Capitolo V

ANALISI DEI GRUPPI

di Andrea Cerioli e Sergio Zani

> «*Dividere in categorie significa considerare, in un gruppo d'individui, non quello che ciascuno ha di proprio, ma quello che ha di comune col gruppo a cui appartiene. Così una classificazione è possibile, una esatta riduzione a generi e specie*»
>
> (Francesco De Sanctis, *Storia della letteratura italiana*).

1. *La classificazione delle unità statistiche*

In questo capitolo ci occuperemo dei metodi statistici per la classificazione delle unità in gruppi omogenei. Questo problema può essere distinto in:

— discriminazione (o analisi discriminante);

— analisi dei gruppi (o *cluster analysis*).

Nel primo caso è noto a priori che le n unità osservate appartengono a due (o più) popolazioni differenti. Per ogni unità si conosce il corrispondente vettore dei valori delle p variabili. L'obiettivo dell'analisi discriminante è quello di stabilire un criterio (basato sulle p variabili) per assegnare correttamente ulteriori unità alla rispettiva popolazione di appartenenza, minimizzando la probabilità degli errori di attribuzione.

Un esempio è fornito dagli n studenti che si sono iscritti ad una facoltà universitaria 6 anni fa. Essi sono distinguibili in due gruppi: coloro che alla data odierna risultano laureati ed i restanti (formati da coloro che hanno abbandonato gli studi e da quelli ancora iscritti alla facoltà). Supponiamo di conoscere per ciascuno degli n studenti un insieme di p variabili: votazione alla maturità, numero di esami so-

stenuti nel primo anno di corso, voto medio negli esami del primo anno, reddito familiare. L'analisi discriminante si propone d'individuare un'opportuna funzione di tali variabili che consenta di distinguere le due popolazioni (laureati e non). Conoscendo quindi i valori delle medesime variabili per un *nuovo* insieme di studenti (ad esempio, quelli iscritti lo scorso anno), è possibile prevedere quelli che si laureeranno entro sei anni dall'immatricolazione, distinguendoli dai restanti, con una ridotta probabilità di errore nell'assegnazione degli studenti alle due popolazioni.

Un secondo esempio, d'interesse economico, è la classificazione delle aziende che richiedono un prestito ad una banca in « aziende solvibili » ed « aziende a rischio ». Sulla base dei dati passati, la banca può distinguere n aziende che hanno ottenuto un prestito in due categorie: quelle che hanno rimborsato regolarmente il medesimo e le altre (che hanno ritardato i pagamenti, non hanno effettuato i rimborsi, etc.). Se per ciascuna di tali aziende la banca conosce un insieme di variabili — desumibili ad esempio dai bilanci — mediante l'analisi discriminante è possibile classificare una nuova azienda che richiede un prestito (e che fornisce i valori delle suddette variabili) nella classe di quelle solvibili oppure di quelle a rischio, minimizzando la probabilità di errore nell'assegnazione. Chiaramente questa conoscenza costituirà la premessa essenziale per la decisione di concedere o meno il prestito.

L'analisi dei gruppi — comunemente indicata con la dizione inglese *cluster analysis* — è invece un metodo tipicamente esplorativo: esso consiste nella ricerca nelle n osservazioni p-dimensionali di gruppi di unità tra loro simili, non sapendo a priori se tali gruppi omogenei esistono effettivamente nel *data set*. Classificazioni di questo tipo assumono una chiara valenza interpretativa solo nei casi in cui nei dati sono realmente presenti delle strutture di gruppo, che vengono individuate dalla metodologia statistica. La *cluster analysis* ha dunque l'obiettivo di riconoscere gruppi che appaiono con « naturalezza » nelle osservazioni (1).

Un classica applicazione della *cluster analysis* è la segmentazione del mercato, che può essere riferita ai tipi di prodotto o alle categorie

---

(1) In letteratura (Kendall, 1966) si distingue la *cluster analysis* dalla tecniche di *dissection*, che si propongono di suddividere un generico insieme in gruppi ottimizzando una funzione obiettivo. « *All collection of entities can be dissected; not all can be classified* » (Cormack, 1971).

di consumatori. Consideriamo, ad esempio, i modelli di automobili in vendita in Italia. Se per ciascuno di essi conosciamo un insieme di variabili (cilindrata, prezzo, peso, velocità massima, etc.), con la *cluster analysis* possiamo suddividere tali modelli in gruppi almeno relativamente omogenei (i «segmenti» del mercato automobilistico) e questa conoscenza è molto importante per impostare le strategie delle aziende produttrici, poiché la concorrenza si esercita quasi esclusivamente tra autovetture del medesimo segmento. Un altro impiego importante della *cluster analysis* è la classificazione dei comuni d'una regione in gruppi omogenei in base ad una pluralità d'indicatori demografici, economici e sociali. Si possono identificare delle aree che presentano analogie di situazioni e quindi richiedono interventi comuni di politica economica.

Data l'impostazione del presente volume, in questo capitolo illustreremo solo l'analisi dei gruppi (2), limitandoci peraltro ai temi essenziali e di maggior interesse applicativo (3).

Vogliamo sottolineare che con la *cluster analysis* si consegue una sorta di riduzione delle dimensioni di R^n: dalle n unità osservate inizialmente (spesso molto numerose), si perviene a g gruppi omogenei ($g \ll n$), con il vantaggio d'una notevole parsimonia nella descrizione e d'una interpretazione più semplice.

2. *Le scelte nell'analisi dei gruppi*

Un medesimo insieme di unità statistiche può essere classificato — com'è ovvio — in molti modi diversi. Si pensi, ad esempio, ai titoli quotati nella Borsa di Milano. Essi possono venire classificati in base al settore di appartenenza, secondo la *performance* realizzata dall'inizio dell'anno, considerando se essi appartengono al MIB 30, al MIDEX o agli altri titoli, etc.

---

(2) Per un'ampia trattazione dell'analisi discriminante rinviamo a McLachlan (1992).

(3) La bibliografia sulla *cluster analysis* è vastissima. Citiamo solo alcuni volumi che possono essere utili al lettore per un approfondimento: Anderberg (1973); Hartigan (1975); Späth (1980; 1985); Jambu et Lebeaux (1983); Jain and Dubes (1988); Kaufman and Rousseeuw (1990); Everitt (1993); Arabie, Hubert and De Soete (1996); Mirkin (1996); Hand (1997); Gordon (1999). Ricordiamo anche che ai temi della classificazione è dedicata — a partire dal 1984 — una rivista specifica, il *Journal of Classification*.

La prima domanda che ci si deve porre di fronte ad un problema classificatorio è allora la seguente: per quale motivo si vuole effettuare una classificazione delle unità considerate?

La definizione precisa dello scopo della classificazione rappresenta dunque un *prius* di tutte le scelte successive che deve effettuare il ricercatore, che elenchiamo nel prosieguo.

i) *Scelta delle variabili*

Dalle finalità assegnate ad una *cluster analysis* discende la scelta delle variabili. A questo riguardo la metodologia statistica è di scarso aiuto: sono le conoscenze specifiche del ricercatore in ordine al problema in esame che possono indirizzare la scelta, che conserva però larghi margini di soggettività. La classificazione dovrebbe fondarsi su tutti gli aspetti che si ritengono importanti per gli scopi prefissati, e questo potrebbe indurre ad ampliare il più possibile le variabili rilevate. Occorre tener presente, tuttavia, che l'aggiunta di variabili con scarso potere di discriminazione tra i gruppi, ovvero con bassa «qualità» dei dati, può peggiorare i risultati, rendendo meno nitida la classificazione ottenuta.

Un criterio, suggerito soprattutto dagli studiosi di scienze naturali, è quello di considerare una pluralità di variabili, in modo tale che l'eliminazione d'una qualsiasi tra esse, ovvero l'aggiunta di un'ulteriore variabile, lasci pressoché inalterati i gruppi individuati. Ad esempio, la classificazione d'un insieme di specie vegetali dovrebbe rimanere all'incirca la medesima quando si toglie o si aggiunge una variabile all'insieme dei p fenomeni considerati (sufficientemente numerosi). In altri termini, una classificazione ragionevole delle unità statistiche non dovrebbe mostrare «sensibilità» eccessiva rispetto a piccoli cambiamenti dell'insieme di variabili sulle quali essa si fonda.

L'analisi delle componenti principali applicata in via preliminare a tutte le variabili disponibili può essere di aiuto nella scelta di quelle da utilizzare nella *cluster analysis*. Se k componenti principali tengono conto d'una percentuale elevata della varianza totale, il ricercatore può effettuare la classificazione direttamente sugli *scores* di tali CP, che costituiscono il «segnale» degli aspetti rilevanti, mentre le restanti componenti rappresentano i residui, cioè il «rumore» (*noise*). In alternativa, il ricercatore potrebbe applicare la *cluster analysis* ad un sottoinsieme delle variabili di partenza, e precisamente solo a quelle più fortemente correlate con le prime k componenti principali, essendo le restanti poco connesse con gli aspetti fondamentali dell'indagine.

Nonostante questi accorgimenti, la scelta delle variabili rimane comunque un problema cruciale dell'analisi, che può condizionare fortemente i risultati della stessa (4).

ii) *Scelta della distanza o dell'indice di similarità*

Date n unità, alle quali corrispondono i vettori p-dimensionali $\mathbf{x}_i$, molti metodi di *cluster analysis* richiedono il calcolo della matrice delle distanze (ovvero degli indici di similarità), che contiene le misure di « prossimità » tra tutte le coppie di unità. Il ricercatore deve quindi scegliere nel caso di variabili quantitative la distanza o l'indice di distanza, e nel caso di fenomeni qualitativi il tipo di indice di similarità. Anche questa scelta condiziona i risultati della classificazione, poiché variando il tipo di distanza non rimane immutato in generale l'ordinamento delle coppie di unità (da quelle tra loro più simili a quelle più diverse) e quindi possono differire anche i gruppi di unità « omogenee ». La scelta della distanza più opportuna in un'applicazione concreta deve basarsi sulle caratteristiche delle singole metriche, ampiamente descritte nel capitolo precedente.

iii) *Scelta del metodo di formazione dei gruppi*

Ci limitiamo per ora a considerare solo i metodi che individuano partizioni dell'insieme delle n unità statistiche (5). L'obiettivo è quello di classificare tali unità in gruppi con le caratteristiche di *coesione interna* (le unità assegnate ad un medesimo gruppo devono essere tra loro simili) e di *separazione esterna* (i gruppi devono essere il più possibile distinti).

I metodi di formazione dei gruppi vengono distinti in gerarchici e non gerarchici.

I metodi gerarchici consentono di ottenere una famiglia di partizioni, con un numero di gruppi da n a 1, partendo da quella banale in cui tutte le unità sono distinte per giungere a quella, pure banale, in cui tutti gli elementi sono riuniti in un unico gruppo (6). I metodi

---

(4) A questo riguardo si vedano: Fowlkes *et al.* (1988); Gnanadesikan *et al.* (1995).

(5) Per la definizione del concetto di partizione, si veda il Vol. I, p. 41.

(6) Tali metodi gerarchici sono di tipo aggregativo, o *bottom up*, poiché partono dal basso, cioè dalle singole unità statistiche, e procedono a riunificazioni successive delle medesime. Vi sono anche metodi chiamati scissori, o *top down*, che par-

non gerarchici forniscono un'unica partizione delle n unità in g gruppi, con g fissato a priori.

Una larga parte di questo capitolo sarà dedicata all'illustrazione particolareggiata di vari metodi gerarchici e non gerarchici di classificazione.

iv) *Criteri di valutazione delle partizioni ottenute ed individuazione del numero ottimo di gruppi*

Dopo aver ricavato la famiglia di partizioni (nei metodi gerarchici) oppure la singola partizione (nei metodi non gerarchici), occorre valutare la classificazione ottenuta per vedere se essa soddisfa le condizioni di coesione interna e di separazione esterna. A livello intuitivo (lasciando l'approfondimento ad un momento successivo), possiamo dire che una partizione è soddisfacente quando la variabilità all'interno dei gruppi individuati è piccola (le unità di ogni gruppo presentano modeste differenze tra loro) ed inoltre i gruppi sono ben distinti l'uno dall'altro.

Per individuare nella famiglia delle $(n-2)$ partizioni non banali ottenute con un metodo gerarchico quella che presenta il numero «ottimo» di gruppi, occorre tener conto che esiste un *trade-off* tra il numero dei gruppi e l'omogeneità all'interno degli stessi: riducendo il numero di gruppi si ottiene una classificazione più sintetica, e quindi generalmente più utile a fini operativi, ma si deve pagare un prezzo in termini di maggiore variabilità nei gruppi, poiché si aggregano unità maggiormente diverse tra loro. La partizione con il numero ottimo di gruppi sarà dunque quella che meglio riesce a contemperare queste opposte esigenze di sintesi delle unità in classi e di coesione interna dei gruppi.

Nei metodi non gerarchici il numero ottimo di gruppi può essere individuato per tentativi, ripetendo più volte la procedura con diversi valori di g, valutando in ogni applicazione la bontà della partizione ottenuta e scegliendo poi quella più soddisfacente.

Le numerose scelte che deve effettuare il ricercatore nella *cluster analysis* introducono elementi di soggettività nei risultati e quindi si prestano a critiche. Occorre precisare, tuttavia, che una classificazione d'un insieme di unità statistiche può ritenersi valida quando essa ri-

---

tono dall'insieme di tutte le unità statistiche ed effettuano successive separazioni in gruppi di tale insieme.

mane almeno approssimativamente stabile al variare degli algoritmi utilizzati per ottenerla, poiché in tal caso essa riflette una struttura realmente presente nei dati multidimensionali, e non è generata semplicemente dalla particolare procedura utilizzata (7). Nelle applicazioni concrete è pertanto consigliabile utilizzare più possibilità per ciascuna delle scelte sopra descritte e confrontare tra loro le varie classificazioni ottenute.

3. *Un criterio esplorativo per la verifica dell'esistenza di gruppi*

L'esistenza d'una reale struttura di gruppo è una condizione fondamentale per un'applicazione «sensata» della *cluster analysis*. Se le osservazioni a disposizione del ricercatore non soddisfano una simile prerogativa, l'analisi effettuata tramite gli algoritmi di classificazione (gerarchici e non gerarchici) illustrati nel presente capitolo rischia infatti di rivelarsi sostanzialmente inutile. Anche il requisito — menzionato nel paragrafo precedente — della stabilità delle conclusioni cui si perviene al variare delle opzioni scelte per ottenerle può essere soddisfatto solo in presenza di dati effettivamente riconducibili a gruppi distinti. Risulta pertanto importante disporre di procedure che consentano di verificare, in via preliminare rispetto all'applicazione della *cluster analysis*, la presenza d'una effettiva struttura di gruppo nei dati multidimensionali.

Un approccio formalizzato al problema, di natura inferenziale, consiste nel sottoporre a verifica l'ipotesi nulla di assenza di struttura nei dati. In particolare, tale ipotesi è abitualmente caratterizzata mediante due assunzioni alternative: il cosiddetto «modello uniforme», secondo cui le n unità sono rappresentate da punti distribuiti uniformemente nello spazio p-dimensionale, ed il «modello unimodale», secondo cui le n osservazioni sono realizzazioni di variabili aleatorie p-dimensionali con la stessa funzione di densità unimodale (8). Un criterio più semplice

---

(7) Si può in tal caso parlare di «stabilità» della *cluster analysis* (Gordon and De Cata, 1988). Un concetto simile — anche se non equivalente — è quello di «validità» della procedura di classificazione, che si riferisce alla sua capacità di mettere in luce le strutture di gruppo realmente presenti nei dati. Per un'ampia rassegna al riguardo si rimanda a Milligan (1996) e Gordon (1999, cap. 7).

(8) Un'esposizione dettagliata dei metodi di verifica di tali modelli esula dagli obiettivi della presente trattazione; per ulteriori approfondimenti si vedano Bock (1985), Hartigan (1985) e le rassegne di Bock (1996a, b) e Gordon (1996a).

e di tipo esplorativo, che consente comunque di individuare l'eventuale esistenza di gruppi e la loro forma approssimativa, si fonda invece sulla costruzione del cosiddetto istogramma p-dimensionale.

Un istogramma p-dimensionale costituisce la generalizzazione al caso di p caratteri del tradizionale concetto di istogramma per una singola variabile (9). Per semplicità d'esposizione, illustriamo dapprima in dettaglio la situazione più semplice, cioè quella bidimensionale (in cui $p = 2$).

i) Sono date n unità statistiche, per ciascuna delle quali si conoscono i valori di due variabili, X_1 e X_2, che supponiamo continue. A partire da tale matrice dei dati si costruisce il corrispondente diagramma di dispersione (*scatterplot*).

ii) Si fissa un'origine x_0 nello spazio bidimensionale per la costruzione dell'istogramma. In particolare, l'origine può essere fissata in corrispondenza del punto che ha per coordinate i valori minimi osservati sulle due variabili.

iii) Al diagramma di dispersione in R^2 si sovrappone, a partire da x_0, una griglia regolare costituita da rettangoli di dimensione $h_1 \times h_2$. Corrispondentemente, si indica con q il numero di elementi della griglia in ciascuna dimensione e con $Q = q \times q$ il numero totale di rettangoli.

iv) Si conta il numero di punti nello *scatterplot* che risultano compresi in ciascun rettangolo. Dopo tale operazione n_j rappresenta la frequenza (assoluta) associata al rettangolo j ($j = 1, \ldots Q$); ovviamente, si ha $\sum_{j=1}^{Q} n_j = n$.

v) L'istogramma bidimensionale è ottenuto dalla rappresentazione delle *densità di frequenza*:

$$d_j = \frac{n_j}{n h_1 h_2} \qquad j = 1, \ldots, Q, \tag{5.1}$$

ove $h_1 h_2$ è l'area di ciascun rettangolo.

Osservazione. È chiaro che le informazioni traibili dalla rappresentazione delle densità di frequenza d_j sono equivalenti a quelle desumibili dalle rispettive frequenze n_j nell'ipotesi, assunta al punto iii), che i rettangoli abbiano tutti la stessa dimensione $h_1 \times h_2$. In caso contrario, la distinzione tra densità e frequenza assoluta diviene essenziale e

(9) Si veda il Vol. I, pp. 80-84.

la (5.1) deve essere definita ponendo al denominatore l'area del corrispondente rettangolo.

I valori calcolati delle densità di frequenza consentono di ottenere utili informazioni sull'eventuale presenza di gruppi nei dati, sulla loro forma ed i rispettivi centroidi. Infatti, ad un *cluster* di osservazioni nella nuvola di punti in R^2 corrispondono tipicamente uno o più rettangoli adiacenti con un elevato valore di d_j. Una visualizzazione immediata dei valori assunti dalle densità di frequenza può essere ottenuta tramite una semplice rappresentazione con istogrammi nello spazio a tre dimensioni, disponibile ad esempio nel programma *S-PLUS*, versione 4.5, utilizzando la sequenza: *Graph - 3D Plot - 3D Bar chart*. In questo caso si rappresentano sugli assi del piano di base i lati della griglia riferiti, rispettivamente, alle variabili X_1 e X_2, mentre sull'asse verticale sono riportati i valori d_j.

Le quantità (5.1) forniscono pertanto lo strumento di base per individuare l'esistenza di zone con un'elevata densità di punti nello spazio bidimensionale — denominate «mode» della corrispondente distribuzione, in analogia con il corrispondente termine della statistica unidimensionale — e quindi di possibili gruppi di unità statistiche (10). Un'applicazione meccanica degli algoritmi di classificazione che non tenga conto di questa informazione preliminare rischia di condurre a conclusioni fuorvianti o quantomeno artificiose. I risultati ottenuti attraverso le metodologie di *cluster analysis* portano infatti ad effettivi vantaggi in termini di conoscenza del problema sottostante solamente quando rispecchiano le caratteristiche d'una popolazione realmente strutturata in gruppi. In caso contrario, viceversa, essi rischiano di essere la mera conseguenza dell'imposizione ai dati d'una struttura che

---

(10) In un'impostazione di tipo inferenziale, l'istogramma è spesso considerato come una stima preliminare — semplice ma alquanto rozza — dell'ignota densità della variabile che si suppone aver generato le osservazioni. Per un'introduzione alle tecniche di stima non parametrica della densità si rimanda, ad esempio, ai testi di Silverman (1986) e Scott (1992); alcuni metodi specifici per l'individuazione delle corrispondenti mode sono stati recentemente proposti da Minnotte, Marchette and Wegman (1998).

Un approccio alternativo, che conduce alla definizione di veri e propri indici di «addensamento locale» nella nuvola di punti (*local clustering statistics*), si fonda invece sulla considerazione esplicita dei vincoli di contiguità spaziale tra i rettangoli in R^2. L'obiettivo di tale impostazione consiste nell'individuare l'esistenza di eventuali gruppi di rettangoli contigui caratterizzati tutti da elevati valori di d_j; per ulteriori approfondimenti al riguardo si veda Cerioli e Zani (1999).

è a loro estranea. L'esame delle densità d_j può suggerire al ricercatore anche indicazioni sul numero di gruppi esistenti, fornendo così una plausibile soluzione ad un problema di grande rilevanza soprattutto nei metodi non gerarchici.

Se $p > 2$, la costruzione dell'istogramma p-dimensionale avviene secondo modalità analoghe ai passi i)-v) illustrati in precedenza, ma con riferimento allo spazio R^p in luogo di R^2. In particolare, se si dispone di n osservazioni su p variabili continue, $X_1, \ldots, X_p$, i vettori (riga) corrispondenti alle singole unità statistiche sono rappresentati tramite punti in R^p, la griglia da sovrapporre allo *scatterplot* p-dimensionale è costituita da iper-rettangoli di dimensione $h_1 \times h_2 \ldots \times h_p$ e la densità di frequenza associata all'iper-rettangolo j diventa

$$d_j = \frac{n_j}{n h_1 h_2 \cdots h_p} \qquad j = 1, \ldots, Q.$$

Nel caso multidimensionale le più comuni rappresentazioni dei dati — quali i grafici in coordinate parallele e le curve di Andrews, illustrati nel capitolo 2 di questo volume — non risultano tuttavia appropriate, in quanto trascurano le relazioni di vicinanza spaziale che esistono tra le unità statistiche (cioè gli iper-rettangoli in R^p). Tali relazioni sono invece fondamentali nel presente contesto, poiché consentono di aggregare iper-rettangoli appartenenti ad una medesima «zona» dello spazio multidimensionale e di confrontare le densità d_j rilevate in corrispondenza di iper-rettangoli ovvero di zone tra loro adiacenti. In simili circostanze risulta dunque preferibile cercare di ottenere una riduzione delle dimensioni, ad esempio attraverso il metodo delle componenti principali, in modo tale da ricondursi comunque al caso $p = 2$. In alternativa, è possibile proporre la rappresentazione delle densità d_j per ciascuna *coppia* di variabili, eventualmente condizionando rispetto ai valori assunti dalle rimanenti (11).

Esempio. La metodologia esplorativa sopra descritta è stata applicata ai dati riportati nella tab. 5.1, che fanno riferimento alla *performance* di 103 fondi comuni d'investimento operanti da almeno tre anni,

---

(11) Nel caso specifico in cui $p = 3$ si può anche ottenere una rappresentazione grafica in R^3, visualizzando una versione discretizzata delle curve di livello associate ai valori delle densità d_j. Tali curve assumono tipicamente la forma di «gusci» nello spazio tridimensionale: si vedano Scott (1992, p. 22); Hyndman (1996); Pison, Struyf and Rousseeuw (1999).

dei quali 56 azionari (con specializzazione Italia) e 47 bilanciati (fonte: *Il Sole-24 Ore*, 7 maggio 1999, p. 36). In particolare, per ciascun fondo sono stati riportati due indicatori di rendimento — uno a breve ed uno a medio termine — ed un indicatore di volatilità. Le variabili prese in esame sono quindi:

X_1 = performance % (a 12 mesi);
X_2 = rendimento medio annuale (a 3 anni);
X_3 = volatilità (a 3 anni).

