ESERCIZIO 1. Vengono riportati di seguito i risultati di una cluster analysis gerarchica.

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	28	,008	0	0	4
2	9	38	,026	0	0	9
3	13	37	,048	0	0	27
4	2	39	,064	1	0	8
5	1	25	,075	0	0	20
6	6	18	,111	0	0	11
7	3	34	,124	0	0	16
8	2	35	,128	4	0	22
9	9	12	,129	2	0	14
10	16	40	,145	0	0	20
11	6	21	,164	6	0	17
12	10	23	,172	0	0	19
13	15	22	,184	0	0	23
14	5	9	,189	0	9	18
15	20	26	,246	0	0	28
16	3	8	,297	7	0	32
17	6	33	,305	11	0	29
18	5	11	,311	14	0	27
19	10	24	,319	12	0	30
20	1	16	,321	5	10	24
21	7	14	,388	0	0	26
22	2	29	,401	8	0	29
23	15	32	,411	13	0	25
24	1	19	,510	20	0	25
25	1	15	,557	24	23	30
26	7	17	,668	21	0	31
27	5	13	,689	18	3	32
28	20	27	,750	15	0	37
29	2	6	,862	22	17	35
30	1	10	1,085	25	19	34
31	7	36	1,189	26	0	33
32	3	5	1,313	16	27	35
33	7	30	1,341	31	0	34
34	1	7	2,040	30	33	36
35	2	3	2,246	29	32	36
36	1	2	2,968	34	35	38
37	4	20	3,281	0	28	38
38	1	4	4,622	36	37	39
39	1	31	6,687	38	0	0

1.a. In quale stadio si fondono i cluster 10 e 24 ?
1.b. Quale è il livello di distanza al quale si fondono i cluster 7 e 14 ?
1.c. Quale è il primo stadio in cui si fondono due gruppi dei quali almeno uno è formato da più di un caso, e quali sono i gruppi in questione ?
1.d. In quale stadio il gruppo 1 che si è formato allo stadio 20 si fonderà con un altro gruppo?

