STATISTICA DESCRITTIVA - SCHEDA N. 6
CLUSTER ANALYSIS

Una tecnica di statistica multivariata: la cluster analysis

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.
Quando si vogliono analizzare *contemporanemente* varie caratteristiche di una popolazione o di un fenomeno, si incontrano difficoltà nelle rappresentazioni grafiche.

Un esempio: le merendine

I dati (2002) riguardano il prezzo medio, il peso e altre informazioni nutrizionali di 16 tipi di merendine al cioccolato.
Sono tratti da: www.sci.usq.edu.au/staff/dunn/Datasets/applications/popular/chocolates.html

	Size (g)	Unit.Pr. ($)	Energy (Kcal 100g)	Protein (%)	Fat (%)	Carbo (%)	Sodium (g 10^{-4})
Dark.Bounty	50.0	1.76	1970	3.1	27.2	53.2	7.5
Bounty	50.0	1.76	2003	4.6	26.5	59.0	11.5
Milo.Bar	40.0	2.88	2057	9.9	23.0	60.9	11.6
Viking	80.0	1.93	1920	5.1	18.4	67.5	22.0
KitKat.White	45.0	2.56	2250	7.2	30.1	59.4	11.0
KitKat.Chunky	78.0	1.79	2186	7.0	28.4	59.7	9.3
Cherry.Ripe	55.0	2.33	1930	3.5	24.5	56.4	4.0
Snickers	60.0	1.62	1980	10.2	22.9	59.9	19.0
Mars	60.0	1.62	1890	4.7	19.5	67.9	16.0
Crunchie	50.0	2.56	2030	5.6	20.4	67.4	25.0
Tim.Tam	40.0	2.75	2180	5.5	26.8	67.3	16.0
Turkish.Delight	55.0	2.33	1623	2.2	9.2	73.3	9.0
Mars.Lite	44.5	2.18	1640	3.7	12.0	77.9	22.0
Dairy.Milk.King	75.0	2.11	2210	8.2	29.8	57.0	11.0
Maltesers	60.0	2.58	1980	8.5	20.6	63.3	13.0
MandMs	42.5	2.78	1970	5.0	20.0	69.0	14.8

Una chilocaloria (Kcal) è la quantità di di calore (energia termica) necessaria per innalzare la temperatura di 1000 grammo di acqua 1°C.
La resa di energia per grammo è la seguente:
Grassi: 1 grammo = 9 calorie Proteine: 1 grammo = 4 calorie Carboidrati: 1 grammo = 4 calorie
Ci sono diversi tipi di merendine al cioccolato? È possibile classificare i cioccolatini in alcuni gruppi logici (merendine dietetiche, merendine molto caloriche, dolcetti, ecc)?

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

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Size	16	55.31	12.96	40.00	44.63	52.50	60.00	80.00
Unit.Pr.	16	2.221	0.435	1.620	1.768	2.255	2.575	2.880
Energy	16	1988.7	177.7	1623.0	1922.5	1980.0	2149.3	2250.0
Protein	16	5.875	2.400	2.200	3.925	5.300	7.950	10.200
Fat	16	22.46	5.97	9.20	19.63	22.95	27.10	30.10
Carbo	16	63.69	6.67	53.20	59.10	62.10	67.80	77.90
Sodium	16	139.2	57.9	40.0	97.3	123.0	182.5	250.0

Se vogliamo invece considerare tutte le variabili assieme, cioè effettuare una analisi multivariata, la principale difficoltà è quella di rappresentare graficamente i dati.

Se le variabili fossero soltanto due si potrebbero fare grafici bidimensionali del tipo a fianco, in cui le unità sperimentali (le merendine) sono rappresentata con punti nel piano: le coordinate di un punto sono i valori delle due variabili rilevate su quella unità.

