizzati da una procedura iterativa che genera una gerarchia nelle partizioni
- *Gli algoritmi di classificazione diretta,* che si basano per la formazione delle classi sulla minimizzazione di una funzione obiettivo assegnata.

- Metodi gerarchici
  - Agglomerativi
  - Divisivi
- Metodi non gerarchici
  - Metodi basati sull'errore quadratico (es. k-means)
  - Metodi basati sui modelli mistura
  - Metodi basati sulla teoria dei grafi
  - Altri metodi

# I metodi non gerarchici

- I metodi di classificazione non gerarchici (a volte detti anche *partitioning methods*) consentono di ottenere un'unica partizione degli $n$ elementi in $g$ gruppi $(g < n)$: l'obiettivo è trovare una classificazione che soddisfi determinati criteri e che sia formata da un numero di gruppi $g$ fissato a priori dal ricercatore.
- Solitamente tali metodi prevedono la specificazione esplicita di una funzione obiettivo, che viene spesso espressa in termini di scomposizione della devianza.

# I metodi non gerarchici

- Il processo di classificazione diventa un problema di ottimizzazione, dove si ricerca la partizione con la maggior omogeneità nei gruppi: questo significa che, almeno teoricamente, è possibile formalizzare il meccanismo di allocazione delle unità ai gruppi.
- Questo non è l'unico vantaggio di questa tipologia di metodi di *clustering* infatti nei metodi non gerarchici viene meno il vincolo che tutte le coppie di unità che risultano tra loro unite ad un determinato livello di aggregazione gerarchica non possono più essere separate a livelli successivi.

# I metodi non gerarchici

- Uno dei punti critici sta nel fatto che il numero di modi in cui è possibile suddividere $n$ unità in $g$ clusters non sovrapposti è notevolmente grande anche per valori piccoli: ciò significa che spesso nella pratica non è possibile elencare tutte le partizioni possibili . Di conseguenza, gli algoritmi di uso operativo, non possono mirare al raggiungimento di un ottimo globale della funzione obiettivo, ma devono necessariamente fare riferimento a criteri meno vincolanti.
- Anche per questi metodi è possibile abbozzare una procedura generale che si può sintetizzare nelle seguenti fasi.

# I metodi non gerarchici: fasi

- Fase 1: si sceglie una classificazione iniziale con un numero di *cluster* prefissato.
- Fase 2: si calcola la variazione nella funzione obiettivo causata dallo spostamento di ciascuna unità dal *cluster* in cui si trova ad un altro e si sceglie per ciascuna unità il *cluster* che garantisce la maggiore omogeneità nei gruppi.
- Fase 3: si ripete la fase 2 finché non viene verificata una regola di arresto prestabilita.

# I metodi non gerarchici

- I metodi non gerarchici hanno una struttura di tipo iterativo, che per un valore $g$ prefissato rende l'algoritmo veloce;
- inoltre non è necessario costruire la matrice delle distanze completa.
- Questi vantaggi rendono questi metodi ideali nel caso di *dataset* con un numero elevato di unità statistiche e nei casi nei quali lo studio vuole evidenziare le caratteristiche dei gruppi e non delle singole unità.

# I metodi non gerarchici: osservazioni

- L'uso di questa famiglia di algoritmi nella prassi applicativa è ampiamente diffuso, in particolare per la possibilità di analizzare grandi masse di dati a "costi" bassi in termini di risorse di hardware e tempo.
- Essi sono caratterizzati da una procedura iterativa, che ammette nelle varie fasi una riallocazione degli elementi già clusterizzati, in modo da consentire un progressivo miglioramento delle partizioni ottenute. Tali metodi assumono che il numero desiderato di gruppi sia fissato a priori; l'ipotesi non è limitante in quanto vi è la possibilità di ripetere l'analisi più volte cambiando il numero di clusters richiesti

# I metodi non gerarchici:osservazioni

- La premessa dell'analisi è la fissazione di tre criteri distinti:
- Un criterio per la scelta dei centri dei clusters iniziali;
- un criterio per allocare gli elementi nei clusters iniziali;
- un criterio per l'uscita dalla procedura iterativa.

