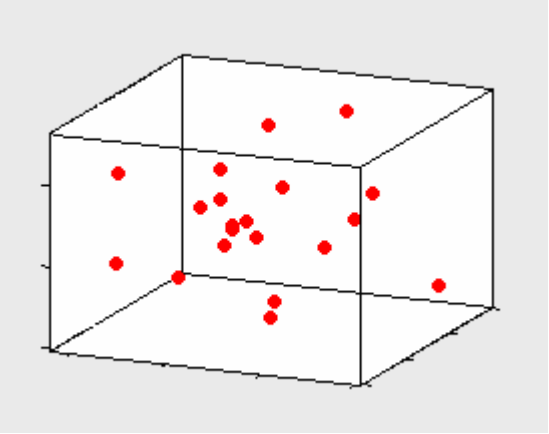
STATISTICA DESCRITTIVA - SCHEDA N. 6
CLUSTER ANALYSIS

La statistica multivariata

Nelle schede precedenti abbiamo visto come si rappresentano e si analizzano una o due variabili alla volta: questo tipo di analisi statistiche sono dette di tipo univariato e bivariato. Spesso però si ha a che fare con insiemi di dati che riguardano un numero maggiore di variabili; in questi casi le prime analisi saranno comunque univariate e bivariate, ma è interessate anche poter studiare eventuali legami fra le variabili tutte insieme.
La principale difficoltà delle analisi multivariate è quella di rappresentare graficamente i dati; in presenza di due variabili abbiamo rappresentato le unità sperimentali con punti nel piano: le coordinate di un punto sono i valori delle due variabili rilevate su quella unità.
Con tre variabili è ancora possibile una tale rappresentazione nello spazio, ma la sua visualizzazione su un foglio (piano) non è univoca.



Le tecniche di analisi multivariata si basano su una generalizzazione e astrazione della rappresentazione dei dati tramite punti in uno spazio a molte dimensioni.
A ciascuna unità sperimentale è quindi associato un punto con molte coordinate.

Indichiamo con p il numero di variabili da analizzare; due punti x e y hanno coordinate: $x = (x_1, x_2, \ldots, x_p)$ e $y = (y_1, y_2, \ldots, y_p)$.
Da un punto di vista matematico questi punti si trattato allo stesso modo dei punti di un piano. Ad esempio la distanza euclidea fra i due punti è:

$$d(x,y) = \sqrt{\sum_{k=1}^{p} (x_k - y_k)^2}$$

ESEMPIO

Consideriamo il seguente insieme di dati, che si riferisce ad un rilevamento ISTAT sui servizi sanitari nelle regioni italiane (i dati sono del 1998). Le variabili su cui effettuare l'analisi sono:
- il numero di medici per 10.000 abitanti;
- il numero di posti letto ospedalieri per 10.000 abitanti;
- il numero di pediatri per 10.000 abitanti di età inferiore ai 14 anni;
- il numero annuo di interventi di pronto soccorso per 1.000 abitanti;
- la percentuale di persone che si sono dichiarate soddisfatte dell'assistenza medica ospedaliera.

I dati sono riportati nella tabella a fianco.

In questo caso le unità sperimentali, cioè i punti, sono le regioni e le variabili sono 5.

Ad esempio il Piemonte ha coordinate:
(8.47, 5.0, 8.38, 413.9, 48.5)
E la Valle d'Aosta ha coordinate:
(8.60, 4.2, 8.68, 298.2, 35.5)

	Regione	medici	posti letto	pediatri	interv PS	soddisfaz
1	Piemonte	8.47	5.0	8.38	413.9	48.5
2	Valle d'Aosta	8.60	4.2	8.68	298.2	35.5
3	Lombardia	8.18	5.5	7.80	386.6	48.5
4	Trent-Alto Adige	6.13	6.2	6.86	423.6	56.8
5	Veneto	7.99	5.3	8.46	454.3	61.5
6	Friuli-Ven. Giulia	8.84	5.7	7.40	393.4	51.8
7	Liguria	8.89	5.6	10.57	399.6	44.6
8	Emilia-Romagna	8.29	5.4	10.54	387.6	50.6
9	Toscana	8.73	4.9	9.64	325.3	43.3
10	Umbria	9.97	4.3	11.33	393.4	41.8
11	Marche	8.36	5.7	9.01	408.8	50.1
12	Lazio	9.19	6.5	9.84	436.2	32.6
13	Abruzzo	8.04	6.5	8.87	515.5	43.3
14	Molise	8.44	5.3	7.21	410.2	27.5
15	Campania	7.69	4.7	5.88	410.5	37.2
16	Puglia	7.97	5.2	8.16	358.7	17.7
17	Basilicata	8.49	4.4	6.70	338.2	34.3
18	Calabria	8.34	4.9	7.79	349.4	26.2
19	Sicilia	7.84	4.2	9.07	439.5	35.7
20	Sardegna	7.79	5.3	8.29	264.0	25.7

La distanza fra le prime due regioni è quindi:

$$\sqrt{(8.47 - 8.60)^2 + (5.0 - 4.2)^2 + (8.38 - 8.68)^2 + (413.9 - 298.2)^2 + (48.5 - 35.5)^2} = 116.431$$

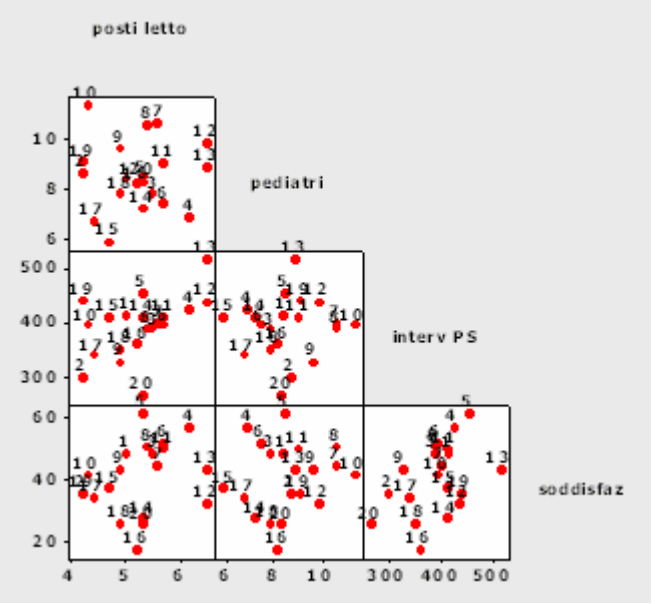
Prima di inizare l'analisi multivariata è opportuno trattare i dati da un punti di vista univariato e bivariato, tramite indici e rappresentazioni grafiche che già conosciamo:

```
Variable      N   Mean  StDev  Minimum     Q1  Median     Q3 Maximum
medici       20  8.312  0.741    6.130  7.975   8.350  8.698   9.970
posti letto  20  5.240  0.692    4.200  4.750   5.300  5.675   6.500
pediatri     20  8.524  1.398    5.880  7.498   8.420  9.498  11.330
interv PS    20  390.3   56.8    264.0  351.7   396.5  421.2   515.5
soddisfaz    20  40.66  11.38    17.70  33.03   42.55  49.70   61.50
```

Matrice di correlazione:

```
             medici  postil. Ped.  interv
PS
postiletto -0.236
pediatri    0.592    0.027
interv PS  -0.113    0.486    0.065
soddisfaz  -0.142    0.284    0.128
0.446
```

Il grafico a fianco viene anche detto matrix plot e consiste in tutte le coppie di grafici bivariati fra le variabili.