L'analisi esplorativa è stata condotta considerando separatamente le possibili coppie di indicatori e costruendo i corrispondenti istogrammi bidimensionali. A tale scopo è stato necessario definire preliminarmente il valore di q, cioè il numero di elementi della griglia in ciascuna dimensione. Si osservi che la scelta del numero di rettangoli è un passo che riveste notevole importanza ai fini di garantire l'efficacia informativa dell'istogramma (Wand, 1997). Infatti, un valore insufficiente di q rischia di «lisciare» eccessivamente la rappresentazione grafica, con conseguente perdita delle informazioni relative ad eventuali addensamenti locali di punti nello spazio p-dimensionale. Viceversa, l'opzione per un valore di q troppo elevato rischia di produrre una rappresentazione dominata dal «rumore», cioè dall'errore casuale, portando così ad individuare un numero eccessivo di mode e quindi di gruppi. Nell'esempio, buoni risultati a fini esplorativi sono stati ottenuti ponendo $q = 10$, in modo tale che il valore medio della frequenza associata a ciascun rettangolo sia approssimativamente uguale a uno:

$$\sum_{j=1}^{Q} n_j/Q \approx 1.$$

Gli istogrammi relativi alle 3 coppie di variabili sono riportati nelle figure 5.1-5.3. Tutti i diagrammi pongono in luce la bimodalità della corrispondente distribuzione, imputabile alle differenti caratteristiche di fondi azionari e bilanciati. L'esistenza di due gruppi è quindi chiaramente visualizzata dalla procedura esplorativa, che fornisce le informazioni essenziali per affrontare correttamente i passi successivi della *cluster analysis*. Una simile indicazione si può ovviamente rivelare cruciale in tutte le situazioni in cui non si conosce a priori la presenza dei gruppi e la loro natura: è questa la prassi nelle analisi reali!

TAB. 5.1. *Tipologia (A = azionario; B = bilanciato) ed indicatori di* performance *per 103 fondi comuni d'investimento operanti almeno dal 1996* (fonte: *Il Sole-24 Ore*, 7 maggio 1999).

Fondo	Tipo	Performance % (a 12 mesi)	Rendimento medio annuale (a 3 anni)	Volatilità (a 3 anni)
Arca az. Italia	A	13.20	29.40	21.30
Aureo previdenza	A	12.10	27.10	20.00
Azimut crescita Italia	A	15.50	32.40	21.70
Azzurro	A	10.60	28.30	21.00
Bn azioni Italia	A	12.40	28.90	21.20
Bpb Tiziano	A	14.60	34.90	20.70
Capitalgest Italia	A	9.60	30.50	21.60
Capitalras	A	8.70	28.30	22.20
Centrale capital	A	14.80	32.90	20.30
Centrale Italia	A	17.30	36.90	22.30
Cliam az. italiane	A	7.70	25.80	21.10
Comit azione	A	12.10	27.60	23.50
Credit Suisse az. Italia	A	13.50	32.50	22.40
Ducato az. Italia	A	5.50	32.40	25.50
Epta az. Italia	A	10.90	32.40	23.00
Euromob. az. Italia	A	15.40	34.60	20.20
F&F gestione Italia	A	11.60	29.90	22.00
F&F lagest az. Italia	A	13.00	27.60	22.40
F&F select Italia	A	10.00	32.50	20.60
Fondersel Italia	A	14.60	40.80	23.40
Fond. picc. e med. imp.	A	3.80	28.70	18.20
Fondersel select Italia	A	15.00	30.90	22.00
Fondinvest Piazza Affari	A	10.70	30.10	21.20
Galileo	A	12.70	31.60	20.40
Genercomit capital	A	11.40	26.10	19.80
Gepocapital	A	9.60	28.00	19.50
Gesfimi Italia	A	14.40	31.60	22.10
Gesticredit borsitalia	A	10.40	29.50	21.60
Gestielle a	A	10.30	34.10	22.90
Gestifondi az. Italia	A	11.50	36.20	23.30
Gestnord Piazza Affari	A	12.60	28.70	21.20
Grifoglabal	A	11.80	25.70	20.10
Imi-Italy	A	12.70	34.30	24.00
Ing azionario	A	9.20	30.80	24.00

(Segue Tab. 5.1)

Interbancaria az.	A	18.30	28.30	19.80
Investire az.	A	11.20	30.40	22.10
Italy stock manag.	A	21.00	36.40	20.40
Mediceo indice Italia	A	9.20	25.40	21.90
Mida az.	A	17.00	45.10	23.20
Oasi az. Italia	A	10.20	30.20	22.40
Oasi italian equity risk	A	15.50	35.70	22.60
Oltremare az.	A	10.90	32.20	22.80
Padano Indice Italia	A	16.70	33.20	23.20
Performance az. Italia	A	5.50	27.50	23.10
Prime Italy	A	12.60	31.00	21.40
Primecapital	A	10.00	27.20	20.00
Primeclub az. Italia	A	11.60	30.00	21.50
Risparmio Italia Crescita	A	10.00	28.80	22.30
Roloitaly	A	10.60	26.40	20.00
RSA Small Cap	A	-0.30	28.00	16.70
SanPaolo Aldebaran Italia	A	13.80	31.40	22.80
SanPaolo azioni	A	24.60	49.10	22.00
Venetoblue	A	12.20	34.60	20.90
Venetoventure	A	-0.40	22.10	15.90
Zecchino	A	15.80	34.00	23.20
Zeta az.	A	14.80	32.10	20.30
Adriatic multi f.	B	7.70	10.80	9.50
Arca bb	B	13.40	18.60	10.70
Arca te	B	15.50	13.00	10.50
Armonia	B	11.40	13.80	8.50
Aureo	B	10.30	19.20	11.60
Azimut bil.	B	6.70	17.20	11.00
Bn bil. Italia	B	9.60	17.20	10.30
Capitalcredi	B	8.70	13.60	8.60
Capitalgest bil.	B	9.20	20.00	12.20
Carifondo blue chips	B	8.10	16.50	12.10
Carifondo libra	B	5.80	22.70	12.50

(Segue Tab. 5.1)

Cisalpino bil.	B	15.20	23.50	13.40
Eptacapital	B	7.90	18.80	12.20
Euromob. capitalfit	B	11.60	23.20	12.70
F&F eurorisparmio	B	9.70	22.80	12.70
F&F professionale	B	6.50	18.20	13.50
Fideuram performance	B	16.20	20.50	12.40
Fondersel	B	10.30	19.80	10.80
Fondicri bil.	B	9.80	16.50	10.30
Fondinvst futuro	B	9.00	19.00	11.20
Fondo alto bil.	B	17.00	34.30	12.70
Fondo centrale	B	7.80	13.20	10.00
Genercomit	B	11.40	19.60	12.00
Genercomit espansione	B	-0.60	10.80	9.10
Geporeinvest	B	9.00	19.70	11.90
Gepoworld	B	7.10	15.00	8.80
Gesfimi int.	B	10.50	14.10	9.30
Gesticredit finanza	B	9.20	15.70	9.30
Giallo	B	7.60	20.20	11.70
Grifocapital	B	8.40	17.10	12.60
Imicapital	B	9.90	16.40	9.10
Imindustria	B	13.20	19.90	12.20
Ing portfolio	B	10.80	31.10	15.50
Intermobiliare fondo	B	12.90	26.60	12.20
Investire bil.	B	11.40	18.00	11.40
Multiras	B	8.30	17.40	12.20
Nagracapital	B	11.90	20.80	12.10
Nordcapital	B	7.40	18.00	10.90
Nordmix	B	8.90	12.40	8.70
Primerend	B	0.70	18.90	12.90
Quadrifoglio int.	B	3.90	19.40	10.80
Rolointernational	B	11.60	17.10	10.30
Rolomix	B	10.50	17.90	12.30
Sanpaolo soluzione 5	B	8.40	16.30	13.30
Venetocapital	B	9.30	21.20	12.30
Visconteo	B	10.30	18.70	10.30
Zeta bil.	B	10.20	21.20	12.10

FIG. 5.1. *Istogrammi bidimensionali per 103 fondi comuni d'investimento riferiti alla coppia di indicatori: Performance % a 12 mesi; Rendimento medio annuale (3 anni).*

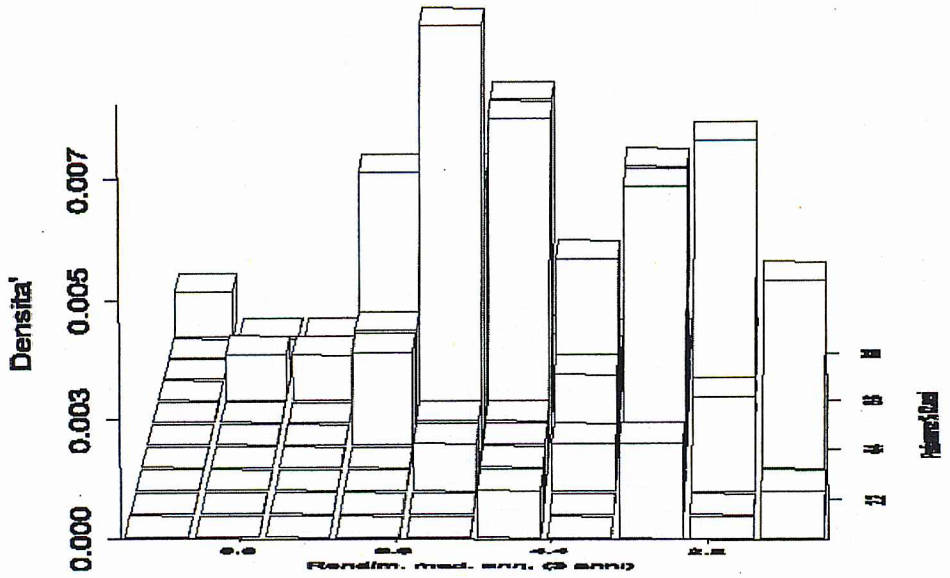


FIG. 5.2. *Istogrammi bidimensionali per 103 fondi comuni d'investimento riferiti alla coppia di indicatori: Performance % a 12 mesi; Volatilità (3 anni).*

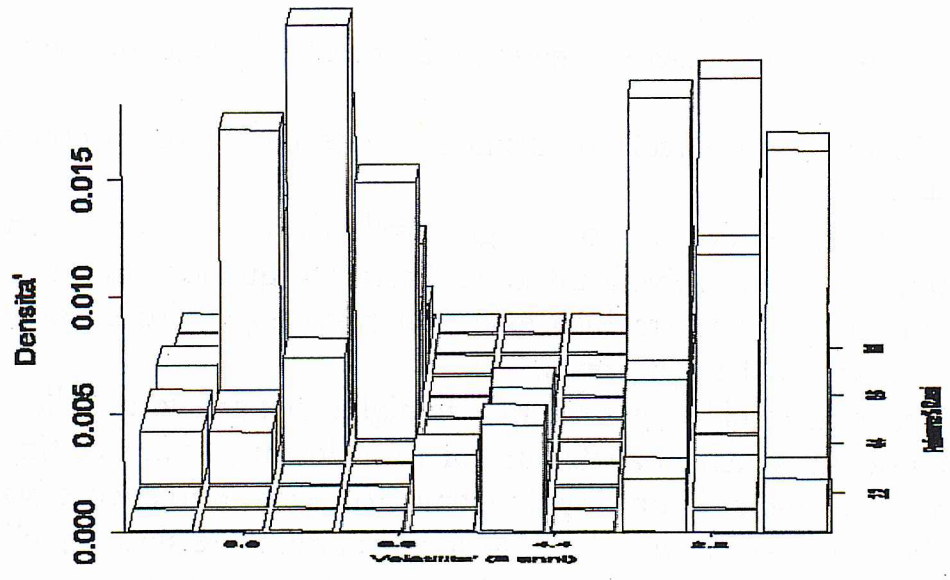
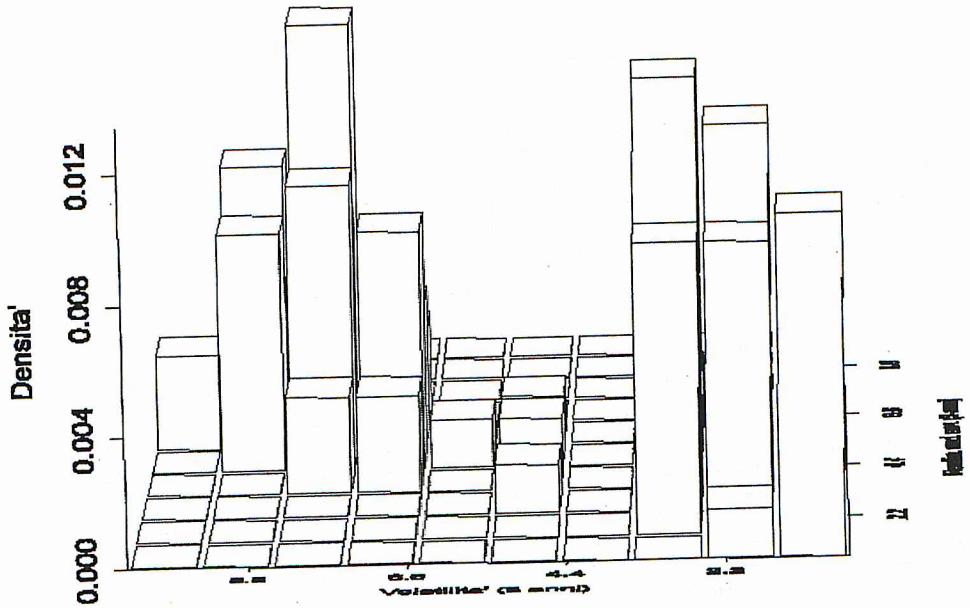


FIG. 5.3. *Istogrammi bidimensionali per 103 fondi comuni d'investimento riferiti alla coppia di indicatori: Rendimento medio annuale (3 anni); Volatilità (3 anni).*



4. *Caratteristiche generali dei metodi gerarchici*

Un metodo di classificazione gerarchico presenta le seguenti caratteristiche:

i) considera tutti i «livelli di distanza», che indicheremo con γ $(0 \leq \gamma < +\infty)$;

ii) i gruppi che si ottengono ad ogni livello di distanza comprendono i gruppi ottenuti ai livelli inferiori. Pertanto, quando due unità (o più unità) si uniscono tra loro esse non possono venire separate nei passi successivi della procedura.

Un metodo gerarchico genera una famiglia di partizioni delle n unità partendo da quella (banale) in cui tutte le unità sono distinte (ogni gruppo è formato da un singolo elemento), ricavando successivamente quelle con $(n-1)$, $(n-2)$ ... gruppi, per giungere sino a quella (banale) in cui tutte le unità sono riunite in un unico gruppo.

Un'ampia classe di metodi gerarchici si fonda sull'impiego iniziale d'una matrice di distanze **D** (o più in generale di indici di prossimità) calcolata per le n unità statistiche. In tal caso la procedura seguita per individuare i gruppi d'elementi (cioè per ottenere le partizioni successive) si articola nelle seguenti fasi:

i) Si individuano nella matrice **D** le due unità con la minore distanza (cioè tra loro più simili) e si riuniscono a formare il primo gruppo. Si ottiene una partizione con $(n-1)$ gruppi di cui $(n-2)$ costituiti da singole unità e l'altro formato da due unità.

ii) Si ricalcola — adottando un certo criterio — la distanza del gruppo ottenuto dagli altri gruppi (eventualmente costituiti da una sola unità), ricavando una nuova matrice delle distanze, con dimensioni diminuite di uno.

iii) Si individua nella nuova matrice delle distanze la coppia di unità (o gruppi) con minore distanza, riunendole in un unico gruppo.

iv) Si ripetono le fasi ii) e iii) sino a quando tutte le unità sono riunite in un solo gruppo.

Le differenze tra i vari metodi gerarchici consistono nel criterio utilizzato per calcolare la distanza tra due gruppi di unità (uno dei quali eventualmente formato da una sola unità).

Osservazione. Se si parte da una matrice di indici di similarità, nella procedura precedente le parole «minore distanza» devono essere sostituite da «maggiore similarità».

4.1. *Definizioni di distanza tra due gruppi e metodi di raggruppamento*

Si considerino due gruppi (*clusters*) C_1 e C_2 e siano n_1 e n_2 le rispettive numerosità. Sono possibili diverse definizioni di distanza tra i due gruppi, che identificano altrettanti metodi gerarchici. Consideriamo per ora i metodi che richiedono esclusivamente la conoscenza della matrice delle distanze.

Metodo del legame singolo (*single linkage*) o del vicino più prossimo (*nearest neighbour*).

La distanza tra due gruppi è definita come il minimo delle $n_1 n_2$ distanze tra ciascuna delle unità d'un gruppo e ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = \min(d_{rs}), \text{ per } r \epsilon C_1, s \epsilon C_2. \tag{5.2}$$

Metodo del legame completo (*complete linkage*) o del vicino più lontano (*furthest neighbour*).

La distanza tra due gruppi è definita come il massimo delle $n_1 n_2$ distanze tra ciascuna delle unità d'un gruppo e ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = \max(d_{rs}), \text{ per } r \epsilon C_1,\ s \epsilon C_2. \tag{5.3}$$

Adottando questo criterio, tutte le distanze tra le unità del primo gruppo e quelle del secondo sono minori o uguali alla distanza tra i due gruppi così definita.

Metodo del legame medio (*average linkage*) tra i gruppi, talvolta indicato anche con l'acronimo UPGMA (*Unweighted Pair-Group Method using arithmetic Averages*).

La distanza tra due gruppi è definita come la media aritmetica delle $n_1 n_2$ distanze tra ciascuna delle unità d'un gruppo e ciascuna delle unità dell'altro gruppo:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_r \sum_s d_{rs} \text{ per } r \epsilon C_1,\ s \epsilon C_2. \tag{5.4}$$

Metodo del legame medio nei gruppi

È una variante del metodo precedente. La distanza tra due gruppi è definita come la media aritmetica delle distanze tra tutte le possibili coppie di unità del nuovo gruppo che si ottiene (considerando pertanto anche le distanze tra le unità incluse nel medesimo gruppo di partenza). Indicando con $m = (n_1 + n_2)$ il numero complessivo di unità dei due gruppi considerati, tale distanza è la seguente:

$$d(C_1, C_2) = \frac{1}{m(m-1)/2} \sum_{r=1}^{m} \sum_{s>r}^{m} d_{rs}. \tag{5.5}$$

Per chiarire le differenze tra questi due metodi, consideriamo un gruppo C_1 formato da tre unità (1, 2, 3) ed un gruppo C_2 formato da due unità (4, 5). Con il metodo del legame medio (tra i gruppi) la distanza tra C_1 e C_2 è uguale alla media delle distanze tra le seguenti 6 coppie di unità: (1, 4); (1, 5); (2, 4); (2, 5); (3, 4); (3, 5). Con il metodo del legame medio nei gruppi la distanza tra C_1 e C_2 è uguale alla media delle distanze calcolate su ciascuna delle combinazioni di due elementi delle 5 unità (che sono in numero di 10) e cioè:

(1, 2); (1, 3); (1, 4); (1, 5); (2, 3); (2, 4); (2, 5); (3, 4); (3, 5); (4, 5).

4.2. *Un'applicazione illustrativa*

Riprendiamo in considerazione la matrice delle distanze della città a blocchi calcolate sulle prime quattro variabili della tab. 4.1, già ripor-

tata nella tab. 4.3. Vogliamo classificare le reti televisive in gruppi sulla base dell'analogia della loro programmazione. Essendo le variabili tutte espresse nella stessa unità di misura (ore di trasmissione), possiamo operare sui valori originari, senza effettuare la standardizzazione.

i) *Metodo del legame singolo*

Consideriamo il metodo del legame singolo e svolgiamo per esteso tutti i passi della procedura. Le due reti tra loro più simili (quelle con distanza minima, uguale a 1494) sono RAIUNO e RAIDUE. Esse vengono quindi riunite a formare il primo gruppo.

A questo punto occorre calcolare la distanza tra il gruppo ottenuto e le restanti unità. Avendo scelto il metodo del legame singolo, tale distanza è definita come il minimo tra la distanza che presenta RAIUNO e quella che presenta RAIDUE dalle altre unità. Ad esempio, la distanza del gruppo (RAIUNO, RAIDUE) da RAITRE è uguale a 1626 (che è il minimo tra 1626 e 2356).

Si può quindi ottenere la nuova matrice delle distanze, di dimensioni 5×5, riportata nella tab. 5.2. In detta matrice il minimo, uguale a 1626, corrisponde alla distanza tra il gruppo (RAIUNO, RAIDUE) e RAITRE, per cui si aggrega quest'ultima rete alle due precedenti, ottenendo la seguente partizione delle 6 unità con 4 gruppi:

(RAIUNO, RAIDUE, RAITRE) (RETE4) (CANALE5) (ITALIA1).

TAB. 5.2. *Matrice delle distanze della città a blocchi tra le reti televisive ottenuta dopo il primo passo della procedura gerarchica con il metodo del legame singolo.*

	(RAIUNO, RAIDUE)	RAITRE	RETE4	CANALE5	ITALIA1
(RAIUNO, RAIDUE)		1626	2285	2921	2916
RAITRE	1626		1869	4373	3230
RETE4	2285	1869		5206	3171
CANALE5	2921	4373	5206		5993
ITALIA1	2916	3230	3171	5993	

A questo punto occorre ricalcolare la distanza tra il gruppo ottenuto e ciascuna delle altre unità, pervenendo ad una matrice di distanze di dimensioni 4×4, riportata nella tab. 5.3. In essa il minimo, uguale a 1869, corrisponde alla distanza tra il gruppo precedente e RETE4, per

cui questa viene aggregata ad esso, ottenendo la seguente partizione con tre gruppi:

(RAIUNO, RAIDUE, RAITRE, RETE4) (CANALE5) (ITALIA1).

La procedura continua in maniera analoga sino a quando tutte le unità sono riunite in un solo gruppo.

TAB. 5.3. *Matrice delle distanze della città a blocchi tra le reti televisive ottenuta dopo il secondo passo della procedura gerarchica con il metodo del legame singolo.*

	(RAIUNO, RAIDUE, RAITRE)	RETE4	CANALE5	ITALIA1
(RAIUNO, RAIDUE, RAITRE)		1869	2921	2916
RETE4	1869		5206	3171
CANALE5	2921	5206		5993
ITALIA1	2916	3171	5993	

ii) *Metodo del legame completo*

Passiamo a considerare il medesimo esempio svolto con il metodo del legame completo. Il primo passo della procedura è identico a quello del legame singolo, per cui si ottiene la partizione con il gruppo (RAIUNO, RAIDUE) e le altre unità distinte. Occorre quindi calcolare la nuova matrice delle distanze di dimensioni 5×5. La distanza del gruppo ottenuto è ora definita come il massimo tra la distanza che presenta RAIUNO e quella che presenta RAIDUE dalle altre unità. Ad esempio, la distanza del gruppo (RAIUNO, RAIDUE) da RAITRE è ora uguale a 2356 (che è il massimo tra 1626 e 2356). Nella matrice 5×5 il minimo, uguale a 1869, corrisponde alla coppia RAITRE e RETE4, per cui queste due reti si aggregano in gruppo, ottenenendo la seguente partizione:

(RAIUNO, RAIDUE) (RAITRE, RETE4) (CANALE5) (ITALIA1).

Si vede immediatamente che a questo passo della procedura i risultati della classificazione gerarchica ottenuti con i due metodi sono diversi, poiché nel legame singolo si ha una partizione con un gruppo di 3 unità e le altre separate tra loro, mentre nel legame completo si ricavano due gruppi di due unità ciascuno, e le altre distinte.

Nello stadio successivo si ricalcola la matrice delle distanze, di dimensioni 4×4, e procedendo in maniera analoga si perviene alla seguente partizione con tre gruppi:

(RAIUNO, RAIDUE, RAITRE, RETE4) (CANALE5) (ITALIA1), che coincide invece con quella ottenuta con il legame singolo.

Questo semplice esempio pone in luce che una classificazione gerarchica d'un insieme d'unità, partendo dalla stessa matrice di distanze, è influenzata dall'algoritmo scelto per la formazione dei gruppi.

Per ottenere la successione gerarchica delle partizioni si può utilizzare il pacchetto SPSS, scegliendo dal menù: *statistica - classificazione - cluster gerarchica*. Sullo schermo compare una finestra nella quale inseriamo dapprima le variabili che intendiamo utilizzare nella classificazione e clicchiamo poi sull'opzione: *metodo*. Nella nuova finestra che appare è possibile scegliere il metodo di formazione dei gruppi, il tipo di distanza, l'eventuale trasformazione delle variabili. Per ottenere la classificazione col metodo del legame singolo prima descritta scegliamo le opzioni seguenti:

metodo di raggruppamento: *del vicino più vicino*;
misura (di distanza): *city-block*;
standardizzazione: *nessuna*.

