Statistica sociale - 13

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

L’individuazione di tipologie

Cluster Analysis - 1

- In questo caso il problema è quello di ***raggruppare*** un insieme di unità statistiche, descritte da un insieme di indicatori (variabili), in un numero ***ridotto*** di ***gruppi*** (***cluster***), che individuino ***tipologie*** di comportamento simili (sempre rispetto a quegli indicatori).
- È la ***Cluster Analysis***; nell'Add si definisce ***Classification Automatique***, per caratterizzare il fatto che la scelta delle tipologie è automatica, ovvero non è legata alla scelta soggettiva del ricercatore, bensì ad ***algoritmi*** formalizzati.

Cluster Analysis - 2

- I ***gruppi*** sono individuati in modo tale che le unità al loro interno siano simili fra di loro e dissimili rispetto a quelle presenti negli altri gruppi (sempre rispetto agli indicatori considerati).
- Nel caso in cui i gruppi (e le loro caratteristiche) siano conosciuti a priori, e quindi l'obiettivo sia quello di assegnare le unità statistiche ad uno di tali gruppi, si utilizzano un'altra strategia statistica: i metodi dell'***Analisi discriminante***.

Cluster Analysis - 3

- Si tratta, quindi, di un'analisi esplorativa, descrittiva, di riduzione della complessità rispetto alle righe della matrice dei dati (unità).
- Pertanto la prima operazione da compiere è quella di calcolare misure di relazione fra le unità (distanze o similarità) e non fra le variabili (frequenze, correlazioni).
- Anche in questo caso le differenze sostanziali sono legate alla natura delle variabili (qualitativa vs. quantitativa): nel primo caso si usano le ***similarità***, nel secondo le ***distanze***.

Cluster Analysis - 4

- Lo sviluppo della CA non è avvenuto nell'ambito di una singola disciplina, pertanto soluzioni simili e metodi analoghi sono contrassegnati con nomi diversi.
- Ciò ha spesso impedito un confronto fra algoritmi elaborati in ambiti disciplinari diversi, contribuendo a una sovrastima del numero dei metodi di cluster disponibili e ha reso difficile una trattazione sistematica di questa "strategia".

Cluster Analysis - 5

- In sintesi le tappe preliminari che il ricercatore deve affrontare per effettuare una classificazione sono:
  - scegliere gli indicatori (le variabili) rispetto alle quali classificare;
  - scegliere come normalizzare (standardizzare) queste variabili;
  - scegliere il peso da attribuire a ciascuna variabile nella classificazione (tenendo conto delle loro relazioni);
  - scegliere quali misure di relazione fra le unità adottare;
  - scegliere il metodo di classificazione da usare.

Cluster Analysis - 5

- Le misure di relazione fra unità sono di due tipi:
  - (dis)similarità;
  - distanze.
- Entrambi i tipi vanno applicati in tre situazioni diverse:
  - relazione fra due unità statistiche;
  - relazione fra un'unità e un gruppo di unità;
  - relazione fra due gruppi di unità.

Le similarità - 1

- I coefficienti di ***similarità*** (associazione) sono misure della relazione esistente fra due unità rispetto a un insieme di "p" caratteri comuni a entrambi.
- In genere questi caratteri (variabili) sono ***qualitativi***.
- Per calcolarli si trattano queste variabili assegnando ad ogni modalità una codifica binaria: "1" presenza; "0" assenza.
- Vediamo la tabella di base per il loro calcolo.

Le similarità - 2

Unità "i"-> / Unità "j"	Presenza	Assenza	Totale
Presenza	(a)	(b)	(a)+(b)
Assenza	(c)	(d)	(c)+(d)
Totale	(a)+(c)	(b)+(d)	(p)

con (p)=(a)+(b)+(c)+(d)

- (a) contemporanea presenza
- (b) presente solo in j
- (c) presente solo in i
- (d) contemporanea assenza (***negative match***)

Le similarità - 3

- Esempi di coefficienti di ***similarità***:
  - concordanza = (a+d)/p
  - Jaccard = a/(a+b+c)
  - Sorenson = 2a/(2a+b+c)
  - Sokal e Sneath 1 = 2(a+d)/{2(a+d)+b+c}
  - Sokal e Sneath 2 = a/{a+2(b+c)}
  - Russel e Rao = a/p

- Il campo di variazione è sempre fra "0", quando "i" e "j" si comportano in modo sempre distinto, e "1" se sempre uguale.

Le similarità - 4

- Calcolate i vari coefficienti di similarità fra queste 3 unità:

Caratteri->	1	2	3	4	5	6	7	8	9	10
Unità “1”	1	0	0	0	1	1	0	0	1	1
Unità “2”	0	0	0	0	1	0	0	1	1	0
Unità “3”	0	0	0	0	0	0	0	1	0	0

- e commentate i risultati.

Le similarità - 5

- Se i caratteri sono quantitativi si può usare il coefficiente di correlazione (Pearson) fra le unità: ma i risultati non sono sempre validi.