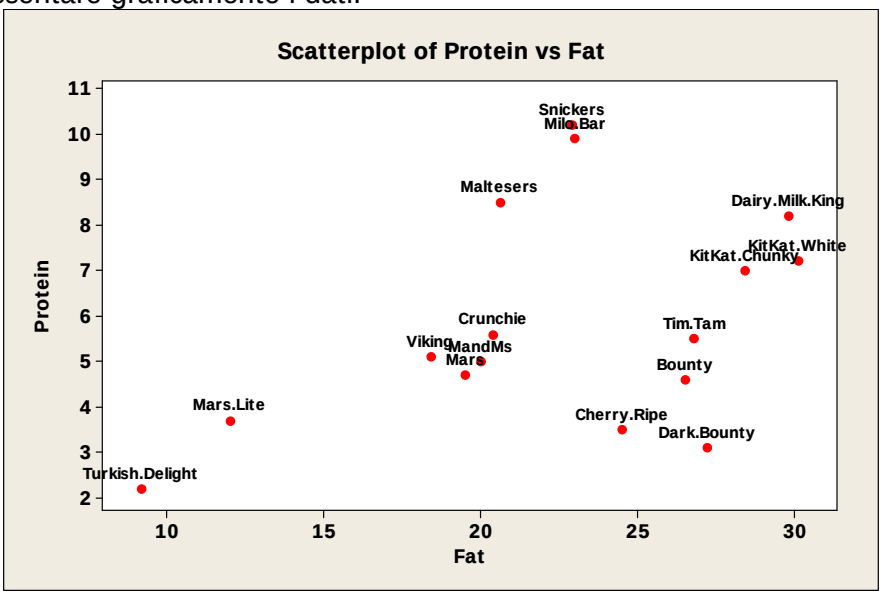


grafico 1 – proteine e grassi

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 una 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.

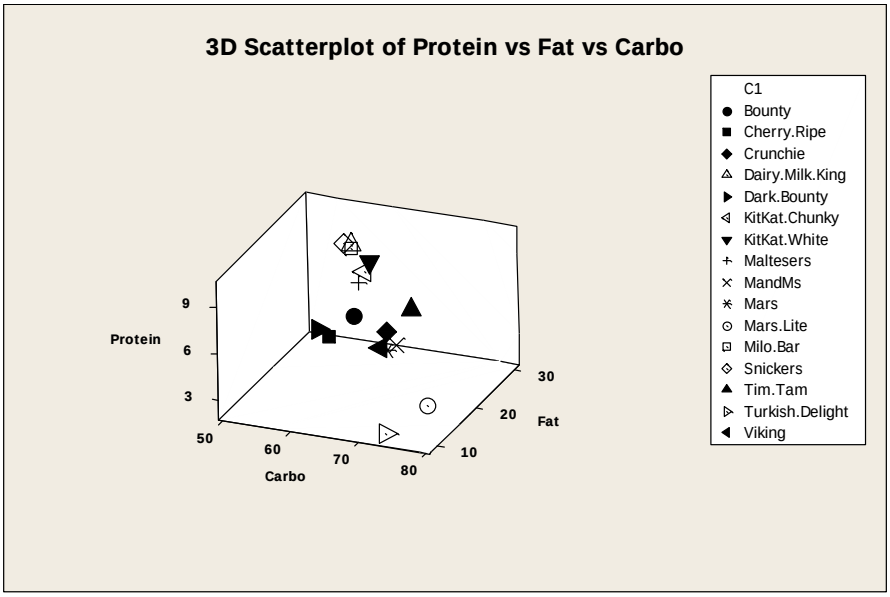


grafico 2 – proteine, grassi e carboidrati

Vogliamo costrure raggruppamenti di merendine in modo tale che le merendine 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**.

ESERCIZIO: Prova a costruire tre o quattro gruppi di merendine basandoti solo sulle proteine e i grassi. Da che cosa sono caratterizzati i gruppi (rispetto a proteine e grassi)?

Vogliamo fare la stessa cosa con tutte le variabili e ottenere dei raggruppamenti basati sulla **distanza fra i punti.**

Procederemo in modo sequenziale aggregando prima i punti più vicini e poi aggregando le classi di punti.

Inzialmente ciascun punto è considerato una singola classe; poi si aggregano punti e classi, 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".

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.

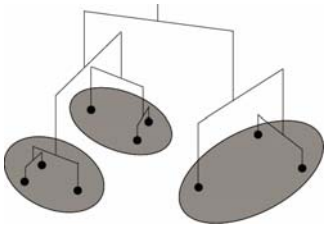


grafico 3 – schema di aggregazione dei punti

Come si calcola la distanza fra punti con p coordinate? In modo analogo a quanto sapete fare con due o tre coordinate.