# I metodi basati sull'errore quadratico: k-means

- Il criterio dell'errore quadratico è la tecnica di *clustering* non gerarchica più intuitiva e più utilizzata. In particolare il metodo delle $k$-medie è il più semplice algoritmo che sfrutta questo criterio ed è implementato nei principali *packages* statistici.
- L'algoritmo è caratterizzato da una procedura iterativa che consiste nei seguenti passi.

# I metodi basati sull'errore quadratico: k-means

- Fase 1. Si scelgono $g$ "poli" (detti anche semi, seeds, o punti origine) iniziali, punti nello spazio $p$-dimensionale che costituiscono i centroidi dei gruppi della partizione iniziale. Questi poli possono essere individuati tramite metodi differenti, generalmente in maniera tale che siano sufficientemente distanti l'uno dall'altro. Quindi si ripartiscono le unità statistiche allocando ciascuna di esse al *cluster* il cui polo risulta più vicino, costituendo una partizione iniziale formata da $g$ gruppi.

# I metodi basati sull'errore quadratico: k-means

- Fase 2. Si calcola per ogni unità la distanza dai centroidi di tutti i $g$ clusters e ogni unità viene assegnata al cluster del centroide più vicino, qualora non vi fosse già allocata. In caso di riallocazione di un'unità si ricalcola il centroide sia del nuovo che del vecchio gruppo di appartenenza.
- Fase 3. Si ripete la fase 2 fino a che l'algoritmo converge, cioè fino a quando non si verifica alcuna riallocazione rispetto all'iterazione precedente.

Oss.

- Le fasi della metodologia illustrata prevedono il calcolo ripetuto della distanza tra ogni unità ed i centroidi dei $g$ *clusters*: per tale operazione viene solitamente utilizzata la distanza euclidea.
- Tale procedura appartiene alla classe di algoritmi di classificazione che adottano la tecnica denominata “ordinamento rispetto al centroide più vicino” (*nearest centroid sorting).*

# Esempio

| | Country | RedMeat | WhitM | Eggs | Milk | Fish | Frveg |
|---|---|---|---|---|---|---|---|
| 1 | Finland | 9.5 | 4.9 | 2.7 | 33.7 | 5.8 | 1.4 |
| 2 | France | 18 | 9.9 | 3.3 | 19.5 | 5.7 | 6.5 |
| 3 | Italy | 9 | 5.1 | 2.9 | 13.7 | 3.4 | 6.7 |
| 4 | Spain | 7.1 | 3.4 | 3.1 | 8.6 | 7 | 7.2 |
| 5 | Sweden | 9.9 | 7.8 | 3.5 | 24.7 | 7.5 | 2 |
| 6 | UK | 17.4 | 5.7 | 4.7 | 20.6 | 4.3 | 3.3 |

The SAS System

La procedura FASTCLUS
Replace=FULL Radius=0 Maxcluster=3 Maxiter=1

Semi iniziali

| Cluster | RedMeat | WhitM | Eggs | Milk | Fish | Frveg |
|---|---|---|---|---|---|---|
| 1 | 9.50000000 | 4.90000000 | 2.70000000 | 33.70000000 | 5.80000000 | 1.40000000 |
| 2 | 18.00000000 | 9.90000000 | 3.30000000 | 19.50000000 | 5.70000000 | 6.50000000 |
| 3 | 7.10000000 | 3.40000000 | 3.10000000 | 8.60000000 | 7.00000000 | 7.20000000 |

Oss. 1 Finlandia 2 Francia 3 Spagna

- calcola le distanze di tutti gli altri paesi (Svezia, Italia, UK) da:

1 Finlandia 2 Francia 3 Spagna

- considera la distanza minima quindi forma i cluster

1 Finlandia Svezia

2 Francia UK

3 Spagna Italia

VIEWTABLE: Sasuser.Clusterng

| | Country | RedMeat | WhitM | Eggs | Milk | Fish | Frveg | Cluster |
|---|---|---|---|---|---|---|---|---|
| 1 | Finland | 9.5 | 4.9 | 2.7 | 33.7 | 5.8 | 1.4 | 1 |
| 2 | France | 18 | 9.9 | 3.3 | 19.5 | 5.7 | 6.5 | 2 |
| 3 | Italy | 9 | 5.1 | 2.9 | 13.7 | 3.4 | 6.7 | 3 |
| 4 | Spain | 7.1 | 3.4 | 3.1 | 8.6 | 7 | 7.2 | 3 |
| 5 | Sweden | 9.9 | 7.8 | 3.5 | 24.7 | 7.5 | 2 | 1 |
| 6 | UK | 17.4 | 5.7 | 4.7 | 20.6 | 4.3 | 3.3 | 2 |

Oss. 1 Finlandia, Svezia 2 Francia UK 3 Spagna Italia

Ricalcola i centri

Oss. 1 Finlandia, Svezia Centroide(RedMeat)=(9.5+9.9)/2=9.7 ecc.

2 Francia UK Centroide(RedMeat)=(18+17.4)/2=17.7 ecc.

3 Spagna Italia Centroide(RedMeat)=(7.1+9)/2=8.05 ecc.

**Medie dei cluster**

| Cluster | RedMeat | WhitM | Eggs | Milk | Fish | Frveg |
|---|---|---|---|---|---|---|
| 1 | 9.70000000 | 6.35000000 | 3.10000000 | 29.20000000 | 6.65000000 | 1.70000000 |
| 2 | 17.70000000 | 7.80000000 | 4.00000000 | 20.05000000 | 5.00000000 | 4.90000000 |
| 3 | 8.05000000 | 4.25000000 | 3.00000000 | 11.15000000 | 5.20000000 | 6.95000000 |

- ricalcola le distanze dei paesi (Finlandia, Svezia, Francia, UK, Spagna, Italia) con i 3 nuovi centroidi

- considera la distanza minima quindi forma i cluster

Oss: Nell’esempio non c’è riallocazione

# Esempio

- base di dati contenente una descrizione di 183 Paesi relativamente a variabili: economiche, politiche, demografiche e religiose.

# Esempio

- *country*: nome con cui viene identificata ciascuna delle Nazioni;
- - *gdp* (*Gross Domestic Product*): Prodotto Interno Lordo della nazione, espressa in migliaia di USD secondo valutazione risalente all'anno 2010;
- - *pres*: classificazione del sistema governativo vigente, a seconda della presenza di una o più cariche governative;
- - *pop*: numero di cittadini residenti;
- - *christians*: proporzione di cittadini di orientamento religioso cristiano;
- - *muslims*: proporzione di cittadini di orientamento religioso islamico;
- - *region*: macroregione di appartenenza.

# Proc cluster (analisi dei cluster gerarchica)

proc cluster data= scmm.nazioni2 method=ward ccc pseudo out=scmm.clustergstd;

var gdp pop muslims christians pres2;

id country;

run;

OSS:

Lavoro su variabili standardizzate

Metodo di Ward

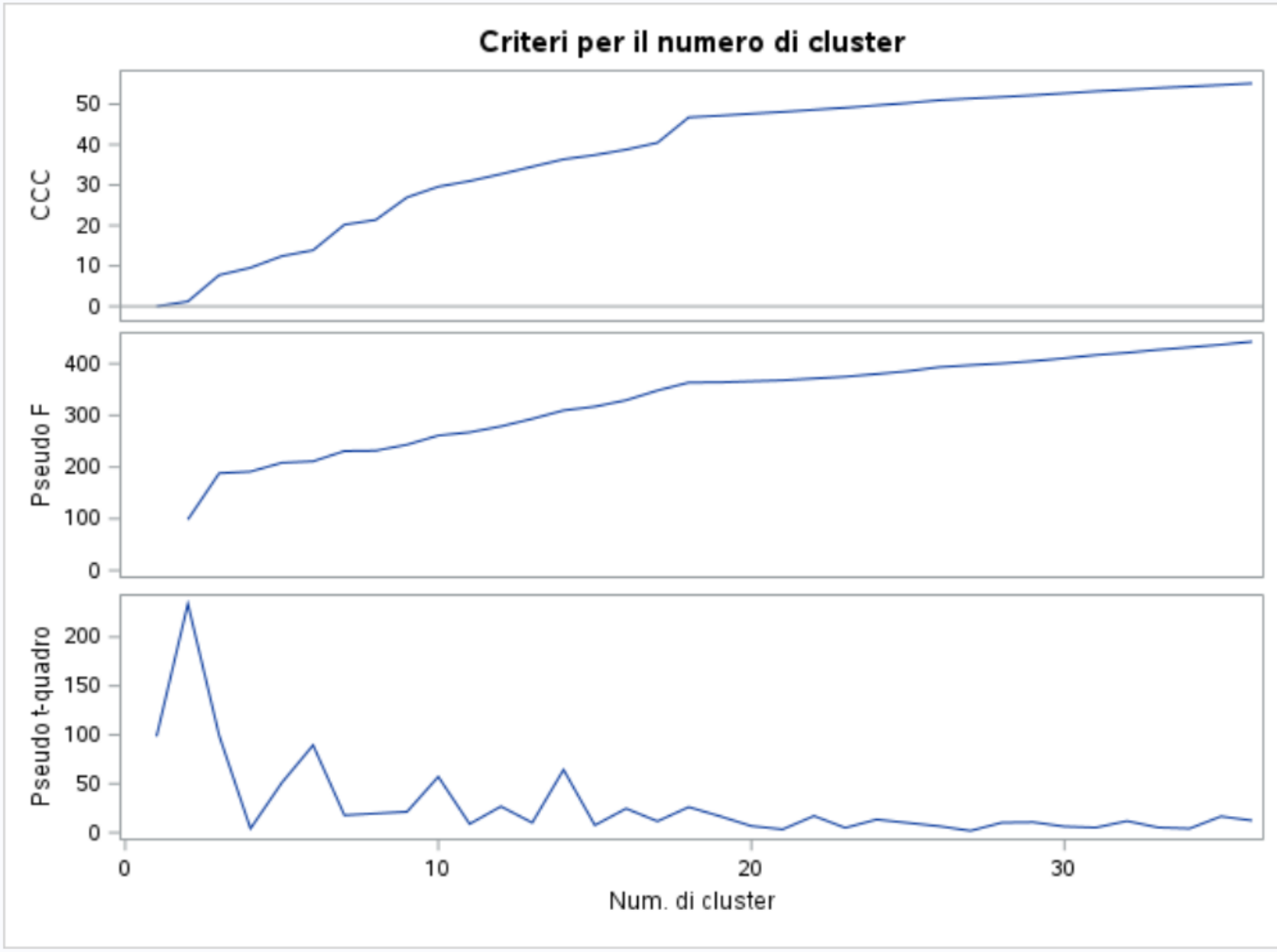
Salvo output su file scmm.clustergstd

Scelgo numero cluster

| Cronologia dei cluster | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Numero di cluster | Cluster uniti | | Freq | R-quadro semiparziale | R-quadro | R-quadro atteso approssimato | Criterio di clusterizzazione cubica | Statistica pseudo F | Pseudo t-quadro | Legame |
| 182 | Tonga | Vanuatu | 2 | 0.0000 | 1.00 | . | . | 84E5 | . | |
| 181 | Comoros | Maldives | 2 | 0.0000 | 1.00 | . | . | 7E5 | . | |

.....