La cluster analysis o l'analisi di aggregazione

In questa scheda vedremo una delle principali e più semplici tecniche di analisi descrittiva multivariata: la cluster analysis o analisi di aggregazione.
Lo scopo della cluster analysis è quello di raggruppare le unità sperimentali in classi secondo criteri di similarità, cioè determinare un certo numero di classi in modo tale che le osservazioni siano il più possibile omogenee all'interno delle classi ed il più possibile disomogenee tra le diverse classi. Il concetto di omogeneità viene specificato in termini di distanza ed esistono diversi criteri che saranno chiariti in seguito.

ESEMPIO – continua

Applicando un metodo di aggregazione che preciseremo in seguito si individuano tre classi, che corrispondono a:

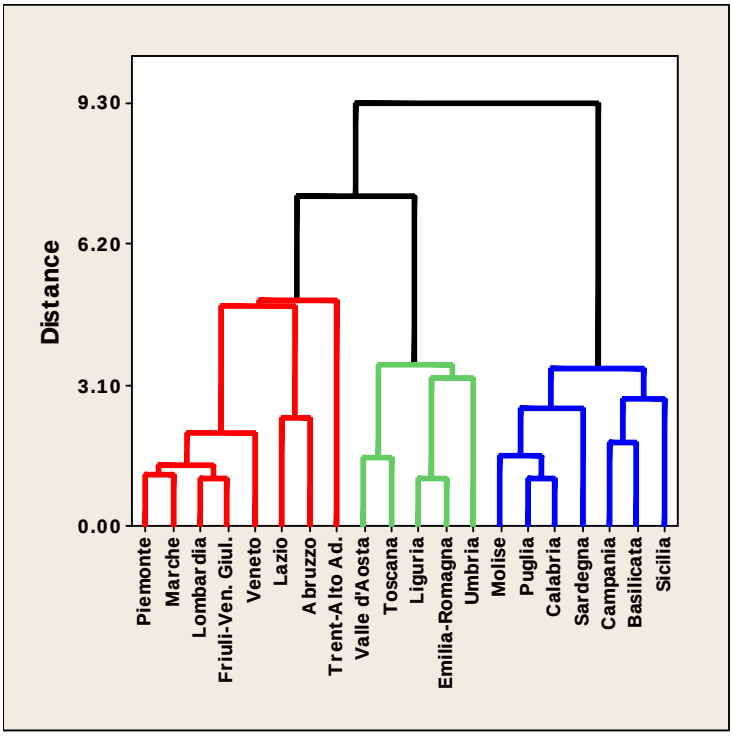
- Piemonte, Lombardia, Trentino A. A., Veneto, Friuli V. G., Marche, Lazio, Abruzzo
- Valle d'Aosta, Liguria, Emilia-Romagna, Toscana, Umbria
- Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna

Come sono stati aggregati i punti? Per dare un significato alle tre classi osserviamo le medie delle variabili nelle tre classi:

classe	N	medici	postiletto	pediatri	interv PS	soddisfaz
1	8	8.150	5.800	8.328	429.0	49.14
2	5	8.896	4.880	10.152	360.8	43.16
3	7	8.080	4.857	7.586	367.2	29.19

Osservando tali medie si può dire che nella terza classe si trovano quelle regioni che hanno valori bassi per tutte le variabili, soprattutto per la soddisfazione, mentre le regioni della seconda classe sono caratterizzate da un alto numero di pediatri e parzialmente di medici.

Una rappresentazione grafica di questo raggruppamento in classi riportata a fianco. Vedremo in seguito come si costruisce.

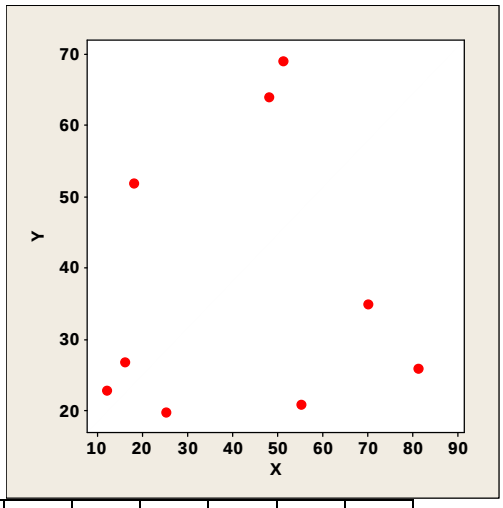


Per poter operare delle aggregazioni fra le unità sperimentali, è necessario in primo luogo precisare alcune questioni relative alle distanze, in particolare alle distanze fra classi di punti.

Inizialmente prenderemo in considerazione un semplice esempio con 9 punti e 2 variabili X e Y in quanto nel caso di due variabili è possibile visualizzare i punti con un grafico bidimensionale; ci serviremo di questo esempio per capire come funzionano i metodi di aggregazione dei punti.

ESEMPIO con 9 punti.
Consideriamo i seguenti dati:

	X	Y
1	12	23
2	25	20
3	16	27
4	81	26
5	55	21
6	70	35
7	18	52
8	48	64
9	51	69

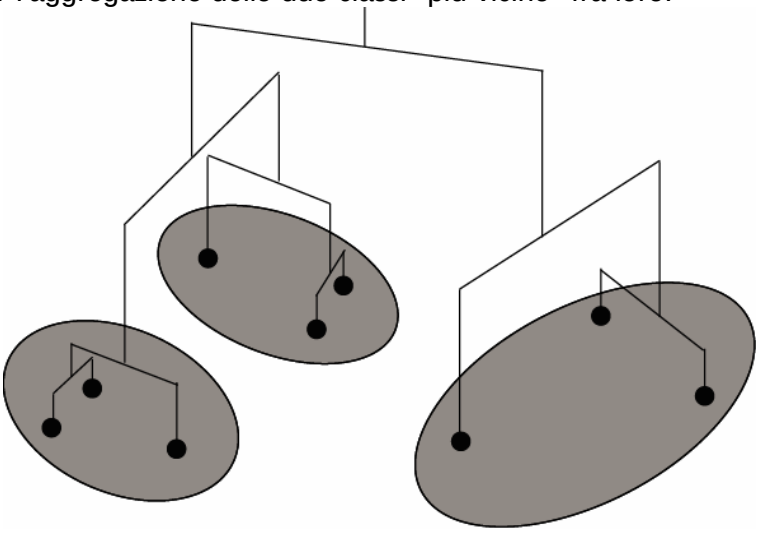


Qui a fianco è riportata la matrice dei quadrati delle distanze euclidee fra i punti (è più comodo usare i quadrati delle distanze).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	178	32	4770	1853	3508	877	2977	3637
2	178	0	130	3172	901	2250	1073	2465	3077
3	32	130	0	4226	1557	2980	629	2393	2989
4	4770	3172	4226	0	701	202	4645	2533	2749
5	1853	901	1557	701	0	421	2330	1898	2320
6	3508	2250	2980	202	421	0	2993	1325	1517
7	877	1073	629	4645	2330	2993	0	1044	1378
8	2977	2465	2393	2533	1898	1325	1044	0	34
9	3637	3077	2989	2749	2320	1517	1378	34	0

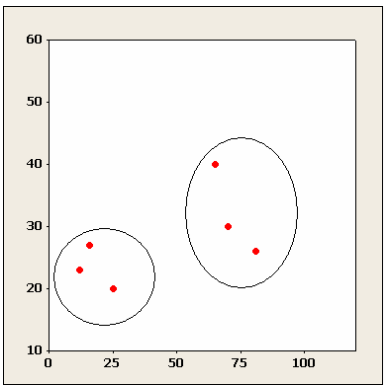
In questa scheda vedremo alcuni metodi di aggregazione gerarchica fra punti o classi basati sulla distanza.
Tali metodi operano essenzialmente nello stesso modo, procedendo sequenzialmente dallo stadio in cui ciascun punto è considerato una singola classe fino allo stadio finale in cui c'è una sola classe. Ad ogni passo il numero delle classi è ridotto di uno per l'aggregazione delle due classi “più vicine” fra loro.
Poiché ad ogni passo le classi sono ottenute dalla fusione di due classi del passo precedente, questi metodi conducono ad una struttura gerarchica per i punti, che può essere visualizzata con un diagramma ad albero, chiamato dendogramma.

