La cluster analysis

1. Introduzione

La cluster analysis consiste in un insieme di tecniche statistiche atte ad individuare gruppi di unità tra loro simili rispetto ad un insieme di caratteri presi in considerazione, e secondo uno specifico criterio. L'obiettivo che ci si pone è sostanzialmente quello di riunire unità tra loro eterogenee in più sottoinsiemi tendenzialmente omogenei e mutuamente esaustivi. Le unità statistiche vengono, in altri termini, suddivise in un certo numero di gruppi a seconda del loro livello di "somiglianza" valutata a partire dai valori che una serie di variabili prescelte assume in ciascuna unità. La *cluster analysis* consente allora di pervenire ai seguenti risultati (Fabbris, 1989) :

- *la generazione di ipotesi di ricerca*, infatti per effettuare una analisi di raggruppamento non è necessario avere in mente alcun modello interpretativo;
- *la riduzione dei dati* in forma (anche grafica) tale da rendere facile la lettura delle informazioni rilevate e parsimoniosa la presentazione dei risultati;
- *ricerca tipologica* per individuare gruppi di unità statistiche con caratteristiche distintive che facciano risaltare la fisionomia del sistema osservato;
- *la costruzioni di sistemi di classificazione automatica* (Jardine e Sibson, 1971);
- *la ricerca di classi omogenee*, dentro le quali si può supporre che i membri siano mutuamente surrogabili (Green et al., 1967).

Vale la pena sottolineare soprattutto il punto 1. Infatti, la *cluster analysis*, a differenza di altre tecniche statistiche multivariate (ad esempio, l'analisi discriminante, che rende possibile la ripartizione di un insieme di individui in gruppi, predeterminati fin dall'inizio della ricerca in base alle diverse modalità assunte da uno o più caratteri), non compie alcuna assunzione "a priori" sulle tipologie fondamentali esistenti che possono caratterizzare il collettivo studiato. In questo caso la tecnica ha un ruolo esplorativo di ricerca di strutture latenti, al fine di desumere la partizione più probabile. La *cluster analysis* è infatti un metodo puramente empirico di classificazione, e come tale, in primo luogo, una tecnica induttiva.
Operativamente, in ambito economico-aziendale, la cluster analysis può essere utilizzata per l'identificazione di gruppi di:

- consumatori o utenti di un certo servizio pubblico sulla base di:
  - comportamento al consumo/utilizzo
  - opinioni sul prodotto/servizio
  - importanza assegnata a varie caratteristiche di un prodotto/servizio (segmentazione del mercato)
- strutture di servizi secondo varie caratteristiche che ne definiscono l'efficienza
- marche di un certo prodotto secondo varie caratteristiche
- aziende secondo caratteristiche legate ai rapporti con l'estero

2. Le fasi dell'applicazione

L'applicazione della cluster analysis si articola in alcune fasi.

1. innanzitutto occorre effettuare la **scelta delle variabili di classificazione**: delle unità osservate. La scelta delle variabili rispecchia essenzialmente le convinzioni e le idee del ricercatore, ed è una operazione che implica un grado molto alto di soggettività: può capitare di non considerare variabili fortemente selettive ed avere quindi una partizione in gruppi "sbagliata"; d'altra parte, l'inclusione di variabili dotate di una elevata capacità discriminante, ma non rilevanti ai fini dell'indagine, può portare a risultati di scarso rilievo pratico. E' da precisare che generalmente le variabili da utilizzare debbono essere espresse nella stessa unità di misura. Se le variabili quantitative da utilizzare nella cluster sono espresse in unità di misura diverse o hanno ordini di grandezza diversi è opportuno standardizzare le variabili.