Con tale sequenza la procedura di SPSS determina la classificazione gerarchica col metodo del legame singolo, partendo dalla matrice delle distanze della città a blocchi. L'output prevede fra l'altro la stampa della matrice delle distanze (opzionale) ed il «programma di agglomerazione» (v. tab. 5.4). Esso presenta gli $(n-1)$ passi o «stadi» della classificazione gerarchica, partendo da quello in cui si uniscono le due unità meno distanti sino all'ultimo in cui tutti gli elementi sono aggregati in un unico gruppo. Nel caso in esame nel primo stadio si uniscono le unità 1 (RAIUNO) e 2 (RAIDUE) ad un livello di distanza (nella tabella chiamato «coefficiente», sia pure impropriamente) uguale a 1494; nel secondo stadio il *cluster* 1, prima ottenuto, si unisce con l'unità 3 (RAITRE), ad un livello di distanza uguale a 1626, etc.

Le due successive colonne della tabella («stadio di formazione del *cluster*») servono per sapere se in quel passo si uniscono tra loro delle unità oppure dei gruppi ottenuti negli stadi precedenti: lo 0 corrisponde alle singole unità (in quanto al passo zero tutte le unità sono distinte tra loro), mentre un numero diverso da 0 segnala che il *cluster* in oggetto (formato da più unità) è quello ottenuto nel passo corrispondente. Ad esempio, la seconda riga della tab. 5.4 nelle colonne intestate «stadio di formazione del *cluster*» indica che si aggregano il *cluster* ottenuto al passo 1 ed un'unità singola. L'ultima colonna («stadio successivo») mostra inoltre a quale passo il gruppo appena formato subirà una nuova aggregazione nella procedura gerarchica.

Ripetiamo la procedura di SPSS con il metodo del legame completo. Nella sequenza precedente cambiamo soltanto il metodo di raggruppamento: *del vicino più lontano*. La parte di output che ora interessa è riportata nella tab. 5.5. Il primo stadio (prima riga) è uguale a quello del metodo del legame singolo. Nel secondo stadio si forma un nuovo gruppo, costituito dalle unità 3 e 4 (cioè RAITRE e RETE4). Nel terzo stadio si uniscono tra loro i due gruppi prima individuati; si comprende che si tratta di gruppi ottenuti in precedenza per il fatto che nelle colonne intestate « stadio di formazione del *cluster* » compaiono 1 e 2, che indicano i passi nei quali essi si erano formati. La procedura di aggregazione prosegue poi sino al quinto stadio, in cui tutte le unità sono riunite in un unico gruppo.

La procedura di SPSS consente di ottenere in maniera analoga la classificazione gerarchica col metodo del legame medio o con altri metodi che vedremo in seguito.

TAB. 5.4. *Programma di agglomerazione di SPSS per la classificazione delle reti televisive con il metodo del legame singolo e la distanza della città a blocchi.*

	Cluster accorpati			Stadio di formazione del cluster		Stadio
Stadio	Cluster 1	Cluster 2	Coefficienti	Cluster 1	Cluster 2	Successivo
1	1	2	1494	0	0	2
2	1	3	1626	1	0	3
3	1	4	1869	2	0	4
4	1	6	2916	3	0	5
5	1	5	2921	4	0	0

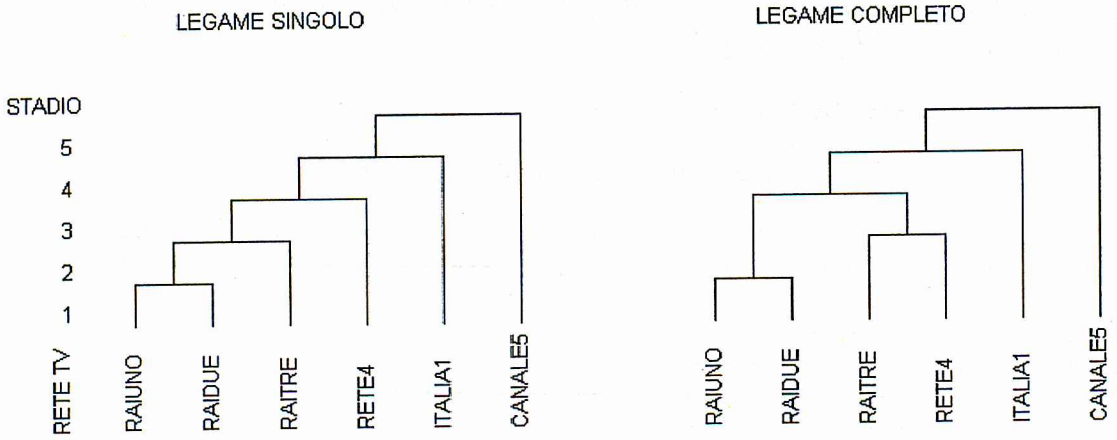
TAB. 5.5. *Programma di agglomerazione di SPSS per la classificazione delle reti televisive con il metodo del legame completo e la distanza della città a blocchi.*

	Cluster accorpati			Stadio di formazione del cluster		Stadio
Stadio	Cluster 1	Cluster 2	Coefficienti	Cluster 1	Cluster 2	Successivo
1	1	2	1494	0	0	3
2	3	4	1869	0	0	3
3	1	3	2581	1	2	4
4	1	6	3230	3	0	5
5	1	5	5993	4	0	0

5. *Il dendrogramma*

La famiglia di partizioni ottenuta con un metodo gerarchico può essere rappresentata graficamente mediante un «albero *n*-dimensionale» (*n-tree*). Nella fig. 5.4 abbiamo riportato l'albero corrispondente alla classificazione delle reti televisive col metodo del legame singolo e quello ricavato col metodo del legame completo.

FIG. 5.4. *Diagrammi ad albero ottenuti nella classificazione delle reti televisive col metodo del legame singolo e del legame completo.*



L'albero *n*-dimensionale pone in evidenza i gruppi che si ottengono ad ogni stadio della classificazione.

Un'informazione più precisa può ottenersi considerando congiuntamente la successione delle partizioni ed i rispettivi livelli di distanza.

Definizione - Si dice *dendrogramma* definito sull'insieme di n unità $a_i \epsilon A$ un'applicazione $D(\gamma) : R_+ \rightarrow \pi(A)$, ove $\pi(A)$ è l'insieme di tutte le partizioni di A e γ è il livello di distanza, che soddisfa le condizioni seguenti:

1) $D(0)$ è la partizione costituita dalle n unità distinte; $D(h)$, per h maggiore d'un opportuno valore di soglia, è la partizione formata da un unico gruppo.

2) Se $h < h'$ la partizione $D(h)$ è uguale o più fine (12) di quella $D(h')$. Pertanto, al crescere del livello di distanza si ottengono parti-

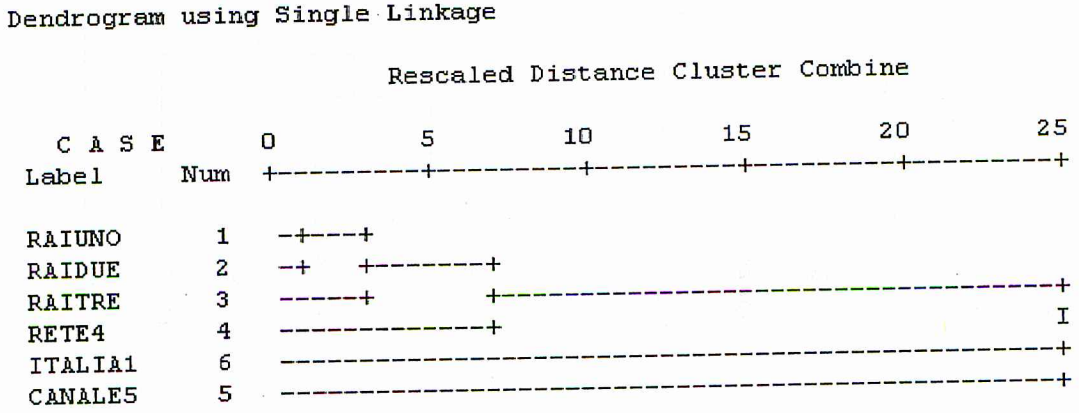
(12) Date due partizioni, P e P^*, definite sul medesimo insieme di elementi, si

zioni con un minor numero di gruppi, costituiti da aggregazioni di gruppi (o unità) ricavati ai livelli inferiori.

3) $D(h+\epsilon) = D(h)$, per $\epsilon > 0$ sufficientemente piccolo. Vi sono cioè degli incrementi del livello di distanza che lasciano immutata la partizione ottenuta. (La funzione è a gradini).

Il dendrogramma (13) può essere rappresentato graficamente in maniera analoga all'albero n-dimensionale, considerando però la famiglia di partizioni in funzione dei livelli di distanza γ anziché in funzione del numero dello stadio della procedura di aggregazione.

FIG. 5.5. *Dendrogramma ottenuto tramite SPSS nella classificazione delle reti televisive col metodo del legame singolo.*



La procedura di SPSS consente di ottenere direttamente il dendrogramma per ogni metodo gerarchico utilizzato, presentandolo però in orizzontale, invece che in verticale come negli esempi degli alberi n-dimensionali. Inoltre, i livelli di distanza non sono quelli originari desunti dalla matrice delle distanze ed utilizzati nella procedura di classificazione gerarchica, bensì sono riscalati nell'intervallo da 0 a 25 (ponendo uguale a 25 il livello di distanza al quale tutte le unità si riuniscono in un gruppo). Questo criterio ha il difetto di rendere meno immediata la lettura d'un singolo dendrogramma, ma ha il grande vantaggio di rendere direttamente comparabili tra loro dei dendrogrammi (per le medesime unità statistiche) ottenuti con metodi diversi, i quali, come si è visto, forniscono dendrogrammi aventi livelli di distanza con differenti campi di variazione.

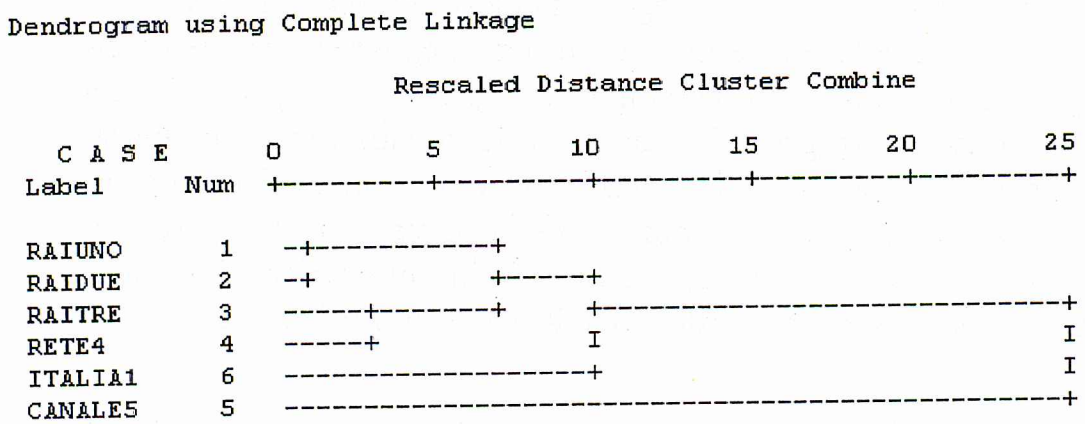
---

dice che P è più fine di P^* se ogni gruppo di P^* è formato da uno o più gruppi di P (ma non viceversa).

(13) La denominazione deriva dal greco $\delta\varepsilon\nu\delta\rho o\nu$, che vuol dire albero.

Esempio. Nella fig. 5.5 e nella fig. 5.6 sono riportati i dendrogrammi dell'esempio precedente, corrispondenti rispettivamente al metodo del legame singolo e del legame completo, ottenuti da SPSS scegliendo l'opzione: *grafici - dendrogramma.* Facciamo notare che con i due criteri i livelli di distanza ai quali si formano i gruppi sono diversi, ad eccezione del primo passo della classificazione.

FIG. 5.6. *Dendrogramma ottenuto tramite SPSS nella classificazione delle reti televisive col metodo del legame completo.*



6. *Alcune proprietà dei metodi gerarchici*

Gli esempi precedenti hanno posto in luce che, partendo da una stessa matrice delle distanze, si possono ottenere classificazioni gerarchiche differenti. Ci si può domandare allora se esiste una classificazione migliore delle altre. La risposta non è univoca, ma il problema può essere impostato in termini di proprietà «desiderabili», di cui dovrebbe godere un metodo di formazione dei gruppi e di verifica di tali proprietà per i vari algoritmi di classificazione. Questo approccio è stato introdotto da Fisher and Van Ness (1971) e ripreso ampiamente in letteratura. In questa sede ci limiteremo a considerare due proprietà che ci sembrano di particolare importanza sul piano logico.

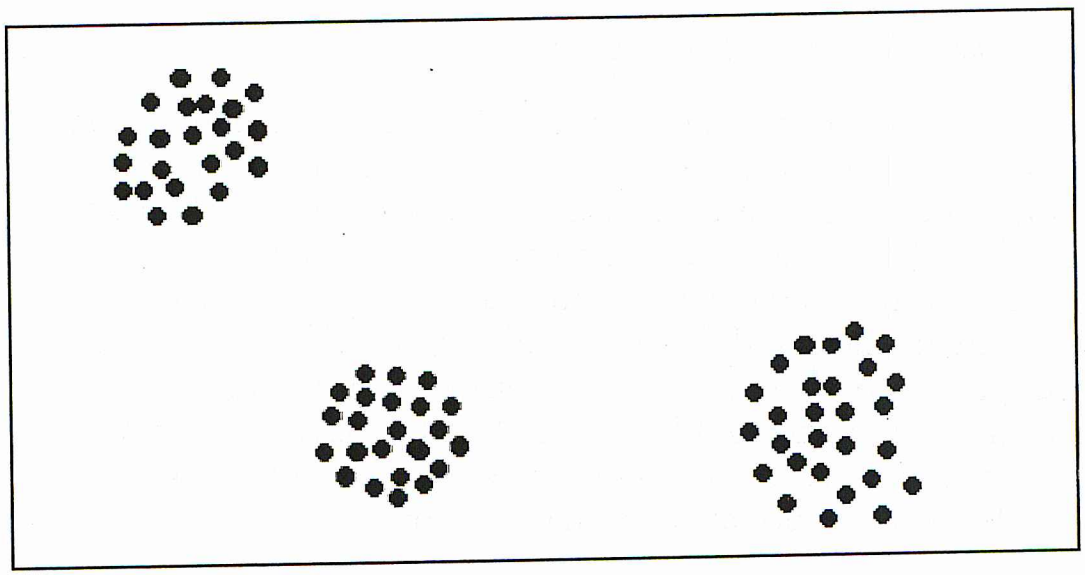
6.1. *La partizione ben strutturata minimale*

Partendo dalla matrice dei dati iniziali, supponiamo di aver ottenuto una matrice delle distanze (d'un certo tipo) tra le unità statistiche. Come si è illustrato nel capitolo 4, il calcolo di tale matrice implica varie

scelte preliminari (tipo di metrica, standardizzazione o meno delle variabili, eventuale ponderazione delle variabili, etc.). Ci domandiamo ora quale sia il massimo grado di «oggettività» che si può conseguire nella formazione dei gruppi di unità, partendo da tale matrice (e quindi considerando ferme le scelte sopra menzionate).

Un criterio che pare molto naturale per determinare gruppi «oggettivi» è quello di richiedere che la massima distanza all'interno dei gruppi considerati sia minore della minima distanza tra i gruppi. In tale evenienza, infatti, i gruppi (eventualmente formati da un solo elemento) corrispondono a sottoinsiemi omogenei nel loro interno e nessun criterio convenzionale deve essere introdotto per assegnare un elemento ad un gruppo. Si veda, a titolo d'esempio, la fig. 5.7, nella quale le unità sono classificate in tre gruppi in base a due variabili: emerge chiaramente che la distanza tra i due punti più lontani in ogni gruppo (il «diametro» del medesimo) è minore della minima distanza tra elementi appartenenti a gruppi diversi. Si può dunque parlare di classificazione «oggettiva» o «naturale» dell'insieme di unità in tre gruppi.

FIG. 5.7. *Insieme d'elementi in R^2, suddiviso in tre gruppi che corrispondono alla partizione ben strutturata.*



Osservazione. Questo criterio d'individuazione di gruppi oggettivi presuppone che alle distanze si attribuisca semplicemente un significato su scala ordinale (v. capitolo 4, n. 5.3), poiché in esso si fa riferimento solo ai valori minimi e massimi delle distanze, i quali conservano il verso

della disuguaglianza per trasformazione monotona crescente delle distanze stesse.

Definizione. Una partizione $P = \{C_1, C_2, \ldots, C_g\}$ d'un insieme di n elementi u_i, per i quali si è definita una distanza d, si dice *ben strutturata* se:

$\max(d_{ij}) < \min(d_{rs})$, per ogni u_i, u_j appartenenti allo stesso gruppo e u_r, u_s appartenenti a gruppi diversi.

Definizione. Si dice partizione *ben strutturata minimale* la partizione ben strutturata con il minor numero di gruppi.

Castagnoli (1978) ha dimostrato il seguente:

Teorema. Per ogni matrice delle distanze esiste una ed una sola partizione ben strutturata minimale.

Esempio. Riprendiamo in esame la matrice delle distanze euclidee standardizzate riportate nella tab. 4.5 e consideriamo la classificazione gerarchica col metodo del legame singolo. Nella tab. 5.6 è riportato il programma di agglomerazione ottenuto con SPSS, scegliendo la sequenza: *statistica - classificazione - cluster gerarchica* e le opzioni: *del vicino più vicino - euclidea - punteggi Z.*

TAB. 5.6. *Programma di agglomerazione di SPSS per la classificazione di 10 aziende alimentari con il metodo del legame singolo e la distanza euclidea sugli scostamenti standardizzati.*

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio Successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	5	0.947	0	0	3
2	2	7	1.007	0	0	7
3	1	4	1.104	1	0	4
4	1	10	1.187	3	0	5
5	1	6	1.362	4	0	6
6	1	3	1.671	5	0	7
7	1	2	1.808	6	2	8
8	1	8	2.179	7	0	9
9	1	9	3.256	8	0	0

Al primo stadio si riuniscono tra loro le unità 1 (Barilla) e 5 (Kraft), mentre le restanti aziende costituiscono gruppi formati da un solo elemento; tale partizione è ovviamente ben strutturata, poiché l'unico gruppo formato da due aziende è quello con la minima distanza. Al secondo passo si riuniscono le unità 2 (Eridania) e 7 (Nestlé), ed anche questa partizione con 8 gruppi è ben strutturata. Al terzo passo l'unità

4 (Galbani) si aggrega al gruppo ottenuto allo stadio 1 (Barilla, Kraft) ed il massimo della distanza all'interno di questo gruppo di tre unità è uguale a 1.245 (la distanza tra Galbani e Kraft); la partizione ottenuta a questo passo non è più ben strutturata, poiché vi sono distanze tra elementi assegnati a gruppi diversi che risultano minori di tale valore (la distanza tra Barilla e Star è uguale a 1.187 e quella tra Kraft e Star a 1.206). Pertanto, la partizione ben strutturata minimale è quella con 8 gruppi individuata nel passo precedente e cioè:

(Barilla, Kraft) (Eridania, Nestlé) ed i restanti 6 gruppi costituiti ciascuno da una singola azienda.

Dal punto di vista operativo, per individuare la partizione ben strutturata minimale ci si può avvalere congiuntamente delle tabelle «programma di agglomerazione» di SPSS ottenute col metodo del legame singolo (che nella colonna intestata «coefficienti» riporta il minimo delle distanze tra i gruppi, o elementi, che si uniscono) e col metodo del legame completo (che indica il massimo della distanza per il gruppo che si forma a quel passo).

Teorema. I metodi del legame singolo, del legame completo e del legame medio ad un certo passo della classificazione gerarchica individuano la partizione ben strutturata minimale.

Dim. La partizione ben strutturata minimale è, per definizione, quella con il minor numero di gruppi nella quale la massima distanza all'interno dei gruppi è minore della minima distanza tra i gruppi. In essa dunque tutte le distanze tra gli elementi d'un generico gruppo sono minori delle distanze tra gli elementi di detto gruppo ed i restanti elementi. Conseguentemente, si perviene alla partizione ben strutturata minimale utilizzando quale operatore per la riunione dei gruppi tra loro sia il minimo, sia il massimo delle distanze, sia assumendone qualunque valore intermedio.

Questo risultato è di notevole rilievo sul piano logico: i diversi metodi considerati individuano tutti la partizione ben strutturata minimale, la quale, come s'è detto, rappresenta la classificazione più sintetica ottenibile in maniera oggettiva. Ne consegue che tali metodi sono tutti logicamente accettabili e si differenziano tra loro con riguardo alle aggregazioni degli elementi e dei gruppi per i quali non è possibile definire *a priori* un criterio oggettivo. Sul piano pratico, tuttavia, la partizione ben strutturata minimale si rivela quasi sempre troppo dispersa (cioè costituita da un numero eccessivo di gruppi), come è accaduto anche nell'esempio precedente (14). Occorrerà allora procedere con le

(14) In qualche caso — come ad esempio quando si è in presenza d'un gruppo

tappe successive d'un metodo gerarchico, il quale, com'è noto, non opera mai scissioni dei gruppi ottenuti nei passi precedenti, ma consente di aggregare ulteriormente tra loro i raggruppamenti che costituiscono la partizione oggettiva individuata.

6.2. *Invarianza per trasformazione monotona delle distanze*

Nel capitolo 4 abbiamo sottolineato che l'interpretazione più ragionevole delle distanze è quella su scala ordinale. Sembra allora naturale richiedere che un metodo di classificazione fornisca i medesimi risultati quando si opera una trasformazione monotona crescente delle distanze.

Definizione. Si dice che un metodo gerarchico di formazione dei gruppi è *invariante per trasformazione monotona crescente* delle distanze quando esso fornisce la medesima successione di partizioni ($n - tree$) per ogni trasformazione monotona crescente delle distanze che compaiono nella matrice **D**.

Teorema. Il metodo del legame singolo ed il metodo del legame completo sono invarianti per trasformazione monotona crescente delle distanze.

Dim. Basta considerare che gli operatori $\min(d_{rs})$ e $\max(d_{rs})$ sui quali si fonda la riunione di due gruppi rispettivamente nei due metodi non mutano il verso delle disuguaglianze a seguito d'una trasformazione monotona crescente delle distanze.

Pertanto, detti metodi forniscono la medesima classificazione gerarchica se, partendo da un'assegnata matrice dei dati, si utilizza ad esempio la matrice delle distanze euclidee, oppure la matrice dei quadrati delle distanze euclidee (indici di distanza) o semplicemente la matrice dei gradi crescenti di tali indici.

Invece, il metodo del legame medio non è invariante per trasformazione monotona crescente delle distanze, poiché l'operatore media aritmetica non gode di tale proprietà.

6.3. *Caratteristiche dei gruppi individuati dai vari metodi*

Il metodo del legame singolo e quello del legame completo soddisfano entrambi sia la proprietà di partizione ben strutturata minimale, sia quella di invarianza per trasformazione monotona, ma consentono d'individuare gruppi con caratteristiche differenti.

---

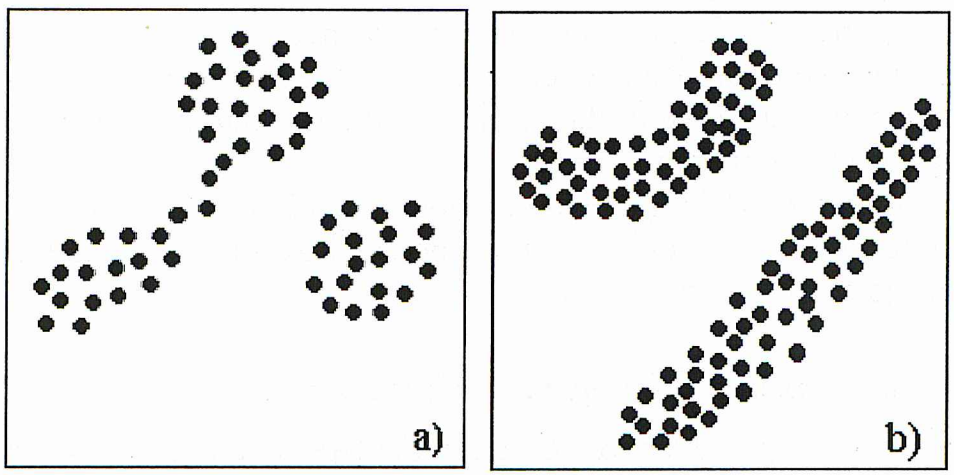
compatto e di uno o più *outliers* molto lontani da esso — la partizione ben strutturata minimale può risultare troppo aggregata.