Cluster di appartenenza

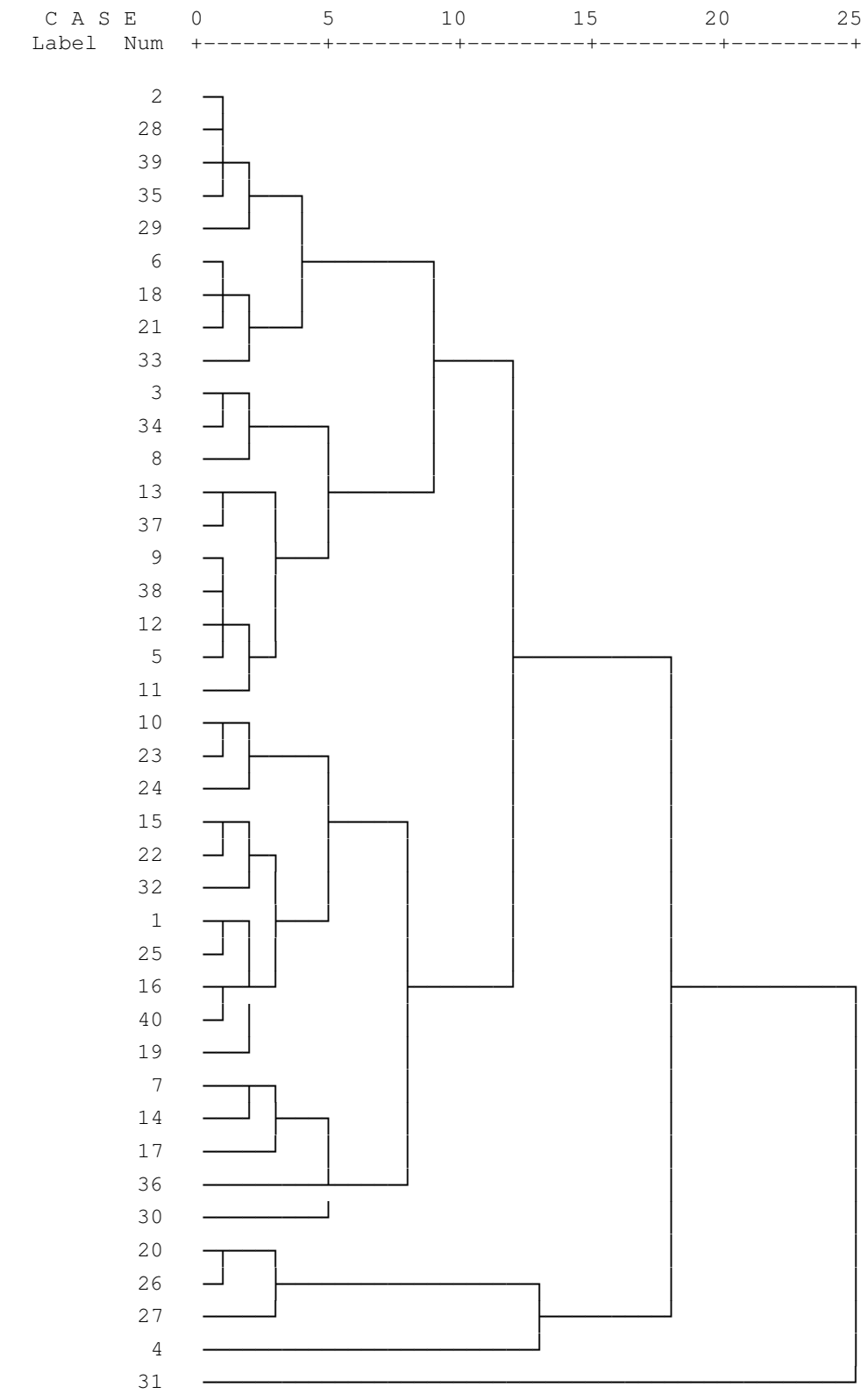
Caso	5 cluster	4 cluster	3 cluster	2 cluster
1	1	1	1	1
2	2	1	1	1
3	2	1	1	1
4	3	2	2	1
5	2	1	1	1
6	2	1	1	1
7	1	1	1	1
8	2	1	1	1
9	2	1	1	1
10	1	1	1	1
11	2	1	1	1
12	2	1	1	1
13	2	1	1	1
14	1	1	1	1
15	1	1	1	1
16	1	1	1	1
17	1	1	1	1
18	2	1	1	1
19	1	1	1	1
20	4	3	2	1
21	2	1	1	1
22	1	1	1	1
23	1	1	1	1
24	1	1	1	1
25	1	1	1	1
26	4	3	2	1
27	4	3	2	1
28	2	1	1	1
29	2	1	1	1
30	1	1	1	1
31	5	4	3	2
32	1	1	1	1
33	2	1	1	1
34	2	1	1	1
35	2	1	1	1
36	1	1	1	1
37	2	1	1	1
38	2	1	1	1
39	2	1	1	1
40	1	1	1	1

1.e. Considerando la soluzione a 4 gruppi, da quali casi sono formati i 4 gruppi così individuati ?
1.f. Considerando il dendrogramma nella pagina successiva, quale o quali soluzioni sarebbe più indicato scegliere volendo ottenere una soluzione allo stesso tempo parsimoniosa (pochi gruppi) e valida (gruppi costituiti possibilmente da soggetti simili)?

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine



ESERCIZIO 2. **Vengono riportati di seguito i risultati di una cluster analysis non gerarchica.**

Cluster di appartenenza

Numero di caso	Cluster	Distanza
1	4	,576
2	4	,391
3	3	,711
4	3	,454
5	4	,436
6	4	,734
7	4	,833
8	4	,415
9	2	,942
10	3	,556
11	1	,577
12	3	,788
13	4	,865
14	3	,592
15	4	,647
16	3	,372
17	4	,597
18	3	,662
19	2	,358
20	1	,524
21	3	,751
22	2	,355
23	4	,479
24	4	,420
25	2	,654
26	3	,361
27	4	,786
28	4	,565
29	3	,697
30	2	,586
31	1	,500
32	3	,388
33	4	,729
34	4	,520
35	3	1,198
36	1	,102
37	3	,310
38	1	1,010
39	3	,518
40	2	1,114
41	4	,395

