electoral behaviour that focuses on the concept of subculture. It is emphasized how the traditional formulation of this hypothesis needs to be redefined, by articulating it in relation to the overall characteristics of the social system in areas of peripheral development where industrialization has been libht and rather spread out: the features of this development model are the ones that enable the subcultural reality to persist and adapt.

Using this framework as a springboard, the discussion turns to possible transformations of the socio-economic bases of the subculture and the effects that an erosion of the subcultural reality, hitherto averted, might have on political and electoral behaviour.

Finally, an area for further research is proposed, namely to pinpoint and verify some of these hypotheses through a study of the political and electoral behaviour of two Tuscan working class groups belonging to two specific production and territorial sectors: the textile sector (e.g. Prato and its surrounding communities) and the metal-working sector (e.g. the Florence industrial belt). Such a research should verify and explain, in particular, the different factors leading to the formation of voting preference in the two environments (even though the election results are similar) in relation to the diversity of political behaviour. Moreover, it would be most interesting to determine whether or not the traditional foundations of these subcultures reveal signs of crisis and if so, the possible evolutions such a crisis may undergo with regard to political representation and voting behaviour.

L'ANALISI DEI GRUPPI: UNA METODOLOGIA PER LO STUDIO DEL COMPORTAMENTO ELETTORALE

(Parte Prima)

di Bruno Chiandotto

Nota di redazione

Nello studio del comportamento elettorale è possibile fare ricorso a svariate tecniche di analisi statistica multivariata; tra queste, un ruolo di indubbio rilievo spetta all'analisi dei gruppi (*cluster analysis*). All'apparente semplicità degli obiettivi di questo tipo di analisi, e che consistono principalmente nella individuazione di « gruppi omogenei » di unità di osservazione, si contrappongono numerose difficoltà concettuali aggravate dalla mancanza di una *teoria generale* che possa guidare il ricercatore nella scelta dei metodi più appropriati allo specifico caso in esame, ponendo così seri ostacoli ad una sua valida utilizzazione in campo operativo.

In questa ottica, il saggio di Bruno Chiandotto intende fornire al lettore un quadro generale introduttivo dello « stato dell'arte » sollevando anche problemi non ancora risolti, o risolti in maniera insoddisfacente, sul piano teorico. Il lavoro è diviso in due parti; nella prima parte, che appare su questo numero dei « Quaderni », l'autore analizza la problematica che sottende alle varie metodiche cercando di dare una sistemazione organica della materia. Nella seconda parte verranno, invece, discusse questioni tecniche e problemi metodologici più direttamente connessi allo studio del comportamento elettorale.

Premessa

Quando si parla di *analisi dei gruppi* o dei *raggruppamenti* (*cluster* o *clustering analysis*) ci si vuol riferire, nella generalità dei casi, all'insieme delle tecniche *statistico-matematiche* utilizzate per la formazione di *gruppi omogenei* di unità di osservazione che risultano essere caratterizzate dalle modalità di più variabili e/o mutabili (1).

L'analisi dei gruppi rientra nell'ambito della più vasta problematica relativa alla *classificazione matematica*, campo questo in cui si rileva una notevole confusione, sia terminologica che concettuale, con riflessi che risultano particolarmente accentuati nel caso specifico dell'analisi dei gruppi. L'insoddisfacente stato di cose va, soprattutto, attribuito all'eccezionale e non coordinata proliferazione di metodi di classificazione cui si è assistito negli ultimi anni e che ha interessato quasi tutti i campi della ricerca scientifica sia teorica che applicata: dalla biologia alla medicina, dalla geologia alla archeologia, dalla psicologia alla sociologia, dall'economia alla politologia, ecc..

Se da un lato l'eccezionale sviluppo che ha interessato questo tipo di analisi sottolinea in modo evidente la sua estrema rilevanza ed attualità, dall'altro, il modo con cui si è realizzato ed espresso impone un lavoro teorico di sistemazione della materia che, prescindendo dagli specifici campi di applicazione, ne chiarisca i fondamenti logici come primo passo verso la costruzione di una *teoria generale* in cui i vari metodi di classificazione proposti possano venire correttamente inquadrati e riconosciuti. In un tale quadro, i diversi metodi dovranno evidenziare le loro proprietà ed i loro limiti, dando ragione delle scelte particolari effettuate da ogni singolo ricercatore nell'ambito di ciascuna specifica disciplina scientifica.

(1) Appartiene allo stesso ambito disciplinare l'insieme delle tecniche volte alla individuazione di *gruppi omogenei* di variabili e/o mutabili. In questo saggio ci occuperemo in modo pressoché esclusivo del problema della formazione di gruppi di unità; comunque, molte delle argomentazioni svolte conservano di validità anche se trasferite al problema della individuazione di gruppi di variabili e/o mutabili. Per quanto concerne poi la distinzione tra variabili e mutabili, si segnala che le prime riguardano i cosidetti caratteri quantitativi mentre le seconde i caratteri qualitativi; una discussione più estesa su questo punto verrà fatta nelle pagine successive.

In attesa di una *teoria generale* (2), mi sembra auspicabile che ogni ricercatore, al momento della presentazione di un proprio contributo su problemi concernenti la classificazione matematica, abbia cura di spendere qualche parola sia sul significato che intende attribuire ai termini usati che sul quadro concettuale cui intende riferirsi.

Le brevi considerazioni sopra fatte non vogliono essere tanto la premessa di un tentativo di sistemazione teorica della materia, quanto la giustificazione di un saggio sostanzialmente metodologico, agli occhi di quel lettore che avrebbe forse preferito una trattazione specificatamente orientata verso il tema elettorale, che costituisce poi il corpo della collana in cui questo saggio compare.

In ogni caso si deve osservare che la finalità più immediata, anche se non esclusiva, dell'analisi classificatoria, e più in particolare dell'analisi dei gruppi, quando viene applicata allo studio del comportamento elettorale è quella della individuazione di *aree politicamente omogenee*. Sarà possibile all'interno di tali *aree* procedere — se si vuole, e più di quanto non sia consentito a livello di aggregati indistinti e/o troppo ampi — alla individuazione dei possibili nessi *causali* tra fattori socio-economici e comportamento elettorale. Inoltre, un'analisi dei mutamenti che possono intervenire, col trascorrere del tempo, nell'ambito delle *aree politicamente omogenee* consente, tra l'altro, di meglio valutare i possibili effetti dell'attività dei vari partiti politici nei confronti dell'elettorato.

Anche se alcune applicazioni dell'analisi dei gruppi verranno brevemente discusse nella parte seconda di questo saggio, va sottolineato ancora una volta, che la presentazione sarà di tipo essenzialmente teorico e verterà sulla problematica generale dell'argomento, senza indulgere nella descrizione di quei particolari metodi di raggruppamento che sono già stati utilizzati nello studio del comportamento elettorale; ciò viene fatto perché si ritiene di facilitare, in tal modo, sia l'interpretazione dei risultati cui le applicazioni stesse hanno dato luogo che il compito dei ricercatori che intendono condurre ulteriori, e magari diverse, applicazioni delle procedure di raggruppamento all'analisi del fenomeno elettorale.

(2) I tentativi fatti in tale direzione, volti cioè alla costruzione di una teoria generale della classificazione matematica, pure se pregevoli, non mi sembra abbiano raggiunto un sufficiente grado di generalità.

Tra gli innumerevoli lavori pubblicati che trattano il tema della classificazione matematica (e più specificatamente dell'analisi dei gruppi) vanno segnalati, per il loro particolare contenuto (tentativi di formazione di una teoria generale, rassegne critiche, esposizioni ampie e sistematiche della materia) quelli di SOKAL e SNEATH (1963), LANCE e WILLIAMS (1967 a, 1967 b), TRYON e BAILEY (1970), CORMACK (1971), JARDINE e SIBSON (1971), ANDERBERG (1973), BENZÉCRI (1973), BIJNEN (1973), LUNETTA (1973), SNEATH e SOKAL (1973), DURAN e ODELL (1974), BAILEY (1975), HARTIGAN (1975), BELLACICCO (1976), DIDAY e SIMON (1976), LIS e SAMBIN (1977), VAN RYZIN ed. - con articoli di SOKAL, KRUSKAL, HARTIGAN, GOOD, MATULA, HUBERT e BAKER, SOLOMON, RAO, FU, ZADEH - (1977), ZANI (1977), RIZZI (1978).