Il metodo del legame singolo presenta il cosiddetto «effetto catena», cioè può riunire in un unico gruppo elementi anche molto distanti in R^p quando tra essi esiste una successione di punti intermedi. Questo effetto è posto in evidenza nella fig. 5.8a, in cui si vedono chiaramente tre nuvole di punti nel piano, ma la presenza di alcune unità (gli anelli della catena) conduce col metodo del legame singolo alla riunione in un unico gruppo delle due nuvole a sinistra ed in alto; tale gruppo non presenta tuttavia coesione interna.

Il metodo del legame completo individua invece gruppi compatti nel loro interno, ma di forma approssimativamente circolare (in R^2), sferica o ipersferica, e nell'esempio precedente trova correttamente i tre gruppi assegnando i punti intermedi in parte all'uno ed in parte all'altro dei due gruppi più prossimi.

L'effetto catena del legame singolo, se da un lato rappresenta uno svantaggio, dall'altro ha il pregio di consentire l'individuazione di gruppi con forme anche molto diverse da quelle ipersferiche (15). Si veda a questo proposito la fig. 5.8b: vi sono due gruppi nel piano, nettamente separati tra loro, l'uno a forma di fagiolo e l'altro di sigaro, che il metodo del legame singolo riesce ad individuare proprio grazie all'effetto catena, mentre il metodo del legame completo non sarebbe in grado di farlo.

FIG. 5.8. *Metodo del legame singolo:* a) *effetto catena;* b) *gruppi di forma non circolare individuabili.*



(15) Il metodo del legame singolo consente di ottenere anche l'albero di minima lunghezza (*minimum spanning tree*), che è il grafo che unisce gli n punti nel piano, in modo che ciascuno di questi sia raggiunto una sola volta, e tale che la lunghezza dei segmenti tra essi sia la minima possibile tra tutte le $n(n-1)/2$ distanze (Gower and Ross, 1969; Gordon, 1999, pp. 53-55).

Sulla base dell'esperienza di molte applicazioni, si può dire tuttavia che spesso il metodo del legame singolo tende ad aggregare successivamente le varie unità al gruppo o ai gruppi ottenuti nei primi passi della classificazione gerarchica, per cui non si coglie con chiarezza il *pattern* sottostante.

Il metodo del legame medio può allora costituire in molte circostanza un compromesso ragionevole, per ottenere gruppi con discreta coesione interna e separazione esterna. Si veda Gordon (1996a) per un approfondimento di questo tema.

7. *Metodi gerarchici che utilizzano anche la matrice dei dati*

Vi sono altri metodi gerarchici che utilizzano anche la matrice dei dati di partenza e non la sola matrice delle distanze.

Metodo del centroide.

La distanza tra due gruppi C_1 e C_2 di numerosità n_1 e n_2 è definita come la distanza (d'un certo tipo) tra i rispettivi centroidi, $\overline{\mathbf{x}}_1$ e $\overline{\mathbf{x}}_2$:

$$d(C_1, C_2) = d(\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2) \tag{5.6}$$

Ovviamente, il calcolo del centroide d'un gruppo di unità richiede la conoscenza dei rispettivi valori delle p variabili, che si possono leggere nella matrice dei dati.

Facciamo notare che, per un principio di coerenza, la distanza tra i centroidi deve essere del medesimo tipo (euclidea, della città a blocchi, etc.) utilizzato nella matrice delle distanze.

Il centroide del nuovo gruppo che si forma può essere calcolato (avvalendosi della proprietà associativa della media aritmetica) in funzione dei centroidi dei due gruppi di partenza:

$$\textit{centroide}\ (C_1 \cup C_2) = \frac{\overline{\mathbf{x}}_1 n_1 + \overline{\mathbf{x}}_2 n_2}{n_1 + n_2}. \tag{5.7}$$

Il metodo del centroide presenta analogie logiche con il metodo del legame medio (tra i gruppi): in quest'ultimo, infatti, si considera la media delle distanze tra le unità dell'uno e dell'altro gruppo, mentre nel metodo del centroide si individua dapprima un «centro» di ciascun gruppo (cioè il vettore che contiene i valori medi delle p variabili per le unità incluse nel gruppo) e poi si misura la distanza tra essi.

Metodo di Ward o della minima devianza.

Nel metodo di Ward si definisce esplicitamente una funzione obiettivo. Dato che lo scopo della classificazione è quello di ottenere gruppi con la maggiore coesione interna, si considera la scomposizione della Devianza totale (T) delle p variabili in Devianza nei gruppi (*Within, W*) e Devianza fra i gruppi (*Between, B*) (16):

$$T = W + B \tag{5.8}$$

ove, data una partizione in g gruppi,

$$T = \sum_{s=1}^{p} \sum_{i=1}^{n} (x_{is} - \bar{x}_s)^2 \tag{5.9}$$

è la Devianza totale delle p variabili, ottenuta come somma delle devianze delle singole variabili rispetto alla corrispondente media generale $\bar{x}_s$;

$$W = \sum_{l=1}^{g} W_l \tag{5.10}$$

è la Devianza nei gruppi, cioè la somma delle devianze di gruppo;

$$W_l = \sum_{s=1}^{p} \sum_{i=1}^{n_l} (x_{is} - \bar{x}_{s,l})^2 \tag{5.11}$$

è la devianza delle p variabili nel gruppo l-esimo (di numerosità n_l e centroide $\bar{\mathbf{x}}_l = [\bar{x}_{1,l}, \ldots, \bar{x}_{p,l}]'$) e

$$B = \sum_{s=1}^{p} \sum_{l=1}^{g} n_l (\bar{x}_{s,l} - \bar{x}_s)^2 \tag{5.12}$$

è la Devianza fra i gruppi, cioè la somma — calcolata su tutte le variabili — delle devianze (ponderate) delle medie di gruppo rispetto alla corrispondente media generale.

---

(16) Per la scomposizione della devianza d'una singola variabile si veda il Vol. I, pp. 126-130.

Ad ogni passo della procedura gerarchica si aggregano tra loro i gruppi (eventualmente costituiti da una singola unità) che comportano il minor incremento della Devianza nei gruppi, cioè che assicurano la maggiore coesione interna possibile.

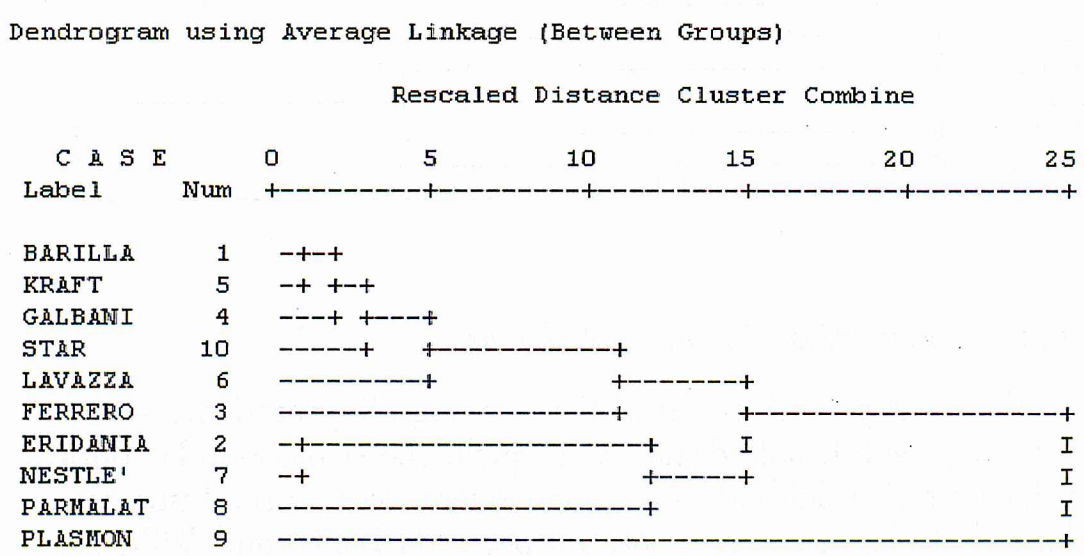
Il metodo di Ward non richiede il calcolo preliminare d'una matrice delle distanze. Tuttavia, si può dimostrare (Seber, 1984, p. 363) che la minimizzazione dell'incremento della devianza ad ogni passo della procedura gerarchica è equivalente all'impiego della seguente distanza tra due gruppi:

$$d(C_1, C_2) = \frac{n_1 n_2}{n_1 + n_2} \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2, \qquad (5.13)$$

che corrisponde al quadrato della distanza euclidea tra i centroidi dei due gruppi, moltiplicato per una quantità che è funzione del numero di unità dei due gruppi.

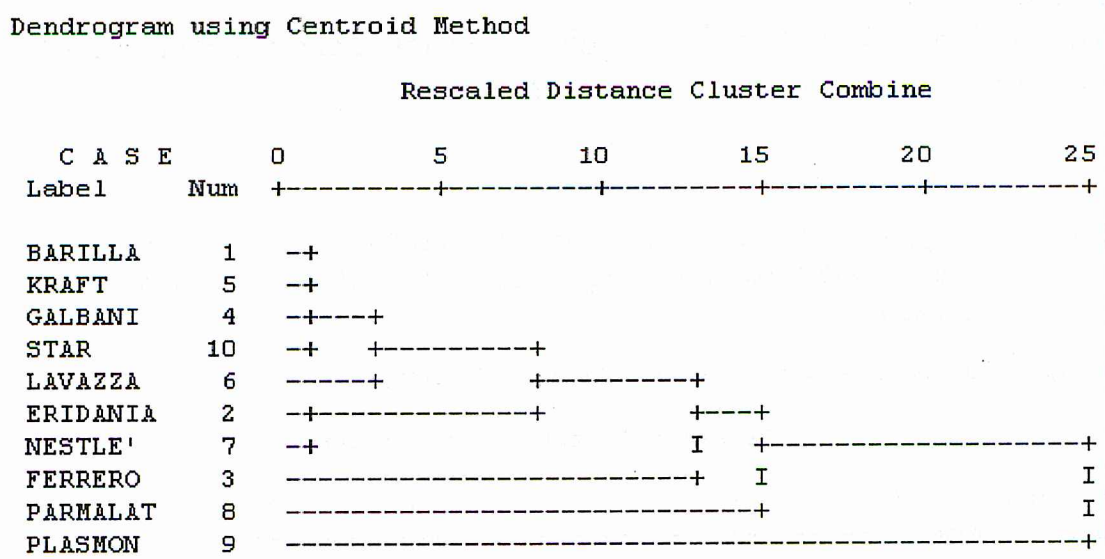
Anche il metodo di Ward può quindi rientrare nello schema generale di formazione dei gruppi prima illustrato.

FIG. 5.9. *Dendrogramma ottenuto tramite SPSS nella classificazione delle aziende alimentari con la distanza euclidea sugli scostamenti standardizzati ed il metodo del legame medio.*



Esempio. Consideriamo la classificazione gerarchica delle 10 aziende alimentari in base alla matrice delle distanze euclidee standardizzate (tab. 4.5), utilizzando successivamente il metodo del legame medio, del centroide e di Ward. Per economia di spazio riportiamo solo i dendrogrammi ottenuti con SPSS (fig. 5.9; fig. 5.10; fig. 5.11). Con i tre diversi metodi s'individuano sempre ad un basso livello di distanza un gruppo di due aziende (Eridania e Nestlé) ed un altro gruppo di 4 aziende (Barilla, Kraft, Galbani, Star), che risultano quindi molto simili tra loro con riferimento alle variabili considerate. Ad un maggior livello di distanza Lavazza si unisce al gruppo precedente formato da 4 aziende. Gli stadi successivi della classificazione non risultano coincidenti per i tre metodi, ma il metodo del legame medio e quello del centroide concordano nel segnalare che Plasmon è l'ultima azienda che si aggrega alle altre, ad un livello di distanza molto elevato, per cui essa può considerarsi un *outlier* in R^5.

FIG. 5.10. *Dendrogramma ottenuto tramite SPSS nella classificazione delle aziende alimentari con la distanza euclidea sugli scostamenti standardizzati ed il metodo del centroide.*

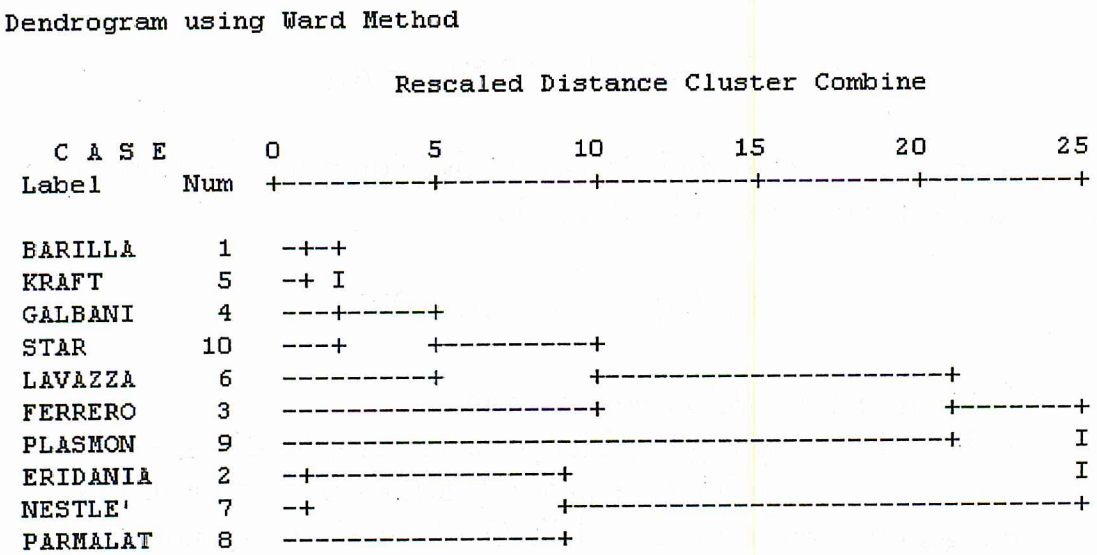


8. *La formula ricorsiva di Lance e Williams*

Gli algoritmi gerarchici descritti nei paragrafi precedenti — sia quelli che utilizzano le sole distanze, sia quelli che richiedono la conoscenza della matrice dei dati — possono essere ricondotti ad una formulazione unitaria, originariamente proposta da Lance and Williams (1967). A tale scopo, si considerino tre *clusters* C_1, C_2 e C_3 (eventual-

mente costituiti da un unico elemento) e si indichi con $d[C_3, (C_1, C_2)]$ la distanza tra il gruppo C_3 ed il gruppo (C_1, C_2), formatosi dalla fusione di C_1 e C_2.

FIG. 5.11. *Dendrogramma ottenuto tramite SPSS nella classificazione delle aziende alimentari con la distanza euclidea sugli scostamenti standardizzati ed il metodo di Ward.*



È importante poter esprimere $d[C_3, (C_1, C_2)]$ in funzione delle distanze tra i *clusters* originari, $d_{12} = d(C_1, C_2)$, $d_{31} = (C_3, C_1)$ e $d_{32} = (C_3, C_2)$, in quanto una simile proprietà consente di aggiornare in modo ricorsivo la matrice delle distanze nei successivi passi della procedura di aggregazione gerarchica. Lance e Williams hanno suggerito di considerare la seguente formula generale:

$$d[C_3, (C_1, C_2)] = \alpha_1 d_{31} + \alpha_2 d_{32} + \beta d_{12} + \gamma |d_{31} - d_{32}|, \quad (5.14)$$

ove α_1, α_2, β e γ sono parametri fissati dal ricercatore.

Una scelta appropriata dei parametri permette di ricavare come casi particolari della (5.14) le espressioni della distanza tra gruppi adottate dai principali metodi gerarchici . Ad esempio, il metodo del legame singolo corrisponde al caso: $\alpha_1 = \alpha_2 = 0.5$, $\beta = 0$ e $\gamma = -0.5$. Infatti, con questi parametri la (5.14) diventa:

$$d[C_3, (C_1, C_2)] = 0.5(d_{31} + d_{32}) - 0.5|d_{31} - d_{32}|.$$

Nell'espressione così ottenuta il valore di $d[C_3, (C_1, C_2)]$ dipende dal segno della quantità $d_{31} - d_{32}$. Se $d_{31} > d_{32}$, si ha infatti $|d_{31} - d_{32}| = d_{31} - d_{32}$ e dunque:

$$d[C_3, (C_1, C_2)] = 0.5d_{32} + 0.5d_{32} = d_{32};$$

viceversa, se $d_{31} < d_{32}$ si ha $|d_{31} - d_{32}| = d_{32} - d_{31}$ e quindi:

$$d[C_3, (C_1, C_2)] = 0.5d_{31} + 0.5d_{31} = d_{31}.$$

Nel caso esaminato la formula (5.14) conduce quindi ad aggiornare la matrice delle distanze secondo il criterio

$$d[C_3, (C_1, C_2)] = \min(d_{31}, d_{32}),$$

corrispondente al metodo del legame singolo.

Anche gli altri algoritmi gerarchici presi in considerazione — legame completo, legame medio, centroide e Ward — possono rientrare nello schema generale rappresentato dalla (5.14), se si selezionano adeguatamente i valori dei parametri α_1, α_2, β e γ. Per ulteriori dettagli si vedano, ad esempio, Everitt (1993, p. 67) e Gordon (1996a).

9. *Criteri di valutazione delle partizioni*

Un algoritmo gerarchico genera una famiglia di partizioni delle n unità di partenza, cioè una successione di n classificazioni di tali unità, con un numero di gruppi via via decrescente da n a 1. Ovviamente, non tutte le partizioni così ottenute forniscono una rappresentazione soddisfacente della realtà. Ad ogni passo della procedura di classificazione gerarchica risulta pertanto importante valutare la «bontà» della corrispondente partizione.

Come osservato più volte, l'obiettivo di fondo della *cluster analysis* consiste nella ricerca di gruppi dotati di coesione interna e di separazione esterna. Un valido criterio di giudizio può quindi essere basato sulla scomposizione della devianza totale delle p variabili riportata nella (5.8). Una «buona» classificazione è infatti tipicamente caratterizzata da una ridotta quota di devianza entro i gruppi (W) e da un elevato valore della devianza fra i gruppi (B). Data una partizione costituita da g gruppi, si considera quindi l'indice:

$$R^2 = 1 - \frac{W}{T} = \frac{B}{T}, \qquad (5.15)$$

ove le quantità W, B e T sono definite, rispettivamente, nelle espressioni (5.10), (5.12) e (5.9).

L'indice R^2 assume valori nell'intervallo $[0; 1]$ e risulta pertanto confrontabile anche con riferimento a partizioni caratterizzate da un differente numero di gruppi, oppure ottenute mediante algoritmi differenti. In particolare, se R^2 è prossimo al valore massimo unitario la corrispondente classificazione può essere ritenuta omogenea, in quanto le unità che appartengono ad un medesimo gruppo sono molto simili tra loro ($W_l \simeq 0$ per ogni $l = 1, \ldots, g$) ed i gruppi sono ben separati ($B \simeq T$). In analogia al coefficiente di determinazione nella retta di regressione, l'indice R^2 misura quindi la quota di variabilità totale nella matrice dei dati (considerando tutte le p variabili) che può essere « spiegata » dalla partizione considerata (17).

Occorre tuttavia ricordare il *trade-off* esistente tra il numero di gruppi e la coesione all'interno degli stessi, già menzionato nel n. 2 di questo capitolo. In particolare, l'indice R^2 assume valori non decrescenti all'aumentare di g. La ricerca del numero « ottimo » di gruppi non può dunque fondarsi sulla semplice massimizzazione del criterio (5.15), che porterebbe a privilegiare la partizione banale formata da n gruppi d'una sola unità (per la quale $R^2 = 1$), ma deve compendiare le esigenze contrapposte di omogeneità interna dei *clusters* e di sintesi della classificazione prescelta.

Una misura alternativa ad R^2 è la cosiddetta *Root-Mean-Square Standard Deviation*, abbreviata con l'acronimo *RMSSTD*. Tale coefficiente considera soltanto la componente della (5.10) che fa riferimento al gruppo formatosi al corrispondente passo della procedura di classificazione gerarchica. In particolare, considerando il passo h-esimo ($h = 2, \ldots n - 1$), l'indice *RMSSTD* è così definito:

---

(17) La similitudine con la valutazione della bontà di adattamento d'un modello di regressione ha portato anche alla definizione di « test F » analoghi a quelli che compaiono in una tabella dell'analisi della varianza, con riferimento alla verifica della significatività sia delle singole variabili sia della partizione nel suo complesso (si veda, ad esempio, Hartigan, 1975, pp. 89-91). A differenza della regressione, tali statistiche presentano tuttavia un significato prevalentemente descrittivo e non possono essere correttamente riferite ad una distribuzione F. Il loro impiego presenta pertanto limitati vantaggi, dal punto di vista operativo, rispetto a quello del semplice indice R^2.

$$RMSSTD = \sqrt{\frac{W_h}{p(n_h - 1)}} \tag{5.16}$$

ove W_h rappresenta la devianza delle p variabili, definita tramite la (5.11), nel *cluster* costituitosi al passo h della procedura e n_h è la corrispondente numerosità. Ovviamente, il calcolo dell'indice *RMSSTD* perde interesse nel caso delle partizioni banali costituite rispettivamente da n gruppi ($h = 1$), in cui la (5.16) non è neppure definita, e da un unico gruppo ($h = n$), in cui $W_h = W = T$. Nei passi intermedi un forte incremento di *RMSSTD* rispetto al passo precedente segnala invece che si sono uniti due gruppi fortemente eterogenei tra loro.

L'indice R^2 ed il coefficiente *RMSSTD* consentono di valutare il grado di coesione interna dei gruppi ottenuti in ciascun passo d'una classificazione gerarchica. Essi forniscono dunque due criteri alternativi per individuare un opportuno «taglio» del corrispondente dendrogramma, cioè per scegliere la partizione giudicata più soddisfacente tra quelle via via ricavate tramite aggregazioni successive. A tale scopo si considerano due passi consecutivi della classificazione, caratterizzati rispettivamente da $(g + 1)$ e g gruppi. Se la partizione più aggregata (con g gruppi) comporta una riduzione modesta di R^2 ed un incremento contenuto di *RMSSTD* si passa ad esaminare la tappa successiva, poiché l'aggregazione in oggetto coinvolge gruppi (elementi) relativamente omogenei tra loro. Se, viceversa, si manifesta un «salto» rilevante nei valori degli indici di bontà della partizione si considera come soddisfacente la classificazione del passo precedente, cioè quella con $(g + 1)$ gruppi.

Osservazione. Come già rilevato con riferimento al calcolo d'una distanza (si veda il capitolo 4, n. 5.1), la scomposizione (5.8) ha significato logico soltanto se tutte le variabili sono espresse nella stessa unità di misura; anche in tale circostanza, essa risulta comunque influenzata da eventuali differenze nell'ordine di grandezza o nella variabilità dei caratteri presi in esame. Nelle analisi reali i coefficienti R^2 e *RMSSTD* presentano pertanto le stesse limitazioni interpretative delle distanze utilizzate per effettuare la *cluster analysis*, potendo essere correttamente impiegati solo nel caso in cui le variabili siano tra loro comparabili oppure siano state preventivamente rese tali tramite un'opportuna trasformazione (quale, ad esempio, il calcolo degli scostamenti standardizzati).

La devianza totale delle p variabili non costituisce l'unico possibile criterio per misurare la variabilità in ambito multidimensionale (capi-

tolo 1, n. 9). Metodi differenti, qualora applicati alle matrici di covarianza nei gruppi e fra i gruppi associate ad una partizione, conducono alla definizione di numerosi coefficienti alternativi rispetto alla (5.15) ed alla (5.16). Per una rassegna degli approcci proposti in letteratura rimandiamo, ad esempio, agli articoli di Milligan and Cooper (1985), in cui sono riportati ben 30 criteri differenti, e di Hardy (1996).

Gli indici di valutazione delle partizioni ottenute ai vari passi dell'algoritmo non sono direttamente disponibili nella procedura *Cluster gerarchica* di SPSS (18). Per tale motivo nell'esempio che segue facciamo invece riferimento al pacchetto SAS, che consente il calcolo sia della (5.15) sia della (5.16) attraverso le opzioni *RSQUARE* e *RMSSTD* della procedura *Cluster*, applicata a partire dalla matrice dei dati originari (19).