In seguito vedremo un modo più preciso per costruire il dendogramma.



Distanza fra classi

Come definiresti la distanza fra due classi di punti?



Prova a fare delle ipotesi per il semplice esempio a fianco.

In genere la distanza fra classi viene definita in due passi: prima si definisce la distanza fra un punto e una classe formata da due punti; successivamente si definisce la distanza fra due classi.

In seguito denoteremo con "classe" anche una classe formata da un singolo punto.
Indichiamo con:

- x, y, z, … , x_i, x_j, … i punti,
- d(x,y) la distanza fra due punti x e y; in genere sarà la distanza euclidea
- C_{xy} la classe ottenuta dal raggruppamento di x e y,
- C_A e C_B due classi con rispettivamente n_A e n_B punti e con baricentri $\bar{x}_A$ e $\bar{x}_B$:

$$\bar{x}_A = \frac{1}{n_A} \sum_{x_i \in A} x_i \qquad \bar{x}_B = \frac{1}{n_B} \sum_{x_i \in B} x_i$$

- $\sum_{x_i \in A}$ la somma su tutti i punti di C_A

Ricordiamo che i punti appartengono a uno spazio a p dimensioni e quindi anche i baricentri sono punti a p dimensioni, le cui coordinate sono i baricentri delle singole variabili.

Considerate l'esempio sulle regioni e verificate le coordinate dei baricentri delle tre classi.

Molti sono i metodi per aggregare le classi, infatti varie sono le possibilità di definire una distanza fra due classi.

- Metodo della *distanza minima* o *single linkage*

La distanza fra la classe formata dai punti x e y e un punto z è definita come:

$$D(C_{xy},z) = \min \{ d(x,z), d(y,z) \}$$

La distanza fra C_A e C_B è la minima distanza fra ogni punto di C_A e ogni punto di C_B

- Metodo della *distanza massima* o *complete linkage*

La distanza fra la classe formata dai punti x e y e un punto z è definita come:

$$D(C_{xy},z) = \max \{ d(x,z), d(y,z) \}$$

La distanza fra C_A e C_B è la massima distanza fra ogni punto di C_A e ogni punto di C_B

- Metodo della *distanza media* o *average linkage*.

La distanza fra la classe formata dai punti x e y e un punto z è definita come:

$$D(C_{xy},z) = \frac{d(x,z) + d(y,z)}{2}$$

La distanza fra C_A e C_B è la distanza media fra coppie di punti, uno in C_A e uno in C_B

- Metodo dei *centroidi*

In generale la distanza fra C_A e C_B è la distanza fra i baricentri C_A e C_B:

$$D(C_{xy},z) = d(\bar{x}_A, \bar{x}_B)$$

Algoritmo di aggregazione

Una volta scelto il metodo di aggregazione fra le classi, i passi dell'algoritmo di aggregazione gerarchica ascendente di n punti sono i seguenti:

passo 1: si costruisce la matrice delle distanze fra gli n punti; si cercano i due punti più vicini e li si aggrega in un'unica classe;

passo s: a partire dalla matrice delle distanze del passo s-1 si costruisce una nuova matrice delle distanze: si ricalcolano le distanze della classe costruita al passo precedente con le altre classi, le altre distanze rimangono inalterate. Si cercano le due classi più vicine e si aggregano in un'unica classe;

passo n-1: si hanno solo due classi che vengono raggruppate nella classe costituita da tutti i punti iniziali.

ESEMPIO con 9 punti – continua

Utilizziamo il quadrato della distanza euclidea e come criterio di aggregazione quello della distanza minima. Qui sotto sono riportati tutti i passi dell'algoritmo.

Osservazione: questa non è propriamente una distanza, non va bene lo stesso per i nostri scopi

Matrice iniziale

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	0	178	32	4770	1853	3508	877	2977	3637
2	178	0	130	3172	901	2250	1073	2465	3077
3	32	130	0	4226	1557	2980	629	2393	2989
4	4770	3172	4226	0	701	202	4645	2533	2749
5	1853	901	1557	701	0	421	2330	1898	2320
6	3508	2250	2980	202	421	0	2993	1325	1517
7	877	1073	629	4645	2330	2993	0	1044	1378
8	2977	2465	2393	2533	1898	1325	1044	0	34
9	3637	3077	2989	2749	2320	1517	1378	34	0

Passo 1:
Si aggregano i punti 1 e 3 nella classe C_1.

		C_1							
		1-3	**2**	**4**	**5**	**6**	**7**	**8**	**9**
C_1	**1-3**	0	130	4226	1557	2980	629	2977	2989
	2	130	0	3172	901	2250	1073	2465	3077
	4	4226	3172	0	701	202	4645	2533	2749
	5	1557	901	701	0	421	2330	1898	2320
	6	2980	2250	202	421	0	2993	1325	1517
	7	629	1073	4645	2330	2993	0	1044	1378
	8	2977	2465	2533	1898	1325	1044	0	34
	9	2989	3077	2749	2320	1517	1378	34	0

Passo 2:
Si aggregano i punti 8 e 9 nella classe C_2.

		C_1						**C_2**
		1-3	**2**	**4**	**5**	**6**	**7**	**8-9**
C_1	**1-3**	0	130	4226	1557	2980	629	2977
	2	130	0	3172	901	2250	1073	2465
	4	4226	3172	0	701	202	4645	2533
	5	1557	901	701	0	421	2330	1898
	6	2980	2250	202	421	0	2993	1325
	7	629	1073	4645	2330	2993	0	1044
C_2	**8-9**	2977	2465	2533	1898	1325	1044	0

		C_3					C_2
		C_1-2	**4**	**5**	**6**	**7**	**8-9**
C_3	**C_1-2**	0	3172	901	2250	629	2465
	4	3172	0	701	202	4645	2533
	5	901	701	0	421	2330	1898
	6	2250	202	421	0	2993	1325
	7	629	4645	2330	2993	0	1044
C_2	**8-9**	2465	2533	1898	1325	1044	0

Passo 3:
Si aggregano il punto 2 e la classe C_1 nella classe C_3, formata quindi da tre punti.

		C_3	C_4			C_2
		C_1-2	**4-6**	**5**	**7**	**8-9**
C_3	**C_1-2**	0	2250	901	629	2393
C_4	**4-6**	2250	0	701	4645	2533
	5	901	701	0	2330	1898
	7	629	4645	2330	0	1044
C_2	**8-9**	2393	2533	1898	1044	0

Passo 4:
Si aggregano i punti 4 e 6 nella classe C_4.