1. INTRODUZIONE

Uno degli obiettivi primari della scienza è certamente quello di procedere ad una *sistemazione ordinata* del *mondo fenomenico*, ma gli eventi che *generano* i *fenomeni* sono spesso delle entità troppo complesse e troppo numerose per poter essere analizzate individualmente, risulterà pertanto indispensabile procedere alla introduzione di uno *schema categoriale*, o *schema classificatorio* o *schema* di *gruppo*, entro il quale le entità stesse possono essere collocate e riconosciute.

Il ricercatore scientifico che dispone di un insieme di *entità osservazionali* (*unità statistiche di osservazione*), e che vuol procedere ad una loro classificazione, si può trovare di fronte ad un *continuum* di situazioni diverse le cui specificità impongono l'uso di strumenti differenziati. Ad un estremo si possono collocare le situazioni in cui esiste già uno schema categoriale perfettamente determinato e nel quale gli attributi essenziali di ciascuna categoria o gruppo sono esattamente definiti. In tali casi, il problema di classificazione si riduce alla *identificazione* di ciascuna unità di osservazione ed alla conseguente *assegnazione* o attribuzione alla rispettiva categoria o gruppo di appartenenza. Il problema di classificazione delineato, che ha la semplice natura di problema di *assegnazione*, può complicarsi a causa di definizioni imperfette dei gruppi, per eventuali sovrapposizioni degli stessi o per la presenza di elementi di variabilità di tipo accidentale nelle unità di osservazione. Un modo per trattare *statisticamente* i problemi che hanno una tale natura è quello di valutare le probabilità di appartenenza di ciascuna unità di osservazione ai vari gruppi, per poi procedere all'attribuzione (*assegnazione*) dell'unità al gruppo d'appartenenza più probabile.

All'estremo opposto, a quello sopra delineato, si possono collocare le situazioni in cui non si dispone di alcuna informazione sulla struttura dei gruppi cui assegnare le varie unità di osservazione che, al limite, potrebbe anche non esistere. I problemi di classificazione emergenti in tali situazioni sono quelli che ci interessano più da vicino in quanto costituiscono l'oggetto specifico della *analisi dei gruppi*.

In una posizione intermedia, rispetto alle due sopra considerate, si ritrovano tutte le situazioni in cui pur esistendo una struttura categoriale di riferimento, questa non è esattamente o completamente definita.

In questi casi, l'elemento distintivo delle analisi (*discriminanti* e di *pattern recognition* (3)) sta nell'uso delle informazioni aggiuntive, fornite dalle unità campionarie di osservazione, per procedere ad una migliore caratterizzazione delle categorie o gruppi ai fini della attribuzione delle unità campionarie ai gruppi stessi.

I problemi che rientrano nell'ambito dell'analisi dei gruppi rappresentano, pertanto, lo stadio meno definito tra quelli cui ci si può imbattere allorquando si intende procedere alla classificazione di unità di osservazione (4). Infatti, in tali situazioni gli unici elementi di cui si può disporre sono le unità di osservazione stesse che risultano individuate dalle modalità (quantitative e/o qualitative) assunte da più elementi caratterizzanti e l'obiettivo che si vuol perseguire è quello della *scoperta* di una eventuale *struttura di gruppo* che, come già detto, potrebbe anche non esistere. Si tratta in sostanza di individuare la *eventuale presenza* di *gruppi naturali* cui poter attribuire le singole unità di osservazione sulla scorta delle informazioni fornite dalle unità stesse e dalle conoscenze a priori, più o meno approfondite, che ogni singolo ricercatore ha sul *fenomeno* oggetto di studio.

La chiarificazione del termine *categoria* o *gruppo naturale* rappresenta l'essenza stessa di tutta la problematica concernente l'analisi dei gruppi.

(3) Trattazioni ampie e sistematiche di questi due tipi di analisi sono state fatte da ANDERSON (1958), COOLEY e LOHNES (1971), FUKUNAGA (1972). Di particolare interesse sono il volume curato da CACOULLOS (1972) (che raccoglie gli atti del convegno « *NATO Advanced Study Institute on Discriminant Analysin and Applications* » svoltosi ad Atene nel 1972) il lavoro di SALVEMINI (1960) e, anche per gli interessanti risvolti applicativi, l'articolo di VITALI (1972).

(4) La distinzione fatta tra *assegnazione*, *analisi discriminante* e *analisi dei gruppi*, come possibile tripartizione del problema generale della *classificazione* matematica, oltre ad essere riduttiva, non è accettata da tutti gli autori, così come non lo è la terminologia adottata. Ad esempio, per tradurre il termine « *cluster analysis* », LUNETTA (1973) e RIZZI (1978) utilizzano la stessa dizione qui usata di « *analisi dei gruppi* », mentre ZANI (1977) adopera la dizione « *analisi classificatoria* », forse influenzato in questo da KENDALL (1966 e 1972) che usa in modo quasi intercambiabile le locuzioni « *classification* » e « *cluster analysis* » distinguendole poi dalla « *pattern recognition* », dalla « *discrimination* » e dalla « *dissection* », osservando, a quest'ultimo proposito, che tutti gli insiemi di unità di osservazione possono essere « *dissected* » ma non tutti possono essere « *classified* ». CORMACK (1971) pur utilizzando, come Kendall, in modo sostanzialmente intercambianile le dizioni « *classification analysis* » e « *cluster analysis* » riconosce esplicitamente, richiamandosi a DAGNELIE (1966), come migliore l'uso dei termini *identificazione* o *assegnazione* per caratterizzare i processi di attribuzione di nuove unità di osservazione a classi o gruppi già ben definiti. L'uso qui riservato al termine *classificazione* e alla dizione *analisi classificatoria* per indicare l'area globale di ricerca ci sembra il più appropriato, così come ci sembra sufficiente la distinzione operata, all'interno di questa, tra problemi di assegnazione, di dicriminazione e di raggruppamento, anche se si riconosce una certa utilità, ma non ai fini della presente trattazione, alla ulteriore distinzione tra problemi di individuazione di *gruppi naturali* (*cluster analysis*) e quelli di *segmentazione* di insiemi di unità di osservazione (*dissection*) a prescindere dalla esistenza o meno di *gruppi naturali*.

2. PROBLEMATICA GENERALE DELL'ANALISI DEI GRUPPI.

Dalle brevi considerazioni introduttive svolte, dovrebbe risultare in modo evidente come i problemi proprio dell'*analisi dei gruppi* siano logicamente antecedenti a quelle della *discriminazione*, e come questi debbano logicamente precedere quelli dell'*assegnazione*. Infatti, fino a quando non si sa (a meno che non si assuma come ipotesi) che le unità di osservazione appartengono a gruppi distinti, non sarà possibile procedere alla risoluzione dei problemi di discriminazione; d'altronde, non è corretto procedere ad una assegnazione delle unità di osservazione ai vari gruppi fino a quando i gruppi stessi non siano ben discriminati ed esattamente definiti.

Nonostante la sequenzialità logica sopra evidenziata si deve osservare come il processo di sviluppo storico dei tre argomenti, almeno dal punto di vista analitico-formale, abbia seguito un ordine inverso. Le ragioni di un tale stato di fatto vanno ricercate nella necessità, già sottolineata nella premessa, per ogni studioso di darsi un insieme di *categorie sistematiche* e *concettuali* che servano come schema di riferimento per l'organizzazione della conoscenza, dando così ragione dello sviluppo dei metodi di identificazione come primo momento nello studio dei problemi di classificazione matematica. Il ritardo nello sviluppo dei metodi di analisi dei gruppi va invece attribuito, in larga misura, alla comparsa soltanto recente nel panorama scientifico di elaboratori elettronici ad alta potenzialità e con grande capacità di memoria; strumenti questi indispensabili per l'applicazione di una qualunque procedura non banale di analisi dei gruppi. Per contro, se gli elaboratori elettronici hanno dato un enorme impulso allo sviluppo dei metodi di analisi dei gruppi, si deve ossevare, purtroppo, come questi, associati alla disponibilità di programmi di calcolo già predisposti, abbiano stimolato un numero non indifferente di ricercatori verso una utilizzazione spesso acritica di procedure di raggruppamento inadeguate.

Il problema della ricerca di *gruppi naturali*, in un insieme di unità di osservazione, richiede il superamento di tutta una serie di fasi d'analisi e di decisioni specifiche la cui complessità esclude ogni possibile ricorso a metodi di risoluzione di tipo meccanicistico.

Prima di passare alla illustrazione, seppure sommaria, delle varie fasi che devono essere superate nella esecuzione di un qualsiasi processo di analisi dei gruppi, risulta opportuna una presentazione algebrico-formale della problematica generale di questo tipo di analisi.