Esempio. A titolo esemplificativo, le precedenti misure sono state ottenute per la classificazione gerarchica — analoga a quella già illustrata nel n. 4.2 — delle prime quattro variabili della tab. 4.1, utilizzando il metodo del legame completo e la distanza euclidea. Il corrispondente output del SAS è riportato nella tab. 5.7, che fornisce informazioni simili a quelle della tab. 5.5; si osservi tuttavia che — a differenza di SPSS — i passi della procedura sono qui indicati secondo un ordine decrescente, in funzione del numero di gruppi (anziché dello stadio di aggregazione). Come si evince dal confronto con la tab. 5.5, la gerarchia di partizioni è identica a quella ottenuta adottando la distanza della città a blocchi: la scelta d'una differente definizione di distanza non ha dunque condotto a variazioni nella sequenza di aggregazioni delle reti televisive.

---

(18) Qualora si volesse impiegare comunque SPSS, occorrerebbe innanzitutto salvare — attraverso la corrispondente opzione — il *cluster* di appartenenza di ciascuna unità statistica ai vari passi della procedura, scegliendo un numero di gruppi compreso tra 2 e $n - 1$. Le quantità che compaiono nella (5.8) dovrebbero poi essere calcolate separatamente per ciascuna partizione, utilizzando ad esempio la sequenza di istruzioni *Statistica - Modello lineare generalizzato - GLM multivariato*, che fornisce gli strumenti dell'analisi della varianza multivariata (Seber, 1984, cap. 9).

(19) Se l'*input* della procedura *Cluster* del SAS non consiste nella matrice dei dati originari ma in una matrice delle distanze — opzione necessaria quando si vuole adottare una metrica non euclidea — gli indici R^2 e *RMSSTD* non possono più essere ottenuti. In tale circostanza, il SAS consente comunque di calcolare una misura di coesione tra i gruppi che richiede soltanto la conoscenza delle distanze tra le unità. Questo indice è detto «Distanza relativa» (*Normalized Distance*) ed è definito come il rapporto tra la distanza dei due gruppi che si uniscono nel corrispondente passo e la media delle distanze. Se esso è maggiore di 1, i gruppi che si uniscono sono più eterogenei della media e quindi la loro aggregazione è sicuramente sconsigliabile.

Le ultime due colonne della tab. 5.7 contengono i valori degli indici R^2 e *RMSSTD* per i successivi passi del processo di agglomerazione. Dai risultati indicati emerge una variazione sensibile di entrambi i coefficienti passando dalla partizione con 4 gruppi a quella seguente con 3 gruppi ed un cambiamento ancora più accentuato considerando la classificazione con soli 2 gruppi. Tali modificazioni testimoniano un peggioramento molto rilevante delle soluzioni via via ottenute: ad esempio, passando da 3 a 2 *clusters*, la quota di variabilità totale delle 4 variabili «spiegata» dalla suddivisione in gruppi si riduce dal 78.9% a meno del 48%.

TAB. 5.7. *Output del programma SAS per la classificazione delle reti televisive con il metodo del legame completo e la distanza euclidea.*

Numero di cluster	Cluster accorpati		Frequenza del nuovo cluster	*RMS STD* del nuovo cluster	R^2
5	RAIUNO	RAIDUE	2	305.326	0.9637
4	RAITRE	RETE4	2	349.958	0.9159
3	CL 5	CL 4	4	425.136	0.7886
2	CL 3	ITALIA1	5	579.592	0.4761
1	CL 2	CANALE5	6	716.228	0.0000

Un «taglio» ragionevole del dendrogramma relativo alla classificazione delle reti televisive con il metodo del legame completo potrebbe quindi essere quello corrispondente alla partizione con 4 gruppi: (RAIUNO, RAIDUE); (RAITRE, RETE4); (CANALE5); (ITALIA1). In alternativa, una soluzione più sintetica — anche se meno omogenea — è costituita dalla successiva aggregazione in cui il primo ed il secondo *cluster* si fondono tra loro. Si noti che informazioni analoghe potevano essere colte — seppure in maniera assai meno evidente — anche dall'esame del «Programma di agglomerazione» di SPSS riportato nella tab. 5.5, osservando i considerevoli aumenti del livello di distanza (in quel caso della città a blocchi) nel passaggio dal secondo al terzo stadio e dal terzo al quarto stadio del procedimento gerarchico.

10. *Confronti tra partizioni*

Date due differenti partizioni d'un medesimo insieme di n unità statistiche, può essere interessante valutare in che misura tali classifica-

zioni differiscono tra loro. Il confronto più naturale è solitamente quello tra partizioni con il medesimo numero di gruppi, ma non si può escludere, in linea di principio, anche il confronto tra partizioni con un diverso numero di gruppi.

In particolare, esempi salienti di partizioni da raffrontare nelle applicazioni della *cluster analysis* possono essere:

i) quelle ottenute partendo da due matrici di distanze di diverso tipo (ad esempio, distanze euclidee e della città a blocchi), applicando alle stesse un medesimo metodo gerarchico ed effettuando il «taglio» del dendrogramma in corrispondenza dello stesso numero di gruppi;

ii) quelle, con il medesimo numero di gruppi, ottenute applicando due diversi algoritmi di classificazione (ad esempio, metodo del legame medio e metodo del centroide);

iii) quelle ottenute con la stessa metodologia di classificazione, per le medesime unità, considerando due differenti insiemi di variabili (ad esempio, per unità territoriali, una «batteria» di indicatori demografici ed una di variabili economiche), oppure lo stesso insieme di variabili in tempi successivi (come avviene solitamente nel caso delle indagini longitudinali - si veda il capitolo 1, n. 5).

Osservazione. Gli indici per il confronto tra partizioni sono dunque lo strumento che consente di misurare numericamente la stabilità d'una classificazione al variare dei caratteri analizzati e delle opzioni scelte per ottenerla (si veda il n. 2). Un campo di applicazione più recente consiste inoltre nello studio dell'*influenza* delle singole unità sui risultati della *cluster analysis*, cioè nella valutazione dell'impatto che ciascun elemento presenta sulla classificazione ottenuta. Tale ricerca avviene raffrontando la partizione di riferimento — calcolata su tutte le n unità statistiche — con quella di ugual numero di gruppi ottenuta classificando $n - 1$ unità ed escludendo di volta in volta l'elemento di cui si vuole misurare l'influenza (20).

La somiglianza tra partizioni può essere misurata a partire dal confronto tra le coppie di unità statistiche (21). (Al riguardo si ri-

---

(20) Per ulteriori dettagli al riguardo rimandiamo al contributo di Gnanadesikan, Kettenring and Landwehr (1977) ed ai più recenti articoli di Jolliffe, Jones and Morgan (1995), Cheng and Milligan (1996) e Cerioli (1999).

(21) Nel caso di algoritmi gerarchici, è possibile anche il confronto tra gli alberi n-dimensionali (o i dendrogrammi) che rappresentano l'intera famiglia delle partizioni ottenute ai successivi passi della procedura di agglomerazione (Gordon, 1996a, n. 9). Un problema connesso è costituito dalla determinazione d'un albero

corda che, dati n elementi, i confronti possibili sono in numero di $\binom{n}{2} = \frac{n(n-1)}{2}$). Se si considera una sola partizione, che indichiamo con P, a ciascuna delle $\binom{n}{2}$ coppie di unità può essere infatti associata una variabile indicatrice che assume il valore:

1 se la coppia è costituita da elementi appartenenti allo stesso gruppo nella partizione P;

0 se la coppia è costituita da elementi appartenenti a gruppi distinti nella partizione P.

Se si considerano due partizioni del medesimo insieme di n elementi, che indichiamo con P e P^*, le corrispondenti variabili indicatrici sono definite in modo analogo. Le *coppie* d'elementi possono allora essere classificate nella seguente tabella 2×2:

$P \setminus P^*$	1	0	*Tot.*
1	c_{11}	c_{10}	$c_{1\cdot}$
0	c_{01}	c_{00}	$c_{0\cdot}$
Tot.	$c_{\cdot 1}$	$c_{\cdot 0}$	$\binom{n}{2}$

ove:

c_{11} = numero di coppie di unità che appartengono allo stesso gruppo sia nella partizione P sia in P^*;

c_{10} = numero di coppie di unità che appartengono allo stesso gruppo nella partizione P ma appartengono a gruppi diversi in P^*;

c_{01} = numero di coppie di unità che appartengono allo stesso gruppo nella partizione P^* ma appartengono a gruppi diversi in P;

c_{00} = numero di coppie di unità che appartengono a gruppi diversi sia nella partizione P sia in P^*.

Si ha inoltre: $c_{1\cdot} = c_{11} + c_{10}$, $c_{\cdot 1} = c_{11} + c_{01}$ ed analogamente per $c_{0\cdot}$ e $c_{\cdot 0}$. Ovviamente, $c_{11} + c_{10} + c_{01} + c_{00} = \binom{n}{2}$.

Si noti che le quantità c_{11} e c_{00} indicano il numero di coppie d'elementi «trattate» in modo analogo nelle due partizioni. Esse forniscono

---

«di consenso», cioè d'una struttura di sintesi che rappresenti l'informazione comune a tutte le classificazioni gerarchiche poste a confronto; per una rassegna recente si veda Lapointe (1998).

pertanto una misura dell'accordo esistente tra P e P^*. Viceversa, le quantità c_{10} e c_{01} corrispondono alle coppie classificate in modo discordante nelle due partizioni e rappresentano il grado di disaccordo tra tali classificazioni.

Definizione. Si dice *indice di Rand* per il confronto tra due partizioni P e P^*, costituite rispettivamente da g e g^* gruppi, l'espressione seguente (Rand, 1971):

$$R_{P;P^*} = \frac{c_{11} + c_{00}}{\binom{n}{2}}. \tag{5.17}$$

L'indice (5.17) assume valore 1 se le due partizioni sono identiche ($P = P^*$) e valore 0 se tutte le coppie d'elementi appartenenti ad un gruppo in una partizione sono assegnate a gruppi diversi nell'altra partizione (come avviene, ad esempio, nel caso estremo in cui P è costituita da un unico gruppo di n unità, mentre P^* è formata da n gruppi d'un elemento ciascuno). L'indice di Rand può pertanto essere interpretato come una vera e propria misura di similarità tra partizioni, mentre il suo complemento ad uno, cioè il coefficiente $1 - R_{P;P^*}$, rappresenta un indice di dissimilarità sull'insieme di tutte le possibili partizioni di n unità (cfr. il capitolo 4).

Osservazione. L'indice $R_{P;P^*}$ può assumere il valore massimo unitario solo se $g = g^*$; in caso contrario si avrà certamente $R_{P;P^*} < 1$, poiché le corrispondenti partizioni non possono coincidere. L'interpretazione della (5.17) come misura di similarità è dunque corretta se si considera come spazio di riferimento quello di tutte le possibili partizioni di n unità (con un numero di gruppi variabile da 1 a n). Essa perde invece significato se lo spazio di riferimento è ristretto alle sole partizioni costituite, rispettivamente, da g e g^* gruppi, con $g \neq g^*$, poiché in tale spazio non è ammesso il caso $P = P^*$.

La definizione (5.17), seppure concettualmente molto semplice, non si dimostra appropriata a fini di calcolo. Innanzitutto, essa richiede l'effettuazione di tutti i confronti a coppie, il cui numero $\binom{n}{2}$ si rivela ben presto disagevole da gestire al crescere dei valori di n. Inoltre, i principali *packages* statistici (quali SPSS e SAS) non forniscono nel loro output alcuna indicazione su tali confronti. Pertanto, se si vuole applicare la (5.17) e non si dispone d'un programma *ad hoc*, occorre preliminarmente tagliare i dendrogrammi delle due classificazioni gerarchi-

che prese in esame, costruire le corrispondenti partizioni P e P^*, ed infine calcolare le quantità c_{11} e c_{00} considerando tutte le possibili coppie d'unità statistiche.

Per fortuna, una procedura di calcolo così laboriosa non è necessaria in pratica, poiché è possibile riscrivere l'indice $R_{P;P^*}$ in funzione degli elementi di P e P^*. A tale scopo, indichiamo con $\{C_1, \ldots, C_g\}$ i *clusters* che costituiscono la partizione P e con $\{C_1^*, \ldots, C_{g^*}^*\}$ quelli che formano P^*. Le corrispondenti numerosità sono date, rispettivamente, da $\{n_1, \ldots, n_g\}$ e $\{n_1^*, \ldots, n_{g^*}^*\}$. Il passo seguente consiste nel prendere in esame il gruppo in cui ciascuna unità statistica è classificata in P e quello cui appartiene in P^*. La sintesi di tali informazioni dà quindi luogo alla seguente tabella a doppia entrata (cfr. il Vol. I, p. 31):

$P \setminus P^*$	C_1^*	$\cdots$	C_c^*	$\cdots$	$C_{g^*}^*$	*Tot.*
C_1	n_{11}	$\cdots$	n_{1c}	$\cdots$	n_{1g^*}	n_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
C_r	n_{r1}	$\cdots$	n_{rc}	$\cdots$	n_{rg^*}	n_r
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
C_g	n_{g1}	$\cdots$	n_{gc}	$\cdots$	n_{gg^*}	n_g
Tot.	n_1^*	$\cdots$	n_c^*	$\cdots$	$n_{g^*}^*$	n

ove n_{rc} rappresenta il numero di *unità statistiche* classificate simultaneamente nel gruppo r-esimo di P e nel gruppo c-esimo di P^* ($r = 1, \ldots, g$; $c = 1, \ldots, g^*$). Ovviamente,

$$n_r = \sum_{c=1}^{g^*} n_{rc}$$

coincide con la numerosità del *cluster* r-esimo della partizione P ($r = 1, \ldots, g$);

$$n_c^* = \sum_{r=1}^{g} n_{rc}$$

coincide con la numerosità del *cluster* c-esimo di P^* $(c = 1, \ldots, g^*)$ e

$$n = \sum_{r=1}^{g} \sum_{c=1}^{g^*} n_{rc}$$

è il numero totale di unità statistiche.

Si può allora dimostrare (Rand, 1971, p. 847) che l'indice $R_{P;P^*}$ è esprimibile nella forma equivalente:

$$R_{P;P^*} = 1 - \frac{\sum_{r=1}^{g}(n_r)^2 + \sum_{c=1}^{g^*}(n_c^*)^2 - 2\sum_{r=1}^{g}\sum_{c=1}^{g^*}(n_{rc})^2}{n(n-1)}. \qquad (5.18)$$

Sebbene meno intuitiva rispetto alla (5.17), la (5.18) presenta un importante vantaggio computazionale: essa non richiede infatti l'effettuazione di alcun confronto a coppie, ma soltanto la conoscenza del *cluster* di appartenenza di ciascuna unità statistica nelle partizioni P e P^*.

Osservazione. La formulazione alternativa dell'indice di Rand fornita dalla (5.18) presenta anche un interessante risvolto teorico. Essa consente infatti di inserire il tema del confronto tra partizioni nell'ambito più ampio dell'analisi delle tabelle di contingenza e della misura dell'associazione (si veda il Vol. I, capitolo VI). In particolare, la tradizionale assunzione d'indipendenza tra due variabili corrisponde — nel presente contesto — all'ipotesi di assegnazione casuale delle n unità statistiche ai gruppi di P e P^*. La statistica $R_{P;P^*}$ può allora essere utilizzata, in un'impostazione di tipo inferenziale, per sottoporre a verifica tale ipotesi nulla e valutare la significatività della discrepanza tra P e P^*.

L'indice di Rand non costituisce l'unica misura proponibile ai fini del confronto tra partizioni. Per una rassegna dei possibili approcci si vedano gli articoli di Fowlkes and Mallows (1983), Hubert and Arabie (1985) e Cerioli (1997).

Esempio. Riprendiamo in esame le classificazioni gerarchiche delle reti televisive ottenute nel n. 4.2 tramite il metodo del legame singolo e quello del legame completo. Nel primo stadio della procedura entrambi gli algoritmi forniscono la medesima partizione; pertanto, l'indice di Rand tra le classificazioni con 5 gruppi ricavate con i due metodi alternativi è pari a 1.

Consideriamo invece le partizioni in quattro gruppi determinate al passo seguente dell'agglomerazione gerarchica, ed indichiamo con P quella relativa al legame singolo e con P^* quella riferita al legame completo. Si ha pertanto:

C_1 = (RAIUNO, RAIDUE, RAITRE), C_2 = (RETE4), C_3 = (CANALE5) e C_4 = (ITALIA1),

mentre

C_1^* = (RAIUNO, RAIDUE), C_2^* = (RAITRE, RETE4), C_3^* = (CANALE5) e C_4^* = (ITALIA1).

La tabella 2×2 ricavata dai $\binom{6}{2} = 15$ confronti a coppie è contenuta nella tab. 5.8, da cui si evince che per una sola coppia le corrispondenti reti televisive — RAIUNO e RAIDUE — sono classificate nello stesso *cluster* sia con il metodo del legame singolo sia con quello del legame completo, mentre 11 coppie sono costituite da reti appartenenti a gruppi diversi in entrambe le classificazioni. L'indice di Rand calcolato attraverso la definizione originaria (5.17) risulta quindi

$$R_{P;P^*} = \frac{1 + 11}{15} = 0.8$$

e segnala un buon accordo tra le due partizioni.

Se si desiderano evitare i confronti a coppie — scelta pressoché obbligata quando n è elevato — occorre invece applicare la formula (5.18). La tabella a doppia entrata ottenuta dalla classificazione (in quattro gruppi) delle reti TV secondo i due algoritmi presi in esame è riportata nella tab. 5.9. Quest'ultima può essere agevolmente costruita con SPSS salvando dapprima il *cluster* di appartenenza per le partizioni da raffrontare (tramite l'opzione *salva* del programma *Cluster gerarchica*, con il prefissato numero di gruppi) e selezionando poi il programma *Tavole di contingenza* dal menù *Riassumi*; in tale applicazione occorre infine indicare nel comando *righe* la variabile che raccoglie i gruppi di P e nel comando *colonne* quella che contiene i gruppi di P^*.

Il calcolo della (5.18) sui dati della tab. 5.9 fornisce il risultato seguente:

$$R_{P;P^*} = 1 - \frac{(3^2 + 1^2 + 1^2 + 1^2) + (2^2 + 2^2 + 1^2 + 1^2) - 2(2^2 + 1^2 + \cdots)}{30} = 0.8,$$

che coincide con quello ottenuto in precedenza. Si noti che l'«etichetta» associata ai diversi gruppi nelle due classificazioni è assolutamente irrilevante ai fini del confronto: ciò che conta è soltanto il fatto che le unità statistiche appartengano o meno allo stesso *cluster* — qualunque sia il suo numero progressivo — in P ed in P^*.

TAB. 5.8. *Classificazione delle coppie di reti televisive in base al* cluster *di appartenenza nelle partizioni di quattro gruppi ottenute tramite il metodo del legame singolo* (P) *ed il metodo del legame completo* (P*).

$P \setminus P^*$	1	0	*Tot.*
1	1	2	3
0	1	11	12
Tot.	2	13	15

TAB. 5.9. *Classificazione delle reti televisive in base al* cluster *di appartenenza nelle partizioni di quattro gruppi ottenute tramite il metodo del legame singolo* (P) *ed il metodo del legame completo* (P*).

$P \setminus P^*$	C_1^*	C_2^*	C_3^*	C_4^*	*Tot.*
C_1	2	1	0	0	3
C_2	0	1	0	0	1
C_3	0	0	1	0	1
C_4	0	0	0	1	1
Tot.	2	2	1	1	6

Nei successivi stadi della procedura aggregativa, le partizioni con tre e due gruppi ottenute attraverso i metodi del legame singolo e del legame completo sono identiche e pertanto i corrispondenti valori dell'indice di Rand risultano pari a 1.

11. *Esempio riepilogativo (I)*

I dati riportati nella tab. 5.10 fanno riferimento ad una pluralità di caratteristiche tecniche (quali le dimensioni, la cilindrata, la velocità massima etc.) ed economiche (il prezzo) per un insieme di 42 autovetture presenti sul mercato italiano nel 1998 (fonte: *Quattroruote*, novembre 1998). Tali variabili sono importanti ai fini d'una identificazione accurata delle peculiarità di ciascun modello, che non potrebbero essere colte nella loro interezza tramite semplici analisi univariate o bivariate. Attraverso l'applicazione della *cluster analysis* all'intera «batteria» di indicatori ($p = 8$), possiamo quindi tentare di suddividere i differenti modelli in gruppi relativamente omogenei, interpretabili come potenziali segmenti del mercato automobilistico.

TAB. 5.10. *Caratteristiche tecniche e prezzo di listino per 42 autovetture presenti sul mercato italiano nel 1998* (fonte: *Quattroruote*, novembre 1998).

N.	Modello	Prezzo (milioni di L.)	Lungh. (cm)	Massa (kg)	Cilindr. (cm^3)	Pot. (HP)	Vel. max (km/h)	Accel. (sec. da 0 a 100 km/h)	Cons. (l. × 100 km)
1	A. 156	40.400	443	1230	1747	106	210	9.30	8.20
2	A. 166	59.900	472	1345	1970	114	213	9.60	9.70
3	ASTRA	27.850	411	1181	1598	74	188	11.50	7.40
4	AUDI 4	49.393	448	1255	1781	92	205	10.50	8.50
5	AUDI 6	70.006	480	1430	2393	121	222	9.10	9.90
6	BMW 320	60.450	447	1356	1991	110	219	9.90	8.90
7	BMW 528	78.650	478	1440	2793	142	236	7.50	9.90
8	BRAVA	30.300	419	1090	1581	76	180	11.50	8.30
9	BRAVO	28.100	403	1050	1581	76	184	11.00	8.20
10	CIVIC	41.800	433	1190	1797	124	223	8.30	8.80
11	CLIO	21.950	377	960	1390	55	170	12.10	6.80
12	COROLLA	25.400	427	1105	1332	63	175	12.50	6.90
13	CORSA	17.600	374	939	973	40	150	18.00	5.80
14	FIAT 500	16.200	323	710	899	29	140	18.00	6.10
15	FIAT 600	15.800	332	730	899	29	140	18.00	6.10
16	FIESTA	21.470	383	1020	1242	55	170	12.70	7.10
17	FOCUS	28.550	415	1077	1596	74	185	11.00	6.80
18	GOLF	32.299	415	1132	1595	74	188	10.90	7.60
19	IBIZA	18.901	385	960	999	37	145	19.40	6.10
20	KA	16.470	362	913	1299	44	155	14.80	5.90
21	LANCIA K	59.000	469	1450	2446	129	218	8.70	10.90
22	LANCIA Y	24.250	372	910	1242	63	177	10.90	6.60
23	LUPO	17.695	353	818	999	37	152	17.90	5.80
24	MAREA	40.850	439	1255	1998	108	207	8.70	9.80
25	MEGANE	32.500	413	1220	1598	66	170	13.70	8.20
26	MERC. 200	61.700	449	1290	1998	100	203	11.00	9.40
27	MERC. 320	101.457	480	1580	3199	165	238	7.70	10.30
28	MICRA	21.640	372	865	1275	55	170	12.00	6.60
29	P. 106	18.300	368	815	954	37	150	19.20	6.20
30	P. 206	23.650	384	950	1360	55	170	13.20	6.60
31	P. 406	38.150	456	1275	1761	81	194	12.50	8.70
32	PANDA	11.650	341	715	899	29	135	19.50	6.70
33	PASSAT	45.355	468	1340	1781	92	206	10.90	8.60
34	POLO	22.447	372	940	999	37	151	18.50	5.90
35	PUNTO 55	21.150	376	865	1108	40	150	16.50	6.50
36	PUNTO 75	23.350	376	895	1242	54	170	12.00	7.10
37	SAXO	17.500	372	825	1124	44	162	15.30	6.50
38	TWINGO	15.950	343	820	1149	43	151	13.40	6.00
39	VECTRA	39.210	448	1314	1799	85	203	11.00	8.20
40	VOLVO 70	58.474	472	1437	1984	132	215	9.30	11.30
41	XANTIA	38.350	452	1264	1761	81	194	11.90	8.70
42	JAGUAR	110.900	502	1710	3996	209	240	7.30	11.90

Le figure 5.12 e 5.13 riportano i dendrogrammi di SPSS ottenuti applicando il metodo del legame medio (fra i gruppi) con la distanza euclidea e con quella della città a blocchi sugli scostamenti standardizzati. Si nota immediatamente che l'impiego delle due metriche produce risultati marcatamente diversi all'ultimo passo della procedura classificatoria: con la distanza euclidea si fondono infatti due ampi gruppi formatisi nei passi precedenti, mentre utilizzando la metrica *city-block* un'unica unità isolata (corrispondente all'autovettura Jaguar) si aggrega al *cluster* costituito da tutte le rimanenti; anche l'incremento nel livello di distanza che si osserva nelle ultime fasi della procedura risulta molto più elevato nel caso della città a blocchi.