		C_3	C_5		C_2
		C_1-2	**C_4-5**	**7**	**8-9**
C_3	**C_1-2**	0	901	629	2393
C_5	**C_4-6**	901	0	4645	2533
	7	629	4645	0	1044
C_2	**8-9**	2393	2533	1044	0

Passo 5:
Si aggregano il punto 5 e la classe C4 nella classe C5, formata quindi da 3 punti.

		C_6	C_5	C_2
		C_3-7	**C_4-5**	**8-9**
C_6	**C_3-7**	0	901	1044
C_5	**C_4-6**	901	0	1325
C_2	**8-9**	1044	1325	0

Passo 6:
Si aggregano il punto 7 e la classe C_3 nella classe C_6, formata quindi da 4 punti.

		C_7	C_2
		C_5-C_6	**8-9**
C_7	**C_5-C_6**	0	1044
C_2	**8-9**	1044	0

Passo 7:
Si aggregano le classi C5 e C6 nella classe C7, formata quindi da 7 punti.

Passo 8:
Si aggregano le ultime due classi C_7 e C_2 nella classe C_8, formata quindi da tutti i 9 punti.

L'aggregazione dei punti può essere sintetizzata nel seguente modo:

Passo 0: {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}
Passo 1: {1, 3}, {2},{4}, {5}, {6}, {7}, {8}, {9}
Passo 2: {1,3}, {2},{4}, {5}, {6}, {7}, {8 , 9}
Passo 3: {1, 2, 3}, {4}, {5}, {6}, {7}, {8 , 9}
Passo 4: {1, 2, 3}, {4, 6}, {5}, {7}, {8 , 9}
Passo 5: {1, 2, 3}, {4, 5 , 6}, {7}, {8 , 9}
Passo 6: {1, 2, 3, 7}, {4, 5, 6}, {8 , 9}
Passo 7: {1, 2, 3, 4, 5, 6 , 7 }, {8 , 9}
Passo 8: {1, 2, 3, 4, 5, 6 , 7 , 8 , 9}

Gerarchia, indice di aggregazione e dendogramma

Al passo iniziale tutti le classi sono formate da un solo punto. Al passo finale c'è una sola classe.
Abbiamo indicato con C_1, C_2, ... , C_{n-1} le classi costruite ai passi 1, 2, ... , n-1. Due classi fra queste o sono disgiunte o una delle due è inclusa nell'altra.
L'algoritmo di aggregazione produce quindi un ordinamento fra le classi costruite, cioè una gerarchia.

È possibile assegnare a ciascuna classe C_1, C_2, ... , C_{n-1} un *indice di aggregazione* a_i corrispondente alla distanza fra le due classi aggregate nella loro costruzione.
Indichiamo con a_1, a_2, ... , a_{n-1} gli indici di aggregazione.
A fianco sono riportati gli indici di aggregazione corrispondenti all'esempio precedente, dove si è utilizzato il metodo della distanza minima.

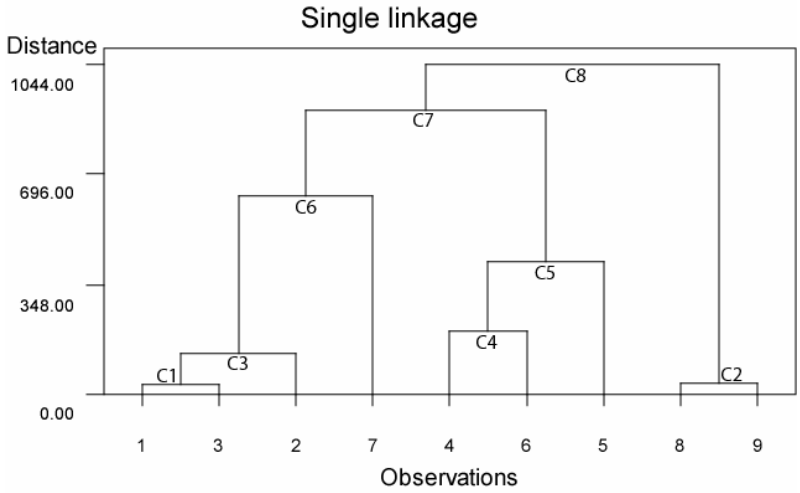
C1	32	{1, 3}
C2	34	{8, 9}
C3	130	{1, 2, 3}
C4	202	{4, 6}
C5	421	{4, 5 , 6}
C6	629	{1, 2, 3, 7}
C7	901	{1, 2, 3, 4, 5, 6, 7 }
C8	1044	{1, 2, 3, 4, 5, 6, 7, 8, 9}

Il processo di aggregazione gerarchica può essere visualizzato con un albero, chiamato *dendogramma* con altezze proporzionali agli indici di aggregazione.

Talvolta è riportato in ordinata un indice di similarità calcolato come: $s_i = \left(1 - \frac{a_i}{d_M}\right) \times 100$ dove d_M è il massimo delle distanze fra i punti disaggregati (al passo iniziale).

A fianco è riportato il dendogramma corrispondente all'esempio con 9 punti.

Sull'asse verticale è indicato l'indice di aggregazione fra le classi.
Al grafico ottenuto con il software Minitab sono stati successivamente aggiunti i nomi delle classi per chiarire i passaggi dell'algoritmo.



"Tagliando" l'albero con una retta orizzontale si ottiene una partizione dell'insieme dei punti tanto più fine quanto più si è vicini alle classi terminali. Ad esempio "tagliando" intorno a 700 si ottiene una partizione in 3 classi.
Il software Minitab produce un output in cui viene esplicitato il processo di aggregazione passo per passo, come è riportato sotto per i dati dell'esempio con 9 punti.

```
Cluster Analysis of Observations: X, Y
Squared Euclidean Distance, Single Linkage

Amalgamation Steps

Step Number of Similarity  Distance  Clusters   New   Number of obs.
     clusters    level      level     joined  cluster in new cluster
  1      8       99.33      32.000     1    3     1           2
  2      7       99.29      34.000     8    9     8           2
  3      6       97.27     130.000     1    2     1           3
  4      5       95.77     202.000     4    6     4           2
  5      4       91.17     421.000     4    5     4           3
  6      3       86.81     629.000     1    7     1           4
  7      2       81.11     901.000     1    4     1           7
  8      1       78.11    1044.000     1    8     1           9
```

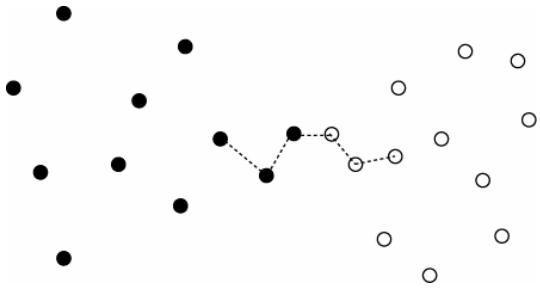
Per i metodi della distanza minima, massima e media la sequenza degli indici di aggregazione $a_{\_1}$, $a_{\_2}$, ... , a_{n-1} è formata da numeri che risultano ordinati in modo crescente:

$$a_{\_1} < a_{\_2} < ... < a_{n-1}$$

Se si effettua l'aggregazione dei punti con il metodo dei centroidi, può verificarsi il caso che la sequenza degli indici di aggregazione non sia crescente, come è evidenziato dall'esempio seguente.