Consideriamo i punti "Dark Bounty" e "Bounty"; le loro coordinate sono:

Dark.Bounty (50.0, 1.76, 1970, 3.1, 27.2, 53.2, 7.5)
Bounty (50.0, 1.76, 2003, 4.6, 26.5, 59.0, 11.5)

Se consideriamo solo le "proteine" e i "grassi" i punti hanno coordinate:

Dark.Bounty (3.1, 27.2)
Bounty (4.6, 26.5)

e la loro distanza nel piano è

$$\sqrt{(3.1 - 4.6)^2 + (27.2 - 26.5)^2} = \sqrt{2.25 + 0.49} = 1.66$$

Consideriamo ora p variabili (nel nostro esempio p = 7); 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 trattano allo stesso modo dei punti di un piano. La distanza euclidea fra i due punti è:

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

La distanza fra "Dark Bounty" e "Bounty" nello spazio a 7 dimensioni è:

$\sqrt{(50-50)^2 + (1.76-1.76)^2 + (1970-2003)^2 + (3.1-4.6)^2 + (27.2-26.5)^2 + (53.2-59.0)^2 + (7.5-11.5)^2} = 33.784$

Qui sotto è riportata la matrice delle distanze dei punti, con i valori arrotondati.

COMPLETATE

	D_B	B	MB	V	KKW	KKC	CR	S	M	C	TT	TD	ML	DMK	Mal	MM
D_Bounty	0	34	88	62	280	218	41	21	83		211	348	332	241	20	20
Bounty	34	0	55	90	247	185	74		114	32	178	381	364	209	26	36
Milo.Bar	88	55	0	143	193	135	128	80		33	123	435	418	157	80	88
Viking	62	90	143	0	332	267	35	64	37	114	263	299	283	291	64	63
KitKat.W	280	247	193	332	0	72	320	271	361	221	71	628	611	50	271	280
KitKat.C	218	185	135	267	72	0	257	207	297	160	40	564	548	24	207	219
Cherry.R		74	128	35	320	257	0	53	44	103	251	308	292	281	52	45
Snickers	21		80	64	271	207	53	0	91	52	201	358	341	231	8	23
Mars	83	114		37	361	297	44	91	0	141	291	267	251	321	90	82
Crunchie	65	32	33	114	221	160	103	52	141	0	151	408	390	183	53	61
Tim.Tam	211	178	123	263	71	40	251	201	291	151	0	558	540	48	201	210
Turkish.D	348	381	435	299	628	564	308	358	267	408	558	0	24	588	357	347
Mars.Lite	332	364	418	283	611	548	292	341	251	390	540	24	0	572	341	330
DairyMilkK	241	209	157	291	50	24	281	231	321	183	48	588	572	0	231	243
Maltesers	20	26	80	64	271	207	52	8	90	53	201	357	341	231	0	21
MandMs	20	36	88	63	280	219	45	23	82	61	210	347	330	243	21	0

Un risultato dell'aggregazione è riportato qui sotto.