Caratteri->	1	2	3	4	5
Unità "1"	-1,0	-0,5	0,0	0,5	1,0
Unità "2"	-1,0	0,0	1,0	2,0	3,0
Unità "3"	-1,0	-0,5	0,0	0,5	1,5

- Commentare i risultati.
- Pertanto per caratteri quantitativi si preferisce usare le misure di ***distanza***!

Le distanze - 1

- Le misure di distanza sono caratterizzate da alcune ***proprietà matematiche***; se d(i,j) è una funzione numerica di coppie di punti, si avrà:

  (a) $d(i,j) >= 0$; $d(i,j) = 0$ se $i = j$;

  (b) $d(i,j) = d(j,i)$;

  (c) $d(i,k) + d(k,j) >= d(i,j)$

  N.B. La (c) è definita "diseguaglianza triangolare" e caratterizza quella tipologia di misure di distanza che si definisce ***metrica***.

  Le (dis)similarità, ad esempio, non rientrano in questa tipologia!

  Vi rientrano invece le più note misure di distanza.

Le distanze - 2

- Distanza ***euclidea***

$$d_{ij} = \{\Sigma_{k=1,p}(x_{ik} - x_{jk})^2\}^{1/2}$$

- x_{ik} è la determinazione del k-simo carattere nella i-esima unità;
- distanza poco adatta a dati grezzi, perché risente dell'unità di misura: meglio prima standardizzare i caratteri.

Le distanze - 3

- Distanza ***city block*** (assoluta)

$$d_{ij} = \Sigma_{k=1,p} |x_{ik} - x_{jk}|$$

- Distanza di ***Minkowski***

$$d_{ij} = \{\Sigma_{k=1,p} |x_{ik} - x_{jk}|^r\}^{1/r}$$

n.b. generalizzazione delle precedenti, per r=1,2,...

Le distanze - 4

- Distanza di ***Mahalanobis***

$$d_{ij}=(\mathbf{x}_i-\mathbf{x}_j)'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}_i-\mathbf{x}_j)$$

- $\mathbf{x}_i$ e $\mathbf{x}_j$ sono i vettori dei punteggi di "i" e "j" in tutti i caratteri considerati; $\boldsymbol{\Sigma}$ è la matrice delle varianze e covarianze fra i caratteri stessi;
- sviluppando il prodotto matriciale si ottiene una distanza euclidea "depurata" delle ridondanze dovute al fatto che i caratteri sono correlati.

Le distanze - 5

- Queste sono distanze fra ***unità***, ma vanno considerate anche le distanze fra ***gruppi*** di unità.

- Il caso di distanza fra una unità e un gruppo si può ricomprendere nel caso di distanze fra due gruppi, di cui uno è costituito da una sola unità.

- In tal caso si possono seguire quattro strategie:

Le distanze – 6

- ***group means*** (ovvero la distanza fra le medie dei gruppi);
- ***nearest neighbour*** (ovvero la distanza fra le loro unità più vicine);
- ***furthest neighbour*** (ovvero la distanza fra le loro unità più lontane);
- ***group average*** (ovvero la media fra tutte le distanze fra tutte le unità di un gruppo e quelle dell'altro).

Per tutte le strategie si possono utilizzare (dis)similarità, distanze metriche e non, a seconda delle caratteristiche delle variabili e del problema.

Tecniche di clustering

- Le tecniche di clustering possono essere classificate a seconda della strategia che si usa per affrontare il problema, che – ricordiamo – è quello di classificare le unità in gruppi, individuati in modo che le unità più simili fra di loro siano nello stesso gruppo.
- Una classificazione (anche questa è un'operazione di clustering) molto ampia è:
  - tecniche ***gerarchiche***;
  - tecniche ***non gerarchiche*** (o di ottimizzazione);
  - tecniche ***miste*** fra le prime due;
  - altre tecniche.

Tecniche gerarchiche

- I gruppi ottenuti sono aggregati in gruppi di ordine superiore (o inferiore), ripetendo il processo per tutti i livelli in modo da formare un "albero di aggregazione" (***dendrogramma***).
- Si parte da "n" gruppi di "1" sola unità e si arriva, attraverso aggregazioni delle unità via, via più simili, ad avere "1" solo gruppo di "n" unità (aggregazione gerarchica ascendente: algoritmi ***aggregativi***).
- Oppure si parte dal gruppo con "n" unità e si arriva agli "n" gruppi con "1" unità (aggregazione gerarchica discendente: algoritmi ***scissori***).

Tecniche non gerarchiche

- I gruppi si formano tramite l'aggregazione ***ottimale*** a "tipologie" prestabilite: si ottimizza una funzione obiettivo delle distanze fra i gruppi, oppure fra le unità all'interno dei gruppi.
- I gruppi sono reciprocamente esclusivi e formano una ***partizione*** dell'insieme delle unità.
- Si parte da una soluzione in cui le "n" unità sono allocate in "r" gruppi (r<<n) e si spostano le unità fra i gruppi al fine di rendere massima la distanza fra i gruppi, o minima quella fra le unità al loro interno.