Se si tiene presente il fatto che la metrica *city-block*, essendo funzione delle differenze in modulo tra i valori, è meno influenzata dalla presenza di eventuali *outliers* (cfr. il capitolo 4, n. 3.1) e tende quindi a porli in luce più chiaramente, il comportamento molto differente dell'unità 20 nelle due analisi può essere interpretato come un'indicazione dell'anomalia di tale unità rispetto alle altre. Pertanto, possiamo considerare la Jaguar come un *outlier* nello spazio multidimensionale costituito dalle 8 variabili prese in esame ed escludere tale modello dalle elaborazioni successive. Ovviamente, ciò non significa che la Jaguar non meriti attenzione da parte del ricercatore o dell'analista di mercato, ma soltanto che essa ha caratteristiche così peculiari da dover essere considerata separatamente rispetto alle altre autovetture: la sua inclusione rischierebbe infatti di distorcere i risultati delle procedure classificatorie riferite alle rimanenti 41 unità.

FIG. 5.12. *Dendrogramma ottenuto tramite SPSS nella classificazione delle autovetture (compresa Jaguar) con la distanza euclidea sugli scostamenti standardizzati ed il metodo del legame medio.*

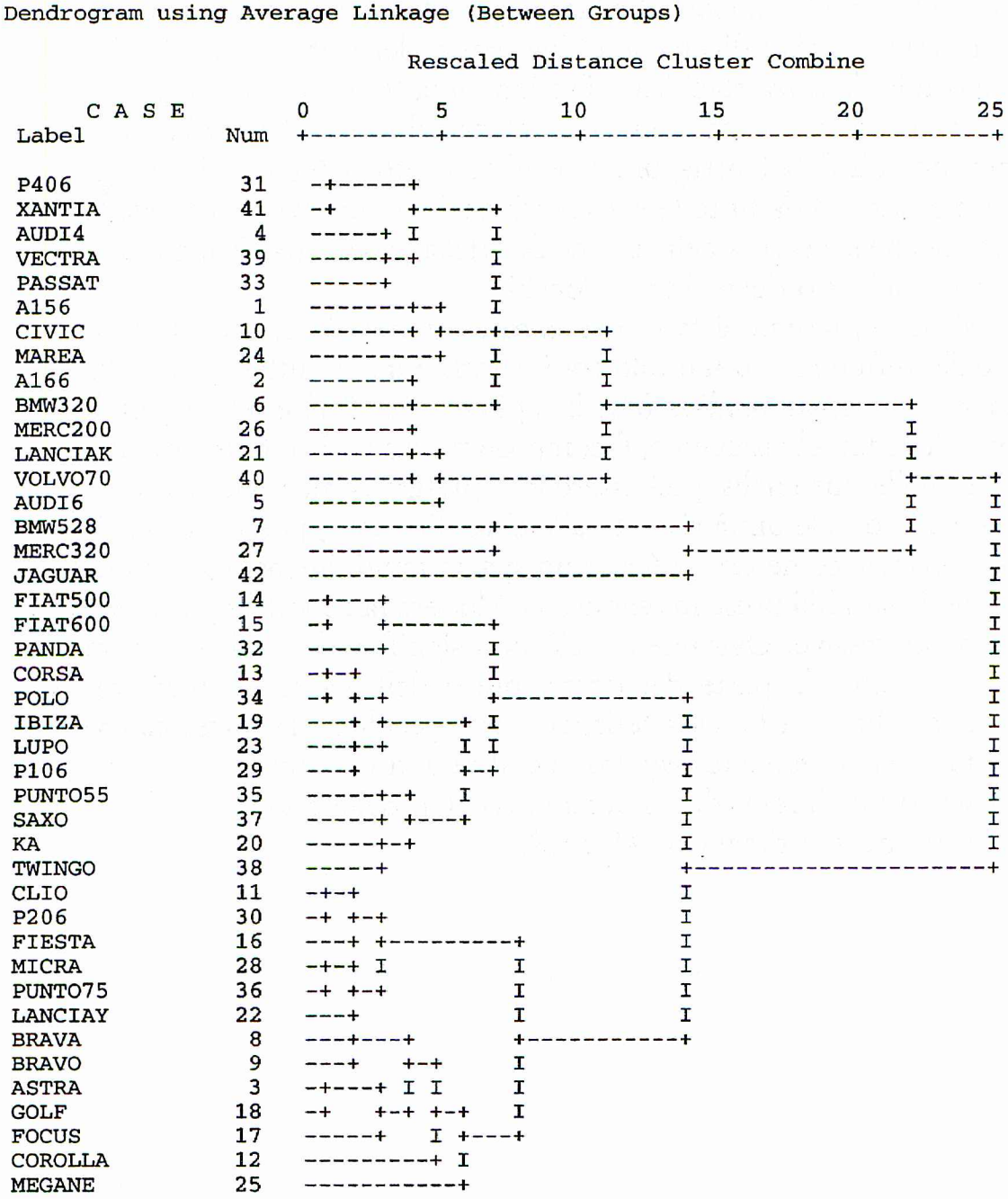
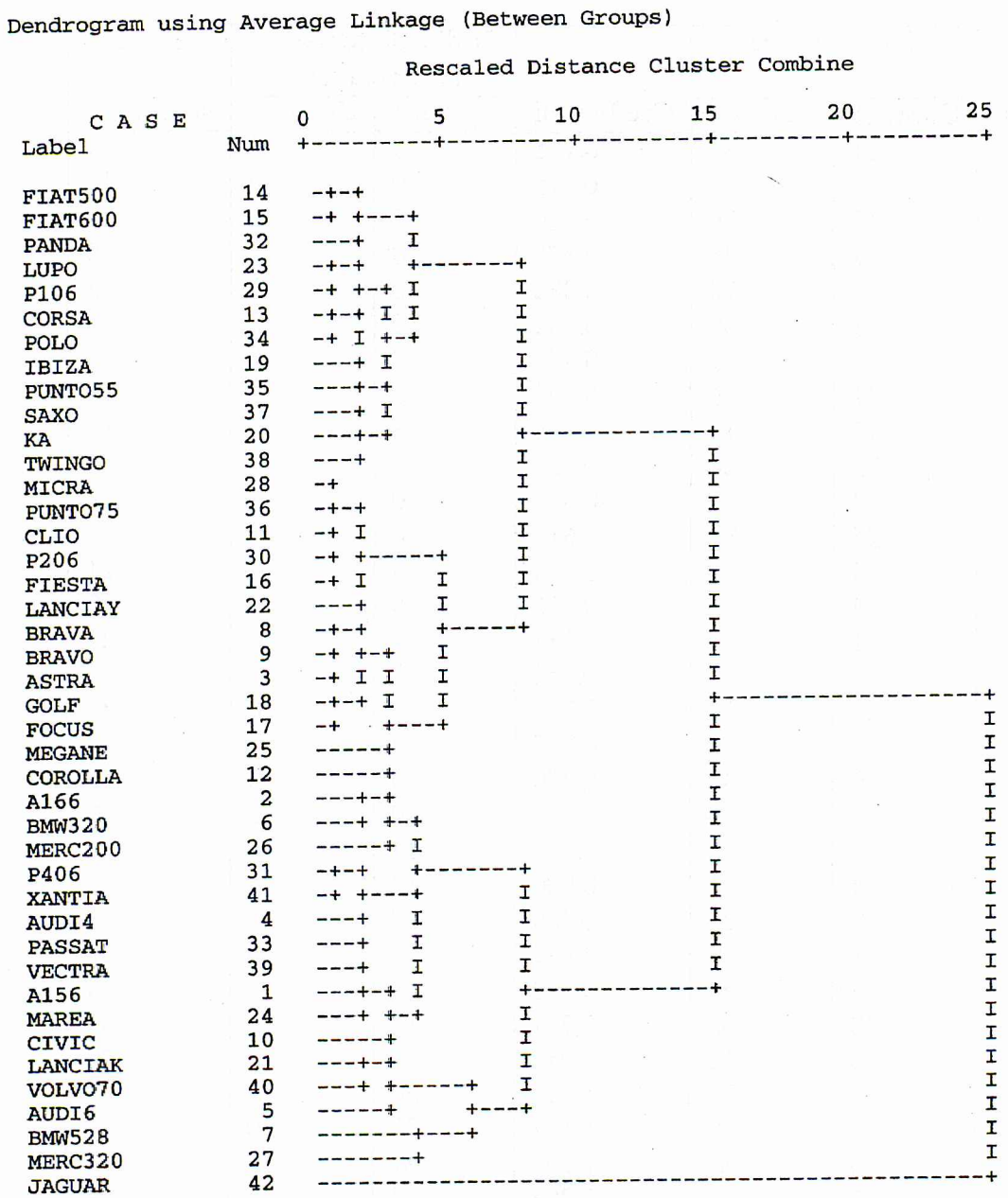


FIG. 5.13. *Dendrogramma ottenuto tramite SPSS nella classificazione delle autovetture (compresa Jaguar) con la distanza* city-block *sugli scostamenti standardizzati ed il metodo del legame medio.*



Il «programma di agglomerazione» di SPSS ottenuto dall'applicazione del metodo del legame medio al *data set* ridotto ($n = 41$) con la distanza euclidea è riportato nella tab. 5.11, mentre il corrispondente dendrogramma è mostrato nella fig. 5.14. In questo caso, l'impiego della metrica della città a blocchi in luogo di quella euclidea produce variazioni del tutto trascurabili sui risultati della procedura gerarchica ed il relativo output è pertanto omesso.

TAB. 5.11. *Programma di agglomerazione di SPSS per la classificazione delle autovetture (esclusa Jaguar) con il metodo del legame medio e la distanza euclidea sugli scostamenti standardizzati.*

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione dei cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	31	41	0.198	0	0	15
2	14	15	0.215	0	0	17
3	13	34	0.310	0	0	9
4	28	36	0.375	0	0	10
5	3	18	0.382	0	0	13
6	11	30	0.388	0	0	8
7	8	9	0.454	0	0	21
8	11	16	0.502	6	0	16
9	13	19	0.524	3	0	18
10	22	28	0.551	0	4	16
11	23	29	0.569	0	0	18
12	35	37	0.613	0	0	26
13	3	17	0.615	5	0	21
14	4	33	0.616	0	0	22
15	31	39	0.633	1	0	22
16	11	22	0.652	8	10	36
17	14	32	0.712	2	0	34
18	13	23	0.765	9	11	31
19	20	38	0.770	0	0	26
20	2	6	0.799	0	0	23
21	3	8	0.800	13	7	27
22	4	31	0.806	14	15	32
23	2	26	0.865	20	0	33
24	1	10	0.889	0	0	28
25	5	21	0.938	0	0	29
26	20	35	0.951	19	12	31
27	3	12	1.086	21	0	30
28	1	24	1.130	24	0	32
29	5	40	1.172	25	0	37
30	3	25	1.198	27	0	36
31	13	20	1.292	18	26	34
32	1	4	1.433	28	22	33
33	1	2	1.512	32	23	37
34	13	14	1.512	31	17	38
35	7	27	1.659	0	0	39
36	3	11	1.670	30	16	38
37	1	5	2.288	33	29	39
38	3	13	2.767	36	34	40
39	1	7	3.656	37	35	40
40	1	3	4.845	39	38	0

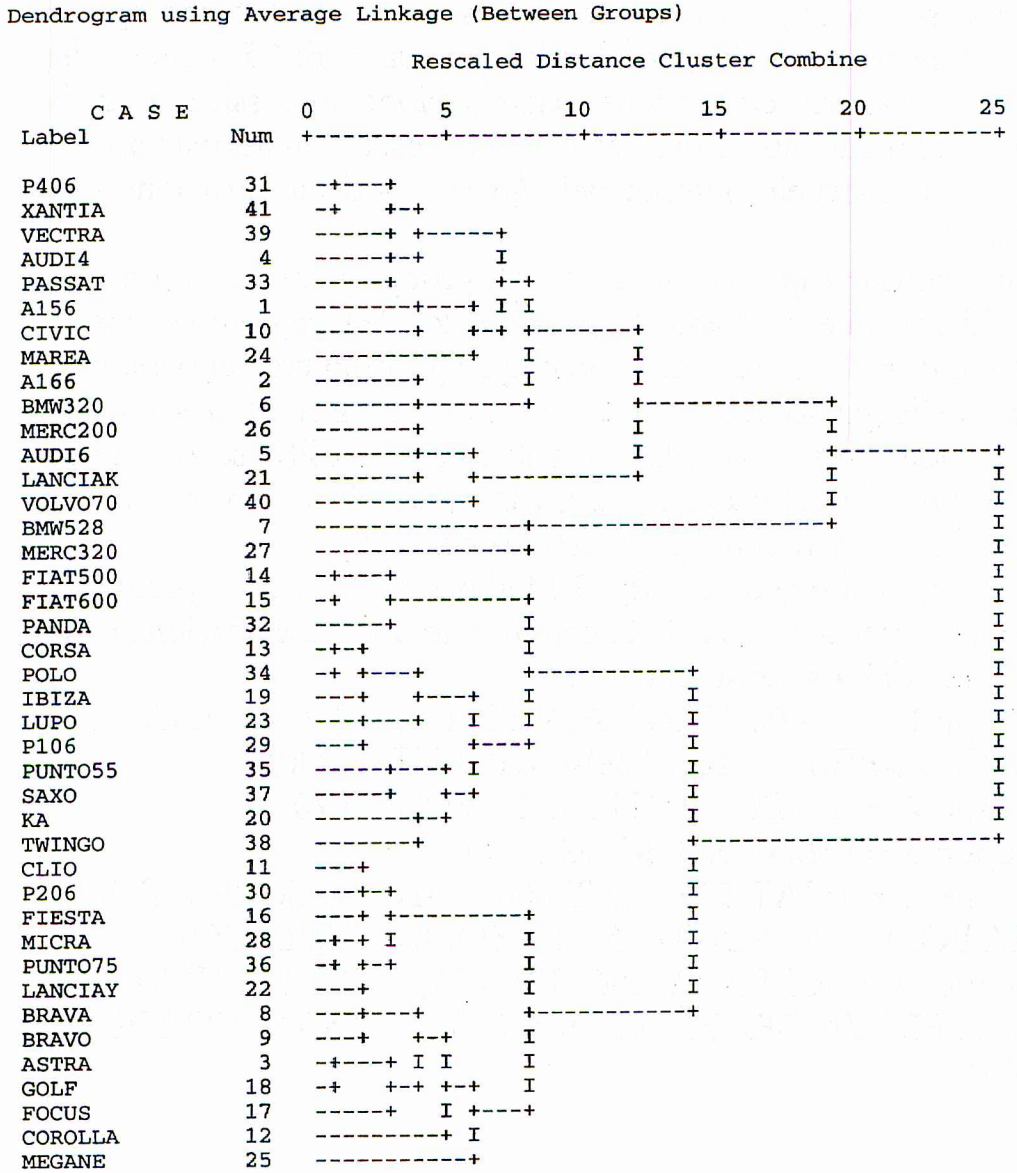
L'esame della famiglia di partizioni ricavate ai successivi passi della procedura di agglomerazione gerarchica pone in luce una «macrostruttura» costituita da 5 *clusters*, ricavata sezionando il dendrogramma allo stadio 36, cui corrisponde un livello di distanza uguale a 1.67 (8.62 se riscalato rispetto al valore massimo di 25, come nella fig. 5.14). La distanza cui avviene l'aggregazione successiva si rivela invece notevolmente superiore, segnalando così l'inadeguatezza — in termini di omogeneità interna dei *clusters* — della partizione con soli 4 gruppi.

Le autovetture appartenenti ai singoli gruppi possono essere agevolmente identificate mediante l'ispezione del dendrogramma, osservando i legami tra le unità statistiche che l'algoritmo crea ai passi precedenti a quello prescelto. Il *cluster* di appartenenza di ciascun elemento può inoltre essere visualizzato nell'output di SPSS selezionando l'opzione *statistiche* del programma *Cluster* gerarchica, ovvero memorizzato nel *file* dei dati come già illustrato nel n. 10.

Nell'esempio, il taglio della fig. 5.14 al passo 36, in corrispondenza d'una distanza riscalata pari a 8.62, conduce ad una classificazione delle auto costituita dai 5 gruppi seguenti:

- gruppo 1 = (P. 406, XANTIA, VECTRA, AUDI 4, PASSAT, A. 156, CIVIC, MAREA, A. 166, BMW 320, MERC. 200);
- gruppo 2 = (AUDI 6, LANCIA K, VOLVO 70);
- gruppo 3 = (BMW 528, MERC. 320);
- gruppo 4 = (FIAT 500, FIAT 600, PANDA, CORSA, POLO, IBIZA, LUPO, P. 106, PUNTO 55, SAXO, KA, TWINGO);
- gruppo 5 = (CLIO, P. 206, FIESTA, MICRA, PUNTO 75, LANCIA Y, BRAVA, BRAVO, ASTRA, GOLF, FOCUS, COROLLA, MEGANE).

FIG. 5.14. *Dendrogramma ottenuto tramite SPSS nella classificazione delle autovetture (esclusa Jaguar) con la distanza euclidea sugli scostamenti standardizzati ed il metodo del legame medio.*



I *clusters* così individuati possono essere considerati rappresentativi di altrettanti (ampi) segmenti del mercato automobilistico. La corrispondente partizione risulta sufficientemente omogenea, in quanto il valore dell'indice R^2 (calcolato separatamente, tramite il programma SAS) è pari a circa 0.90. Si osservi che la prossimità nella sequenza secondo cui le unità statistiche sono rappresentate nel dendrogramma non indica necessariamente la somiglianza tra esse: ad esempio, i modelli Mercedes 320 e Fiat 500 appaiono di seguito nella fig. 5.14 ma sono agli estremi opposti del mercato automobilistico, poiché appar-

tengono a gruppi distinti che si fondono soltanto all'ultimo passo della sequenza gerarchica.

Se si desidera una classificazione meno aggregata delle autovetture, può essere ragionevole un taglio del dendrogramma al passo 32 oppure a quello immediatamente precedente, in corrispondenza d'una partizione di 9 o 10 gruppi. In particolare, nella prima di tali soluzioni (cui è associato un valore di R^2 uguale a circa 0.95), il gruppo 1 precedentemente individuato risulta scisso in:

- gruppo 1a = (P. 406, XANTIA, VECTRA, AUDI 4, PASSAT, A. 156, CIVIC, MAREA);
- gruppo 1b = (A. 166, BMW 320, MERC. 200);

il *cluster* 4, corrispondente alle utilitarie, è suddiviso in:

- gruppo 4a = (FIAT 500, FIAT 600, PANDA);
- gruppo 4b = (CORSA, POLO, IBIZA, LUPO, P. 106, PUNTO 55, SAXO, KA, TWINGO);

ed il gruppo 5, contenente autovetture di fascia intermedia, risulta ora frazionato in:

- gruppo 5a = (CLIO, P. 206, FIESTA, MICRA, PUNTO 75, LANCIA Y);
- gruppo 5b = (BRAVA, BRAVO, ASTRA, GOLF, FOCUS, COROLLA, MEGANE).

Il gruppo 3 rimane invece invariato, mentre le autovetture di cilindrata maggiore (BMW 528 e MERC. 320) sono ora isolate e formano ciascuna un *cluster* a sé (22).

Si noti che le unità che appartengono al medesimo gruppo in questa partizione (più fine) sono classificate insieme anche nella partizione (più aggregata) formata da 5 gruppi, come discende dalle proprietà dei metodi gerarchici.

In conclusione, i risultati ottenuti nel presente esempio (tramite il metodo del legame medio) forniscono alcune potenziali classificazioni delle 41 autovetture prese in esame, con $g = 5$, $g = 9$, oppure $g = 10$. Tali soluzioni appaiono tutte ragionevoli, essendo costituite da *clusters* sufficientemente omogenei al proprio interno, e forniscono — seppure ad un livello di aggregazione differente — criteri alternativi per una possibile segmentazione del mercato automobilistico. La scelta

---

(22) Questa classificazione più analitica è molto simile a quella usualmente considerata per il mercato automobilistico, che per le berline distingue i segmenti: A (mini), B (piccole), C (compatte), D (medie), E (superiori), F (grandi berline).

finale d'una sola di esse dipenderà dunque dagli obiettivi specifici del ricercatore e dalla sua volontà di privilegiare la coesione interna dei gruppi oppure la sintesi della partizione finale.

12. *Metodi non gerarchici di classificazione*

I metodi non gerarchici di classificazione — a volte detti anche *partitioning methods* — si propongono di ottenere una sola partizione degli n elementi in g gruppi ($g < n$), con g scelto a priori dal ricercatore. La ricerca d'un unico raggruppamento costituisce dunque l'elemento distintivo rispetto agli algoritmi gerarchici esaminati in precedenza: in luogo della famiglia di partizioni associate ai successivi livelli di distanza e con un numero di *clusters* via via decrescente da n fino a 1, le procedure non gerarchiche mirano infatti a conseguire una classificazione che soddisfi determinati criteri d'ottimalità e che sia costituita da un numero di gruppi prefissato (pari a g).

I possibili vantaggi d'un simile approccio sono di varia natura. Innanzitutto, è possibile formalizzare — almeno in via di principio — il meccanismo di allocazione delle unità ai gruppi attraverso la specificazione esplicita d'una funzione obiettivo. Date le finalità della *cluster analysis* tale funzione è solitamente espressa in termini della scomposizione della devianza totale riportata nella formula (5.8). Il processo di classificazione si riconduce quindi ad un problema di ottimizzazione, cioè alla ricerca della partizione cui corrisponde la massima coesione nei gruppi, secondo il criterio prescelto (23).

Inoltre, variando il numero di gruppi, viene meno il vincolo che tutte le coppie di unità che risultano tra loro unite ad un determinato livello di aggregazione gerarchica non possono più essere separate ai livelli successivi (si veda il n. 4). Per ogni valore di g, infatti, l'algoritmo non gerarchico classifica ogni elemento esclusivamente sulla base del criterio prescelto ed i risultati ottenuti possono essere diversi al variare del numero di gruppi. Ciò consente di superare i potenziali inconvenienti dovuti ad una fusione « errata » di unità eterogenee nei passi iniziali d'una procedura gerarchica.

---

(23) Per un'introduzione alle funzioni obiettivo proposte in letteratura si vedano, ad esempio, Seber (1984, pp. 380-384) e Everitt (1993, cap. 5); per una trattazione più approfondita rinviamo invece a Späth (1985).

Purtroppo, il numero di modi in cui è possibile suddividere n elementi in g gruppi non sovrapposti si rivela incredibilmente grande anche per valori modesti di n e di g. Ad esempio, Everitt (1993, p. 93) riporta che vi sono oltre 45 miliardi di potenziali partizioni anche nel caso molto semplice in cui $n = 20$ e $g = 4$, mentre tale numero sale a circa 6.9×10^{17} se $n = 25$ e $g = 8$. È dunque chiaro che l'enumerazione completa di tutte le possibili partizioni non risulta quasi mai perseguibile nelle applicazioni di concreto interesse, anche disponendo d'un calcolatore molto potente. Di conseguenza, gli algoritmi di uso operativo non possono mirare al raggiungimento d'un ottimo globale della funzione obiettivo, ma devono necessariamente fare riferimento a criteri meno vincolanti.

In particolare, come vedremo in dettaglio nel n. 12.1 con riferimento al cosiddetto metodo delle k-medie, gli algoritmi di classificazione non gerarchici si fondano solitamente sull'esecuzione d'una procedura iterativa che può essere schematizzata nelle seguenti fasi:

i) scelta d'una classificazione iniziale delle n unità con un numero di gruppi prefissato;

ii) calcolo della variazione nella funzione obiettivo causata dallo spostamento di ciascun elemento dal gruppo di appartenenza ad un altro ed allocazione di ogni unità al *cluster* che garantisce il miglioramento più elevato nella coesione interna dei gruppi;

iii) iterazione del passo precedente finché non viene soddisfatta una regola di arresto.