2. Fissate le variabili, il passo successivo è la **scelta di una misura della dissomiglianza esistente fra le unità statistiche**. I caratteri rilevati possono essere espressi in quattro distinte scale di misura: nominali, ordinali, per intervalli e per rapporti. I caratteri qualitativi possono essere misurati solo con riferimento alle prime due, mentre le variabili ammettono scale di qualunque tipo. Nel caso di caratteri quantitativi possono essere utilizzati vari tipi di indici di distanza (Hartigan, 1975) di cui i più utilizzati sono:
   (i) la distanza euclidea, che corrisponde al concetto geometrico di distanza nello spazio multidimensionale:
   $$_2d_{hk} = \left\{ \sum_{v=1}^{p} w_v (x_{hv} - x_{kv})^2 \right\}^{1/2}$$
   dove x_{hv}, x_{kv} sono le coordinate dei due punti P_h e P_k nello spazio cartesiano sulla variabile x_v e w_v è il peso attribuito alla variabile
   (ii) il quadrato della distanza euclidea qualora si voglia dare un peso progressivamente maggiore agli oggetti che stanno oltre una certa distanza;
   (iii) la distanza assoluta (o city-block o distanza di Manhattan) è semplicemente la differenza media fra le dimensioni:
   $$_1d_{hk} = \sum | x_{hv} - x_{kv} | w_v$$
   conigliata in generale quando le variabili di classificazione sono su scala ordinale;
   (iv) la distanza di Chebychev può essere appropriata nei casi in cui si voglia definire due oggetti come "differenti" se essi sono diversi in ciascuna delle dimensioni:
   $$_c d_{hk} = \max | x_{hv} - x_{kv} |;$$
   (v) la distanza di Mahalanobis è quella che invece tiene conto anche delle interdipendenze esistenti tra le p variabili utilizzate ridimensionando il peso delle variabili portatrici di informazioni eccedenti, già fornite da altre. Quando le variabili originarie sono correlate tra loro è improprio utilizzare la distanza euclidea, mentre è pertinente l'uso della statistica proposta da Mahalanobis data dalla forma quadratica:
   $$D^2_{hk} = (x_h - x_k)' W^{-1} (x_h - x_k) \quad \text{con } h \neq k = 1,...,n$$

dove X_h e X_k sono i vettori con le osservazioni sugli individui h e k e W è la matrice di varianze-covarianze tra le variabili osservate. Occorre fare attenzione nell'uso di questa distanza, infatti se sussiste collinearità tra le variabili, la matrice non è invertibile, e anche se, pur non essendovi collinearità, è presente una forte intercorrelazione, errori di misura o di calcolo possono condurre a pesanti distorsioni nei risultati.

3. Una volta scelta la misura di dissomiglianza, si pone il problema di procedere alla **scelta di un idoneo algoritmo di raggruppamento** delle unità osservate. La distinzione che normalmente viene proposta è fra
   - *metodi gerarchici* che conducono ad un insieme di gruppi ordinabili secondo livelli crescenti, con un numero di gruppi da n ad 1;
   - *metodi non gerarchici*. forniscono un'unica partizione delle n unità in g gruppi, e g deve essere specificato a priori.

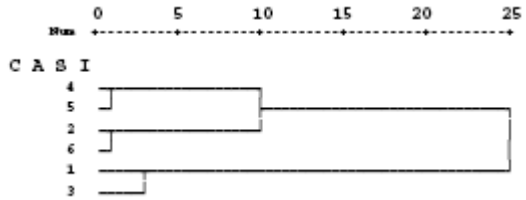
4. Valutazione della partizione ottenuta e scelta del numero ottimale di gruppi.

5. Interpretazione dei risultati ottenuti (connotazione dei gruppi).

3 Algoritmi di raggruppamento

3.1 Metodi gerarchici

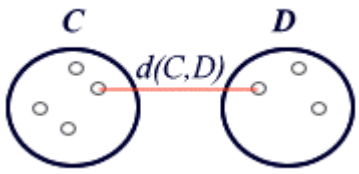
I metodi gerarchici (Johnson, 1967; Everitt 1979) si affiancano ad una situazione in cui si hanno n grappoli di una sola unità per giungere, attraverso successive fusioni dei grappoli meno distanti tra di loro, ad una situazione in cui si ha un solo grappolo che contiene tutte le n unità. Il prodotto finale dei metodi gerarchici non è, quindi, una singola partizione delle n unità, ma una serie di partizioni che possono essere rappresentate graficamente attraverso un "dendogramma" o "diagramma ad albero" nel quale sull'asse delle ordinate viene riportato il livello di distanza, mentre sull'asse delle ascisse vengono riportate le singole unità. Ogni ramo del diagramma (linea verticale) corrisponde ad un grappolo. La linea di congiunzione (orizzontale) di due o più rami individua il livello di distanza al quale i grappoli si fondono. I metodi gerarchici si distinguono per il modo in cui, dopo la p-esima fusione, vengono calcolate le distanze tra il nuovo grappolo ed i rimanenti. Gli algoritmi gerarchici proposti in letteratura (metodo del legame singolo, metodo del legame completo, metodo del legame medio, metodo del centroide, metodo di Ward, solo per ricordarne alcuni) si differenziano unicamente per il diverso criterio che regola la valutazione delle distanze tra i gruppi ai fini delle aggregazioni in serie.