Sia

$$U = \{U_1, U_2, \ldots, U_i, \ldots, U_n\}$$

un insieme di n unità di osservazione relative ad una determinata popolazione *P*,

$$C = \{C_1, C_2, \ldots, Cj, \ldots, C_k\}$$

un insieme di k caratteri (aventi natura quantitativa e/o qualitativa) osservabili su ciascuna unità e si indichi con x_{ij} ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$) la modalità assunta dal j-esimo carattere nella i-esima unità di osservazione. In tali condizioni ciascuna unità U_i resta individuata da un vettore a k dimensioni

$$X_i = [x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ik}] \quad (i = 1, 2, \ldots, n).$$

L'insieme degli n vettori X_i riceve la sua rappresentazione matriciale nei termini seguenti

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_i \\ \cdot \\ \cdot \\ X_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2k} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ik} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nk} \end{bmatrix}$$

Utilizzando i dati contenuti nella matrice X, il problema di analisi dei gruppi si risolve nella ricerca di una *partizione* (5) dell'insieme U (6) in m ($m \leq n$, intero) sottoinsiemi (*gruppi* o *cluster*) in modo tale che le unità appartenenti ad uno stesso sottoinsieme (gruppo) siano il più possibile *omogenee* tra loro. In altri termini, il problema di analisi dei gruppi si sostanzia nella suddivisione di P in m sottopopolazioni $P_1, P_2, \ldots, P_m$ in modo tale che ciascuna unità appartenga ad una sola sottopopolazione.

Appare evidente come abbia senso parlare di analisi dei gruppi soltanto nei casi in cui sia ipotizzabile a priori l'esistenza di sub-popolazioni (*gruppi naturali*) di P che, nondimeno, le procedure applicate potrebbero anche non evidenziare o fornire elementi di giudizio insufficienti a sostegno di una tale supposizione.

(5) In questo saggio verranno esaminati soltanto metodi di analisi dei gruppi che determinano delle *partizioni* dell'insieme U; in modo tale cioè, che ciascuna unità di osservazione appartenga, al termine del processo, ad uno ed uno solo sottoinsieme di U. Non si considerano pertanto tutti quei metodi che prevedono la sovrapposizione tra sottoinsiemi. Chi fosse interessato a tali metodi, detti di « *clumping* » o di « *overlapping* », si può riferire a JARDINE e SIBSON (1971).

(6) In questa sede verrà esaminato, come già detto, soltanto il caso in cui si voglia procedere alla formazione di gruppi di unità di osservazione; considerazioni analoghe, *mutatis mutandi*, possono essere fatte riguardo al problema della formazione di gruppi di variabili. In alcune situazioni potrà anche interessare una analisi dei gruppi combinata, considerando cioè simultaneamente sia unità che caratteri (variabili e/o mutabili). Su questo ultimo punto il lettore interessato potrà consultare utilmente il volume di HARTIGAN (1975).

Affinché sia possibile sviluppare l'analisi dei gruppi come procedura automatica di classificazione, risulta necessario introdurre una definizione di *gruppo* più rigorosa o, quantomeno, operativamente valida. Un modo per perseguire una tale finalità è quello di introdurre una funzione obiettivo $g(.)$ che faccia corrispondere ad ogni possibile partizione dell'insieme delle unità di osservazione un numero reale: la migliore partizione è quella che *massimizza* (*minimizza*) la funzione obiettivo. In termini algebrici, se con

P_{α_i} dove gli α_i ($1 = 1, 2, \ldots, n$) possono assumere i valori interi da 1 a m

si indica la sub-popolazione (gruppo) alla quale viene attribuita la i-esima unità di osservazione, ogni partizione π risulterà espressa da un vettore

$$\pi = [P_{\alpha_1}, P_{\alpha_2}, \ldots, P_{\alpha_i}, \ldots, P_{\alpha_n}]$$

e la funzione obbiettivo assume la forma

$$g(.) = g(\pi\ ;\ X) = g(P_{\alpha_1}, P_{\alpha_2}, \ldots, P_{\alpha_n}\ ;\ X_1, X_2, \ldots, X_n).$$

La migliore partizione π_o è quella che realizza la condizione

$$g(\pi_o\ ;\ X) = \max_{\pi}\ g(\pi\ ;\ X)$$

o, alternativamente,

$$g(\pi_o\ ;\ X) = \min_{\pi}\ g(\pi\ ;\ X)$$

dove il problema si risolve nella individuazione del massimo o del minimo della funzione obbiettivo, a seconda della particolare specificazione della funzione stessa (7).

La ricerca del massimo (del minimo) della funzione $g(\pi\ ;\ X)$ implica il computo di tutte le possibili partizioni (configurazioni classificatorie) π; lavoro questo generalmente impraticabile, o comunque molto oneroso anche nel caso in cui si abbia a che fare con popolazioni di modeste dimensioni (8). Sarà pertanto necessario introdurre un *algoritmo* di raggruppamento che pur riducendo il computo delle possibili configurazioni classificatorie, e quindi dei valori assumibili dalla funzione obbiettivo, consenta, se non la indivi-

(7) Quello esposto costituisce un particolare approccio alla analisi dei gruppi, il cosidetto *approccio statistico* o *approccio decisionale*.

(8) Il numero delle possibili partizioni di un insieme di n unità in m gruppi, per $m = 1, 2, \ldots, n$, è espresso dalla formula

$$\sum_{m=1}^{n} \frac{1}{m!} \sum_{h=0}^{m} (-1)^{m-h} \binom{m}{h} h^n$$

che da un valore elevatissimo anche per valori di n piuttosto bassi; ad esempio per $n = 25$ si ottiene un numero superiore a 4×10^{18}.

duazione della migliore partizione π_o rispetto a $g(\pi ; X)$, quantomeno l'individuazione di una partizione π_o prossima a quella ottima.

Dalle considerazioni sopra svolte dovrebbe emergere in modo abbastanza evidente come la specificazione analitica della funzione obbiettivo comporti, generalmente, l'introduzione di una qualche *misura* capace di esprimere il *grado di omogeneità* tra le diverse unità di osservazione. Infatti, lo scopo ultimo dell'analisi dei gruppi è, come già detto, proprio quello di formare dei gruppi che contengano unità di osservazione il più possibile *simili* (*omogenee*) tra loro, purché ne risulti una sufficiente *diversità* (*disomogeneità*) tra le unità appartenenti a gruppi distinti. Naturalmente, un tale scopo può essere perseguito sia operando con *misure* di *similarità* tra unità, ed in questo caso si dovrà procedere ad una massimizzazione della funzione obbiettivo, sia facendo uso di *misure* di *diversità*, il che comporta, generalmente, una minimizzazione della funzione obbiettivo.

Nella sostanza, un problema di analisi dei gruppi si risolve attraverso la scelta di una opportuna *funzione obbiettivo* (*criterio di raggruppamento*) e di un efficiente *algoritmo classificatorio*. Per poter operare tali scelte in modo soddisfacente è necessario risolvere tutta una serie di problemi collaterali, ma non per questo meno importanti, che vanno dalla scelta delle unità di osservazione alla interpretazione dei risultati ottenuti, alla luce delle proprietà e limiti del criterio e dell'algoritmo di raggruppamento utilizzati.

3. Fasi del processo di analisi dei gruppi.

Pur operando ad un livello di generalità abbastanza spinto, in relazione ad ogni singola analisi dei gruppi si possono individuare almeno otto diversi tipi di problema da affrontare e risolvere. Si tratta di problemi relativi alla:

a) scelta delle unità di osservazione;
b) scelta degli elementi (mutabili e/o variabili) caratterizzanti ciascuna unità di osservazione;
c) omogeneizzazione delle scale di misura utilizzate per esprimere i vari caratteri considerati;
d) scelta della misura di similarità o diversità tra unità di osservazione;
e) definizione del numero di gruppi che si vogliono formare;
f) scelta del criterio di raggruppamento;
g) scelta dell'algoritmo di classificazione;
h) interpretazione dei risultati ottenuti.

I problemi richiamati, pure se si presentano generalmente al ricercatore nell'ordine elencato, non vanno e non possono essere risolti isolatamente ed in progressione. Si tratta di una serie di problemi alla cui soluzione si potrà pervenire soddisfacentemente soltanto se si opera in modo pressoché simultaneo, attraverso un processo di adattamento di tipo recursivo che comporti la revisione continua delle decisioni prese negli stadi precedenti del processo in atto.

a) *Scelta delle unità di osservazione*

Il problema della scelta delle unità di osservazione sulle quali operare una analisi dei gruppi sembrerebbe, a prima vista, non sussistere in quanto le unità stesse dovrebbero rappresentare un dato del problema. In realtà ci sono almeno tre situazioni in cui la scelta delle unità di osservazione assume rilevanza.