Il ricorso ad una struttura di calcolo di tipo iterativo, per un solo valore di g, rende gli algoritmi non gerarchici assai veloci e non richiede la determinazione preliminare della matrice delle distanze tra le unità (di dimensioni $n \times n$). Tali metodi risultano quindi particolarmente adatti alle situazioni in cui n è molto elevato (ad esempio, $n > 1000$) (24): in simili circostanze, l'implementazione d'una procedura gerarchica tradizionale comporterebbe infatti tempi d'esecuzione assai superiori e richiederebbe un elaboratore con una capacità di me-

(24) *Data sets* di dimensioni così grandi sono ormai frequenti in numerosi settori d'applicazione: si vedano, ad esempio, Eddy, Mockus and Oue (1996) e Hruschka and Natter (1999). In simili casi una metodologia alternativa all'analisi classificatoria può essere costituita dalle reti neurali, che saranno illustrate nel successivo capitolo 8. Entrambe le tecniche — *cluster analysis* e reti neurali — costituiscono strumenti essenziali di quell'approccio esplorativo allo studio di grandi masse di dati denominato *data mining*, già menzionato nel capitolo 1.

moria centrale molto più elevata per il salvataggio della matrice delle distanze.

Le tecniche non gerarchiche sono solitamente utilizzate anche nelle applicazioni in cui n è dell'ordine di qualche centinaia di unità, poiché — a causa dell'elevato numero di aggregazioni — il dendrogramma associato ad un algoritmo gerarchico risulterebbe comunque di difficile lettura ed interpretazione. Inoltre, esse possono essere preferibili quando l'interesse della ricerca consiste nella caratterizzazione delle peculiarità dei gruppi (ad esempio, tramite il calcolo dei centroidi e delle misure di variabilità multidimensionale dei *clusters*), piuttosto che nello studio del comportamento delle singole unità nei successivi passi dell'agglomerazione gerarchica.

A fronte delle potenzialità indicate, i metodi di classificazione non gerarchici presentano tuttavia anche alcuni inconvenienti, legati soprattutto alla necessità di definire preventivamente:

i) il valore di g, cioè il numero di gruppi della partizione «ottima» che si vuole ricavare;

ii) la configurazione di partenza dei *clusters*, necessaria per inizializzare l'algoritmo iterativo di classificazione.

Tali problemi saranno affrontati nel successivo n. 12.2, dopo l'illustrazione d'un particolare algoritmo non gerarchico: quello delle k-medie.

12.1. *Il metodo delle* k-*medie*

Il metodo delle k-medie costituisce — seppure con alcune varianti — l'algoritmo di classificazione non gerarchico di uso più comune ed è implementato nei principali *packages* statistici (tra i quali SPSS e SAS, che utilizzeremo nelle applicazioni). Seguendo la notazione del presente capitolo, in cui g rappresenta il numero di gruppi, esso dovrebbe essere indicato più coerentemente come «metodo delle g-medie»; per uniformità con la letteratura esistente continueremo tuttavia ad adottare la denominazione convenzionale, che è quella richiamata anche da SPSS.

Tale algoritmo conduce ad una classificazione delle unità statistiche in g gruppi distinti, con g fissato *a priori*, tramite una procedura iterativa che consiste dei passi seguenti.

1) Si scelgono g «poli» iniziali (detti anche semi, *seeds*, oppure punti origine), cioè g punti nello spazio p-dimensionale che costituiscono i centroidi dei *clusters* nella partizione iniziale. I poli possono essere individuati attraverso criteri differenti (illustrati nel n. 12.2), gene-

ralmente in modo tale che essi siano abbastanza distanti tra loro. Si costruisce quindi la partizione iniziale, costituita da g gruppi, allocando ciascuna unità al *cluster* il cui polo risulta il più vicino.

2) Per ogni elemento si calcola la distanza dai centroidi dei g gruppi: se la distanza minima non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, l'unità è riassegnata al *cluster* corrispondente al centroide più vicino. In caso di riallocazione di un'unità, si ricalcola il centroide sia del nuovo sia del vecchio gruppo di appartenenza.

3) Si ripete il passo 2 sino a quando non è raggiunta la convergenza dell'algoritmo, cioè finché non si verifica alcuna modificazione dei poli — e quindi dei gruppi — rispetto all'iterazione precedente.

In alternativa, se si vogliono ridurre i tempi di calcolo, la condizione di arresto prevista al punto 3 della procedura può essere sostituita da una regola meno restrittiva, che prevede l'interruzione della procedura in ciascuno dei seguenti casi:

3*a*) convergenza dell'algoritmo;

3*b*) distanza tra ciascun centroide calcolato nell'iterazione corrente ed il corrispondente centroide nell'iterazione precedente non superiore ad una soglia prefissata;

3*c*) raggiungimento del numero massimo di iterazioni prescelto.

Per ottenere una partizione con un diverso numero di gruppi, ad esempio g^*, è necessario ripetere tutti i passi del procedimento, partendo dalla fase 1 e sostituendo g con g^*.

Le successive fasi della metodologia illustrata richiedono il calcolo ripetuto della distanza tra ciascun punto ed i centroidi dei g gruppi: tale procedura appartiene dunque alla classe di algoritmi di classificazione che adottano la tecnica denominata « ordinamento rispetto al centroide più vicino » (*nearest centroid sorting*). La metrica utilizzata nel calcolo di queste distanze è solitamente quella euclidea, in quanto essa garantisce la convergenza della procedura iterativa (Anderberg, 1973, p. 166) (25). Pertanto, all'iterazione t, la distanza tra l'unità i-esima ed il centroide del gruppo l ($i = 1, \ldots, n$; $l = 1, \ldots, g$) è data da:

---

(25) Se si volesse adottare, ad esempio, la metrica della città a blocchi, occorrerebbe calcolare la distanza tra ciascun elemento ed i vettori p-dimensionali che hanno per coordinate le mediane (anziché le medie) delle variabili in ciascun gruppo. Tali vettori costituiscono una versione robusta dei centroidi, in quanto non sono influenzati dall'eventuale presenza di valori anomali, e sono stati chiamati « medoidi » (*medoids*) da Kaufman and Rousseeuw (1990).

$$d(\mathbf{x}_i, \bar{\mathbf{x}}_l^{(t)}) = \sqrt{\sum_{s=1}^{p} \left(x_{is} - \bar{x}_{s,l}^{(t)}\right)^2}, \tag{5.19}$$

ove $\bar{\mathbf{x}}_l^{(t)} = [\bar{x}_{1,l}^{(t)}, \ldots, \bar{x}_{p,l}^{(t)}]'$ rappresenta il centroide del gruppo l calcolato all'iterazione t.

Dalla (5.19) risulta chiaro che il metodo delle k-medie — con l'impiego della distanza euclidea — ha come obiettivo implicito la ricerca della partizione (con g *clusters*) che soddisfa un criterio di coesione interna fondato sulla Devianza nei gruppi, cioè sulla minimizzazione della quantità W riportata nella (5.10). Una naturale misura della bontà della soluzione ottenuta con tale algoritmo è quindi costituita dall'indice R^2 illustrato nel n. 9 (si veda l'espressione (5.15)). Nel presente contesto non risulta invece applicabile l'indice *RMSSTD*, poiché esso richiede la determinazione d'una successione gerarchica di partizioni.

Al pari del metodo di Ward, considerato nel n. 7, anche quello delle k-medie produce tendenzialmente *clusters* di forma sferica e di dimensioni abbastanza simili tra loro. La relazione con la scomposizione (5.8) della Devianza totale ha consentito inoltre lo sviluppo d'una complessa teoria inferenziale concernente il metodo delle k-medie, nonché la predisposizione di veri e propri tests di significatività per la verifica dell'esistenza di gruppi distinti (26).

Osservazione. La procedura iterativa su cui si fonda il metodo delle k-medie presenta anche un potenziale inconveniente, spesso sottaciuto nelle applicazioni: la classificazione finale può essere influenzata dall'ordine in cui sono elencate le unità statistiche nella matrice dei dati. Tale caratteristica — chiaramente insoddisfacente, poiché il numero progressivo con cui sono rappresentati gli elementi non ha alcuna rilevanza ai fini dell'analisi — è infatti attribuibile al ricalcolo dei centroidi di gruppo ad ogni allocazione (Anderberg, pp. 162-163).

È possibile adottare una versione alternativa dell'algoritmo delle k-medie, in cui i centroidi dei *clusters* sono ricalcolati una sola volta in

---

(26) Si vedano Hartigan (1975) e Bock (1985). In un'impostazione di tipo più generale, è possibile anche pensare all'algoritmo classificatorio come ad uno strumento per la stima dei centroidi (ignoti) delle g popolazioni da cui si suppone siano state estratte casualmente le n osservazioni. Per ulteriori approfondimenti con riferimento al metodo delle k-medie, rimandiamo ai lavori di Hartigan (1978), Pollard (1981) e Cuesta-Albertos, Gordaliza and Matrán (1997), nonché alla rassegna di Bock (1996a, n. 4).

ciascuna iterazione, e cioè soltanto dopo avere assegnato tutte le unità al gruppo più vicino. Anche questa opzione non consente tuttavia di risolvere definitivamente l'inconveniente sopra descritto, in quanto l'ordine degli elementi può comunque modificare i risultati tramite l'influenza eventualmente esercitata sulla scelta della partizione iniziale al passo 1 della procedura iterativa.

L'effettiva sensibilità della classificazione al modo in cui è formata la lista delle unità risulta di difficile valutazione in termini generali. Tale aspetto dovrebbe però essere sempre tenuto presente nelle applicazioni, ad esempio attraverso analisi ripetute partendo da differenti sequenze iniziali. In particolare, un'esecuzione «cieca» del metodo delle k-medie può portare a risultati instabili nei casi seguenti:

i) se nei dati non esiste una chiara struttura di gruppo, con *clusters* ben separati;

ii) se si è interessati ad indagare caratteristiche «locali» dei dati, quali l'influenza di singole osservazioni o la presenza di valori anomali;

iii) se n è molto piccolo.

L'algoritmo delle k-medie è implementato sia in SPSS, tramite la procedura *Cluster k-medie* (detta anche *Quick cluster*) selezionabile dal menù *Classificazione*, sia in SAS, tramite la procedura *Fastclus*. Il nome attribuito a tali strumenti fa intuire che si tratta di algoritmi assai veloci e quindi applicabili anche a *data sets* di dimensioni elevate, soprattutto quando g è piccolo rispetto a n. Entrambi i pacchetti prevedono la metrica euclidea (5.19) come unica possibilità per il calcolo della distanza tra unità e centroidi di gruppo.

12.2. *La scelta della partizione iniziale*

L'impossibilità pratica, nelle applicazioni di concreto interesse, di enumerare l'insieme completo delle partizioni di n elementi in g gruppi distinti fa sì che la classificazione «ottima» cui si perviene tramite il metodo delle k-medie possa in realtà corrispondere ad un minimo locale della funzione obiettivo (che si fonda, lo ricordiamo, sul calcolo della Devianza nei gruppi prodotti dalla classificazione). In tali situazioni la soluzione ottenuta può dipendere dalla configurazione di partenza, scelta nella fase 1 di avvio del procedimento iterativo. Ai fini d'una corretta interpretazione dei risultati conseguiti, assume pertanto grande rilievo una valutazione accurata dei criteri utilizzati per pervenire alla partizione iniziale: in particolare, gli aspetti cruciali — che esamineremo nel prosieguo — sono quelli concernenti la scelta del nu-

mero di gruppi e dei semi iniziali. Si noti che molte delle considerazioni svolte in questo paragrafo hanno validità più generale rispetto al solo metodo delle *k*-medie, potendosi applicare a qualunque algoritmo di tipo non gerarchico.

i) *Scelta del numero di gruppi*

Il numero di gruppi, *g*, deve essere fissato *a priori* dal ricercatore, prima dell'applicazione d'un algoritmo non gerarchico di classificazione. Il criterio probabilmente più diffuso per la determinazione pratica del numero di gruppi consiste nell'esecuzione ripetuta dell'analisi con differenti valori di *g*, nella successiva valutazione — ad esempio tramite l'indice R^2 — della bontà delle partizioni così ottenute e nella selezione finale di quella ritenuta più soddisfacente (tenendo presente anche l'obiettivo di sintesi della classificazione prescelta).

Un simile modo di procedere, seppure molto semplice ed intuitivo, presenta però anche alcuni inconvenienti. Innanzitutto, da un punto di vista teorico, non esiste alcuna garanzia che il ricercatore sia in grado d'individuare una soluzione «ottima» tra tutte le partizioni ottenute (al variare di *g*), né tantomeno che gli indici di coesione calcolati per classificazioni alternative siano significativamente diversi tra loro (cioè non differiscano per il solo effetto di eventuali fluttuazioni campionarie). Per risolvere questi problemi, con riferimento al metodo delle *k*-medie, sono applicabili le procedure inferenziali ed i tests di significatività menzionati nel n. 12.1; una trattazione dettagliata di tali metodologie esula tuttavia dagli obiettivi del presente volume e per ulteriori approfondimenti rimandiamo alla bibliografia citata in precedenza.

Secondariamente, ma con importanti risvolti operativi, il numero dei possibili valori di *g* non deve essere troppo elevato, poiché in caso contrario verrebbero meno alcune delle principali prerogative derivanti dall'uso d'un algoritmo non gerarchico: la rapidità dei calcoli e la chiarezza dei risultati. Elaborare ed interpretare una lunga sequenza di partizioni non porterebbe infatti a vantaggi pratici significativi rispetto all'implementazione d'un metodo gerarchico.

Se *n* non è molto grande, una possibile soluzione consiste nel far precedere l'analisi non gerarchica da una di tipo gerarchico, al fine di ottenere un'indicazione preliminare sul valore di *g*. Il taglio del dendrogramma, secondo i criteri illustrati nel n. 9, può infatti fornire utili informazioni sulla struttura dei dati e sul numero di gruppi effettivamente presenti. Una volta stabilito un *range* di valori «ragionevoli»

per g, si può quindi determinare la partizione « ottima » per ciascuno di essi, tramite l'applicazione della procedura non gerarchica prescelta.

Un approccio alternativo è invece rappresentato dall'impiego del criterio esplorativo descritto nel n. 3, che consente di visualizzare le eventuali mode della distribuzione multidimensionale dei dati e, conseguentemente, la presenza ed il numero di gruppi di osservazioni. Tale soluzione appare preferibile alla precedente in quanto più generale ed immediata: essa non soffre infatti delle limitazioni inerenti all'applicazione d'un algoritmo gerarchico, potendo essere utilizzata senza problemi anche per valori elevati di n, e si fonda su una rappresentazione grafica estremamente semplice e comprensibile anche da parte di utenti non esperti.

Osservazione. I risultati ottenuti con il metodo delle k-medie, nella versione standard che adotta la metrica euclidea (5.19), sono soggetti a notevoli distorsioni se nei dati sono presenti più valori anomali (Cerioli, 1998). Un uso attento dell'algoritmo può però consentire l'identificazione di tali *outliers*. Infatti, l'implementazione della procedura con un valore elevato di g fa solitamente sì che le unità non anomale si concentrino in pochi gruppi (quelli più numerosi), mentre gli *outliers* (che hanno caratteristiche molto diverse dalle rimanenti osservazioni) rimangono da soli nella classificazione, formando gruppi isolati costituiti da un unico o eventualmente da pochissimi elementi.

ii) *Scelta dei poli della partizione iniziale*

Una volta determinato il numero di gruppi, occorre individuare i poli che costituiscono i centroidi dei *clusters* nella partizione iniziale. Un criterio molto semplice e poco dispendioso consiste nel prendere come semi iniziali le prime g osservazioni (p-dimensionali) dell'insieme dei dati. Un metodo leggermente più formalizzato conduce invece ad ottenere i poli tramite l'estrazione d'un campione casuale (oppure sistematico) di g unità dalle n che costituiscono il *data set*.

Entrambe le regole indicate non risultano però molto soddisfacenti, poiché non sono in grado di garantire che i semi individuati siano effettivamente rappresentativi dell'intera nuvola di punti nello spazio p-dimensionale. Tale requisito riveste invece notevole importanza nelle applicazioni reali, in quanto consente di migliorare significativamente le proprietà di convergenza degli algoritmi non gerarchici. Esso aumenta infatti la capacità dei metodi di fornire in tempi rapidi una classificazione finale prossima alla soluzione ottima (in termini di omoge-

neità interna dei *clusters*) e di individuare le strutture di gruppo effettivamente presenti nei dati (Milligan, 1996).

La rappresentatività dei poli è abitualmente perseguita mediante la ricerca di osservazioni sufficientemente spaziate in R^p. Ad esempio, la procedura *Fastclus* del SAS effettua un test preliminare sui dati di partenza in modo tale da individuare g elementi la cui distanza reciproca (secondo la metrica euclidea) sia non inferiore ad una soglia prefissata dal ricercatore; in alternativa, il programma esegue una serie di verifiche volte a garantire che la distanza tra i semi prescelti sia comunque superiore a quella calcolata tra i poli stessi e le rimanenti unità. L'obiettivo di tali accorgimenti è quello di far sì che nella partizione iniziale i centroidi risultino ben separati tra loro ed ogni osservazione sia — se possibile — relativamente prossima ad uno di essi.

Osservazione. Tutte le metodologie di selezione dei semi iniziali sopra descritte, ad eccezione di quella fondata sul campionamento casuale, dipendono dall'ordine in cui sono elencate le unità statistiche nel *data set.* Esse soffrono pertanto dell'inconveniente — già illustrato nel n. 12.1 con riferimento al metodo delle k-medie — di condurre a risultati potenzialmente diversi a seconda del modo in cui è formata la lista degli elementi e devono essere applicate con cautela in tutti i casi in cui si vogliono studiare caratteristiche «locali» dei dati, quali l'influenza di ciascuna osservazione o la presenza di *outliers* multivariati (Cerioli, 1998, 1999).

La tecnica esplorativa illustrata nel n. 3 fornisce un semplice criterio alternativo per la scelta dei poli. Dopo avere determinato il numero di gruppi (o un insieme di valori ammissibili per g) attraverso l'esame dei valori assunti dalle densità di frequenza, è infatti possibile definire i semi iniziali come i centroidi calcolati sulle sole osservazioni comprese nei rettangoli che individuano le mode di tale distribuzione.

Similmente, nell'ipotesi di impiego preliminare d'un metodo gerarchico per la specificazione di g, i poli della partizione iniziale possono essere posti uguali ai centroidi dei *clusters* ottenuti dal taglio del corrispondente dendrogramma (27).

---

(27) La *performance* di questo approccio è stata studiata in modo dettagliato da Milligan (1980), con riferimento alle differenti versioni del metodo delle k-medie. Un'ulteriore procedura per l'individuazione «razionale» dei poli è stata proposta da Mineo (1985).

12.3. *Un'applicazione illustrativa*

Consideriamo nuovamente, a titolo illustrativo, il problema della classificazione delle reti televisive sulla base delle prime quattro variabili riportate nella tab. 4.1. Dal momento che $n = 6$, l'applicazione in questo caso d'un algoritmo non gerarchico non costituisce una scelta appropriata da parte del ricercatore; tuttavia, il presente esempio ci consente di delineare con maggiore dettaglio l'impiego del metodo delle k-medie tramite il *package* SPSS. Un'applicazione ad un *data set* di numerosità più elevata sarà invece presa in esame nel successivo n. 13.

In SPSS l'algoritmo delle k-medie è disponibile tramite il menù: *statistica - classificazione - cluster k-medie.* Sullo schermo appare una finestra nella quale inseriamo le variabili da utilizzare nella classificazione. Si noti che, al contrario di quanto avviene nella procedura di *cluster* gerarchica, non compaiono qui le opzioni riferite alla standardizzazione dei caratteri. L'eventuale calcolo degli scostamenti standardizzati deve quindi essere effettuato preventivamente (ad esempio, tramite il menù: *statistica - descrittive - riassumi*, con l'opzione: *salva valori standardizzati come variabili*). Similmente, la distanza adottata è necessariamente quella euclidea.

La finestra associata al metodo delle k-medie richiede la fissazione del numero di gruppi. Nell'esempio, l'applicazione preliminare d'un algoritmo gerarchico (quello del legame completo) ha reso plausibile una partizione costituita da tre oppure quattro gruppi (si veda il n. 9). Esaminiamo in dettaglio la soluzione più sintetica e scegliamo dunque l'opzione:

numero di *cluster*: 3.

Cliccando sul tasto *opzioni* della precedente finestra ed effettuando le corrispondenti selezioni, la procedura di SPSS consente inoltre di rappresentare i centroidi dei gruppi della partizione iniziale, la tabella dell'analisi della varianza su cui si basa il calcolo dell'indice (5.15) ed il *cluster* di appartenenza di ciascuna unità statistica nella classificazione finale. Quest'ultima informazione può anche essere salvata nel *file* dei dati, per eventuali analisi successive, al pari della distanza di ogni osservazione dal centroide del corrispondente gruppo. Infine, in presenza di indicazioni *a priori* sulle caratteristiche della partizione iniziale, è possibile far leggere al programma i vettori p-dimensionali che ne definiscono i poli tramite l'opzione: *centri.*

Una parte dell'output ottenuto con riferimento ai dati delle reti televisive è riportato nelle tabelle 5.12 - 5.14. In particolare, la tab. 5.12

indica la « cronologia » delle iterazioni, cioè il numero di passi che sono stati necessari per raggiungere la convergenza dell'algoritmo. Da essa si rileva che il tempo richiesto per giungere alla classificazione finale è stato estremamente ridotto, poiché i centri dei *clusters* sono rimasti invariati già alla seconda iterazione.

TAB. 5.12. *Cronologia delle iterazioni per la classificazione delle reti televisive tramite la procedura* k-*medie di SPSS* (g=3).

Iterazione	Modifiche ai centri dei cluster		
	1	2	3
1	923.817	0.000	0.000
2	0.000	0.000	0.000

La tab. 5.13 mostra il *cluster* di appartenenza di ciascuna rete nella classificazione finale con tre gruppi, oltre alla distanza (euclidea) dal corrispondente centroide. Nei gruppi formati da una sola unità tale distanza risulta ovviamente uguale a zero, in quanto il centro coincide con l'unico elemento del *cluster*. Le coordinate dei centroidi sono inoltre riportate nella tab. 5.14.

TAB. 5.13. Cluster *di appartenenza e distanza dal corrispondente centroide per ciascuna rete televisiva nella partizione con 3 gruppi ottenuta tramite la procedura* k-*medie di SPSS.*

RETE	Cluster	Distanza
RAIUNO	1	616.049
RAIDUE	1	804.011
RAITRE	1	538.053
RETE4	1	923.817
CANALE5	3	0.000
ITALIA1	2	0.000

Si noti che in questo semplice esempio la classificazione « ottima » con tre gruppi ricavata tramite l'algoritmo delle *k*-medie coincide con quella precedentemente ottenuta attraverso il metodo del legame com-

pleto, con il medesimo numero di gruppi (cfr. il n. 4.2). Tale corrispondenza non vale però in generale e non risulta solitamente verificata nelle applicazioni a *data sets* più complessi.

TAB. 5.14. *Centroidi dei* clusters *finali ottenuti nella classificazione delle reti televisive tramite la procedura* k-*medie di SPSS* (g=3).

	Cluster		
	1	2	3
FILM	1349	1167	582
TELEFILM	1153	3119	1193
VARIETÀ	1071	795	2166
NEWS	1438	1261	3372

La soluzione alternativa costituita da quattro gruppi può essere ottenuta in modo analogo, ripetendo il procedimento con l'opzione:

numero di *cluster*: 4.

13. *Esempio riepilogativo (II)*

Riprendiamo in esame i dati riportati nella tab. 5.1 e riguardanti la *performance* di 103 fondi comuni d'investimento ($p = 3$). Dall'indagine preliminare condotta nel n. 3 è emersa la bimodalità della corrispondente distribuzione e conseguentemente l'esistenza di due gruppi ben separati. Cerchiamo quindi di delineare le caratteristiche di tali *clusters* tramite l'applicazione d'un algoritmo non gerarchico che consenta di determinare la partizione «ottima» con $g = 2$. A tale scopo applichiamo l'algoritmo delle *k*-medie, dapprima tramite SPSS e poi con SAS.

Ad uno stadio preliminare rispetto all'analisi vera e propria, si pone il quesito se standardizzare o meno le variabili. I tre caratteri considerati sono in questo caso espressi in percentuale, tramite numeri puri, e sono quindi direttamente confrontabili tra loro: il calcolo delle distanze (5.19) assume quindi significato anche se effettuato sui valori originari. La standardizzazione avrebbe l'ulteriore conseguenza di uniformare la media e lo scostamento quadratico medio di ciascuna variabile, eliminando così l'effetto della diversa variabilità e del differente ordine di grandezza. Nel presente esempio, finalizzato alla valutazione della *per-*

formance dei singoli fondi, appare tuttavia ragionevole attribuire — seppure implicitamente — un'importanza maggiore ai due indicatori di medio periodo (X_2 e X_3), che sono quelli con il valor medio più elevato. Pertanto, nelle elaborazioni seguenti non è stata effettuata alcuna standardizzazione e gli algoritmi classificatori sono stati applicati ai dati originari della tab. 5.1.