1. **Metodo del legame singolo**, [Single-Linkage] - Vicino più prossimo [Nearest Neighbor]: in questo caso la distanza tra i gruppi è posta pari alla più piccola delle

distanze istituibili a due a due tra tutti gli elementi dei due gruppi: se C e D sono due gruppi, la loro distanza, secondo questo metodo, è definita come la più piccola (il minimo) tra tutte le $n_1 n_2$ distanze che si possono calcolare tra ciascuna unità i di C e ciascuna unità j di D:

$$d(C,D) = \min(d_{ij}),\ \forall i \in C, \forall j \in D.$$

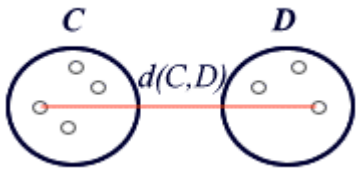


L'adozione di questo algoritmo per la composizione dei gruppi evidenzia in maniera netta e decisamente più accentuata rispetto agli altri due algoritmi tutte le similitudini e somiglianze tra gli elementi: privilegia la differenza tra i gruppi piuttosto che l'omogeneità degli elementi di ogni gruppo. il dendrogramma costruito su questa matrice ha i rami molto più corti ed è più compatto: questo proprio perché vengono valorizzate le somiglianze

2. **Metodo del legame completo**, [Complete-Linkage] - Vicino più lontano [Furthest Neighbor]: secondo questo metodo si considera la maggiore delle distanze istituibili a due a due tra tutti gli elementi dei due gruppi: avremo quindi:

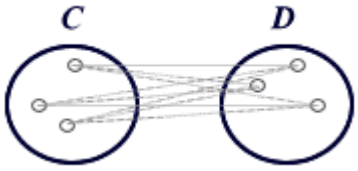
$$d(C,D) = \max(d_{ij})\ \ \forall i \in C, \forall j \in D,$$

evidentemente si uniscono i due gruppi che presentano la più piccola distanza così definita.



Questo algoritmo di aggregazione evidenzia in maniera netta le differenze tra elementi: privilegia l'omogeneità tra gli elementi del gruppo a scapito della differenziazione netta tra gruppi. il dendrogramma costruito su questa matrice ha i rami molto più lunghi, i gruppi (e soprattutto i rami) si formano a distanze maggiori. in uno stesso range di valori, rispetto al legame singolo, gli elementi sono molto meno compatti e più diluiti.

3. **Metodo del legame medio**, [Average-Linkage]: si tratta del valore medio aritmetico di tutte le distanze tra gli elementi;



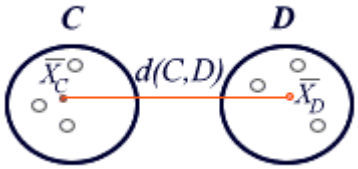
$$d(C,D) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=2}^{n_2} d_{ij} \ , \ \forall i \in C, \forall j \in D \ ,$$

Si uniscono i due gruppi che presentano la più piccola distanza così definita

L'adozione di questo algoritmo per la composizione dei gruppi semplifica notevolmente la composizione dell'albero costruito con l'algoritmo completo, mente rispetto a quello costruito sull'algoritmo singolo rappresenta una movimentazione e differenziazione. Essendo basato sulla media delle distanze, i risultati sono più attendibili e i gruppi risultano più omogenei e ben differenziati tra di loro.