Il primo caso è quello in cui le unità di osservazione non riguardano l'intera popolazione, rispetto alla quale si vuol individuare una struttura di gruppo, ma soltanto un campione di questa [9]; si dovrà allora aver cura di controllare l'eventuale assenza di unità che potrebbero risultare rilevanti ai fini dell'analisi di raggruppamento, onde evitare che pur esistendo una struttura di gruppo questa non venga evidenziata a causa della omissione di particolari unità di osservazione.

In una diversa situazione può accadere, invece, che l'insieme delle unità di osservazione costituiscono una popolazione la cui unitarietà sia determinata non già da fattori di ordine logico ma da fatti contingenti che nulla hanno a che vedere con la effettiva popolazione rispetto alla quale si vuol individuare una struttura di gruppo. In tali casi si dovrà procedere alla eliminazione di tutte quelle unità che risultino spurie rispetto al problema che s'intende risolvere.

Una terza situazione nella quale assume rilevanza il problema della scelta delle unità di osservazione è quella in cui si ha a che fare con un insieme di unità logicamente unitario, rispetto al quale può o meno esistere una struttura di gruppo, mentre si ha ragione di credere che tale struttura esiste certamente nell'ambito di un suo sottoinsieme. In tali casi può risultare conveniente limitare l'analisi dei gruppi al solo sottoinsieme di unità di osservazione [10].

b) *Scelta delle variabili e/o mutabili*

Risulta evidente che ogni qual volta ci si trova di fronte ad un problema di formazione di *gruppi omogenei* di unità di osservazione, *l'omogeneità* è riferita ai fattori che caratterizzano le unità stesse; infatti, se le unità di osserva-

[9] I problemi di natura inferenziale, cioè i problemi di estensione dei risultati dal campione alla popolazione, non verranno trattati in questa sede.

[10] Si è parlato di eventuale limitazione dell'analisi sia perché non è sempre facile per il ricercatore procedere, a priori, ad una corretta individuazione del sottoinsieme di interesse su cui operare, sia perché esistono, come verrà meglio chiarito in seguito, specifici metodi di raggruppamento che consentono l'esclusione a posteriori di quelle unità che si ritengono non appartenenti alla struttura.

zione risultano *omogenee*, non lo sono in quanto tali ma perché sono caratterizzate da livelli analoghi delle variabili e/o mutabili che le identificano.

Di qui l'enorme rilevanza di una scelta appropriata di tali elementi caratterizzanti, che deve essere fatta in modo tale: *a*) da includere le variabili e/o mutabili con elevato *potere discriminante* rispetto alla *omogeneità* di interesse; *b*) da escludere invece quelle variabili e/o mutabili che, pur possedendo un elevato potere discriminante, non sono utili ai fini della specifica analisi di raggruppamento in atto.

Una volta individuate le variabili e/o mutabili in grando di caratterizzare in modo soddisfacente le unità di osservazione può sorgere l'ulteriore problema della loro ponderazione; può accadere cioè che delle variabili e/o mutabili pure se rilevanti, ai fini della formazione dei gruppi, lo siano meno di altre. In tali situazioni risulterà opportuna una riduzione della loro influenza, attraverso una appropriata ponderazione, in modo da evitare che la struttura di gruppo emergente risenta troppo del loro condizionamento. Alla ponderazione delle variabili (e/o mutabili) si dovrà ricorrere, soprattutto, nei casi in cui le variabili (e/o mutabili) ritenute meno rilevanti possiedono un elevato potere discriminante.

Un ulteriore fatto collegato alla scelta delle variabili (e/o mutabili) è quello della *ponderazione implicita*. Il problema della ponderazione implicita si presenta tutte le volte che, nell'analisi dei gruppi, vengono considerate simultaneamente più variabili (e/o mutabili) che, pur avendo diversa natura, hanno comportamento analogo in quanto caratterizzanti uno stesso aspetto di variabilità o mutabilità. Si vedrà nelle pagine successive come sia possibile eliminare eventuali ponderazioni implicite presenti ma non volute.

c) *Omogeneizzazione delle scale di misura*

S'è detto che per esprimere il *grado di omogeneità* tra unità di osservazione ci si basa, generalmente, su indici di similarità o di diversità che risultano dalla sintesi dei singoli confronti tra le corrispettive modalità assunte dalle variabili (e/o mutabili) in due unità distinte. Evidentemente, se le unità sono caratterizzate sia da variabili che da mutabili la possibilità di sintesi diventa molto difficile se non impossibile; inoltre il problema della sintesi può risultare di difficile soluzione anche quando si ha a che fare con unità caratterizzate da sole variabili se queste sono espresse in unità di misura diverse.

Per rendere possibile la sintesi dei confronti sui singoli caratteri quando ci si trova ad operare in situazioni come quelle sopra delineate, si procede, nella maggioranza dei casi, attraverso un lavoro di omogeneizzazione, sia in termini di scala che di unità di misura, delle variabili e delle mutabili.

d) *Scelta della misura di similarità o diversità*

Si è già avuto modo di sottolineare come molti metodi di analisi dei gruppi prevedano l'uso di opportuni indici di similarità o diversità, ed è fa-

cile intuire anche l'esistenza di una serie molto numerosa di misure che possono rispondere allo scopo. Su alcune misure, tra le innumerevoli proposte, ci intratterremo nelle pagine successive [11]; qui basta osservare che tutte le misure di similarità o diversità interagiscono sempre con il criterio di raggruppamento adottato e che queste, insieme ai caratteri scelti per identificare le unità di osservazione, danno un significato operativo alla dizione « *associazione naturale* » e quindi « *gruppo naturale* ».

e) *Definizione del numero dei gruppi*

Uno dei problemi operativi più delicati che ci si trova ad affrontare quando si opera una analisi di raggruppamento è quello relativo alla definizione del numero dei gruppi da formare.

In seguito si avrà occasione di evidenziare come alcuni metodi di raggruppamento, in genere quelli che rientrano nella categoria dei cosidetti metodi di tipo gerarchico, non richiedono una fissazione a priori del numero di gruppi, sussistendo, per il ricercatore, la possibilità di stabilire a posteriori il numero di gruppi più soddisfacente attraverso l'osservazione della *configurazione completa* della *struttura di gruppo* (da n gruppi formati da una sola unità di osservazione ad un solo gruppo contenente tutte le unità).

In altri casi, e ciò accade generalmente quando si applicano metodi di raggruppamento di tipo non gerarchico, il numero dei gruppi da formare deve essere fissato a priori. In queste situazioni, il miglior modo di procedere è spesso quello di agire per tentativi, ipotizzando diverse configurazioni iniziali, in quanto a numerosità, per poi operare la scelta definitiva a posteriori in base all'osservazione dei diversi risultati ottenuti. Un tale modo di procedere si rende necessario tutte le volte che il ricercatore, e sono casi non infrequenti, opera nella mancanza quasi assoluta di informazioni sulla struttura di gruppo entro la quale collocare le unità di osservazione.

Va detto inoltre che in letteratura sono presenti numerose proposte metodologiche per la determinazione meccanica del numero dei gruppi, si tratta però di procedure che non possono essere applicate senza una preventiva ed accurata valutazione delle loro proprietà e dei loro limiti.

f) *Scelta del criterio di raggruppamento*

Come già accennato in precedenza, la serie dei criteri di raggruppamento che sono stati proposti ed usati, nei più diversi campi della indagine scientifica, è numerosissima. In taluni specifici ambiti di ricerca la scelta del criterio di raggruppamento può risultare molto semplice ed univocamente determinata, in genere, però, la scelta non è affatto semplice ed immediata

[11] La discussione sarà comunque necessariamente limitata. Il lettore che intende approfondire l'argomento può consultare i lavori di GOODMAN e KRUSKAL (1954, 1959), BENEDETTI (1959), SNEATH e SOKAL (1963), LETI (1978).

coesistendo vari criteri, ad un primo esame tutti plausibili ed utilizzabili, che possono dar luogo a dei risultati diversificati ed a volte contraddittori. In tali situazioni, tenendo presente che la scelta del criterio di raggruppamento implica anche la definizione implicita del *concetto di gruppo naturale*, la via migliore da seguire è forse quella di non limitarsi ad una scelta aprioristica di un solo criterio, ma quella di procedere alla applicazione di più criteri, rivelatisi logicamente accettabili ad un esame delle loro proprietà e limiti, rimandando la scelta definitiva del criterio al momento dell'analisi dei risultati ottenuti quando si dispone di un insieme più vasto di informazioni a supporto del processo decisionale stesso.