Tecniche miste

- Le tecniche “gerarchiche” sono più accurate, ma si applicano con difficoltà quando la dimensione del collettivo è ampia.
- In questi casi si applicano quelle “non gerarchiche”; queste, peraltro, sono più rozze e risentono pesantemente della scelta iniziale del numero dei gruppi.
- La strategia, in questo caso, può essere migliorata introducendo nella fase finale algoritmi di tipo gerarchico.
- Si parla in questo caso di “tecniche ***miste***”.

Altre tecniche

- Come detto, la Cluster Analysis è applicata in modo diverso in molte differenti aree scientifiche. Questo lascia spazio a problematiche specifiche, risolte con tecniche "ad hoc".
- Si pensi alla necessità di dover consentire che le unità non appartengano a un solo gruppo, oppure che vi possano appartenere "*probabilisticamente*".
- Oppure a quella di dover classificare un numero imprecisato di unità, in cui si cercano gli addensamenti ("densità").
- Queste strategie non si inquadrano precisamente nelle tipologie precedenti, per cui si parla di *altre* tecniche.

Gerachiche vs. Non gerarchiche

- Le tecniche gerarchiche originano “n-2” aggregazioni (o scissioni), ovvero “n-2” partizioni a livelli successivamente decrescenti (crescenti) di distanza; non comportano la scelta preventiva del numero dei gruppi che è fatta a posteriori sulla base di una funzione delle distanze.
- Le tecniche non gerarchiche forniscono una sola partizione in “r” gruppi, pertanto la scelta di “r” è a priori e soggettiva.