Altri metodi di aggregazione gerarchica

Molti dei metodi di aggregazione secondo la distanza hanno il difetto di produrre un "effetto catena"' come quello rappresentato a fianco; in particolare questo è il caso del metodo della distanza minima.



Metodi gerarchici alternativi a quelli basati sulla distanza consistono nel ricercare a ciascun passo una aggregazione di classi in modo che la varianza all'interno della classe sia minima.
Il principale e il più usato fra questi è il Metodo di Ward; consiste nel raggruppare ad ogni passo due classi C_A di peso n_A e C_B di peso n_B che rendono minima la perdita di varianza fra le classi. Si dimostra che questo criterio corrisponde a unire due classi che hanno minima, non la distanza come negli algoritmi precedenti, ma la distanza pesata fra i baricentri, dove il peso è: $\frac{n_A n_B}{n_A + n_B}$ e l'indice di aggregazione fra due classi è:

$$\frac{n_A n_B}{n_A + n_B}\ d(\bar{x}_A, \bar{x}_B)$$

Come determinare il numero di classi

L'individuazione del numero delle classi finali è, in genere, non univoca, a meno che non si abbiano motivi a priori per la sua determinazione.
In genere si effettua una prima analisi senza specificare il numero di classi finali; si osservano quindi il dendogramma e la variazione a ogni passo degli indici di aggregazione (o di dissimilarità): il passo in cui i valori cambiano in modo significativo può identificare un buon punto per "tagliare" il dendogramma.
Si può poi effettuare una seconda analisi specificando il numero di classi e stabilire se i raggruppamenti hanno un qualche senso per i dati in esame.
Per determinare il numero di classi si possono inoltre utilizzare alcune statistiche relative alle classi individuate che sono fornite dai software statistici, come per esempio la numerosità delle classi, la distanza fra i centri delle classi, l'inerzia interna delle classi e le distanze medie e massime dei punti di ciascuna classe dal loro centro, a seconda che l'obiettivo sia di individuare classi di circa uguale numerosità, uguale varianza, uguale distanza fra i centri, e così via.

Standardizzazione delle variabili

Il tipo di aggregazione dei punti può essere influenzato dalla dispersione delle singole variabili: se una variabile ha una varianza molto alta, la nuvola di punti si allunga nella direzione dell'asse su cui è rappresentata: questo talvolta può influenzare involontariamente il risultato.
Se si vuole fare un'analisi che prescinda dalla dispersione di ciascuna variabile si devono preventivamente standardizzare i dati, cioè rendere tutte le variabili di media nulla e varianza unitaria. La “centratura” delle variabili in realtà non è necessaria per il problema che stiamo considerando.
Talvolta, per uniformare la dispersione delle variabili, si divide ciascuna variabile, invece che per la propria standard deviation, per il proprio “range” (cioè l'ampiezza dell'intervallo su cui assume valori la variabile).

ESEMPIO

Consideriamo nuovamente i dati relativi ad alcuni aspetti della Sanità pubblica affrontato nell'esempio iniziale.
L'analisi riportata precedentemente era stata effettuata con le variabili standardizzate.
Le medie e le standard deviation delle variabili su tutte le regioni sono riportate a fianco.
Si può osservare che gli interventi di pronto soccorso e la soddisfazione hanno una standard deviation molto più alta rispetto alle altre variabili.

Variable	Mean	StDev
Medici	8.312	0.741
Postiletto	5.240	0.692
Pediatri	8.524	1.398
interv_PS	390.3	56.8
soddisfaz	40.66	11.38

Se i dati non vengono standardizzati le aggregazioni in tre gruppi sono le seguenti:

- Piemonte, Lombardia, Trentino A. A., Friuli V. G., Liguria, Emilia-Romagna, Umbria, Marche, Molise, Campania
- Valle d'Aosta, Toscana, Puglia, Basilicata, Calabria, Sardegna
- Veneto, Lazio, Abruzzo, Sicilia

Si può osservare ad esempio che nel primo gruppo non compaiono più le regioni Veneto, Lazio e Abruzzo che hanno valori molto lontani dal baricentro per le variabili interventi di pronto soccorso e la soddisfazione. Viceversa compaiono nel primo gruppo di questa seconda analisi le regioni Liguria, Emilia-Romagna, Umbria, Molise, Campania che hanno valori vicini al baricentro per le due variabili precedenti ma lontani dal baricentro per le altre variabili.

Come interpretare i risultati

Abbiamo detto che per determinare il numero di classi si possono utilizzare anche alcune statistiche relative alle classi individuate che sono fornite dai software statistici, come per esempio la numerosità delle classi, la distanza fra i centri delle classi, l'inerzia interna delle classi e le distanze medie e massime dei punti di ciascuna classe dal loro centro. Queste informazioni permettono anche di interpretare i riusltati ottenuti.

Consideriamo nuovamente l'esempio sugli aspetti sanitari nelle regioni italiane.
Avendo scelto una partizione in 3 classi si ha:

```
Final Partition
Number of clusters: 3
                          Within   Average   Maximum
                         cluster  distance  distance
            Number of   sum of      from      from
          observations  squares  centroid  centroid
Cluster1             8  26.1088   1.64228   3.05570
Cluster2             5  11.9566   1.47736   1.97578
Cluster3             7  15.9042   1.42365   2.05276


Cluster Centroids
                                                Grand
Variable      Cluster1  Cluster2  Cluster3   centroid
medici       -0.218731   0.78851  -0.31324  0.0000000
posti letto   0.809712  -0.52053  -0.55358  0.0000000
pediatri     -0.140562   1.16456  -0.67118  0.0000000
interv PS     0.681297  -0.51988  -0.40729  0.0000000
soddisfaz     0.744677   0.21960  -1.00792  0.0000000


Distances Between Cluster Centroids

          Cluster1  Cluster2  Cluster3
Cluster1   0.00000   2.49117   2.53094
Cluster2   2.49117   0.00000   2.47071
Cluster3   2.53094   2.47071   0.00000
```

Esaminiamo nei dettagli le tre tabelle.

- La prima fornisce informazioni sulla numerosità e la "compattezza" di ciascuna classe; in particolare:
    - Within cluster sum of square: è la somma dei quadrati delle distanze dei punti della classe dal baricentro della classe. Dividendo tale valore per la numerosità della classe si ha la varianza dei punti della classe: classi con varianza maggiore sono più disperse. Nell'esempio la prima classe è la più dispersa con una varianza di 3.26.
    - Average distance from centroid: è la media delle distanze dei punti della classe dal baricentro della classe; fornisce informazioni simili al precedente.
    - Maximum distance from centroid: permette di individuare se nella classe ci sono punti anomali. Nell'esempio la prima classe è quella con maggiore distanza.
- La seconda fornisce le medie delle variabili di ciascuna classe; permette di interpretare le classi in base alle variabili oggetto di studio (attenzione le variabili sono standardizzate). Nell'esempio si ottengono gli stessi commenti già riportati a pagina 3 e cioè che nella terza classe si trovano quelle regioni che hanno valori bassi per tutte le variabili, soprattutto per la soddisfazione, mentre le regioni della seconda classe sono caratterizzate da un alto numero di pediatri e parzialmente di medici.
- La terza fornisce le distanze fra i baricentri delle classi; in questo esempio le tre classi sono ugualmente distanti dal baricentro.