Numero di casi in ogni cluster

Cluster	1	5,000
	2	6,000
	3	14,000
	4	16,000
Validi		41,000
Mancanti		,000

Distanze tra i centri dei cluster finali

Cluster	1	2	3	4
1		1,248	2,212	1,373
2	1,248		2,031	1,040
3	2,212	2,031		1,265
4	1,373	1,040	1,265	

Centri dei cluster finali

	Cluster			
	1	2	3	4
SEE_EM98	3,45	2,86	4,32	3,86
SEE_NE98	2,76	2,80	4,16	2,99
SEE_PO98	3,00	4,10	4,47	4,29

ANOVA

	Cluster		Errore			
	Media dei quadrati	df	Media dei quadrati	df	F	Sig.
SEE_EM98	3,251	3	,123	37	26,404	,000
SEE_NE98	4,930	3	,140	37	35,179	,000
SEE_PO98	2,789	3	,187	37	14,941	,000

I test F devono essere utilizzati solo per motivi descrittivi poiché i cluster sono stati scelti per ottimizzare le differenze tra i casi in diversi cluster. I livelli di significatività osservati non sono perciò corretti e non possono quindi essere interpretati come test dell'ipotesi che le medie dei cluster siano uguali.

2.a. Quanti sono i casi classificati nei 4 cluster ?

2.b. In quale cluster è stato classificato il caso 25 ?

2.c. Quale è la distanza euclidea tra tale caso e il centroide del cluster nel quale è stato classificato?

2.d. Quali sono i 2 cluster più vicini e quali i 2 più lontani?

2.e. Quali sono i casi che rappresentano meglio ciascuno dei 4 gruppi individuati e perché?

2.f. Quali sono le coordinate dei centroidi nella soluzione finale ?

2.g. Quali sono le variabili che differenziano meglio i cluster e perché ?

SOLUZIONI

ESERCIZIO 1

1.a. In quale stadio si fondono i cluster 10 e 24 ?
Bisogna esaminare la sezione "Cluster accorpati" e individuare quando i due cluster indicati compaiono entrambi o nella colonna Cluster 1 o nella colonna Cluster 2. Questo si verifica nello stadio 19 (prima colonna). Quindi i due cluster si fondono nello stadio 19.

1.b. Quale è il livello di distanza al quale si fondono i cluster 7 e 14 ?
Bisogna anche in questo caso esaminare la sezione "Cluster accorpati" e individuare quando i due cluster indicati compaiono entrambi o nella colonna Cluster 1 o nella colonna Cluster 2. I due cluster si fondono allo stadio 21 ad un livello di distanza pari a .388 (come appare nella colonna "Coefficienti").

1.c. Quale è il primo stadio in cui si fondono due gruppi dei quali almeno uno è formato da più di un caso, e quali sono i gruppi in questione ?
Per rispondere a questo quesito bisogna ricordare che nel Programma di Agglomerazione SPSS riporta nelle colonne "Stadio di formazione del cluster" lo stadio al quale si è formato il cluster. Dove compare il valore 0 vuol dire che il cluster è composto da un solo caso (ovvero, si è formato allo stadio 0, quando abbiamo tanti gruppi quanti sono i casi). Se invece compare un numero intero maggiore di 0, quel numero rappresenta lo stadio al quale il cluster relativo si è formato. Fino allo stadio 3 incluso, tutti cluster che si fondono si sono formati allo stadio 0, quindi sono composti da 1 solo caso (come risulta dalla sezione "Stadio di formazione del cluster". Allo stadio 4 invece si fondono il cluster 2, che allo stadio 1 si era fuso con il cluster 28, e che quindi è composto da due casi, con il cluster 39 che è composto da un solo caso.

1.d. In quale stadio il gruppo 1 che si è formato allo stadio 20 si fonderà con un altro gruppo?
Bisogna considerare la colonna "Stadio successivo" del programma di agglomerazione. In corrispondenza della riga relativa allo stadio 20 troviamo il numero "24". Vuol dire che il gruppo che si è formato nello stadio 20 dall'aggregazione del cluster 1 e del cluster 16 e che viene etichettato come cluster 1 si fonderà con un nuovo cluster allo stadio 24, ma fino a tale stadio rimarrà invariato.