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 18 | CL61 | CL32 | 24 | 0.0016 | .974 | .845 | 46.7 | 363 | 26.1 | |
| 17 | CL29 | CL27 | 11 | 0.0029 | .971 | .839 | 40.5 | 348 | 11.6 | |
| 16 | CL18 | CL26 | 29 | 0.0038 | .967 | .833 | 38.8 | 329 | 24.6 | |
| 15 | CL42 | CL21 | 10 | 0.0038 | .964 | .827 | 37.4 | 317 | 7.5 | |
| 14 | CL35 | CL28 | 56 | 0.0038 | .960 | .820 | 36.4 | 310 | 64.3 | |
| 13 | CL17 | CL25 | 13 | 0.0058 | .954 | .812 | 34.5 | 293 | 10.1 | |
| 12 | CL48 | CL16 | 35 | 0.0067 | .947 | .803 | 32.7 | 279 | 26.7 | |
| 11 | CL15 | Japan | 11 | 0.0076 | .940 | .793 | 31.0 | 267 | 8.8 | |
| 10 | CL22 | CL19 | 40 | 0.0081 | .931 | .782 | 29.6 | 261 | 57.2 | |
| 9 | CL20 | CL13 | 27 | 0.0135 | .918 | .770 | 27.0 | 243 | 21.3 | |
| 8 | China | India | 2 | 0.0153 | .903 | .755 | 21.4 | 232 | . | |
| 7 | CL9 | CL30 | 38 | 0.0153 | .887 | .737 | 20.2 | 231 | 17.8 | |
| 6 | CL14 | CL11 | 67 | 0.0309 | .856 | .715 | 13.9 | 211 | 89.4 | |
| 5 | CL10 | CL6 | 107 | 0.0325 | .824 | .685 | 12.4 | 208 | 50.9 | |
| 4 | CL8 | United States | 3 | 0.0618 | .762 | .640 | 9.55 | 191 | 4.0 | |
| 3 | CL12 | CL7 | 73 | 0.0858 | .676 | .567 | 7.73 | 188 | 99.2 | |
| 2 | CL3 | CL5 | 180 | 0.3242 | .352 | .330 | 1.21 | 98.4 | 234 | |
| 1 | CL2 | CL4 | 183 | 0.3522 | .000 | .000 | 0.00 | . | 98.4 | |


Criteri per il numero di cluster
CCC
50
40
30
20
10
0
Pseudo F
400
300
200
100
0
Pseudo t-quadro
200
150
100
50
0
0
10
20
30
Num. di cluster

# Proc fastclus (analisi dei cluster non gerarchica)

proc fastclus data= scmm.nazioni2 maxclusters=10 out=scmm.clusterstdng;

var gdp pop muslims christians pres2;

id country;

run;

OSS:

Lavoro su variabili standardizzate

Salvo output su file scmm.clusterstdng

Utilizzo numero cluster dalla cluster gerarchica

**La procedura FASTCLUS**
**Replace=FULL Radius=0 Maxclusters=10 Maxiter=1**

| Semi iniziali | | | | | |
|---|---|---|---|---|---|
| **Cluster** | **gdp** | **pop** | **muslims** | **christians** | **pres2** |
| **1** | -0.238178249 | -0.099047204 | 2.126344779 | -1.478106720 | 0.000000000 |
| **2** | -0.233089794 | 0.377538476 | 0.276712155 | -0.034003177 | 0.000000000 |
| **3** | 0.216421708 | -0.102521977 | -0.683728823 | -1.309143966 | 0.000000000 |
| **4** | -0.253324789 | -0.227898226 | -0.695128716 | 1.088015115 | 1.000000000 |
| **5** | 9.382989739 | 2.019039538 | -0.649529144 | 0.736889390 | 1.000000000 |
| **6** | 0.419862243 | 1.522166543 | 0.843856827 | -1.126980996 | 1.000000000 |
| **7** | 2.658544355 | 0.664081289 | -0.692278742 | -1.390985300 | 0.000000000 |
| **8** | 1.810129665 | 0.330491339 | -0.566879921 | 0.494005431 | 0.000000000 |
| **9** | 7.544339438 | 9.552948587 | -0.652379117 | -1.258983148 | 0.000000000 |
| **10** | 2.854447893 | 8.392967693 | -0.347431982 | -1.319704138 | 0.000000000 |

| Riepilogo dei cluster | | | | | | |
|---|---|---|---|---|---|---|
| Cluster | Frequenza | Deviazione std RMS | Distanza massima da seme a osservazione | Raggio superato | Cluster più vicino | Distanza tra centroidi dei cluster |
| 1 | 39 | 0.2911 | 1.2104 | | 3 | 2.0032 |
| 2 | 23 | 0.3198 | 1.0981 | | 3 | 1.1409 |
| 3 | 15 | 0.2960 | 1.4193 | | 2 | 1.1409 |
| 4 | 90 | 0.2636 | 1.0776 | | 2 | 1.2089 |
| 5 | 1 | . | 0 | | 7 | 7.2515 |
| 6 | 2 | 0.4055 | 1.2822 | | 2 | 1.8372 |
| 7 | 1 | . | 0 | | 8 | 2.5382 |
| 8 | 10 | 0.3647 | 1.2731 | | 4 | 1.4117 |
| 9 | 1 | . | 0 | | 10 | 4.8412 |
| 10 | 1 | . | 0 | | 9 | 4.8412 |