4. **Metodo del centroide**, vanno determinati i vettori contenenti i valori medi delle *p* variabili in tutti gruppi (centroidi), e le distanze tra i gruppi viene assunta pari alla distanza tra i rispettivi centroidi. Se $\overline{X}_C$ e $\overline{X}_D$ sono i centroidi avremo:

$$d(C,D) = d(\overline{X}_C, \overline{X}_D)$$



5. **Metodo di Ward** differisce in parte dai precedenti, in quanto suggerisce di riunire, ad ogni tappa del processo, i due gruppi dalla cui fusione deriva il minimo incremento possibile della devianza "entro".

$$DEV_T = \sum_{s=1}^{p} \sum_{i=1}^{n} (x_{is} - \overline{x}_s)^2 = \sum_{i=1}^{n} \sum_{s=1}^{p} (x_{is} - \overline{x}_s)^2$$

dove $\overline{x}_s$ è la media della variabile s con riferimento all'intero collettivo. Data una partizione in g gruppi, tale devianza può essere scomposta in:

$$DEV_{IN} = \sum_{k=1}^{g} \sum_{s=1}^{p} \sum_{i=1}^{n_k} (x_{is} - \overline{x}_{s,k})^2$$

che è la devianza entro i gruppi riferita alle p variabili con riferimento al gruppo k, dove $\overline{x}_{s,k}$ è la media della variabile s con riferimento al gruppo k;

$$DEV_{OUT} = \sum_{s=1}^{p} \sum_{k=1}^{g} (\overline{x}_{s,k} - \overline{x}_s)^2 \ n_k$$

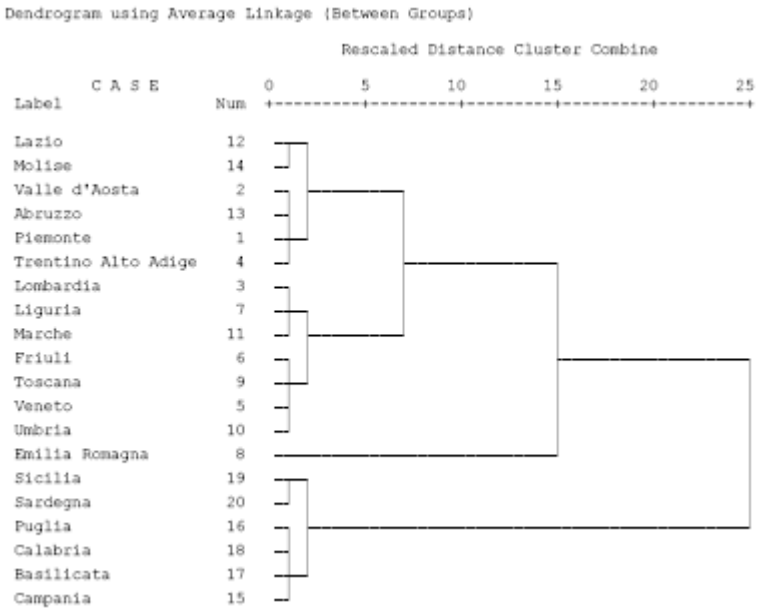
che è la devianza tra i gruppi.
Come noto $DEV_T = DEV_{IN} + DEV_{OUT}$
Nel passare da k+1 a k gruppi (aggregazione) DEV_{IN} aumenta, mentre ovviamente DEV_{OUT} diminuisce. Ad ogni passo metodo di Ward si aggregano tra loro quei gruppi per cui vi è il minor incremento della devianza entro i gruppi.

Scelta del numero di gruppi

Nel caso di una cluster gerarchica la scelta del numero di cluster può essere effettuata utilizzando in primo luogo la distanza di fusione.