g) *Scelta dell'algoritmo di classificazione*

Si è già avuto modo di affermare che l'esame di tutte le possibili partizioni di un insieme di unità di osservazione è, molto spesso, un obiettivo al di fuori della portata dei ricercatori, vista l'enorme massa di calcoli che esso implica. Sarà pertanto necessario, nella generalità dei casi, far ricorso a degli algoritmi di classificazione in grado di ridurre la mole dei calcoli e che consentano al tempo stesso, se non l'individuazione della migliore partizione, quanto meno, la derivazione di una soluzione abbastanza prossima a quella ottima. Evidentemente, rappresentando l'algoritmo un semplice ausilio tecnico per pervenire alla formazione dei gruppi, la sua scelta dovrà essere fatta per ultima in ordine di tempo, cioè dopo aver proceduto alla scelta delle unità di osservazione, delle variabili e/o mutabili (opportunamente omogeneizzate ed eventualmente ponderata), della misura di similarità o diversità e del criterio di raggruppamento.

h) *Interpretazione dei risultati*

Superate le fasi sopra elencate resta la fase, più delicata ed importante, della interpretazione dei i risultati ottenuti. Ma problemi connessi a questa fase sono direttamente e strettamente dipendenti dalla natura dello specifico fenomeno oggetto di analisi, se ne potrà quindi discutere soltanto nella parte seconda del saggio quando si procederà alla illustrazione di alcune applicazioni dell'analisi dei gruppi allo studio del comportamento elettorale.

Nei paragrafi successivi si discutera quasi esclusivamente di quegli aspetti dell'analisi dei gruppi di più preminente interesse metodologico; in particolare, si tratterà dei problemi indicati ai punti c), d), f) e g), dell'elenco sopra riportato.

4. Scale di misura.

Quasi tutte le procedure di analisi dei gruppi, per poter essere applicate, richiedono la omogeneità tra le scale di misura adottate per esprimere le modalità dei caratteri rilevati su ciascuna unità di osservazione. Generalmente accade che, come si è avuto modo di sottolineare più volte, la caratterizzazione delle unità di di osservazione avviene considerando sia aspetti quantitativi che aspetti qualitativi. Un modo per affrontare tali situazioni, sempreché non si preferisca ricorrere a procedure particolari di raggruppamento che consentono la trattazione congiunta di variabili e mutabili ([12]), è quello di individuare una particolare scala cui riferirsi operando poi le necessarie trasformazioni fino al raggiungimento della omogeneità tra tutte le variabili e mutabili considerate.

Per quanto concerne le scale di misura si deve osservare che esse vengono usualmente distinte secondo quattro diversi livelli, a seconda dell'ammontare di informazioni richiesto per la loro definizione. In tale ottica di parla di *scale nominali*, *scale ordinali*, *scale di intervallo* e *scale razionali* o *di rapporto*.

La *scala razionale* o *di rapporto* rappresenta il più alto livello di misurazione, ha uno zero assoluto (non arbitrario) e possiede una unità di misura di tipo fisico scelta come elemento comune di riferimento.

La *scala di intervallo* si differenzia dalla scala di rapporto in quanto, pur possedendo una unità di misura di tipo fisico, lo zero in essa contenuto ha natura arbitraria.

La *scala ordinale* consente un ordinamento delle unità di osservazione in funzione dell'entità posseduta di un certo carattere senza che sia possibile però stabilire la misura dell'ammontare del carattere posseduto.

La *scala nominale* costituisce il più basso livello di misurazione. Sotto il profilo formale le scale nominali possiedono unicamente le proprietà di *simmetria* e di *transitività*; da ciò deriva che relativamente alle unità di osservazione classificate secondo una scala nominale si potrà soltanto stabilire se sono uguali o diverse.

L'edificio teorico della *quantificazione* nella forma sopra schematizzata, che è proprio della cosidetta *scuola anglosassone*, appare, (come ha giustamente osservato Herzel (1974)) per ricchezza di contenuto semantico, inferiore a quello adottato dalla *scuola statistica italiana*.

Come è noto, nell'ambito di questa scuola viene introdotta una distinzione preliminare tra *variabili* (caratteri quantitativi) e *mutabili* (caratteri qualitativi). Le *variabili* vengono a loro volta distinte in *intensive* ed *estensive*, mentre le *mutabili* si distinguono in *sconnesse*, *cicliche* e *rettilinee ordinate*; le mutabili rettilinee vengono ulteriormente distinte in *metriche* e *metrizzabili* ([13]).

([12]) Si tratta in genere, di procedure che operano a stadi successivi attraverso analisi di raggruppamento fatte, alternativamente, considerando le sole variabili o le sole mutabili. Utilizzando l'indice proposto da Gower (1972), per la misura della diversità tra unità di osservazione, è possibile trattare simultaneamente variabili e mutabili.

([13]) Chi fosse interessato ad un approfondimento di questo tema si può riferire a Gini (1951), Castellano (1965).

Nonostante la maggiore ricchezza di contenuto dello schema proposto dalla scuola italiana, in questo saggio ci si riferirà in larga misura allo schema anglosassone, risultando quest'ultimo sufficientemente esplicativo ai fini che si vogliono qui perseguire. Conviene comunque osservare che il concetto di variabile (intensiva e estensiva) coincide con quello di scala ad intervallo e scala di rapporto, il concetto di mutabile sconnessa coincide con quello di scala nominale, quello di mutabile rettilinea con quello di scala ordinale, mentre le mutabili cicliche non trovano riscontro né nelle scale ordinali né in quelle nominali; infatti, rispetto alle scale nominali, che possiedono le sole proprietà di simmetria e transitività, le mutabili cicliche possiedono anche la proprietà dell'ordinamento che è però del tutto arbitrario e convenzionale, cosa questa che non avviene invece, almeno in uguale misura, nelle scale ordinali.

Tornando al problema della omogeneizzazione delle scale di misura, e tenendo presente che per la loro definizione occorre utilizzare un diverso ammontare di informazioni, risulta evidente che una qualsiasi trasformazione di scala implicherà necessariamente: *a*) una perdita di informazioni, nei casi in cui si passa da una scala di livello superiore ad una di livello inferiore (ad esempio da una scala di intervallo ad una scala ordinale); *b*) la necessità di informazioni aggiuntive, rispetto a quelle utilizzate per la costruzione di una particolare scala, nei casi in cui si voglia passare da una scala di livello inferiore ad una di livello superiore (ad esempio da una scala nominale ad una scala ordinale).

Quando ci si trova ad operare su variabili e mutabili, uno dei metodi di omogeneizzazione più frequentemente usati, che è in molti casi l'unico possibile, è quello di esprimere tutti i caratteri attraverso una scala di misura ordinale, sostituendo alle *modalità quantitative* i rispettivi *ranghi* ed introducendo un *ordinamento* tra le *modalità qualitative*.

In ogni caso, si deve tener presente che qualunque processo di omogeneizzazione delle scale, che implichi una loro trasformazione, deve essere attuato con la massima cautela, onde evitare effetti negativi troppo marcati sulla formazione dei gruppi dovuti a perdita di informazioni rilevanti o all'inserimento di informazioni aggiuntive non appropriate.

5. Misure di similarità e di diversità.

Come già detto più volte, un elevato numero di criteri di analisi dei gruppi presuppongono la preventiva introduzione di un opportuno indice capace di esprimere numericamente il grado di similarità, o di diversità, esistente tra coppie di unità di osservazione caratterizzate dalla modalità di più fattori.

Per quanto riguarda gli indici di similarità e di diversità si deve innanzi

tutto osservare che, sotto il profilo formale, essi possono essere visti quali *funzioni* che hanno come dominio l'insieme $U \times U$ (il prodotto cartesiano definito sull'insieme delle unità di osservazione U) e come rango l'asse reale. In particolare, qui ci si limiterà a considerare le funzioni il cui rango è il semiasse reale positivo, nel caso degli indici di diversità, e l'intervallo unitario, nel caso degli indici di similarità.