Algoritmo tipo di clustering

1- Scelta delle variabili da utilizzare

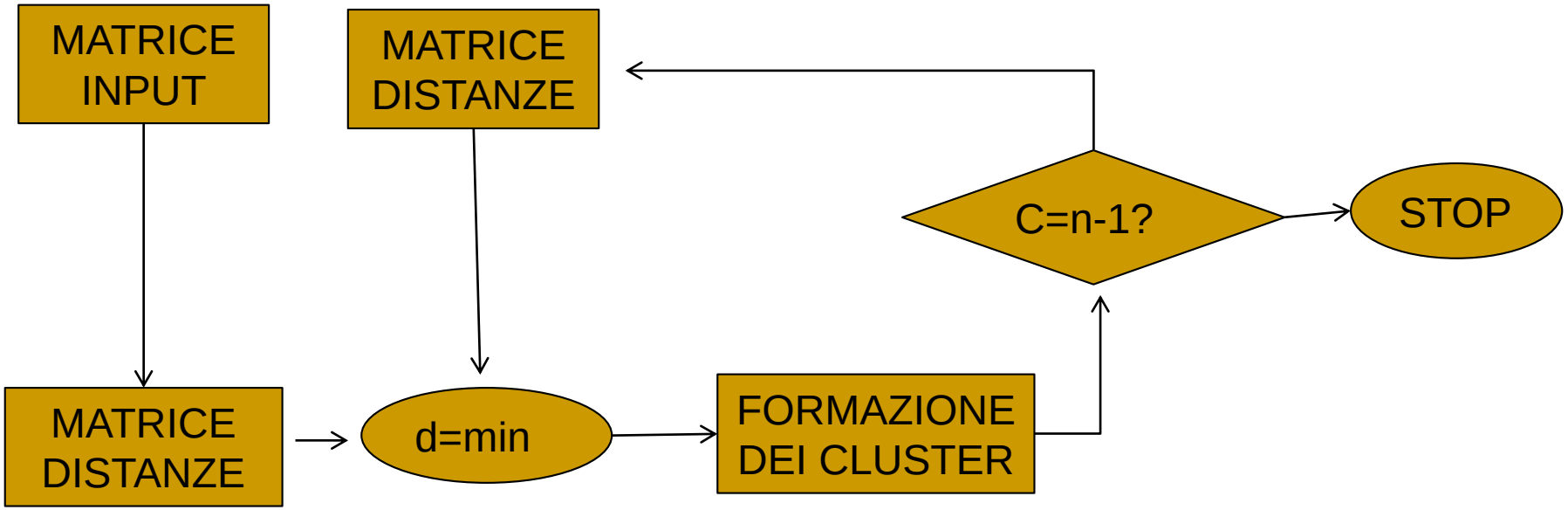
2- Scelta fra distanze e similarità

3- Scelta della tecnica di clustering

4- Scelta della tecnica di rappresentazione

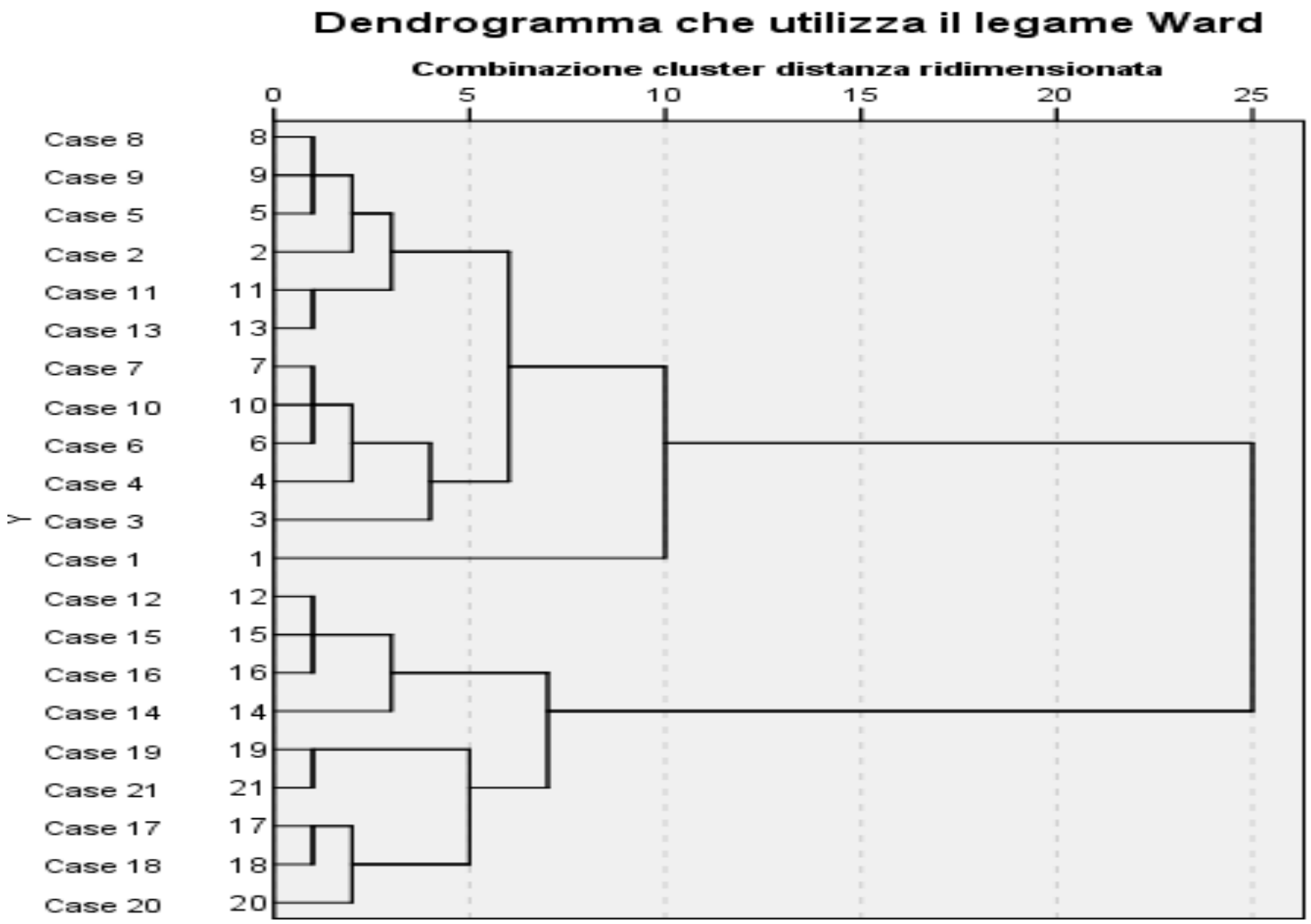
5- Scelta delle modalità di analisi dei gruppi

Algoritmo gerarchico



Il *dendrogramma* - 1

- Questa sequenza è ben rappresentata graficamente da un albero di classificazione, chiamato ***dendrogramma***, come nella figura che segue.
- Sull'asse delle ascisse è rappresentato il livello di aggregazione, che corrisponde alla misura della distanza alla quale le unità e/o i gruppi sono stati aggregati.
- Il passaggio finale è quello di *tagliare* l'albero a un certo livello: i rami tagliati individuano i gruppi, che sono costituiti dai rami che cadono (ognuno dei quali sarà associato al ramo da cui deriva).


Dendrogramma che utilizza il legame Ward
Combinazione cluster distanza ridimensionata
0
5
10
15
20
25
Case 8
Case 9
Case 5
Case 2
Case 11
Case 13
Case 7
Case 10
Case 6
Case 4
Case 3
Case 1
Case 12
Case 15
Case 16
Case 14
Case 19
Case 21
Case 17
Case 18
Case 20
8
9
5
2
11
13
7
10
6
4
3
1
12
15
16
14
19
21
17
18
20
Y