1.e. Considerando la soluzione a 4 gruppi, da quali casi sono formati i 4 gruppi così individuati ?
Bisogna considerare il numero che compare nella colonna "4 cluster". Il cluster 1 è formato dai casi: da 1 a 3, da 5 a 19, da 21 a 25, da 28 a 30, e da 32 a 40; il cluster 2 è formato solo dal caso 4; il cluster 3 è formato dai casi 20, 26 e 27; il cluster 4 è formato dal solo caso 31.

1.f. Considerando il dendrogramma nella pagina successiva, quale o quali soluzioni sarebbe più indicato scegliere volendo ottenere una soluzione allo stesso tempo parsimoniosa (pochi gruppi) e valida (gruppi costituiti possibilmente da soggetti simili)?
Con un taglio del dendrogramma in prossimità del livello di distanza 10 si ottiene una soluzione che individua 5 gruppi così formati:
- due gruppi piuttosto numerosi (gruppo 1: dal caso 2 al caso 11; gruppo 2: dal caso 10 al caso 30),
- un gruppo più piccolo formato dai casi 20, 26 e 27
- due gruppi formati solo da un caso, il 4 e il 31.

Scegliere una soluzione a 4 gruppi porterebbe alla formazione di un gruppo che accorpa i due cluster più consistenti.

ESERCIZIO 2

2.a. Quanti sono i casi classificati nei 4 cluster ?
Bisogna considerare la tabella "Numero di casi in ogni cluster", dalla quale si evince che il cluster 1 è formato da 5 soggetti, il 2 da 6, il 3 da 14 e il 4 da 16.

2.b. In quale cluster è stato classificato il caso 25 ?
Bisogna considerare la tabella "Cluster di appartenenza" ed in particolare la colonna "Cluster", dalla quale si evince che il caso 25 è stato classificato nel cluster 2.

2.c. Quale è la distanza euclidea tra tale caso e il centroide del cluster nel quale è stato classificato?
Sempre nella tabella "Cluster di appartenenza" ma questa volta nella colonna "Distanza", viene riportata la distanza del caso 25 dal centroide del cluster nel quale è stato classificato: essa è pari a .654.

2.d. Quali sono i 2 cluster più vicini e quali i 2 più lontani?
Bisogna considerare la tabella "Distanze tra i centri dei cluster finali": i due cluster più vicini sono quelli che hanno la distanza più piccola, e sono il 2 e il 4. I due cluster più lontani sono quelli che hanno la distanza più grande, e sono l'1 e il 3.

2.e. Quali sono i casi che rappresentano meglio ciascuno dei 4 gruppi individuati e perché?
Bisogna considerare la tabella "Cluster di appartenenza" ed in particolare la colonna "Distanza". I casi più vicini ai centroidi dei cluster (ovvero quelli che hanno il valore della distanza più basso) sono quelli che rappresentano meglio i 4 gruppi individuati. Essi in particolare sono: il caso 36 per il cluster 1, il caso 22 per il cluster 2, il caso 37 per il cluster 3, e il caso 2 per il cluster 4.

2.f. Quali sono le coordinate dei centroidi nella soluzione finale ?
Si tratta dei valori riportati nella tabella "Centri dei cluster finali", dunque: 3.45, 2.76, 3.00 per il cluster 1; 2.86, 2.80, 4.10 per il cluster 2; 4.32, 4.16, 4.47 per il cluster 3; 3.86, 2.99, 4.29 per il cluster 4.

2.g. Quali sono le variabili che differenziano meglio i cluster e perché ?
Si tratta di esaminare i risultati presentati nella tabella ANOVA. Le variabili che differenziano meglio i gruppi sono quelle che presentano un livello di significatività più basso ovvero un valore della F più elevato. Nella tabella di questo esercizio, poiché tutte le variabili presentano lo stesso livello di probabilità, possiamo considerarle tutte ugualmente discriminative, segnalando comunque che la variabile SEE-NE98 presenta un valore della statistica F più elevato rispetto a quello delle altre variabili.