Variabili binarie e variabili ordinali – distanza Manhattan

Quando i dati non sono quantitativi, le distanze fra punti in generale perdono di significato.
Si possono però introdurre degli indici di dissimiglianza che operano sulle codifiche numeriche dei dati qualitativi.
Consideriamo anzitutto il caso di *variabili binarie*.
Possiamo adottare come indice di dissimilarità fra due punti x e y il numero di discordanze dei risultati delle p variabili binarie considerate.
Ad esempio se si codificano i valori assunti con 0 e 1 e se x=(0,1,1,0,0,1) e y=(1,1,1,0,0,0), allora vi sono due discordanze e l'indice di dissimilarità è 2.

Questo indice corrisponde alla cosiddetta distanza Manhattan nel caso in cui le codifiche numeriche delle variabili sono appunto 0 e 1:

$$d(x,y) = \sum_{k=1}^{p} |x_k - y_k|$$

Il nome di questa distanza, Manhattan o City block, è suggerito dalla struttura rettangolare (o a blocchi) di molte città statunitensi, in particolare New York; se si misura la distanza fra due punti della città con il minimo percorso stradale necessario per andare da uno all'altro, a Manhattan, la distanza fra due punti corrisponde proprio alla somma della distanza fra i punti in una direzione e la distanza dei punti nella direzione ad essa perpendicolare.

Approfondimento

Una distanza è una funzione che associa a ciascuna coppia di punti un numero reale: tale che, per ogni scelta di punti x, y, z si abbia:
(i) $d(x,y) \geq 0$ e $d(x,y)=0$ se e solo se $x=y$
(ii) $d(x,y)=d(y,x)$
(iii) $d(x,y) \leq d(x,z)+d(y,z)$ (questa è detta relazione triangolare)

La stessa distanza si anche usare nel caso di variabili *ordinali* quando i livelli sono codificati con numeri che mantengono l'ordinamento.
Se le codifiche numeriche delle p variabili assumono valori molto diversi fra loro allora, come nel caso delle variabili quantitative, si pone il problema di una loro "omogeneizzazione"; questa può essere ottenuta dividendo ogni variabile per il *range* dei valori assunti. Questo è ad esempio il caso in cui una variabile assume valori 100, 150, 200 e un'altra valori 0, 1, 2 e 3.
Nel caso in cui si voglia analizzare un inseme di dati con variabili di differente tipo (quantitative, ordinali, binarie) è usuale utilizzare in tutti i casi la distanza Manhattan dopo aver preventivamente resa omogenea la dispersione delle variabili, dividendo i dati di ciascuna di esse per il corrispondente "range".

Aggregazione delle variabili

Lo scopo principale dell'analisi di aggregazione è quello di formare raggruppamenti delle unità sperimentali. è però possibile utilizzare le stesse tecniche viste per i punti per aggregare le variabili anche allo scopo di individuare le variabili responsabili delle aggregazioni delle unità sperimentali.
Gli algoritmi visti per le unità sperimentali vengono applicati alle variabili aggregando di volta in volta i due punti (variabili) con coefficiente di correlazione *massimo*.

Ciò ha una giustificazione teorica che esponiamo brevemente.
Anzitutto si considerano come punti da aggregare le variabili *standardizzate.* I punti appartengono a uno spazio a n dimensioni. Consideriamo il quadrato della distanza euclidea fra due punti-variabili:

$$d(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} x_i y_i = 2n(1 - \rho(x,y))$$

La distanza usata nel caso dei punti-variabili è quindi $1 - \rho(x,y)$ e la distanza minima fra due punti si ha in corrispondenza del massimo del coefficiente di correlazione.

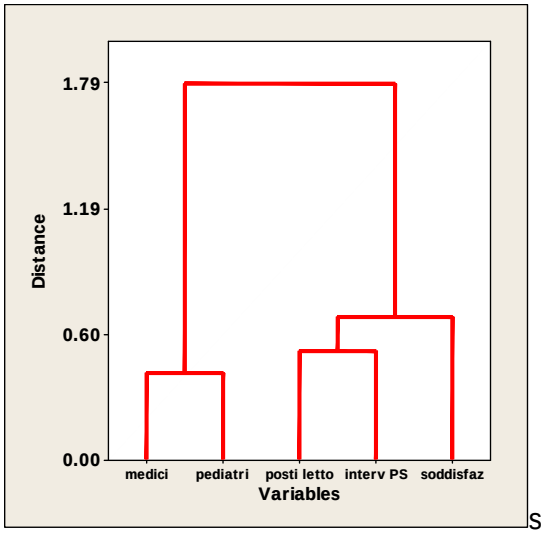
ESEMPIO

Consideriamo nuovamente l'esempio iniziale relativo ad alcuni aspetti della sanità pubblica nelle regioni italiane.

A fianco è riportata la matrice delle distanze fra le variabili.

	medici	postil.	pediatri	intPS	soddis.
medici	0.00	1.24	0.41	1.11	1.14
postiletto	1.24	0.00	0.97	0.51	0.72
pediatri	0.41	0.97	0.00	0.94	0.87
interv_PS	1.11	0.51	0.94	0.00	0.55
soddisfaz	1.14	0.72	0.87	0.55	0.00

Se si effettua la aggregazione dei punti-variabile usando lo stesso metodo dei punti-unità, cioè il metodo di Ward, si ottiene il seguente dendogramma.



S

ESERCIZI

ESERCIZIO 1)

Si condierino le seguenti rilevazioni di due variabili quantitative su 4 unità sperimentali:

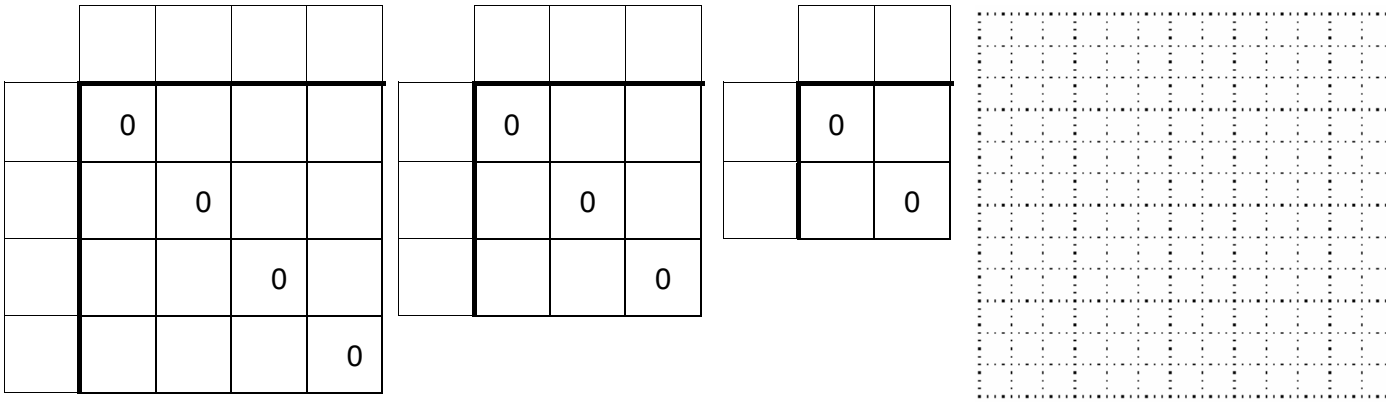
X	Y
2	4
4	2
5	1
3	4

A) Disegnare il grafico di dispersione dei punti.