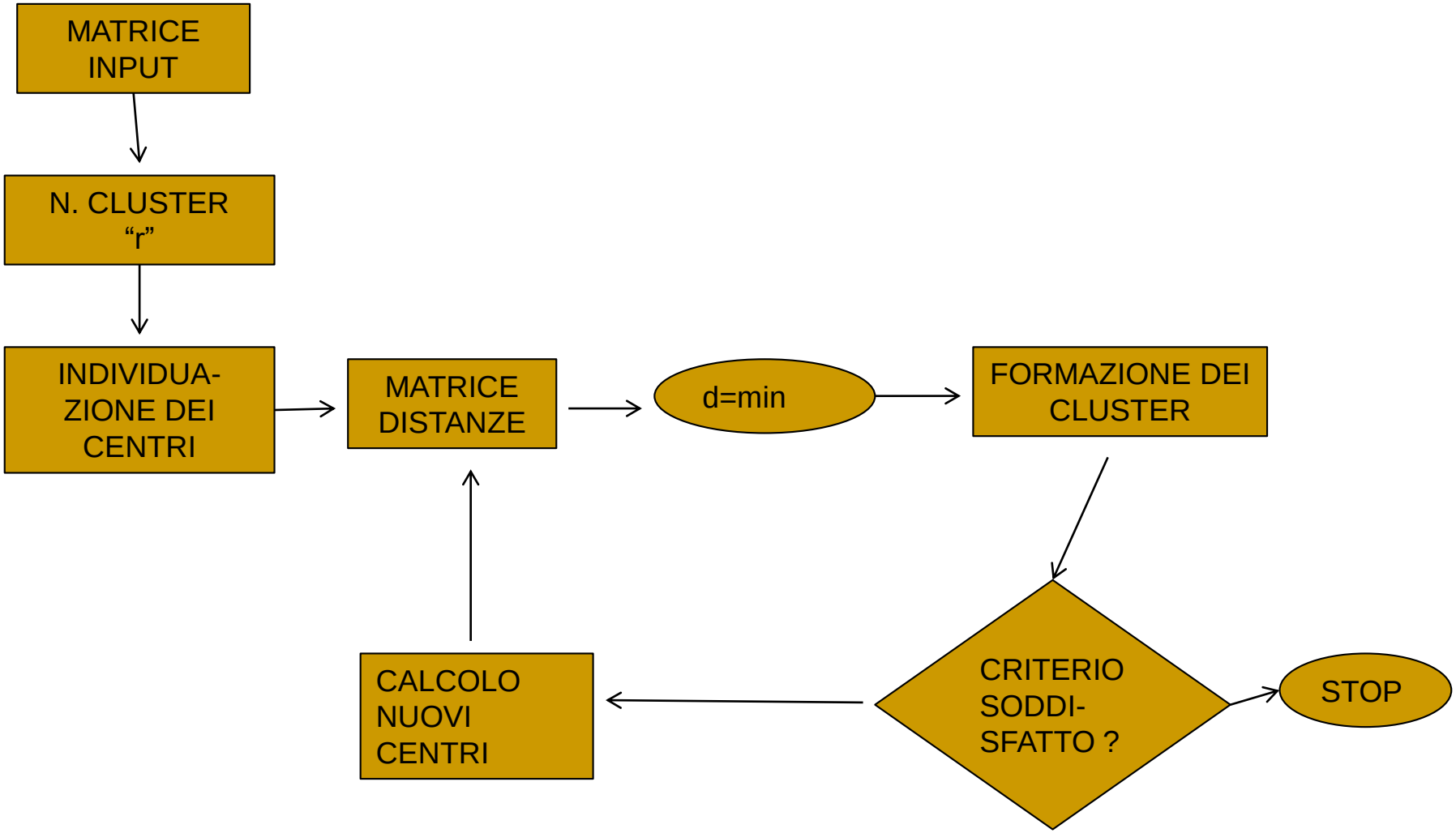
Il *dendrogramma* - 2

- La scelta dei rami da tagliare può essere effettuata con l'aiuto di ***indici*** che misurano il livello di aggregazione, ma si basa fondamentalmente sulla valutazione soggettiva del ricercatore sul significato dei gruppi individuati.
- Gli ***indici*** permettono un'analisi costi/benefici, ottimizzando il rapporto fra guadagno in sintesi (meno gruppi) e perdita di informazione, che si ha comunque mettendo insieme unità diverse.

Il *dendrogramma* - 3

- Un limite della strategia gerarchica è quello di non consentire più a unità, o gruppi, che si sono uniti, di recuperare la propria specificità.
- Pertanto, non sono possibili procedure di iterazione che tendano a ottimizzare l'appartenenza dei singoli a una tipologia, che invece è la caratteristica più interessante dei metodi non gerarchici.

Algoritmo non gerarchico



Tecniche non gerarchiche - 1

- La scelta delle unità tipo che rappresentano i centri dei gruppi può essere fatta soggettivamente o tramite estrazione casuale di una o più unità del collettivo (nel caso di più unità si considera il loro baricentro).
- Il criterio che deve essere soddisfatto per misurare la *qualità* della partizione può, ad esempio, ottimizzare l'*inerzia* tra i gruppi (*between*) spiegata dalla classificazione: l'inerzia si scompone in inerzia *between* (*inter*) e inerzia *within* (*intra*).