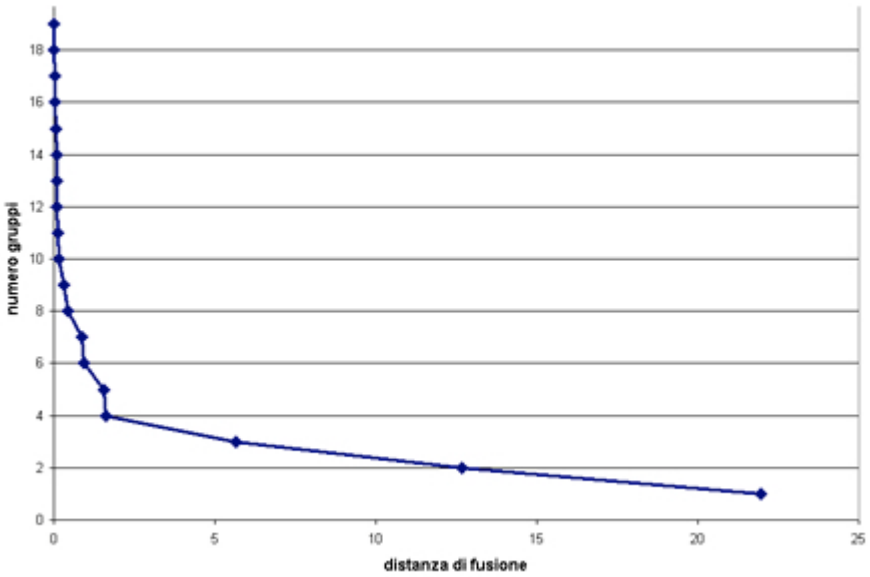
Nell'esempio sono stati utilizzati gli incidenti stradali e ed il numero di morti e feriti per regione rapportati alla popolazione.

Stadio	Cluster accorpati		Distanza Fusione	Incrementi relativi dist.fusione	Distanza riscalata
	Cluster 1	Cluster 2			
1	12	14	0.010		
2	2	13	0.010	0.051	0.012
3	6	9	0.027	1.631	0.030
4	1	2	0.029	0.090	0.033
5	3	7	0.061	1.103	0.070
6	16	18	0.084	0.366	0.095
7	19	20	0.099	0.181	0.112
8	3	11	0.099	0.001	0.112
9	5	10	0.126	0.271	0.143
10	1	4	0.159	0.262	0.181
11	16	17	0.318	1.008	0.362
12	5	6	0.447	0.406	0.509
13	15	16	0.867	0.938	0.987
14	1	12	0.925	0.066	1.053
15	3	5	1.567	0.695	1.784
16	15	19	1.615	0.030	1.839
17	1	3	5.669	2.511	6.455
18	1	8	12.665	1.234	14.420
19	1	15	21.958	0.734	25.000



La distanza di fusione, in termini di distanza riscalata può essere facilmente evinta dall'osservazione del dendrogramma: se nel passaggio da K gruppi a K+1 si registra un forte incremento della distanza di fusione si deve "tagliare" a K gruppi.

Sempre utilizzando la distanza di fusione si possono utilizzare gli scree plot, grafici in cui viene posto in ordinata il numero di gruppi ed in ascissa la distanza di fusione:



Il grafico, in questo caso, suggerisce di mantenere in suddivisione in 4 gruppi. Infatti, nel passaggio a 3 gruppi si registra un consistente incremento della distanza di fusione.

Per valutare l'entità dell'incremento della distanza di fusione si può ricorrere all'incremento relativo della distanza di fusione:

$$\delta_k = \left(d_k - d_{k+1}\right)/d_{k+1}$$

verrà scelto il K per δ_k è massimo.

In generale questi metodi di scelta del numero di gruppi si basano comunque su una osservazione dei dati alla ricerca di una loro discontinuità, e questo può risultare una procedura azzardata e soggettiva. Sono stati proposti vari test e fra questi pseudo-F:

$$F_k = \frac{Dev_{OUT}/(k-1)}{DEV_{IN}/(n-k)} \quad \forall \quad k = 1,\ldots,g$$

L'applicazione di metodi gerarchici, come tutte le tecniche statistiche, reca con sé limiti e vantaggi. I metodi gerarchici presuppongono indirettamente una regola classificatoria sottostante più o meno rispettata dalle unità, nella quale esse rientrano progressivamente. Ovviamente, se nel nostro contesto una tale regola non può essere ipotizzata, l'adozione di metodi gerarchici è abbastanza limitata in quanto può generare tipologie errate.

I vantaggi sono invece legati al fatto che permette di studiare il processo che porta i profili contabili aziendali ad assimilarsi tra loro.