B) Calcolare il baricentro dei punti e indicarlo nella rappresentazione grafica

C) Effettuare una cluster analysis utilizzando la distanza euclidea al quadrato e il metodo del Complete linkage (max).
   a) scrivere la matrice delle distanze iniziale e a ciascun passo dell'aggregazione
   b) indicare gli indici di aggregazione
   c) disegnare approssimativamente il dendogramma.

Passo1 Passo 2



a_1 = a_2 = a_3 =

ESERCIZIO 2)

I dati, già analizzati in una scheda precedente, riguardano atleti che praticano sport di fondo; sono rilevati l'età, il peso, il consumo di ossigeno, il tempo di percorrenza di un fissato tragitto di corsa, le pulsazioni cardiache al minuto da fermo e le pulsazioni medie e massime durante la corsa. La cluster analysis fornisce i seguenti risultati.

```
Final Partition
Number of clusters: 3

                        Within   Average   Maximum
                       cluster  distance  distance
            Number of   sum of      from      from
          observations  squares  centroid  centroid
Cluster1            15  72.9893   2.11503   3.25867
Cluster2             4  14.1313   1.85435   2.20535
Cluster3            12  42.5261   1.76859   3.13582

Cluster Centroids

Variable   Cluster1  Cluster2   Cluster3  Grand centroid
eta       -0.002063  -1.37724   0.461660       0.0000000
peso      -0.108844   0.45332  -0.015050       0.0000000
ossigeno  -0.657454   1.60936   0.285363       0.0000000
tempo      0.591896  -1.38288  -0.278909       0.0000000
pulsferm   0.334458  -0.84673  -0.135830       0.0000000
pulsmed    0.717406  -0.20924  -0.827010       0.0000000
pulsmax    0.650270   0.29744  -0.911986      -0.0000000
Distances Between Cluster Centroids
```

	Cluster1	Cluster2	Cluster3
Cluster1	0.00000	3.69103	2.63022
Cluster2	3.69103	0.00000	2.98704
Cluster3	2.63022	2.98704	0.00000

a) Commentare
b) A quale cluster potrebbero appartenere i due atleti con le seguenti rilevazioni *standardizzate*? Perchè?

eta	peso	ossigeno	tempo	pulsferm	pulsmed	pulsmax	
-0.51376	-1.32010	-0.49215	0.38480	-0.32176	0.61986	0.24288	CLUSTER _____
1.21321	0.68145	0.84081	-0.18461	-0.45300	-0.35556	-0.41185	CLUSTER _____

ESERCIZIO 3)

Viene effettuata una cluster analysis sulle seguenti variabili, rilevate su 31 Stati

a) Percentuale di superficie irrigata
b) Densità di popolazione
c) Percentuale di popolazione al di sotto dei 14 anni
d) Speranza di vita alla nascita
e) Percentuale di alfabetismo
f) Tasso di disoccupazione
g) Numero IPServer per milioni di persone
h) Numero di TV per persona
i) Chilometri di ferrovie sul totale di superficie
j) Numero di aeroporti sul totale di superficie

Cluster Analysis of Observations: Area, Irrigated, Population, Under,14, ...

Standardized Variables, Euclidean Distance, Ward Linkage

Final Partition
Number of clusters: 4

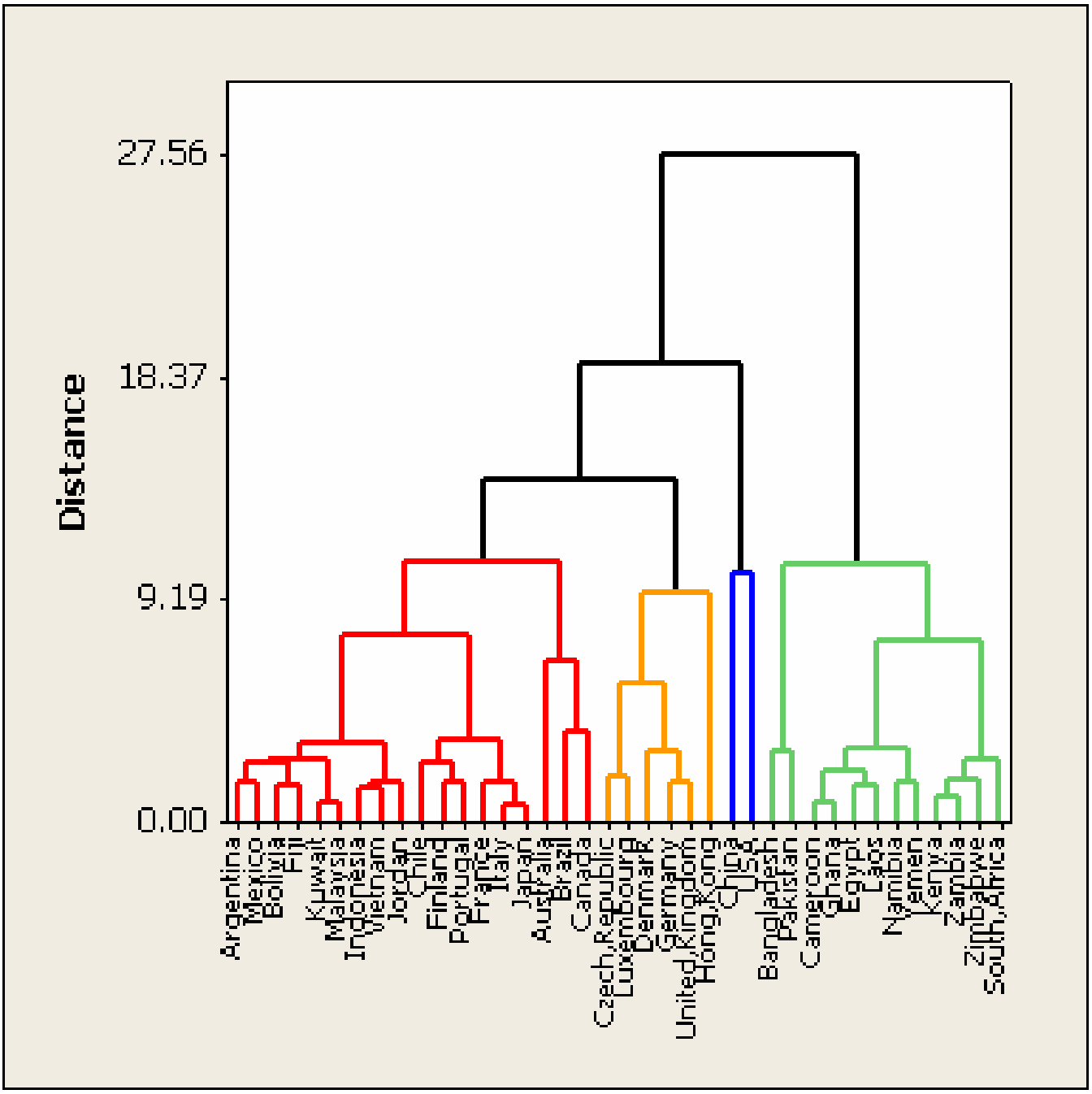
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	18	105.813	2.16245	5.81430
Cluster2	12	67.655	2.18479	4.33060
Cluster3	2	52.588	5.12778	5.12778
Cluster4	6	56.890	2.77860	5.51680