Riportiamo inizialmente l'output di SPSS. Le tabelle 5.15 e 5.16 forniscono rispettivamente i centroidi dei *clusters* iniziali, calcolati automaticamente dal programma secondo le metodologie illustrate nel n. 12.2, e la « cronologia » delle iterazioni necessarie per giungere alla partizione « ottima » con due gruppi. Anche in questa circostanza la convergenza dell'algoritmo è stata sufficientemente rapida e non è stato necessario alcun intervento volto ad interrompere anzitempo la procedura iterativa.

La classificazione finale è costituita da due gruppi di numerosità pari a 57 e 46 unità (si veda la tab. 5.17); le coordinate dei corrispondenti centroidi sono riportate nella tab. 5.18. Quest'ultima consente d'individuare agevolmente le peculiarità dei fondi appartenenti ai singoli *clusters*: il primo gruppo risulta infatti caratterizzato da un elevato rendimento — sia a breve sia a medio termine — e da un'alta volatilità, mentre il secondo mostra prestazioni più modeste — soprattutto nel medio periodo — ma anche una volatilità contenuta. La distanza (euclidea) tra tali centroidi è fornita anch'essa nella tab. 5.17.

L'indicazione del *cluster* di appartenenza (non riportata nel testo per esigenze di spazio) consente di identificare gli elementi effettivamente classificati in ciascun gruppo. In particolare, quasi tutti i fondi azionari sono assegnati al gruppo 1, mentre quelli bilanciati sono attribuiti quasi esclusivamente al gruppo 2. Si noti che soltanto in tre casi la classificazione ottenuta non rispecchia la tipologia della corrispondente unità statistica: si tratta del fondo « Venetoventure » (azionario ma assegnato al *cluster* 2) e dei fondi « Fondo alto bil. » e « Ing portfolio » (bilanciati, ma attribuiti al *cluster* 1). Tali unità possono dunque essere considerate « atipiche » dal punto di vista della *performance*, in quanto presentano caratteristiche più simili a quelle dei fondi con una tipologia differente dalla propria.

Per quanto concerne la bontà della partizione ottenuta (con $g = 2$), SPSS fornisce la tabella dell'analisi della varianza su cui si fonda il calcolo dell'indice (5.15). Informazioni più dettagliate possono invece essere ottenute tramite la procedura *Fastclus* del programma SAS, il cui output è sintetizzato nella tab. 5.19. In particolare, essa contiene nell'ultima riga i valori delle seguenti quantità (si

vedano le formule (5.9) e (5.10)), calcolate sul complesso di tutte le variabili (28):

$$Total\ STD = \sqrt{\frac{T}{p(n-1)}}$$

e

$$Within\ STD = \sqrt{\frac{W}{p(n-g)}}$$

nonché del coefficiente R^2 definito nella (5.15). Nelle righe precedenti sono invece presentate le rispettive componenti riferite ai singoli caratteri.

TAB. 5.15. *Centroidi dei* clusters *iniziali nella classificazione dei fondi d'investimento tramite la procedura* k-*medie di SPSS* (g=2).

	Cluster	
	1	2
Performance % (12 mesi)	24.60	-0.60
Rendim. med. ann. (3 anni)	49.10	10.80
Volatilità (3 anni)	22.00	9.10

TAB. 5.16. *Cronologia delle iterazioni per la classificazione dei fondi d'investimento tramite la procedura* k-*medie di SPSS* (g=2).

	Modifiche ai centri dei cluster	
Iterazione	1	2
1	19.711	14.164
2	1.508	2.537
3	0.176	0.280
4	0.247	0.313
5	0.000	0.000

(28) Il termine *STD* è un acronimo di *standard deviation*. I denominatori di T e W sono determinati in base ai rispettivi «gradi di libertà» nella teoria dell'analisi della varianza multivariata (Seber, 1984, cap. 9).

TAB. 5.17. *Numero di elementi in ciascuno dei* clusters *ottenuti nella classificazione dei fondi d'investimento tramite la procedura* k-*medie di SPSS* (g=2) *e distanze tra i rispettivi centroidi.*

	Numero di elementi	Distanza tra i centroidi di gruppo
Cluster 1	57	17.015
Cluster 2	46	17.015

TAB. 5.18. *Centroidi dei* clusters *finali ottenuti nella classificazione dei fondi d'investimento tramite la procedura* k-*medie di SPSS* (g=2).

	Cluster	
	1	2
Performance % (12 mesi)	12.2491	9.1609
Rendim. med. ann. (3 anni)	31.4807	18.1000
Volatilità (3 anni)	21.3596	11.3130

TAB. 5.19. *Statistiche delle variabili ed indici di bontà della partizione ottenuta nella classificazione dei fondi d'investimento tramite la procedura* Fastclus *di SAS* (g=2).

	Total STD	*Within STD*	R^2
Performance % (12 mesi)	4.0373	3.7493	0.1460
Rend. med. ann. (3 anni)	7.7760	3.9923	0.7390
Volatilità (3 anni)	5.3502	1.8625	0.8800
Totale	5.9271	3.3399	0.6856

La tab. 5.19 rivela una suddivisione in gruppi abbastanza soddisfacente (R^2 prossimo al 70%), anche se quasi 1/3 della devianza totale dei tre indicatori appare «non spiegato» dalla classificazione ottenuta. I due *clusters*, pur essendo dotati d'una certa coesione, non risultano quindi completamente omogenei al proprio interno, soprattutto per quanto concerne la *performance* a 12 mesi. Questo indicatore presenta infatti una devianza nei gruppi assai accentuata, cui corrisponde un valore dell'indice R^2 inferiore al 15%.

Poiché i gruppi sono costituiti da fondi quasi esclusivamente del medesimo tipo, si può comunque affermare che tale eterogeneità è in-

trinseca alle caratteristiche dei fondi considerati (azionari e bilanciati) e non è imputabile al metodo di classificazione adottato. Per aumentare il valore dell'indice R^2 occorrerebbe pertanto ripetere l'analisi con un numero di gruppi superiore ($g > 2$), in modo da scindere ciascuna tipologia di fondi in sottogruppi più omogenei.

In conclusione, nell'esempio proposto l'algoritmo delle k-medie è stato in grado d'individuare correttamente la struttura di gruppo presente nei dati e già indicata in via preliminare dalle figure 5.1-5.3, separando quasi perfettamente i fondi azionari da quelli bilanciati. In aggiunta, l'applicazione di tale metodo ha consentito di valutare quantitativamente le caratteristiche di ciascun gruppo, attraverso il calcolo del rispettivo centroide e delle statistiche di coesione interna, riferite sia ai singoli indicatori sia al loro insieme.

14. *Cenni ad altri metodi di classificazione*

I metodi di classificazione illustrati in precedenza conducevano ad individuare una partizione o una gerarchia di partizioni, per cui un'unità era assegnata ad uno ed un solo gruppo. In alcuni problemi concreti, tuttavia, si possono riscontrare unità statistiche che presentano caratteristiche intermedie tra quelle di due (o più) gruppi, per cui l'assegnazione ad uno ed uno solo di essi non rispecchia fedelmente la struttura dei dati. Ad esempio, con riferimento alla classificazione delle autovetture, si pensi ad un modello che per le sue peculiarità si collochi tra due segmenti del mercato. Una prima soluzione del problema può essere quella di creare nuovi gruppi (nell'esempio dei sottosegmenti del mercato), ciascuno costituito da una (o poche) unità. Questo criterio può condurre però ad incrementare notevolmente il numero dei gruppi, facendo perdere la visione di sintesi, che è uno degli obiettivi della *cluster analysis*.

Un approccio alternativo consiste nel proporre classificazioni non più basate su partizioni, e precisamente:

i) metodi che generano gruppi parzialmente sovrapposti (*overlapping classification*);

ii) metodi di classificazione sfocata, che si basano sulla teoria dei *fuzzy sets*.

Un altro problema riguarda l'introduzione di vincoli nella classificazione. In tal caso, date n unità, solo un sottoinsieme delle possibili partizioni costituisce lo spazio delle soluzioni ammissibili (ad esempio,

quelle in cui il numero massimo di elementi inseriti in un generico gruppo non supera una certa soglia).

In questo paragrafo faremo un breve cenno a detti metodi di classificazione, che possono suggerire al lettore alcune linee di approfondimento del tema.

14.1. *Classificazioni con sovrapposizioni*

Si dicono metodi di *clumping* quelli che originano classificazioni d'un insieme d'unità in gruppi non disgiunti (tali cioè che un elemento possa appartenere contemporaneamente a due o più gruppi) (Gordon, 1981, p. 54). Ad essi è stata dedicata in letteratura un'attenzione molto minore rispetto a quella dei metodi che individuano partizioni. Le ragioni sono da ricercarsi nelle incertezze alle quali può dare luogo l'attribuzione d'una medesima unità a due o più gruppi. Se le sovrapposizioni sono eccessive, l'interpretazione dei risultati è oscura ed al limite può rendere inutile la classificazione proposta. Occorre dunque introdurre opportuni vincoli per limitare le sovrapposizioni consentite. Ad esempio, nel cosiddetto metodo B_k ($k = 1, 2, 3, \ldots$) di Jardine and Sibson (1971, p. 65), il numero massimo di unità che possono sovrapporsi in ciascuna coppia di gruppi è uguale a $(k - 1)$; se il numero di unità in comune è maggiore, i due gruppi sono fusi in uno solo. (Per $k = 1$ tale metodo coincide con il legame singolo).

I metodi di *clumping* possono essere d'un certo interesse nella classificazione di zone d'un territorio: le unità che appartengono contemporaneamente a due gruppi rappresentano gli elementi «di cerniera» tra le aree omogenee individuate, con caratteristiche intermedie a quelle dei gruppi corrispondenti (Zani, 1993).

14.2. *Classificazioni sfocate*

In un insieme sfocato (*fuzzy set*) l'appartenenza d'un elemento all'insieme stesso è definita da una funzione che assume valori (detti «gradi di appartenenza») nell'intervallo [0, 1]. Un valore uguale a 1 della funzione di appartenenza indica che l'elemento appartiene pienamente all'insieme, un valore uguale a 0 che non vi appartiene ed un valore intermedio che esso vi appartiene «parzialmente» (29).

---

(29) La teoria dei *fuzzy sets* è stato introdotta nel 1965 da L.A. Zadeh. Per una

Nei problemi di classificazione la teoria dei *fuzzy sets* è utile quando il problema in esame suggerisce come più ragionevole e naturale una suddivisione delle n unità in maniera tale che almeno alcune di esse possano appartenere «in parte» a due (o più) gruppi. Si pensi, ad esempio, ad un'indagine di mercato sulle intenzioni di acquisto d'un certo bene durevole entro l'anno successivo, condotta mediante una serie di domande indirette. In base alle risposte ottenute si potranno individuare il gruppo dei potenziali acquirenti, il gruppo di coloro che certamente non acquisteranno, ma anche un gruppo intermedio di incerti, che presentano, in quote che possono variare da individuo a individuo, caratteristiche in parte dell'uno ed in parte dell'altro dei due gruppi precedenti.

Nella classificazione sfocata si assume che due unità siano tra loro tanto più simili quanto più prossimo ad uno è il valore della loro funzione di appartenenza al medesimo gruppo (Zadeh, 1977). In questi metodi assume un ruolo centrale il concetto di g-partizione sfocata.

Definizione. Si dice g-partizione sfocata di n unità statistiche la seguente matrice, di dimensioni $n \times g$:

$$\mathbf{U} = [u_{il}]$$

ove u_{il} è il valore della funzione di appartenenza dell'unità i-esima al gruppo l-esimo ($i = 1, \ldots n$; $l = 1, \ldots g$), soggetto ai vincoli:

$$0 \leq u_{il} \leq 1$$

$$\sum_{l=1}^{g} u_{il} = 1.$$

Gli algoritmi di *fuzzy clustering* si propongono di ottenere, in base alle p variabili rilevate, una g-partizione sfocata che ottimizzi un certo criterio (metodi non gerarchici) oppure una successione di dette partizioni (Bezdek, 1981; Zani, 1988; Trauwaert *et al.*, 1991).

«*Some fuzzy clustering are more fuzzy than others*» affermano Kaufman and Rousseeuw (1990), per indicare che il grado di sfocatura può essere più o meno accentuato. Per ottenere una valutazione quan-

---

trattazione con specifico riferimento al suo impiego in statistica rinviamo ai volumi di Zimmermann (1985), Bandemer and Näther (1992).

titativa si utilizza abitualmente il seguente indice, chiamato *partition coefficient*:

$$F = \sum_{i=1}^{n} \sum_{l=1}^{g} (u_{il})^2 / n$$

che assume valori nell'intervallo $[1/g, 1]$ e risulta uguale al minimo quando ciascuna unità presenta distribuzione uniforme del grado di appartenenza ai diversi gruppi, e uguale a 1 in presenza d'una partizione propria (cioè non sfocata). Dato che l'interpretazione dei risultati diventa via via meno chiara al crescere del numero dei gruppi ed al crescere del grado di sfocatura dei medesimi, s'introducono abitualmente dei vincoli sulle soluzioni ammissibili, analogamente a quanto detto per le classificazioni con sovrapposizioni.

14.3. *Classificazioni vincolate*

Col termine « classificazione vincolata » si indicano i metodi per ottenere gruppi di unità che risultino non soltanto simili tra loro, con riguardo alle variabili rilevate, ma anche che soddisfino ulteriori condizioni imposte dal ricercatore. L'introduzione di vincoli rappresenta un modo per restringere le soluzioni ammissibili a quelle che presentano significato nel concreto problema in esame (Gordon, 1996b; 1999, pp. 115-121).

Un tipo di vincolo che assume grande rilievo nell'analisi di dati territoriali o spaziali è quello di contiguità (in un senso da definire) tra le unità assegnate ad ogni gruppo. Detto vincolo appare essenziale in molti problemi di analisi del territorio, in cui le unità (comuni d'una regione, quartieri d'una città, ...) d'un gruppo devono essere non solo simili, ma anche tra loro vicine per poter generare « zone omogenee » di concreto interesse.

La classificazione di dati spaziali (come ed esempio i *pixel* delle immagini telerilevate) rappresenta uno dei campi più attuali della ricerca metodologica ed applicata con riferimento alla *cluster analysis*, ma per il suo carattere specialistico non può trovare posto in questo testo. Per un'introduzione all'argomento rinviamo a Zani (1993).

RIFERIMENTI BIBLIOGRAFICI

ANDERBERG, M. R. (1973), *Cluster Analysis for Applications*, Academic Press, New York.

ARABIE, P., HUBERT, L. J. and DE SOETE, G. (eds.) (1996), *Clustering and Classification*, World Scientific, Singapore.

BANDEMER, H. and NÄTHER, W. (1992), *Fuzzy Data Analysis*, Kluwer, Dordrecht.

BEZDEK, J. C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

BOCK, H. H. (1985), On Some Significance Tests in Cluster Analysis, *Journal of Classification*, vol. 2, pp. 77-108.

BOCK, H. H. (1996a), Probability Models and Hypothesis Testing in Partitioning Cluster Analysis, in: ARABIE, P. *et al.* (eds.), *Clustering and Classification*, World Scientific, Singapore, pp. 377-453.

BOCK, H. H. (1996b), Probabilistic models in cluster analysis, *Computational Statistics and Data Analysis*, vol. 23, pp. 5-28.

CASTAGNOLI, E. (1978), Un'osservazione sull'analisi classificatoria, in: *Due temi di analisi statistica multivariata*, Facoltà di Scienze statistiche, demografiche ed attuariali, CLEUP, Padova.

CERIOLI, A. (1997), Comparing Three Partitions: An Inferential Approach Based on Multi-way Contingency Tables, *Communications in Statistics — Theory and Methods*, vol. 26, pp. 2457-2471.

CERIOLI, A. (1998), A New Method for Detecting Influential Observations in Nonhierarchical Cluster Analysis, in: RIZZI, A., VICHI, M. and BOCK, H-H. (eds.), *Advances in Data Science and Classification*, Springer-Verlag, Berlin, pp. 15-20.

CERIOLI, A. (1999), Measuring the Influence of Individual Observations and Variables in Cluster Analysis, in: VICHI, M. and OPITZ, O. (eds.), *Classification and Data Analysis*, Springer-Verlag, Berlin, pp. 3-10.

CERIOLI, A. e ZANI, S. (1999), Exploratory Methods for Detecting High Density Regions in Cluster Analysis, in: *CLADAG99, Book of Short Papers*, Roma, 1999, pp. 193-196.

CHENG, R. and MILLIGAN, G. W. (1996), Measuring the Influence of Individual Data Points in a Cluster Analysis, *Journal of Classification*, vol. 13, pp. 315-335.

CORMACK, R. M. (1971), A Review of Classification, *Journal of the Royal Statistical Society*, series A, vol. 154, pp. 321-353.

CUESTA-ALBERTOS, J. A., GORDALIZA, A. and MATRAN, C. (1997), Trimmed *k*-means: an attempt to robustify quantizers, *The Annals of Statistics*, vol. 25, pp. 553-576.

EDDY, W. F., MOCKUS, A. and OUE, S. (1996), Approximate single linkage cluster analysis of large data sets in high-dimensional spaces, *Computational Statistics and Data Analysis*, vol. 23, pp. 29-43.

EVERITT, B. S. (1993), *Cluster Analysis*, 3rd edition, Edward Arnold, London.

FISHER, L. and VAN NESS, J. W. (1971), Admissible clustering procedures, *Biometrika*, vol. 58, pp. 91-104.

FOWLKES, E. B., GNANADESIKAN, R. and KETTENRING, J. R. (1988), Variable selection in clustering, *Journal of Classification*, vol. 5, pp. 205-228.

FOWLKES, E. B. and MALLOWS, C. L. (1983), A Method for Comparing Two Hierarchical Clusterings, *Journal of the American Statistical Association*, vol. 78, pp. 553-584.

GNANADESIKAN, R., KETTENRING, J. R. and LANDWHER, J. M. (1977), Interpreting and Assessing the Results of Cluster Analyses, *Bulletin of the International Statistical Institute*, vol. 47, pp. 451-463.

GNANADESIKAN, R., KETTENRING, J. R. and TSAO, S. L. (1995), Weighting and selection of variables for cluster analysis, *Journal of Classification*, vol. 12, pp. 113-136.

GORDON, A. D. (1981), *Classification*, Chapman and Hall, London. 2nd edition, 1999.

GORDON, A. D. (1996a), Hierarchical Classification, in: ARABIE, P. *et al.* (eds.), *Clustering and Classification*, World Scientific, Singapore, pp. 65- 121.

GORDON, A. D. (1996b), A survey of constrained classification, *Computational Statistics and Data Analysis*, vol. 21, pp. 17-29.

GORDON, A. D. and DE CATA, A. (1988), Stability and influence in sum of squares clustering, *Metron*, vol. 46, pp. 347-360.

GOWER, J. C. and ROSS, G. J. S. (1969), Minimum spanning trees and single linkage cluster analysis, *Applied Statistics*, vol. 18, pp. 54-64.

HAND, D. J. (1997), *Construction and Assessment of Classification Rules*, Wiley, Chichester.

HARDY, A. (1996), On the number of clusters, *Computational Statistics and Data Analysis*, vol. 23, pp. 83-96.

HARTIGAN, J. A. (1975), *Clustering Algorithms*, Wiley, New York.

HARTIGAN, J. A. (1978), Asymptotic distributions for clustering criteria, *The Annals of Statistics*, vol. 6, pp. 117-131.

HARTIGAN, J. A. (1985), Statistical Theory in Clustering, *Journal of Classification*, vol. 2, pp. 63-76.

HRUSCHKA, H. and NATTER, M. (1999), Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation, *European Journal of Operational Research*, vol. 114, pp. 346-353.

HUBERT, L. and ARABIE, P. (1985), Comparing Partitions, *Journal of Classification*, vol. 2, pp. 193-218.

HYNDMAN, R. J. (1996), Computing and Graphing Highest Density Regions, *The American Statistician*, vol. 50, pp. 120-126.

JAIN, A. K. and DUBES, R. (1988), *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs.

JAMBU, M. et LEBEAUX, M. O. (1983), *Cluster Analysis and Data Analysis*, North Holland, Amsterdam.

JARDINE, N. and SIBSON, R. (1971), *Mathematical Taxonomy*, Wiley, London.
JOLLIFFE, I. T., JONES, B. and MORGAN, B. J. T. (1995), Identifying influential observations in hierarchical cluster analysis, *Journal of Applied Statistics*, vol. 22, pp. 61-80.
KAUFMAN, L. and ROUSSEEUW, P. J. (1990), *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York.
KENDALL, M. G. (1966), Discrimination and Classification, in: KRISHNAIAH, P. R. (ed.), *Multivariate Analysis*, Academic Press, New York, pp. 165-185.
LANCE, G. N. and WILLIAMS, W. T. (1967), A general theory of classificatory sorting strategies: 1. Hierarchical systems, *Computer Journal*, vol. 9, pp. 373-380.
LAPOINTE, F. J. (1998), For Consensus (With Branch Lengths), in: RIZZI, A., VICHI, M. and BOCK H-H. (eds.), *Advances in Data Science and Classification*, Springer-Verlag, Berlin, pp. 73-80.
MCLACHLAN, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.
MILLIGAN, G. W. (1980), An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, *Psychometrika*, vol. 45, pp. 325-342.
MILLIGAN, G. W. (1996), Clustering Validation: Results and Implications for Applied Analysis, in: ARABIE, P. *et al.* (eds.), *Clustering and Classification*, World Scientific, Singapore, pp. 341- 375.
MILLIGAN, G. W. and COOPER, M. C. (1985), An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, vol. 50, pp. 159-179.
MINEO, A. (1985), A new criterion for the choice of seed points for a nearest centroid cluster analysis, *Rivista di Statistica Applicata*, vol. 18, pp. 191-198.
MINNOTTE, M. C., MARCHETTE, D. J. and WEGMAN, E. J. (1998), The Bumpy Road to the Mode Forest, *Journal of Computational and Graphical Statistics*, vol. 7, pp. 239-251.
MIRKIN, B. (1996), *Mathematical Classification and Clustering*, Kluwer, Dordrecht.
PISON, G., STRUYF, A. and ROUSSEEUW, P. J. (1999), Displaying a clustering with CLUSPLOT, *Computational Statistics and Data Analysis*, vol. 30, pp. 381-392.
POLLARD, D. (1981), Strong consistency of *k*-means clustering, *The Annals of Statistics*, vol. 9, pp. 135-140.
RAND, W. M. (1971), Objective Criteria for the Evaluation of Clustering Methods, *Journal of the American Statistical Association*, vol. 66, pp. 846-850.
SCOTT, D. W. (1992), *Multivariate Density Estimation*, Wiley, New York.
SEBER, G. A. F. (1984), *Multivariate Observations*, Wiley, New York.
SILVERMAN, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
SPÄTH, H. (1980), *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Ellis Horwood, Chichester.
SPÄTH, H. (1985), *Cluster Dissection and Analysis*, Ellis Horwood, Chichester.

TRAUWAERT, E., KAUFMAN, L. and ROUSSEEUW, P. (1991), Fuzzy clustering algorithms based on maximum likelihood principle, *Fuzzy Sets and Systems*, vol. 42, pp. 213-227.

WAND, M. P. (1997), Data-Based Choice of Histogram Bin Width, *The American Statistician*, vol. 51, pp. 59-64; correction, vol. 53, p. 174.

ZADEH, L. A. (1977), Fuzzy sets and their application to pattern classification and clustering, in: VAN RYZIN, J. (ed.), *Classification and Clustering*, Academic Press, New York.

ZANI, S. (1988), Un metodo di classificazione sfocata, in: DIANA, G., PROVASI, C. e VEDALDI, R. *Metodi statistici per la tecnologia e l'analisi dei dati multidimensionali*, Università degli Studi di Padova, pp. 281-288.

ZANI, S. (1993), Classificazione di unità territoriali e spaziali, in: ZANI, S., a cura di, *Metodi statistici per le analisi territoriali*, Franco Angeli, Milano, pp. 93-121.

ZIMMERMANN, H. J. (1985), *Fuzzy Sets Theory and its Applications*, Kluwer, Dordrecht.