Tecniche non gerarchiche - 2

- Nonostante i molti passaggi questa strategia può risentire di una scelta non congrua dei centri iniziali, specialmente quando questa è casuale, e qui si inserisce la soluzione prevista da uno dei metodi misti: i ***gruppi stabili***.

elle unità
ità" della
migliore
à questa
iù potente
mpio, per
10 milioni
con una

egazione
uire in "r"
enti nella

distanze

npio (cfr.
distanza
2" gruppi:

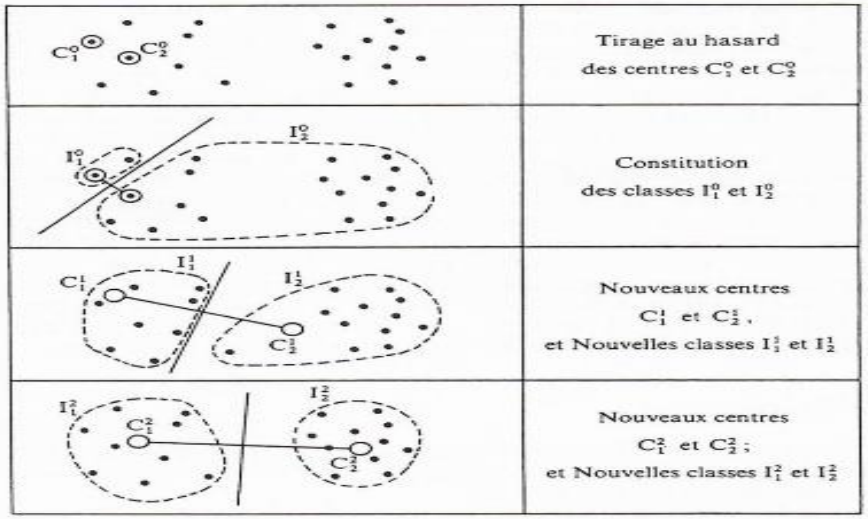
vengono
unità sono
listanti: si
denti non
gruppi: si
ne saranno

e unità più
lcolano i
inché due
he.
mero delle
razioni; si
iterazioni
l crescere
re in modo
iterazioni,
enuta.
ta, per un
alla scelta

eguire due

strade:

Fig. 5.6 - Esempio nello spazio a due dimensioni ($\mathbb{R}^2$) con due classi



Da: *Traitment des Données Statistique*, Dunod, Paris, 1982, p. 400

I gruppi stabili - 1

- Si effettuano più partizioni con la strategia non gerarchica, prevedendo lo stesso numero di gruppi, ad esempio con tre partizioni e cinque gruppi si ottengono 5^3 potenziali *gruppi stabili*.

- I gruppi stabili sono caratterizzati dal fatto che le unità che vi sono inserite hanno avuto sempre lo stesso comportamento, indipendentemente dalla scelta dei centri iniziali.

I gruppi stabili - 2

- Alcuni di questi saranno vuoti, mentre quelli con maggiore numerosità possono essere considerati i gruppi finali, aggregando in un gruppo residuo quelli meno numerosi.
- Oppure – recuperando la logica delle tecniche miste – si può applicare la strategia gerarchica alle nuove unità costituite dai gruppi stabili (ponderati col numero di unità originarie ad essi assegnate).

Un caso di studio: BES Salute - 1

- Come caso di studio riprendiamo in considerazione la matrice di dati usata per l'applicazione dell'ACP, ovvero gli indicatori relativi all'area Salute del BES.
- Sono stati eliminati dall'analisi i 3 indicatori che avevano una bassa comunalità, ovvero il «tasso di mortalità per accidenti nei trasporti», «quello per malattie del sistema nervoso» e quello dei «fumatori dai 14 anni in poi»; si era visto come, effettivamente, questi tre aspetti fossero poco legati al funzionamento di un sistema sanitario.

Un caso di studio: BES Salute - 2

- Per questa analisi, poiché le unità sono solo 21 (19 regioni e 2 province a statuto speciale), è preferibile usare una strategia gerarchica; la metrica usata è quella euclidea, lasciando le distanze al quadrato, sia per le distanze fra le unità, che fra i gruppi (*metodo di Ward*).

- Poiché gli indici sono calcolati con unità di misura diverse, ad essi è applicata la procedura di standardizzazione, in modo che tutti abbiano lo stesso peso nella procedura di clustering.

Un caso di studio: BES Salute - 3

- Seguendo i passi descritti nel paragrafo precedente si procede all'aggregazione (agglomerazione) delle unità, unendo via via le unità/gruppi fino a riformare un gruppo unico di 21 unità.
- Nell'output che segue si possono valutare i singoli passi (stadi).