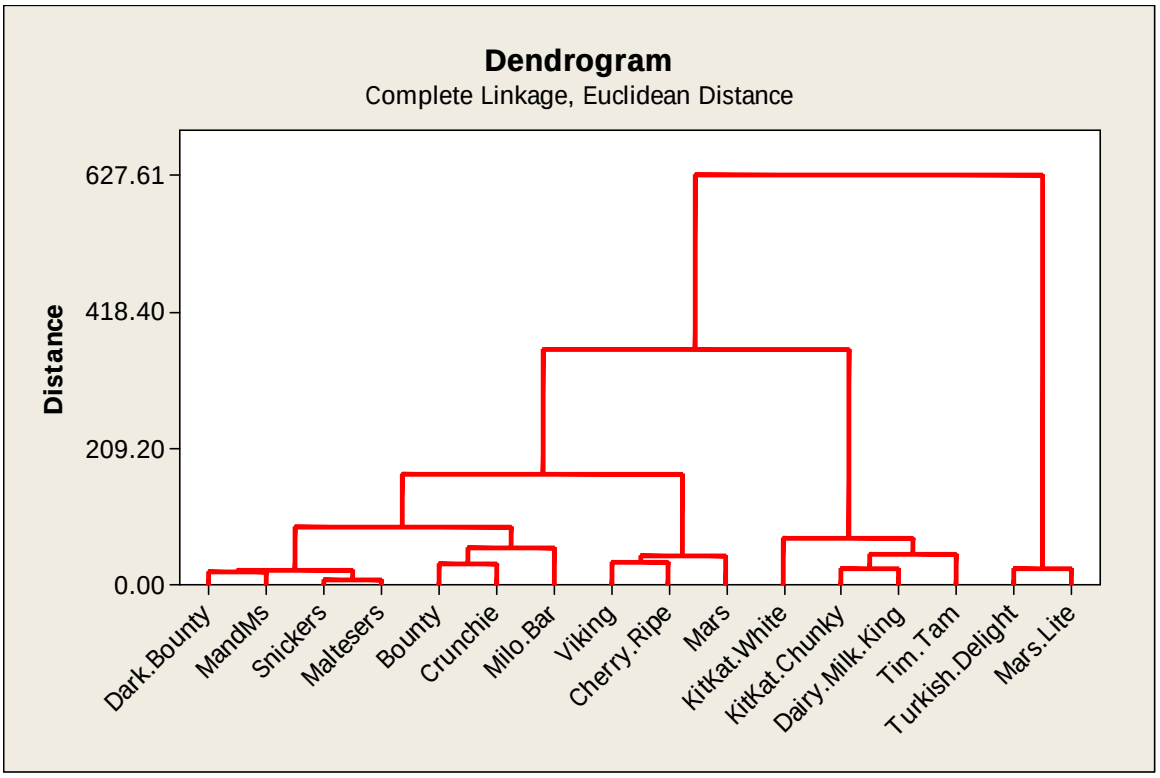


grafico 4 – dendogramma delle merendine

Dobbiamo capire:

1. Come è stato costruito
2. Come si possono individuare delle classi di merendine
3. Da che cosa sono caratterizzate le classi di merendine per quanto riguarda le proteine, i grassi, i carboidrati, ecc.

ESEMPIO con 6 punti

Per comprendere meglio l'algoritmo di agreggazione consideriamo solo le prime 5 merendine

Qui a fianco è riportata la matrice delle distanze euclidee fra i punti.

	D_B	B	MB	V	KKW	KKC
D_Bounty	0	34	88	62	280	218
Bounty	34	0	55	90	247	185
Milo.Bar	88	55	0	143	193	135
Viking	62	90	143	0	332	267
KitKat.W	280	247	193	332	0	72
KitKat.C	218	185	135	267	72	0

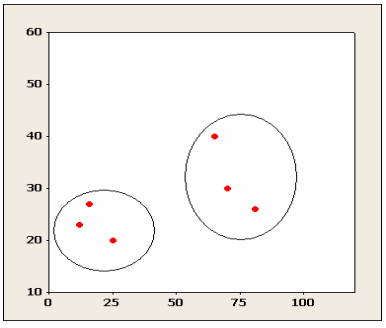
PRIMO PASSO DELL'ALGORITMO

Si aggregano i due punti più vicini e si costruisce una classe formata da questi.

Bisogna poi definire la distanza fra questa classe e gli altri punti.

Distanza fra classi

ESERCIZIO: Come definiresti la distanza fra due classi di punti? Prova a fare delle ipotesi per il semplice esempio a fianco.



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
- 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.

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 6 punti – continua

Come criterio di aggregazione quello della distanza massima.

Qui sotto sono riportati tutti i passi dell'algoritmo.

Matrice iniziale

	D_B	B	MB	V	KKW	KKC
D_Bounty	0	34	88	62	280	218
Bounty	34	0	55	90	247	185
Milo.Bar	88	55	0	143	193	135
Viking	62	90	143	0	332	267
KitKat.W	280	247	193	332	0	72
KitKat.C	218	185	135	267	72	0

Passo 1

Si aggregano D_B e B
Si calcolano le distanze fra la nuova classe e gli altri punti
Le altre distanze rimangono invariate

		C1				
		D_B B	MB	V	KKW	KKC
C1	D_Bounty Bounty	0	**88**	**90**	**280**	**218**
	Milo.Bar	**88**	0	143	193	135
	Viking	**90**	143	0	332	267
	KitKat.W	**280**	193	332	0	72
	KitKat.C	**218**	135	267	72	0

Passo 2

Si aggregano KKW e KKC
Si calcolano le distanze fra la nuova classe e gli altri punti
Le altre distanze rimangono invariate

ESERCIZIO: completa

		C1			C2
		D_B B	MB	V	KKW
C1	D_Bounty Bounty	0	88	90	
	Milo.Bar	88	0	143	
	Viking	90	143	0	
C2	KitKat.W KitKat.C				0

Passo 3

Si aggregano la classe {D_B, B} con V
Si calcolano le distanze fra la nuova classe e gli altri punti
Le altre distanze rimangono invariate

		C3		C2
C3		0		
			0	
C2				0

Passo 4

Completa

Passo 5

Tutti i punti sono riuniti in un'unica classe.