Partendo dalle premesse sopra fatte, si può definire un in*dice di similarità* come qualsiasi funzione (a valori reali $s(.) = \{S: U \times U \rightarrow I_1^+\}$) che soddisfa alle seguenti condizioni

$$\begin{array}{lll} \text{(I)} & O \leq s(U_i, U_j) < 1; & \forall\ U_i \neq U_j \in U \\ \text{(II)} & s(U_i, U_i) = 1; & \forall\ U_i \in U \\ \text{(III)} & s(U_i, U_j) = s(U_j, U_i); & \forall\ U_i, U_j \in U \end{array}$$

La quantità

$$s(U_i, U_j) = s(X_i, X_j) = s_{ij} = s_{ji}$$

(si ricorda che $X_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ è il vettore che identifica la i-esima unità di osservazione U_i) rappresenta pertanto la misura del grado di similarità tra la i-esima e la j-esima unità di osservazione [14].

[14] Nelle pagine successive verranno trattati soltanto alcuni indici di diversità tra quelli più comunemente usati. Chi fosse interessato ad una trattazione più estesa ed approfondita dell'argomento può riferirsi alla letteratura specialistica citata in precedenza. Per comodità del lettore si riporta, comunque, un prospetto dove sono indicate le formule di alcuni indici di similarità per unità di osservazione caratterizzate in modo dicotomico, cioè da unità individuate da vettori contenenti soltanto degli 0 e degli 1 (rispettivamente per evidenziare l'assenza o la presenza di una particolare modalità in relazione a ciascun carattere).

Indice di similarità $s(X_i, X_j)$	Riferimenti bibliografici (ripresi da Duran e Odell (1974))
$\frac{N_{IJ}}{k}$	Russel e Rao (1940)
$\frac{N_{IJ}}{N_{IJ} + N_{iJ} + N_{Ij}}$	Jaccard (1908), Sneath (1957)
$\frac{2\ N_{IJ}}{2N_{IJ} + N_{iJ} + N_{Ij}}$	Dice (1945), Solomon (1970)
$\frac{N_{IJ} + N_{ij}}{k + N_{iJ} + N_{Ij}}$	Rogers e Tanimoto (1960)

I simboli riportati nel prospetto vanno interpretati secondo lo schema seguente:

Si può definire un *indice di diversità* come qualsiasi funzione (a valori reali $d(.) = \{d : U \times U \to R_1^+\}$) che soddisfa alle seguenti condizioni

(I) $\quad d(U_i, U_j) \geq 0; \qquad \forall\ U_i \neq U_j \in U$

(II) $\quad d(U_i, U_i) = 0; \qquad \forall\ U_i \in U$

(III) $\quad d(U_i, U_j) = d(U_j, U_i); \quad \forall\ U_i, U_j \in U$

La quantità

$$d(U_i, U_j) = d(X_i, X_j) = d_{ij} = d_{ji}$$

rappresenta pertanto la misura del grado di diversità tra la i-esima e la j-esima unità di osservazione.

Le definizioni di indici di similarità e di diversità introdotte hanno un carattere molto generale; evidentemente è possibile definire degli indici che non soddisfano le condizioni elencate, o anche introdurre ulteriori condizioni capaci di caratterizzare in modo più specifico gli indici stessi.

Relativamente agli indici di diversità (per gli indici di similarità valgono considerazioni analoghe) si possono introdurre le seguenti ulteriori condizioni

(IV) $\quad d(U_i, U_j) = 0 \Rightarrow U_i = U_j; \ \forall\ U_i, U_j \in U$

(V) $\quad d(U_i, U_j) \leq d(U_i, U_h) + d(U_h, U_j); \ \forall\ U_i, U_j \in U_h \in U$

Le cinque condizioni elencate, il cui soddisfacimento può essere imposto ad un indice di diversità, presentano peculiarità su cui conviene soffermare un attimo l'attenzione.

Sulle prime tre condizioni, dette rispettivamente di *positività*, di *identità* e di *simmetria*, non vale la pena dilungarsi in quanto sono di significato evidente; basta osservare che, usualmente, gli indici di diversità che soddisfano tali condizioni vengono detti *semimetrici* (ISM).

La IV condizione, detta di *nullità*, dice che la misura di diversità tra due unità di osservazione si annulla se e solo se le due unità sono uguali, cioè se assumono le stesse modalità in corrispondenza degli stessi caratteri.

Frequenze assolute di accoppiamento tra le unità di osservazione di tipo dicotomico U_I e U_J

X_j \ X_i	$= 0$	$= 1$	
$= 0$	N_{ij}	N_{Ij}	N_j
$= 1$	N_{iJ}	N_{IJ}	N_J
	N_i	N_I	

La V condizione, detta di *triangolarità*, dice che la misura della diversità tra due unità di osservazione non può mai superare la somma delle misure di diversità tra ciascuna delle due unità ed una terza unità di osservazione.

Gli indici di diversità che, oltre a soddisfare le condizioni di positività, identità e simmetria (indici semimetrici), soddisfano anche la condizione di triangolarità vengono detti *metrici* (IM), ed in tali casi si preferisce parlare di *distanza* piuttosto che di diversità ([15]).

Pure se le cinque condizioni elencate sono di portata del tutto generale (prescindono cioè dal tipo di scala di misura adottata per esprimere i caratteri che identificano le varie unità di osservazione), quando si parla di *indici metrici*, o più semplicemente di *metriche* o di *distanze*, ci si riferisce usualmente ad unità di osservazione i cui caratteri distintivi sono espressi, quanto meno, con scale di intervallo, ci si riferisce cioè a delle unità U_i i cui vettori associati $X_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ possono essere visti come punti di uno spazio euclideo a k dimensioni. In tali situazioni, uno degli indici di distanza più comunemente usati, almeno in alcune specificazioni particolari, è la cosidetta *metrica di Minkowski*.

Se con x_{ih} e x_{jh} si indicano i valori reali assunti dall'h-esima variabile (carattere), rispettivamente nella i-esima e nella j-esima unità di osservazione, la *distanza*, secondo Minkowski, tra le unità U_i e U_j è espressa dalla relazione

$$d(U_i, U_j) = d_r(X_i, X_j) = \left[\sum_{h=1}^{k} | x_{ih} - x_{jh} |^r \right]^{1/r}$$

Ponendo $r = 1$ si ottiene la cosidetta *metrica city-block* (JOHNSON a WALL (1969)), o metrica L_1, o norma L_1

$$d_1(U_i, U_j) = d_1(X_i, X_j) = \sum_{h=1}^{k} | x_{ih} - x_{jh} |$$

Per $r = 2$ si ottiene la notissima *distanza euclidea* o metrica L_2

$$d_2(U_i, U_j) = d_2(X_i, X_j) = \sqrt{\sum_{h=1}^{k} (x_{ih} - x_{jh})} .$$

Calcolando il limite per $r \to +\infty$ si ottiene la *distanza* di *Chebychev*, o metrica L_∞, o norma-sup

$$d(U_i, U_j) = d(X_i, X_j) = \max_{h=1,2,\ldots,k} \{| x_{ih} - x_{jh} |\} \ ([16])$$

([15]) Due ulteriori condizioni, che non sono state qui considerate e che possono risultare di interesse, sono la cosiddetta condizione *ultrametrica* e la condizione di *coerenza*. Il lettore che volesse approfondire questo punto può consultare ZANI (1975). Nel lavoro di questo autore vengono svolte anche alcune considerazioni sulle relazioni esistenti tra le varie condizioni cui gli indici di diversità possono soddisfare.

([16]) Tra le varie specificazioni della metrica di Minkowski vale la seguente relazione di disuguaglianza (in uno spazio euclidico a k dimensioni)

$$d_r(X_i, X_j) \leq d_s(X_i, X_j) \ \forall\ X_i, X_j \in R_k$$

se e solo se $r \geq s$.

Naturalmente è possibile considerare altre particolari specifiche della metrica di Minkowski; si tratta, per lo più, di espressioni d'interesse teorico presentandosi nella realtà pochissimi casi in cui sussistono le condizioni per una loro applicabilità.

Vi sono almeno tre caratteristiche, piuttosto restrittive, delle metriche del tipo Minkowski da tener presenti ogni qual volta s'intenda procedere ad una loro utilizzazione:

a) la dipendenza diretta dall'unità di misura adottata per esprimere le variabili d'interesse;

b) la linearità dell'influenza delle variabili stesse;

c) le variabili vengono trattate come se fossero tra loro indipendenti.