| Statistiche per variabili | | | | |
|---|---|---|---|---|
| Variabile | STD totale | Entro STD | R-quadro | RSQ/(1-RSQ) |
| gdp | 1.00000 | 0.15372 | 0.977539 | 43.522369 |
| pop | 1.00000 | 0.20267 | 0.960957 | 24.612779 |
| muslims | 1.00000 | 0.28469 | 0.922961 | 11.980410 |
| christians | 1.00000 | 0.26553 | 0.932979 | 13.920785 |
| pres2 | 0.45733 | 0.44195 | 0.112322 | 0.126535 |
| OVER-ALL | 0.91751 | 0.28691 | 0.907054 | 9.758969 |

| | |
|---|---|
| Stat. pseudo F = | 187.59 |

| | |
|---|---|
| R-quadro atteso globale approssimato = | 0.68302 |

| | |
|---|---|
| Criterio di clust. cubica = | 37.020 |

**WARNING: i due valori indicati sopra non sono validi per variabili correlate.**

| Medie dei cluster | | | | | |
|---|---|---|---|---|---|
| Cluster | gdp | pop | muslims | christians | pres2 |
| 1 | -0.182794337 | -0.084973887 | 1.780767256 | -1.344006073 | 0.179487179 |
| 2 | -0.256396993 | -0.169960051 | -0.043600054 | -0.191143130 | 0.217391304 |
| 3 | -0.147848974 | -0.116900075 | -0.218233196 | -1.301927848 | 0.066666667 |
| 4 | -0.199675558 | -0.195437530 | -0.644335860 | 0.839997738 | 0.400000000 |
| 5 | 9.382989739 | 2.019039538 | -0.649529144 | 0.736889390 | 1.000000000 |
| 6 | 0.152406723 | 1.186984580 | 0.672858433 | -0.682133744 | 1.000000000 |
| 7 | 2.658544355 | 0.664081289 | -0.692278742 | -1.390985300 | 0.000000000 |
| 8 | 1.046950991 | 0.356293529 | -0.618749433 | 0.533870081 | 0.200000000 |
| 9 | 7.544339438 | 9.552948587 | -0.652379117 | -1.258983148 | 0.000000000 |
| 10 | 2.854447893 | 8.392967693 | -0.347431982 | -1.319704138 | 0.000000000 |

| Deviazioni standard dei cluster | | | | | |
|---|---|---|---|---|---|
| Cluster | gdp | pop | muslims | christians | pres2 |
| 1 | 0.1484596197 | 0.2968324691 | 0.3551738789 | 0.1903657858 | 0.3887764120 |
| 2 | 0.0357459115 | 0.1417155844 | 0.4385851460 | 0.3460570607 | 0.4217411678 |
| 3 | 0.1563722559 | 0.1870402105 | 0.5416027585 | 0.1361404826 | 0.2581988897 |
| 4 | 0.1275271351 | 0.1201089104 | 0.0925624817 | 0.2557355795 | 0.4926424964 |
| 5 | . | . | . | . | . |
| 6 | 0.3782392238 | 0.4740188786 | 0.2418282474 | 0.6291090174 | 0.0000000000 |
| 7 | . | . | . | . | . |
| 8 | 0.3785018694 | 0.3825986858 | 0.0868896550 | 0.4359055820 | 0.4216370214 |
| 9 | . | . | . | . | . |
| 10 | . | . | . | . | . |

# I metodi basati sulla teoria dei grafi

- La teoria dei grafi costituisce uno strumento molto potente, in quanto, pur nella relativa semplicità della nozione di grafo, consente di descrivere in maniera strutturata e formalizzata matematicamente problemi e situazioni correnti di una certa complessità. In questo caso i nodi del grafo corrispondono alle unità e gli archi rappresentano la prossimità tra ogni coppia di punti.
- Il *clustering identification via connectivity kernels* (CLICK), l'algoritmo più conosciuto in questo ambito prevede come primo step la costruzione del MST (*minimal spanning tree*) dei dati; quindi si eliminano gli archi di maggior lunghezza per generare i cluster.