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

Passo 0: {D_B}, {B}, {MB}, {V}, {KKW}, {KKC}
Passo 1: {D_B, B}, {MB}, {V}, {KKW}, {KKC}
Passo 2: {D_B, B}, {MB}, {V}, {KKW, KKC}
Passo 3: {D_B, B, MB}, {V}, {KKW, KKC}
Passo 4: {D_B, B, MB, V}, {KKW, KKC}
Passo 5: {D_B, B, MB, V, KKW, KKC}

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.

A fianco sono riportati gli indici di aggregazione corrispondenti all'esempio precedente, dove si è utilizzato il metodo della distanza minima.

C1	34	{D_B, B}
C2	72	{KKW, KKC}
C3	88	{D_B, B, MB}
C4	143	{D_B, B, MB, V}
C5	332	

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

A fianco è riportato il dendogramma corrispondente all'esempio con 6 punti.

Sull'asse verticale è indicato l'indice di aggregazione fra le classi.

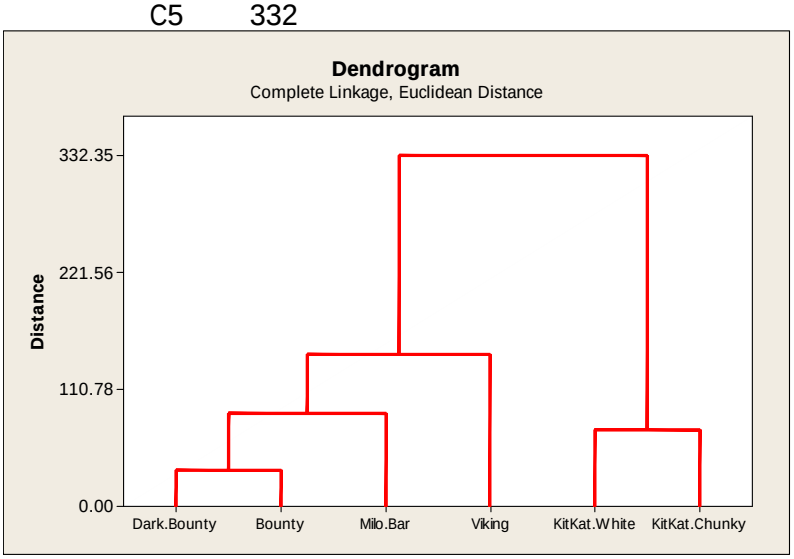


grafico 6 – dendogramma su 6 merendine

"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 200 si ottiene una partizione in 2 classi.

Ritorniamo all'esempio con tutte le merendine.

Osserviamo il grafico 4.
In **quante classi** possiamo suddividere le merendine? Scrivi i nomi delle merendine di ciascuna classe:

_____

_____

_____

_____

Come interpretiamo le classi?

Le medie delle variabili suddivise nelle tre classi sono riportate qui sotto

C	N	Size	Unit.Pr.	Energy	Protein	Fat	Carbo	Sodium
1	10	54.75	2.182	1973.0	6.020	22.300	62.45	14.44
2	4	59.50	2.303	2206.5	6.975	28.775	60.85	11.82
3	2	49.75	2.2550	1631.5	2.950	10.60	75.60	15.50

I massimi (M) e i minimi per ogni variabile sono:

C	N	Size	Unit.Pr.	Energy	Protein	Fat	Carbo	Sodium
1	10		m					
2	4	M	M	M	M	M	m	m
3	2	m		m	m	m	M	M

Quindi i tre gruppi di merendine sono caratterizzati da:

GRUPPO 1: _____

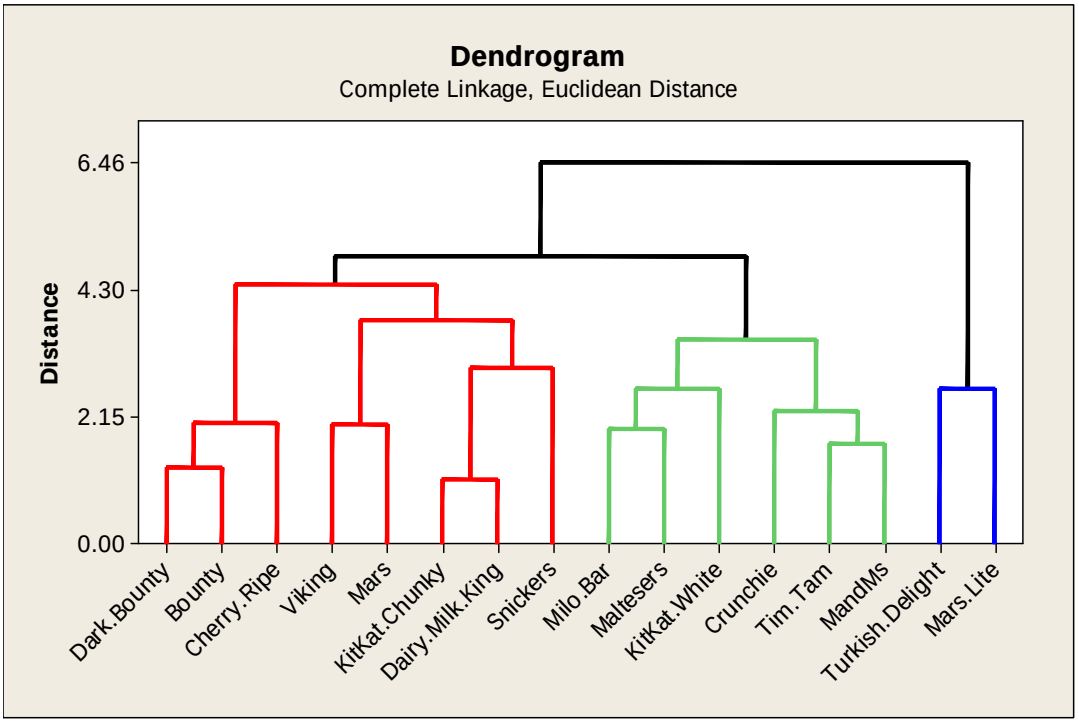
GRUPPO 2: _____

GRUPPO 3: _____

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.

Il dendogramma delle merendine con le variabili standardizzate è il seguente:



Vediamo da che cosa sono caratterizzati i tre gruppi di merendine eseminando la media di ciascuna variabile nei tre cluster. I baricentri dei cluster forniti da Minitab sono i seguenti.

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3
Size	0.631813	-0.69933	-0.42925
Unit.Pr.	-0.818591	1.06560	0.07755
Energy	0.126296	0.50178	-2.01053
Protein	-0.031252	0.44794	-1.21882
Fat	0.367660	0.17213	-1.98704
Carbo	-0.542242	0.12830	1.78406
Sodium	-0.238552	0.22704	0.27309

Completare la seguente tabella scrivendo per ogni variabile in quale cluster ha la media maggiore e la media minore. Se si decide di avere più cluster aggiungere colonne.

	Cluster 1	Cluster 2	Cluster 3
Size			
Unit.Pr.			
Energy			
Protein			
Fat			
Carbo			
Sodium			

Da che cosa sono quindi caratterizzati i raggruppamenti ottenuti?

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.

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:

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

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.

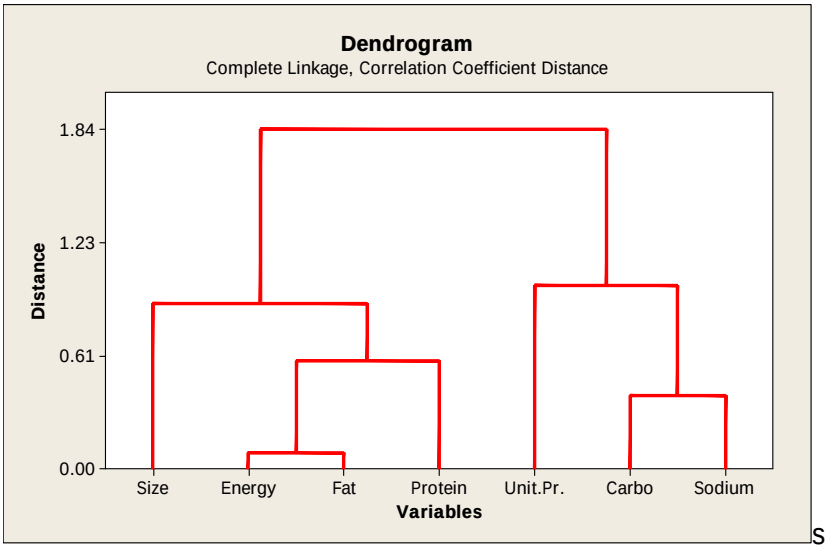
ESEMPIO - continua

Consideriamo nuovamente l'esempio relativo alle merendine.