Programma di agglomerazione

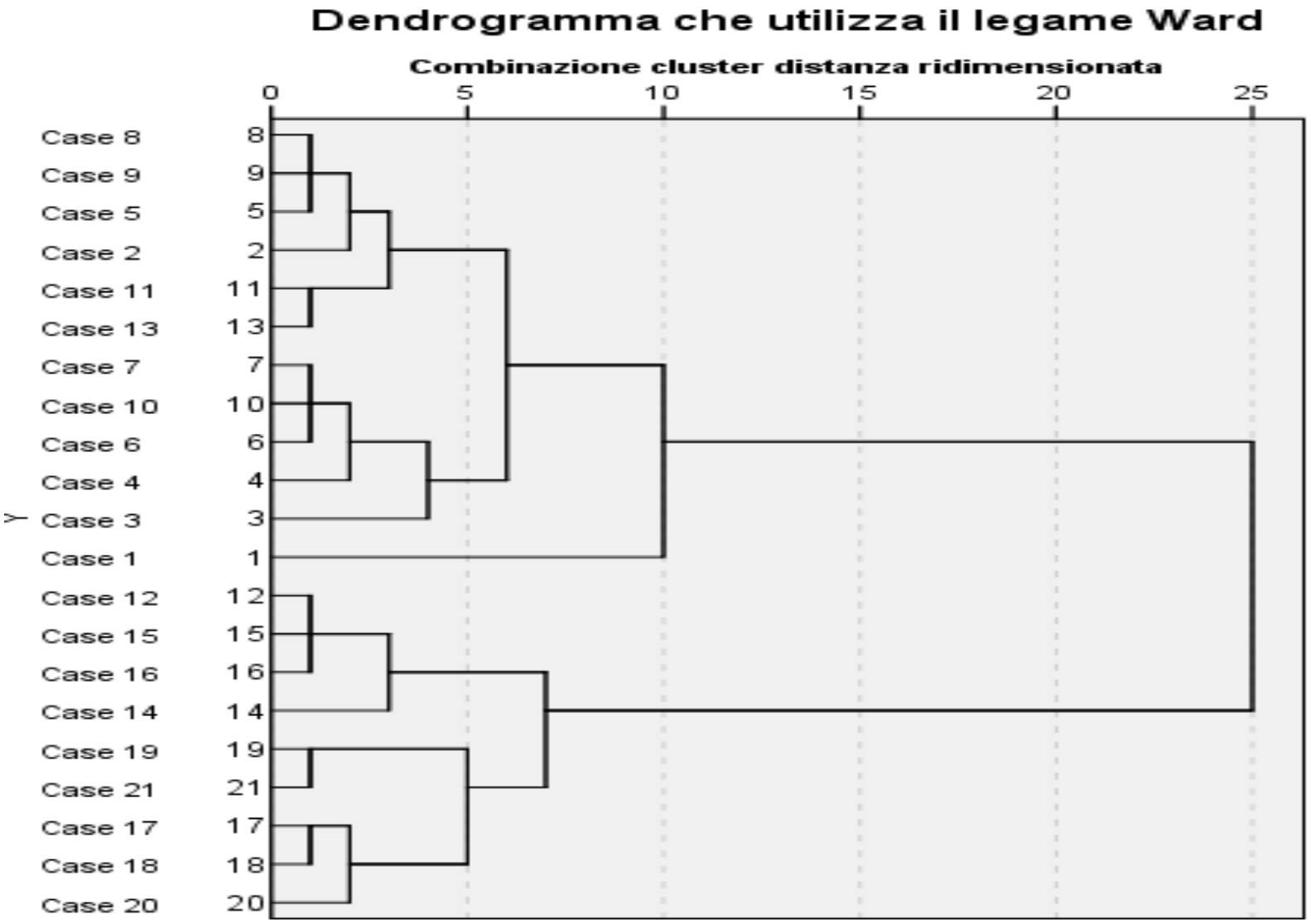
Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	9	1,803	0	0	3
2	11	13	3,838	0	0	14
3	5	8	6,135	0	1	12
4	7	10	8,914	0	0	6
5	12	15	12,157	0	0	7
6	6	7	15,565	0	4	10
7	12	16	20,188	5	0	13
8	17	18	24,929	0	0	11
9	19	21	29,676	0	0	16
10	4	6	36,060	0	6	15
11	17	20	43,164	8	0	16
12	2	5	50,370	0	3	14
13	12	14	60,598	7	0	18
14	2	11	71,880	12	2	17
15	3	4	88,295	0	10	17
16	17	19	106,550	11	9	18
17	2	3	127,345	14	15	19
18	12	17	151,410	13	16	20
19	1	2	186,336	0	17	20
20	1	12	280,000	19	18	0

Un caso di studio: BES Salute - 4

- Al primo stadio le unità più vicine (con coefficiente di distanza di Ward pari a 1,803) sono la 8 (Toscana) e la 9 (Emilia Romagna);
  - al gruppo che si costituisce con la fusione delle due unità viene mantenuto il numero 8 come codice;
- al secondo stadio (coefficiente 3,838) si uniscono la 11 (Marche) e la 13 (Umbria);
- al terzo stadio la 5 (Veneto) e la 8 (il gruppo Toscana ed Emilia Romagna);
- e così via.

Un caso di studio: BES Salute - 5

- I coefficienti di distanza aumentano sempre di più, segnalando il fatto che si stanno unendo unità o gruppi fortemente disomogenei.
- Se osserviamo il dendrogramma relativo, sull'asse delle ascisse è individuata una scala che riporta la *distanza ridimensionata*, ovvero una trasformazione normalizzata del coefficiente di distanza.