# I metodi basati sulla teoria dei grafi

- Un algoritmo sviluppatosi recentemente, ma molto utilizzato, è Chameleon: questo algoritmo procede all'eliminazione di un arco se entrambi i vertici non sono compresi tra i $k$ punti più vicini relativi ad entrambi.
- Altri algoritmi utilizzati sono *Chameleon*, *Delaunay triangulation graph* (DTG), *highly connected subgraphs* (HCS), *cluster affinity search technique* (CAST).

# I metodi basati sui modelli mistura

- L'assunzione di fondo di queste tipologie di *clustering* è che i dati siano generati da diverse distribuzioni dello stesso tipo e l'obiettivo è di identificare i parametri di ognuna e il loro numero; solitamente si assume che le componenti individuali della densità mistura siano Normali (Jain, 1999). La definizione di questi metodi si basa sul concetto di modello mistura: per questo si utilizza il termine *model-based clustering*. L'approccio più tradizionale prevede di ottenere, in maniera iterativa, una stima di massima verosimiglianza dei vettori dei parametri delle densità; più recentemente viene utilizzato l'algoritmo EM (Expectation Maximization).

# Altri metodi

- Molto utilizzati in ambito spaziale sono gli algoritmi basati sulla densità: questi considerano i *clusters* come regioni di spazio ad alta densità, separate tra loro da regioni a bassa densità.
- Gli algoritmi basati sulla densità analizzano la densità attorno ad ogni osservazione e la classificano come "sufficientemente densa" se il numero di osservazioni prese.
- L'algoritmo più utilizzato è il Density Based Spatial Clustering of Applications with Noise (DBSCAN), che separa in diversi *clusters* le regioni con densità sufficientemente elevata: questo avviene andando ad osservare per ogni punto lo spazio determinato da un raggio fissato, per vedere se i punti presenti in tale area superano il numero imposto a priori come soglia. (Han, 2001).

# Altri metodi

- Un altro metodo è il Density-based Clustering (DENCLUE): questo algoritmo si basa sull'idea che ogni punto possa essere modellato utilizzando una funzione matematica chiamata "*influence function*", che descrive l'impatto dell'osservazione sui punti vicini. Sommando queste funzioni è possibile trovare la densità dello spazio dei punti ed i *clusters* possono essere trovati determinando matematicamente i massimi locali della funzione di densità così costruita.

# Altri metodi

- Vi sono metodi basati sulle tecniche di ricerca combinatoria, che vedono il *clustering* come una sorta di problema di ottimizzazione: queste tecniche hanno come obiettivo principale la ricerca dell'ottimo (o di una sua approssimazione) in un problema di ottimizzazione combinatoria.
- Considerato un *dataset* di punti $\boldsymbol{x}_j \in \mathbb{R}^d, j = 1, \dots, N$, l'algoritmo di *clustering* ha lo scopo di organizzare questi punti in $K$ gruppi $\{C_1, \dots, C_K\}$ in modo da ottimizzare una qualche funzione obiettivo.
- Con le tecniche di ricerca utilizzate solitamente nella ricerca combinatoria non viene garantita l'ottimalità della partizione ottenuta, per cui si ricorre a metodi di ricerca più complessi, quali i *genetically guided algorithm* (GGA), il *tabu search* (TS) *clustering* e il *deterministic annealing* (DA) *clustering* (Xu, 2005

# Altri metodi

- Esistono anche metodi basati sulle reti neurali, modelli matematici/informatici di calcolo basati sulle reti neurali biologiche. Gli algoritmi più utilizzati in questo campo sono: *learning vector quantization* (LVQ) e *self-organizing feature map* (SOFM).

- NB: Si rimanda a (Han, 2001) e Xu (2005) per una rassegna su i principali metodi di clustering.