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Grand centroid
Area	0.124156	-0.36517	2.64946	-0.52528	0.0000000
Irrigated	-0.191882	-0.13455	3.60560	-0.35713	-0.0000000
Population	-0.141646	-0.20368	3.31222	-0.27177	-0.0000000
Under,14	-0.291678	1.10003	-0.59913	-1.12531	0.0000000
Life,expectancy	0.512856	-1.24210	0.54172	0.76505	0.0000000
Literacy,Rate	0.463919	-1.17303	0.34636	0.83884	0.0000000
Unemployment	-0.404092	1.02364	-0.57810	-0.64231	-0.0000000
ISPs/million	-0.097244	-0.44400	1.30924	0.74331	-0.0000000
Tvs/person	0.279623	-0.45824	0.10532	0.04250	0.0000000
Railways	-0.077078	-0.36571	3.38510	-0.16570	0.0000000
Airports	-0.064626	-0.25692	2.82919	-0.23533	0.0000000
Irr%	-0.123894	0.22101	0.13383	-0.11495	-0.0000000
Dens_pop	-0.183890	-0.14838	-0.18752	0.91093	-0.0000000
Dens_rail	-0.184123	-0.56711	-0.29789	1.78589	-0.0000000
Dens_airp	-0.221732	-0.55582	0.16302	1.72248	-0.0000000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	0.00000	3.33838	7.44738	3.32885
Cluster2	3.33838	0.00000	8.60023	5.43569
Cluster3	7.44738	8.60023	0.00000	8.34068
Cluster4	3.32885	5.43569	8.34068	0.00000



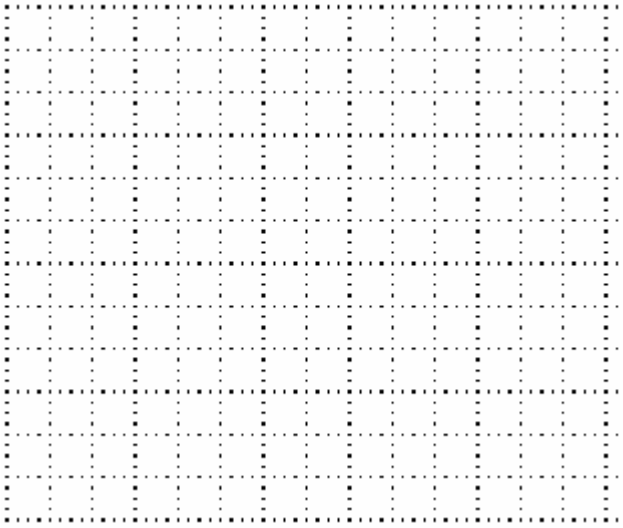
COMMENTARE DETTAGLIATAMENTE, spiegando in particolare da quali variabili sono caratterizzati i cluster.

ESERCIZIO 4)

Si condierino le seguenti rilevazioni di due variabili quantitative su 5 unità sperimentali:

	X	Y
P1	1	1
P2	2	3
P3	5	7
P4	6	5
P5	6	7

a) Disegnare il grafico di dispersione dei punti.



b) Scrivere la matrice delle distanze iniziale utilizzando la distanza euclidea al quadrato.

	0				
		0			
			0		
				0	
					0

Si considerino i seguenti quattro metodi di aggregazione delle classi:

1. Complete linkage (massimo)
2. Avereage linkage (media)
3. Single linkage (minimo)
4. Centroidi

c) Quali punti vengono aggregati al primo passo con i 4 metodi?

Al penultimo passo di aggregazione con tutti i metodi i punti risultano aggregati nelle seguenti due classi:

C1={P1, P2} C2={P3, P4, P5}.

d) Calcolare i baricentri delle due classi.

e) Calcolare la distanza fra C1 e C2 con i quattro metodi.

ESERCIZIO 5)

Viene effettuata una cluster analysis sulle seguenti variabili, rilevate in ciascuna regione italiana. I dati riguardano il 2003 e sono tratti dal sito:
http://www.istat.it/agricoltura/datiagri/fiori/fiori.htm

1. Piante da vaso con fiori coltivate in serra
2. Piante da vaso con fiori coltivate in piena aria
3. Piante da vaso con solo foglie coltivate in serra
4. Altre piante da vaso coltivate in serra
5. Altre piante da vaso coltivate in piena aria
6. Superfici adibite a serra per la coltivazione di fiori recisi
7. Produzione di fiori recisi coltivati in serra
8. Superfici aperete adibite alla coltivazione di fiori recisi
9. Produzione di fiori recisi coltivati in piena aria

I risultati sono i seguenti.

```
Final Partition Number of clusters: 3

                         Within   Average   Maximum
                        cluster  distance  distance
            Number of    sum of      from      from
```

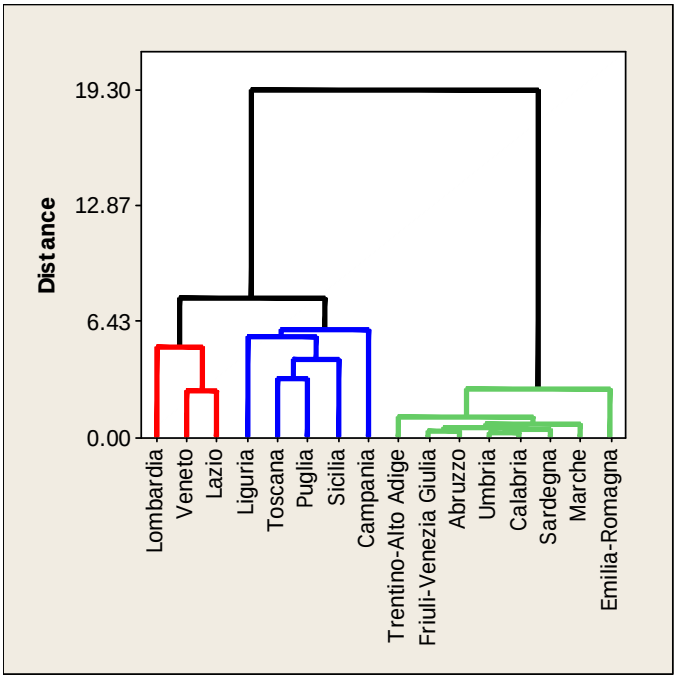
	observations	squares	centroid	centroid
Cluster1	3	14.5751	2.16250	2.76548
Cluster2	8	3.5045	0.53716	1.50354
Cluster3	5	46.5572	2.99969	3.72019

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
PV fiore serra	1.67147	-0.579321	-0.07597	0.0000000
PV fiore aria	0.75241	-0.663903	0.61080	0.0000000
PV foglia serra	0.33237	-0.656307	0.85067	0.0000000
PV foglia aria	-0.23004	-0.521028	0.97167	-0.0000000
Altre PV serra	1.10083	-0.560431	0.23619	0.0000000
Altre PV aria	1.02541	-0.462714	0.12510	0.0000000
Sup FR serra	-0.14360	-0.666903	1.15320	0.0000000
Prod FR serra	-0.30587	-0.610163	1.15978	0.0000000
Sup FR aria	0.08124	-0.752988	1.15604	0.0000000
Prod FR aria	-0.37414	-0.584139	1.15910	-0.0000000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0.00000	3.77027	3.69869
Cluster2	3.77027	0.00000	4.52617
Cluster3	3.69869	4.52617	0.00000



Commentare dettagliatamene indicando in particolare da che cosa sono caratterizzati i tre cluster.