Dendrogramma che utilizza il legame Ward
Combinazione cluster distanza ridimensionata
0
5
10
15
20
25
Case 8
Case 9
Case 5
Case 2
Case 11
Case 13
Case 7
Case 10
Case 6
Case 4
Case 3
Case 1
Case 12
Case 15
Case 16
Case 14
Case 19
Case 21
Case 17
Case 18
Case 20
8
9
5
2
11
13
7
10
6
4
3
1
12
15
16
14
19
21
17
18
20
Y

Un caso di studio: BES Salute - 6

- Se noi tagliamo l'albero quando questa raggiunge la soglia di 5 restano 5 rami: questa potrebbe essere una soluzione che combina il vantaggio di ridurre il numero dei gruppi (massima sintesi) e di non mettere insieme gruppi troppo disomogenei fra di loro (minima perdita di informazione).
- Ovviamente si tratta di un punto di equilibrio fra i due criteri, perché la massima sintesi si avrebbe con 1 solo gruppo e la minima perdita di informazione con 21 gruppi!

Un caso di studio: BES Salute - 7

- Quali sono le unità che compongono i cinque gruppi che sono individuati dalla soluzione scelta?
- Nel programma di agglomerazione i gruppi ancora in gioco sono quelli non ancora uniti al 17 stadio, ovvero
  - il 2 (unione con 5, 8 e 9 e con 11 e 13);
  - il 3 (unione con 4 e 6, con 7 e 10);
  - il 12 (unione con 15, 16 e 14);
  - il 17 (unione con 18, 20 e 19 con 21);
  - l'1 (che presenta una sua specificità ed è il più distante da tutti gli altri, tanto è vero che si unisce solo al penultimo stadio).

Un caso di studio: BES Salute - 8

- Le stesse considerazioni si possono fare, con minor accuratezza, osservando il dendrogramma.
- Riprendendo la corrispondenza con le regioni, si individuano le appartenenze ai cinque gruppi:
  - gruppo A – Toscana, Emilia Romagna, Veneto, Marche, Umbria e Provincia di Trento;
  - gruppo B – Lombardia, Piemonte, Liguria cui si uniscono le regioni autonome Friuli Venezia Giulia e poi Valle d'Aosta;
  - gruppo C – la sola Provincia di Bolzano;
  - gruppo D – Sardegna, Lazio, Abruzzi e Molise;
  - gruppo E – Basilicata, Puglia e Calabria con Sicilia e Campania.

Un caso di studio: BES Salute - 9

- Nel prosieguo dell'analisi si aggregano le regioni del Centro Nord (gruppi A e B); poi quelle del Mezzogiorno (gruppi D ed E); poi la Provincia di Bolzano si aggrega al Centro Nord; infine nell'ultimo stadio tutte le unità confluiscono in un unico gruppo.
- Il commento ai risultati è piuttosto facile, poiché la Cluster Analysis fotografa piuttosto bene le diverse tipologie della situazione della Sanità nel nostro paese, diviso in due tra Centro Nord e Mezzogiorno, con diverse specificità nelle due macro zone e l'anomalia della Provincia autonoma di Bolzano.

La *cluster membership* - 1

- Questo commento è possibile perché il numero di unità è ridotto e, pertanto, la semplice descrizione dell'appartenenza ai cluster ci consente di caratterizzarli.
- Se il numero di unità è grande (ad esempio se considerassimo gli oltre 8.000 comuni italiani, oppure gli oltre 50.000 cittadini che hanno risposto all'Indagine Multiscopo "Tempo libero e cultura"), allora non avrebbe molto senso elencare l'appartenenza ai gruppi.

La *cluster membership* - 2

- Forse nel primo caso potrebbe essere interessante individuare in quali cluster si collochino i grandi centri metropolitani (come Roma, Milano, Torino, Napoli), o altre unità caratterizzate da storia o caratteristiche peculiari (Firenze, Venezia, Lecce, Assisi, Caserta e così via).
- Nel secondo sarebbe interessante vedere in quale gruppo prevalgano gli uomini, i giovani, i laureati, in quale altro le donne, gli anziani, chi risiede nell'Italia insulare e così via.

La *cluster membership* - 3

- In tal caso per descrivere le caratteristiche dei gruppi il primo passo è quello di registrare nella matrice dei dati una nuova variabile: il gruppo di appartenenza (*cluster membership*).
- È una variabile qualitativa non ordinabile (in quanto la strategia di classificazione non dà una gerarchia ai gruppi ottenuti), che possiamo elaborare congiuntamente con le altre variabili presenti nella matrice.

La *cluster membership* – 4

- Le modalità di analisi dipendono quindi dalla variabile con la quale vogliamo caratterizzare i gruppi:
  - se è qualitativa useremo la strategia di analisi delle tabelle doppie;
  - se è quantitativa il confronto fra medie;

La *cluster membership* - 5

- se è una variabile che rappresenta una caratteristica strutturale (genere, età e così via) allora la *cluster membership* sarà considerata come variabile dipendente;
- se è una variabile che descrive le modalità di comportamento o di atteggiamento (pratica di uno sport, durata dell'allenamento, motivazioni per la pratica e così via) allora la *cluster membership* sarà considerata come variabile indipendente.