3.2 Metodi non gerarchici

Questi metodi sono caratterizzati da un procedimento che mira a ripartire direttamente le n unità in r grappoli, fornendo come prodotto finale una sola partizione delle n unità (Andemberg, 1973; Matthews, 1979). Supposto che, *a priori*, sia stato fissato il numero dei gruppi in cui si vuole ripartire il collettivo di partenza, le procedure non gerarchiche si articolano sostanzialmente in due fasi:

- la determinazione di una partizione iniziale degli *n* individui in G gruppi;
- spostamento successivo delle unità tra i G gruppi, in modo da ottenere la partizione che meglio risponde ai concetti di omogeneità interna ai gruppi e di eterogeneità tra gli stessi.

L'individuazione della partizione ottimale comporterebbe a rigore l'esame di tutte le possibili assegnazioni distinte degli *n* individui a G gruppi. Poichè un'operazione di questo genere determina una grande mole di calcoli, le procedure non gerarchiche propongono di risolvere il problema attraverso una strategia di raggruppamento che richiede la valutazione solo di un numero accettabile di possibili partizioni alternative. In pratica, una volta scelta la partizione iniziale, si procede a riallocare le unità in esame tra i diversi gruppi in modo da ottimizzare la prefissata funzione obiettivo. Gli algoritmi di tipo non gerarchico, quali ad esempio quelli di McQueen (detto anche delle k medie) e di Forgy, procedono, data una prima partizione, a riallocare le unità al gruppo con centroide più vicino, fino a che per nessuna unità si verifica che sia minima la distanza rispetto al centroide di un gruppo diverso da quello a cui essa appartiene. Una tale procedura minimizza implicitamente la devianza entro i gruppi, relativamente alle p variabili. Poiché l'applicazione di questi metodi presuppone l'individuazione a priori di una partizione iniziale, è evidente che la soluzione trovata risulta in qualche misura subordinata a tale scelta, e può quindi essere considerata alla stregua di un punto di ottimo locale.

In sintesi si segue una procedura iterativa secondo le fasi:

1. Scelta di g centri $(c_1, c_2, ..., c_h, ..., c_g)$
2. Raggruppamento delle unità intorno ai k centri in modo che il gruppo delle unità associate a c_h è costituita dall'insieme delle unità più vicine a c_h che a qualsiasi altro centro.
3. Calcolo dei centroidi dei g gruppi così ottenuti
4. Calcolo della distanza di ogni elemento da ogni centroide: se la distanza minima non è ottenuta in corrispondenza del centrode del gruppo di appartenenza, allora l'unità è riallocata al gruppo che corrisponde al centroide più vicino
5. Ricalcolo dei centroidi
6. Iterazione dei passi 4. e 5. fino a che i centri non subiscono ulteriori modifiche rispetto alla iterazione precedente

Come misura di distanza tra l'unità i ed il centroide viene normalmente utilizzata la distanza euclidea in quanto garantisce la convergenza della procedura iterativa.

L'indicatore della validità della partizione è invece fornito da

$$R^2 = \frac{DEV_{OUT}}{DEV_T}$$

I vantaggi dei metodi non gerarchici sono costituiti principalmente dalla velocità di esecuzione dei calcoli e dall'estrema libertà che viene lasciata alle unità di raggrupparsi e allontanarsi (negli algoritmi gerarchici se due unità vengono fuse all'inizio, rimangono unite fino alla fine della procedura). Non si suppone, in altri termini, che vi sia una tassonomia delle tipologie alla quale le unità siano costrette più o meno ad adeguarsi.
Tuttavia, quest'ultimo aspetto, se da un lato costituisce un vantaggio, dall'altro pone la necessità di specificare delle ipotesi *a priori* sulla struttura del collettivo. Questo costituisce un limite perché costringe a provare più ipotesi alternative. Una soluzione a questo tipo problema può consistere nel: i) applicare un algoritmo di tipo gerarchico; ii) scegliere un intervallo di valori "ragionevoli" per g; iii) applicare l'algoritmo di tipo non gerarchico per ognuno dei valori così individuati; iv) scegliere la soluzione ottimale attraverso R^2.

Tra i problemi che si possono avere nell'utilizzo di algoritmi non gerarchici sono da sottolineare:

- la classificazione finale può essere influenzata dalla scelta iniziale dei poli; occorre quindi porre attenzione all'ordine delle unità
- si possono ottenere soluzioni instabili in presenza di valori anomali, numerosità insufficiente o qualora non sussista una struttura in gruppi nei dati.