Queste tre caratteristiche assumono, pertanto, la veste di vincoli che in molti casi si rivelano inaccettabili. Può, ad esempio, accadere: *a*) di dover sommare litri a chili, cosa questa invero poco sensata, oppure di dover trattare variabili espresse nella stessa unità di misura ma con livelli molto differenziati, in tali casi le differenze riscontrate tra le variabili ad elevato valore incideranno in modo rilevante nel determinare la misura della distanza tra unità e ciò potrebbe risultare non opportuno; *b*) di sapere che l'effetto di una particolare variabile sulla *omogeneità* non è di tipo lineare, ed in tali situazioni si dovrebbe utilizzare l'informazione modificando in modo opportuno la formula della distanza; *c*) può accadere infine, come generalmente avviene, che le variabili non risultino tra loro indipendenti, il considerarle tali introdurrebbe notevoli elementi di distorsione (come l'introduzione di una ponderazione implicita) nel calcolo della distanza tra unità e, soprattutto, nella misura del grado di omogeneità nei gruppi.

Un modo per trattare i problemi che sorgono quando le variabili sono espresse con diversa unità di misura, e/o hanno campi di variabilità molto differenziati, è quello di procedere alla introduzione di un processo di omogeneizzazione delle variabili stesse da attuarsi attraverso una loro opportuna ponderazione. Se con $p_h (\geq 0)$ si indica il peso da attribuire alla h-esima variabile, la metrica di Minkowski assume la forma

$$d_r^* (X_i , X_j) = \left[\sum_{h=1}^{k} (p_h \mid x_{ih} - x_{jh} \mid)^r \right]^{1/r}$$

Se poi si considera quest'ultima espressione, si osserva come la ponderazione delle variabili possa ridursi effettivamente in un processo di omogeneizzazione delle stesse. Infatti se si pone, ad esempio,

$$p_h = \frac{1}{\sqrt{\sum_{i=1}^{n} (x_{ih} - \bar{x}_h)^2 / n}} = \frac{1}{\sigma_h}$$

dove $\bar{x}_h = \frac{1}{n} \sum_{i=1}^{n} x_{ih}$, si vede immediatamente come tutte le variabili risultino indipendenti dall'unità di misura originariamente utilizzata in quanto espresse dalle nuove unità di misura rappresentata dagli scostamenti quadratici medi σ_h $(h = 1, 2, \ldots, k)$. Osservando inoltre che

$$d_r^* (X_i , X_j) = \left[\sum_{h=1}^{k} \left(\frac{1}{\sigma_h} \mid x_{ih} - x_{jh} \mid \right)^r \right]^{1/r} =$$

$$= \left(\sum_{h=1}^{k} \left| \frac{x_{ih} - \bar{x}_h}{\sigma_h} - \frac{x_{jh} - \bar{x}_h}{\sigma_h} \right|^r \right)^{1/r}$$

si vede anche come la ponderazione, in questo caso, si risolva in un processo di omogeneizzazione delle variabili attraverso standardizzazione, tutte le variabili trasformate avranno cioè stessa media nulla e stessa varianza pari ad uno [17].

Lo scopo principale del processo di ponderazione sopra illustrato è quello di procedere alla omogeneizzazione delle variabili in quanto a unità di misura e campo di variabilità. Può accadere però, come si è già avuto modo di sottolineare in precedenza, che il ricercatore sia interessato ad una attribuzione di pesi in senso proprio; può accadere cioè di dover trattare delle variabili *ritenute* più *significative* o *rappresentative* di altre avendo queste un più elevato potere *discriminatorio* nei confronti del *concetto* di *associazione naturale* tra unità di osservazione. In tali casi, non esistendo in genere validi elementi di informazione sulla *struttura* di gruppo esistente nella popolazione, si dovrà procedere ad una attribuzione soggettiva di pesi, eventualmente da rivedere in funzione dei risultati ottenuti. Pure nella sua soggettività, un tale modo di procedere è l'unico scientificamente valido risultando tuttaffatto significative le procedure di ponderazione meccanica sino ad oggi suggerite.

Natura sostanzialmente soggettiva ha anche l'eventuale introduzione di non linearità nei confronti di variabili specifiche, pure se in questo caso è possibile il ricorso ad espedienti tecnici particolari che facilitano il compito del ricercatore. Disponendo di informazioni sufficienti in merito alla possibile influenza di tipo non lineare delle variabili, la metrica di Minkowski assume la forma

$$d_r^{**} (X_i , X_j) = \left[\sum_{h=1}^{k} (p_h \mid f_h (x_{ih}) - f_h (x_{jh}) \mid)^r \right]^{1/r}$$

dove le funzioni $f_h (.)$ dovranno essere opportunatamente specificate.

Situazione del tutto diversa, rispetto alle due sopra prospettate, è quella che si presenta quando si deve procedere alla eliminazione di eventuali ponderazioni implicite presenti nei dati, ma non volute.

Si è già detto che ci si prova in presenza di ponderazioni implicite tutte

[17] Un diverso sistema di pesi potrebbe essere quello basato sul campo di variazione

$$p_h = \frac{1}{\max_{i,j} \mid x_{ih} - x_{jh} \mid}$$

le volte che le variabili risultano tra loro interrelate. In tali casi sarà possibile, attraverso un processo meccanico di trasformazione, passare dall'insieme delle k variabili interrelate ad un diverso insieme di s ($s \leq k$) variabili linearmente indipendenti (*ortogonali*).

Uno dei metodi più comunemente usati per la derivazione di un insieme di variabili ortogonali, partendo da un insieme dato, è quello delle *componenti principali*.

Le variabili ortogonali che si ottengono applicando il metodo delle componenti principali risultano essere delle combinazioni lineari — che soddisfano certe proprietà — delle variabili originarie.

In particolare, se con Y_j ($j = 1, 2, \ldots, s$) si rappresenta la nuova variabile (j-esima componente principale) essa resta definita dalla formula

$$Y_j = \sum_{h=1}^{k} X\, a_{jh} = X\, A_j$$

dove i vettori A_j, di dimensione k x 1, sono tali che

$$A'_i\, A_j = \begin{cases} 1 & \text{per } i = j \\ 0 & \text{per } i \neq j \end{cases}$$

inoltre: la prima componente principale si ottiene dalla combinazione lineare normalizzata delle variabili originarie che spiega il massimo della varianza totale presente nei dati ([18]) e sarà pertanto la distribuzione Y_1, tra tutte quelle ottenibili come combinazioni lineari normalizzate delle distribuzioni X_j, che ha varianza più elevata; la seconda componente principale è data dalla combinazione lineare normalizzata delle variabili originarie, tra tutte quelle che risultano ortogonali alla combinazione che ha dato luogo alla prima componente principale, che spiega il massimo della varianza totale residua, ecc. Le proprietà di ortogonalità e di massimizzazione della varianza spiegata definiscono in modo univoco le componenti principali.

Si vede chiaramente come la trasformazione in componenti principali si riduca in effetti ad una proiezione dello spazio originario a k dimensioni in uno spazio ortogonale a s dimensioni, dove il valore di s dipende dalla caratteristica o rango della matrice X. Tenuto conto poi del modo con cui le componenti principali vengono ottenute si vede anche come il ricercatore possa, qualora lo ritenga opportuno, limitarsi a considerare soltanto un sottoinsieme di componenti principali, ad esempio le prima r ($r < s$) se queste spiegano una quota della varianza totale sufficientemente elevata.

([18]) La varianza totale presente nei dati originari è definita come somma delle varianze $\sigma_h^2 \left[\sigma_h^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ih} - \bar{x}_h)^2\right]$ relative alle k variabili x_i

$$\text{Var}(T) = \text{Var}(x) = \sum_{h=1}^{k} \text{Var}(x_h) = \sum_{h=1}^{k} \sigma^2{}_h$$

Il lettore interessato ad un approfondimento della conoscenza sul metodo delle componenti principali può consultare, tra gli altri, ANDERSON (1958) o MORRISON (1967).

L'applicazione del metodo delle componenti principali implica il computo di combinazioni delle variabili originarie; evidentemente se le variabili sono espresse con unità di misura diversa, o hanno campi di variabilità molto differenziati, le combinazioni stesse potrebbero avere scarso significato, potrebbe allora risultare conveniente procedere ad preventiva omogeneizzazione delle variabili originarie rapportandole, ad esempio, al proprio scostamento quadratico medio. Le soluzioni che si ottengono applicando il metodo delle componenti principali alle variabili originarie differiscono da quelle che si ottengono considerando le variabili relativizzate con lo scostamento quadratico medio.

Infatti, nel primo caso la forma quadratica da massimizzare è caratterizzata dalla matrice delle varianze e covarianze, mentre nel secondo caso la matrice caratteristica è quella dei coeficienti di correlazione. Questo fatto evidenzia anche la non invarianza del metodo delle componenti principali rispetto a cambiamenti di unità di misura.