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

	Size	Unit.Pr.	Energy	Protein	Fat	Carbo	Sodium
Size	0.00	1.57	0.89	0.85	0.90	1.20	1.01
Unit.Pr.	1.57	0.00	0.84	0.86	1.08	0.76	0.99
Energy	0.89	0.84	0.00	0.43	0.08	1.64	1.18
Protein	0.85	0.86	0.43	0.00	0.59	1.35	0.88
Fat	0.90	1.08	0.08	0.59	0.00	1.84	1.39
Carbo	1.20	0.76	1.64	1.35	1.84	0.00	0.40
Sodium	1.01	0.99	1.18	0.88	1.39	0.40	0.00

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



Commentare.

ESERCIZI

ESERCIZIO 1)

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

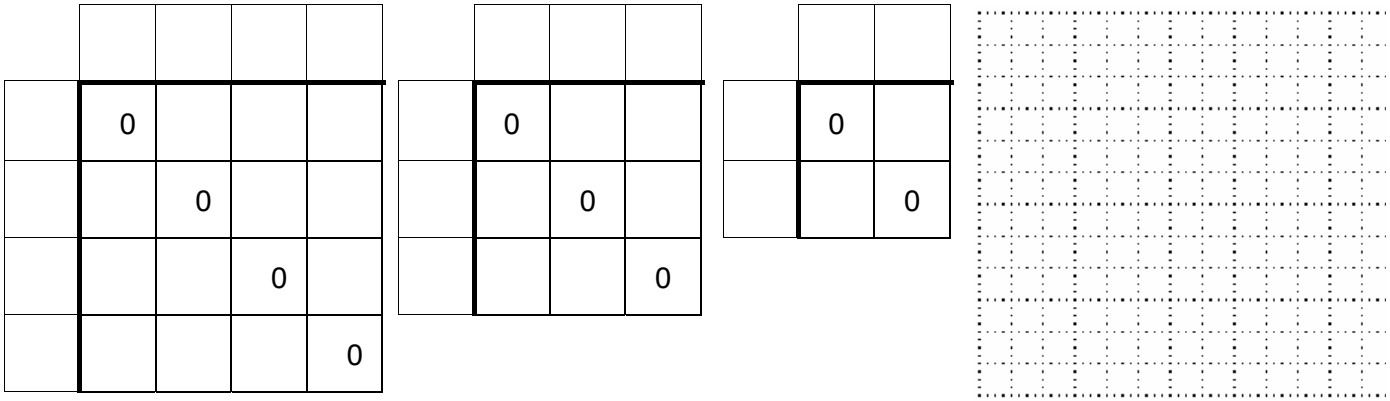
X	Y
2	4
4	2
5	1
3	4

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 il metodo della distanza media.
   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.

Cluster Centroids

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

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

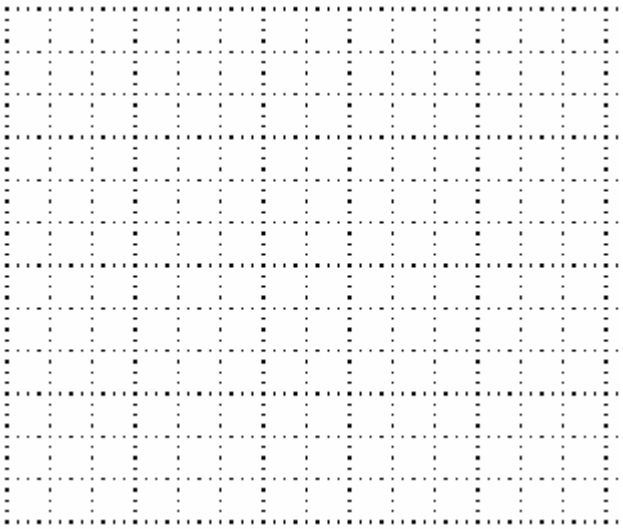
eta	peso	ossigeno	tempo	pulsferm	pulsmed	pulsmax	
-0.51376	-1.32010	-0.49215	0.38480	-0.32176	0.61986	0.24288	CLUSTER _____
1.21321	0.68145	0.84081	-0.18461	-0.45300	-0.35556	-0.41185	CLUSTER _____

ESERCIZIO 3

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.