4. Considerazione sulla scelta dell'algoritmo

Vari studi (Rand, 1971; Ohsumi, 1980) indicano che strategie di raggruppamento differenti conducono spesso a risultati non dissimili. Da altri studi, in molte applicazioni pratiche, i risultati sembrano dipendere molto strettamente non solo dalla strategia seguita, ma anche dalle opinioni utilizzate nell'analisi; ad esempio, si ottengono risultati diversi secondo che i dati siano standardizzati, oppure no. Comunque, i criteri di scelta fra i due tipi di algoritmo non sono ancora stati sufficientemente esplorati e nella letteratura vi sono posizioni molto diversificate. Alcuni sostengono che strategie di raggruppamento diverse conducono a risultati simili (Rand, 1971) mentre altri evidenziano casi concreti di forte divergenza (Everitt, 1979; Fabbris, 1983).
Per quanto riguarda la valutazione della procedura di clustering, si considera la proposta di valutazione di Silvestri e Hill (1964). I criteri suggeriti dai due autori comprendono: (i) l'*oggettività*, per la quale ricercatori che lavorano indipendentemente sullo stesso insieme di dati devono giungere agli stessi risultati; (ii) la *stabilità* dei risultati della classificazione operando su dati equivalenti. Nonostante non sia proprio la stessa cosa, Jones e Needham (1968) assimilano questo ultimo criterio alla proprietà di *robustezza* contro la presenza di errori nei dati; (iii) la capacità *predittiva* delle variabili su un nuovo insieme di dati. Nella pratica la ricerca si concentra sulla classe di metodi che si dimostrano più insensibili a piccole variazioni nei dati analizzati. Ad esempio, si ritiene importante per un metodo se, sottraendo un individuo dall'analisi, l'albero o i gruppi formati cambiano poco o punto. Oppure se, ripetendo l'analisi senza una intera diramazione del dendrogramma,

la struttura degli altri rami resta invariata o quasi. Jardine e Sibson (1971) parlano a questo proposito di condizioni di unitarietà (*fitting togheter*).
Per grandi linee, si può dire che, se si cercano gruppi di unità statistiche caratterizzate da alta omogeneità interne (nel senso di strettezza dei legami tra entità appartenenti a un gruppo), le tecniche gerarchiche sono meno efficaci delle tecniche non gerarchiche.

Bibliografia:

AA.VV (1983), *New trends in Data Analysis and Applications*, North Holland.

Andemberg M. (1973), *Cluster analysis for applications*, New York Academic Press.

Everitt B.S. (1979), *Unresolved problems in cluster analysis*, Biometrics.

Fabbris L. (1983), *Analisi esplorativa di dati multidimensionali*, Cleup editore.

Wegman E.J. (9172), *Non-parametric Probability Density Estimation* in: A survey on available methods, Technometrics.

Green P.E., Frank R.E., Robinson P.J. (1967), *Cluster Analysis in text market selection*, Management science.

Hartigan J.A. (1975), *Clustering Algorithms*, Wiley.

Jardine N., Sibson R. (1971), *Mathematical taxonomy*, Wiley, London.

Jones K.S., Needham R.M. (1968), *Automatic term classification and retrieval*, Inform.

Storage.

Johonson S.C.(1967), *Hierarchical clustering schemes*, Psycometrika.

Kendall M. (1975), *Multivariate analysis*, Charles Griffin & Company, London.

Matthews A. (1979), *Standardization of measures prior to clustering*, Biometric.

Morrison D.F. (1967), *Measurement problems in cluster analysis*, Management science.

Morrison D.F. (1976), *Multivariate statistical methods*, Mc Graw Hill.

Oshumi N. (1980), *Evaluation procedure of agglomerative hierarchical clustering methods by fuzzy relations* in: Data Analysis and Informatics, Diday et al. (a cura di), North Holland.

Rand W.M. (1971), *Objective criteria for the evaluation of clustering methods*, J.A.S.A.

Silvestri L., Hill I.R. (1964), *Some problems of the taxometric approach* in: Phenetic and Phylogenetic Classification, Heywood V.H. e Mc Neil J. (a cura di), Systematic Association Londra.