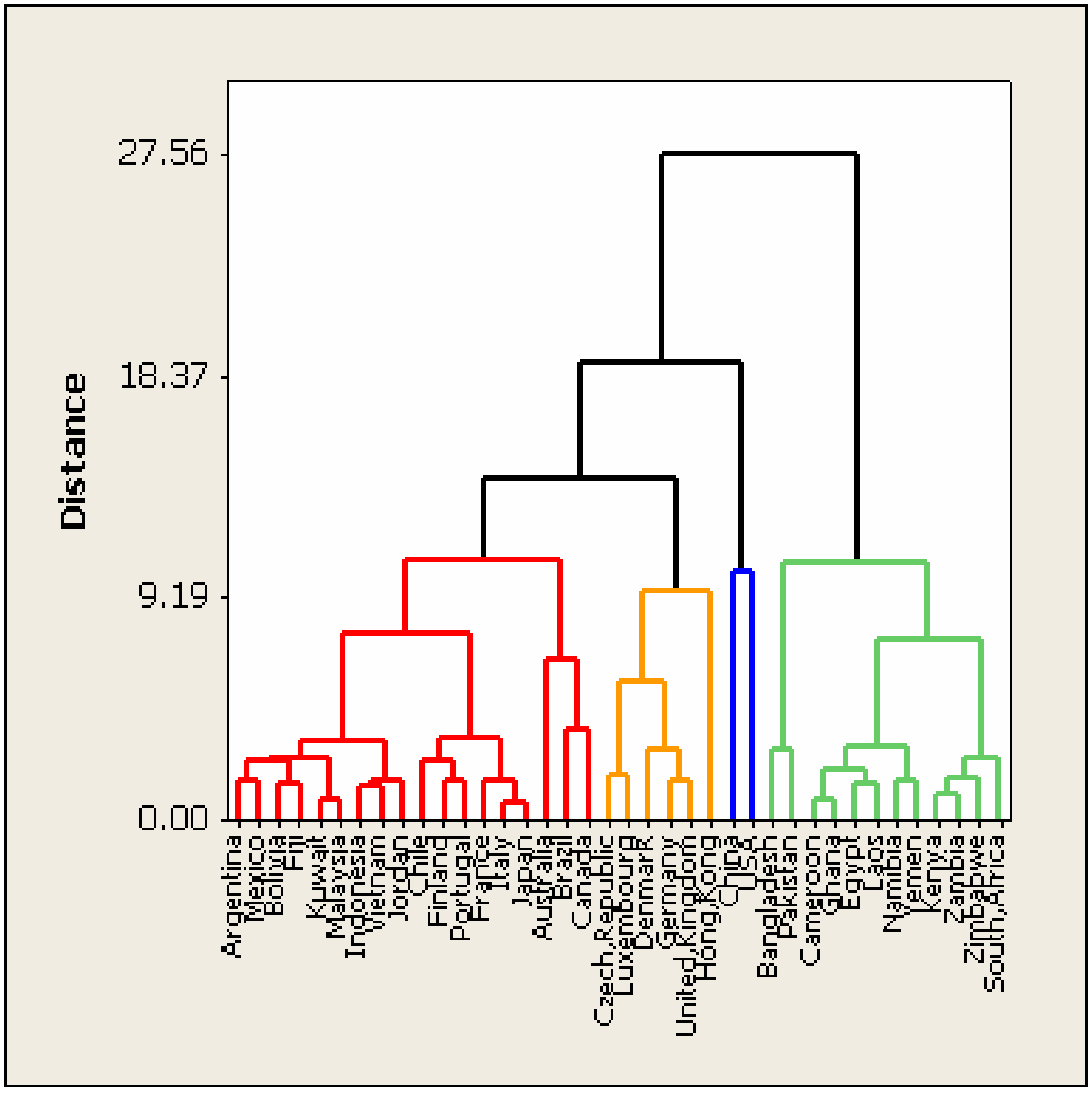
ESERCIZIO 4

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 Centroids

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



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

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.

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

Da che cosa sono caratterizzati i tre cluster?