Il ricercatore nell'applicare un tale metodo dovrà pertanto decidere preventivamente sul tipo di misura che intende adottare per esprimere le variabili.

Una volta calcolate le componenti principali Y_i, il ricercatore potrà, se lo ritiene opportuno, utilizzare la metrica di Minkowski, nella formulazione classica $d_r(Y_i, Y_j)$ o in quella ponderata $d_r^*(Y_i, Y_j)$, per misurare la *distanza* tra unità di osservazioni, considerando tutte le s componenti principali o soltanto le prime r.

Tra gli indici di distanza esaminati, quelli più comunemente usati sono:

1) Il quadrato della distanza euclidea calcolata sui valori originari assunti dalle variabili x_h

$$d_2^2(X_i, X_j) = \sum_{h=1}^{k} (x_{ih} - x_{jh})^2$$

2) Il quadrato della distanza euclidea calcolata sui valori standardizzati

$$d_2^{*2}(X_i, X_j) = \sum_{h=1}^{k} \left(\frac{x_{ih} - \bar{x}_h}{\sigma_h} - \frac{x_{jh} - \bar{x}_h}{\sigma_h} \right)^2$$

3) Il quadrato della distanza euclidea calcolata sui valori assunti dalle componenti principali (tutte o in parte) ottenute dalle variabili originarie

$$\delta_2^2(Y_i, Y_j) = \sum_{h=1}^{s} (y_{ih} - y_{jh})^2$$

4) Il quadrato della distanza euclidea calcolata sui valori assunti dalle variabili relativizzate

$$\delta_2^{*2}(Y_i^*, Y_j^*) = \sum_{h=1}^{s} (y^*{}_{ih} - y^*{}_{jh})^2$$

Si è discusso in precedenza sulle ragioni che possono consigliare la scelta di indici di distanza calcolati sulle componenti principali ([19]), esse risiedono, soprattutto, nella necessità di eliminare eventuali ponderazioni implicite presenti nei dati ma non volute.

Si è aggiunto però che tali indici non sono invarianti rispetto a cambiamenti dell'unità di misura imponendo così dei problemi di scelta che non sono sempre di facile soluzione. Può accadere cioè che il ricercatore non disponga di elementi di informazione sufficienti per poter operare una valida scelta tra il calcolo delle componenti principali sulle variabili originarie o quello sulle variabili relativizzate (non va dimenticato, peraltro, che un tale problema sussiste anche nel caso in cui le distanze vengano calcolate sulle variabili trasformate in componenti principali).

Per evitare un tale problema è stato proposto il ricorso ad un particolare indice di diversità, la cosidetta *distanza euclidea generalizzata o distanza di Mahalanobis*

$$\Delta^2 (X_i , X_j) = (X_i - X_j)\, W^{-1} (X_i - X_j)'$$

dove W^{-1} rappresenta l'inversa della matrice delle varianze e covarianze delle variabili x_h.

L'indice $\Delta^2 (X_i , X_j)$, pur essendo invariante rispetto ad una qualsiasi trasformazione lineare non singolare delle variabili originarie x_h, consente la eliminazione di ponderazioni implicite non volute attraverso una compensazione della correlazione esistente tra le variabili stesse ([20]).

([19]) Una interessante applicazione dell'analisi dei gruppi basata sul calcolo delle componenti principali, e limitata all'uso delle sole due prime componenti, rispetto a 14 variabili originarie, è stata fatta da GREEN, FRANK e ROBINSON (1967). Altre analisi dei gruppi basate su un numero ridotto di componenti principali sono state condotte da BIGGERI (1977), ZANI e SICURI (1978), CHIANDOTTO, GHILARDI e LEONI (1978).

([20]) Per inserire una ponderazione esplicita delle variabili anche nel caso in cui s'intenda utilizzare la *distanza di Mahalanobis* si dovrà fare ricorso ad una matrice diagonale del tipo

$$P = \begin{bmatrix} p_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 & \dots & 0 \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ 0 & 0 & \dots & p_h & \dots & 0 \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ 0 & 0 & \dots & 0 & \dots & p_k \end{bmatrix}$$

(dove p_h ($h = 1, 2, \dots, k$) rappresenta il peso — importanza — che si vuol attribuire alla h-esima variabile) e calcolare la distanza tra l'unità i-esima e l'unità j-esima mediante la formula

$$\Delta^2 (X_i X_j) = (X_i - X_j)\, P\, W^{-1}\, P\, (X_i - X_j)'$$

Conviene osservare comunque che le distanze del tipo δ, conservano, rispetto alle distanze del tipo Δ, il vantaggio della possibilità di riduzione della dimensione dello spazio di riferimento, anche se ciò comporta necessariamente una perdita di informazioni; fatto questo che acquista una rilevanza del tutto particolare nei casi in cui risulti sufficiente l'uso delle sole due prime componenti principali. Infatti, la limitazione alle due prime componenti principali, sempreché queste riescono a spiegare una quota abbastanza elevata della *varianza totale*, permette una rappresentazione grafica dei punti di osservazione facilitando così, in modo notevole, la risoluzione di alcuni problemi cui si è avuto occasione di accennare in precedenza, quali quello della fissazione del numero dei gruppi.

Gli indici di diversità, o in particolare di distanza, fin qui considerati hanno tutti, almeno in origine, natura metrica. All'argomento è stato dato ampio spazio, sia per la sua rilevanza intrinseca, sia perché ha consentito una discussione abbastanza dettagliata dei problemi di ordine generale che sempre si presentano al momento della scelta dell'indice più appropriato per la misura della *omogeneità* tra unità di osservazione.

Evidentemente, anche nel caso (ed è quello su cui si è soffermata l'attenzione) in cui si abbia a che fare con unità individuate da caratteri espressi con scale di intervallo o di rapporto può risultare conveniente il ricorso a misure che non hanno natura metrica quali, ad esempio, l'indice (LANCE e WILLIAMS (1966))

$$D (X_i , X_j) = \frac{\sum_{h=1}^{k} | x_{ih} - x_{jh} |}{\sum_{h=1}^{k} (x_{ih} + x_{jh})}$$

od anche

$$D^* (X_i , X_j) = \sum_{h=1}^{k} \frac{| x_{ih} - x_{jh} |}{(x_{ih} + x_{ih})}$$

Si controlla facilmente come questi due indici non soddisfino la condizione di triangolarità, risultando modificati i denominatori delle espressioni che li definiscono al variare delle unità poste a confronto. Il significato d'attribuire a tali indici (il cui uso dovrebbe essere limitato ai soli casi in cui le variabili assumono valori maggiori o uguali a zero se si vuole rispettare, quanto meno, la condizione di non negatività) è abbastanza ovvio il numeratore coincide con la metrica L_1, mentre il denominatore può essere inteso come elemento inverso di ponderazione, totale delle unità nel primo indice, individuale per le singole variabili nel secondo indice.

Altre misure di similarità o diversità, tra unità di osservazione, non metriche sono quelle basate sul concetto di *entropia* e quelle che si rifanno a indici statistici di associazione ormai classici nell'ambito statistico, quali le

misure di correlazione e quelle che si ricollegano, più o meno direttamente, al χ^2 [21].

A conclusione di questo paragrafo, dedicato alle misure di similarità e di diversità, conviene riprendere alcune annotazioni già fatte nel corso della esposizione e svolgere alcune brevi considerazioni aggiuntive di carattere generale sull'argomento [22].

Va innanzi tutto sottolineato che i risultati di una qualunque analisi dei gruppi dipendono strettamente dall'indice usato per misurare l'*omogeneità* tra unità di osservazione: uno stesso criterio di raggruppamento basato su differenti indici di similarità o diversità può fornire dei risultati completamente diversi. Qualunque ricercatore che intende effettuare un'analisi dei gruppi deve essere cosciente di un tale fatto, così come deve essere cosciente del fatto che non esistono regole generali di comportamento cui si possa riferire per operare la scelta dell'indice più appropriato [23]. Scelta che viene invece suggerita dallo specifico problema di raggruppamento in esame e dai fini che attraverso il raggruppamento stesso si vogliono perseguire, ed è, inoltre, fortemente condizionata dal tipo di variabili e/o mutabili considerate. Nonostante le osservazioni fatte, è possibile raggruppare, a seconda della natura e dell'utilizzo, in quattro grandi categorie gli indici di similarità e di diversità più frequentemente usati:

1) Misure del tipo *distanza* (la metrica di Minkowski appartiene a questa categoria), utilizzate soprattutto quando si ha a che fare con caratteri espressi in scale di rapporto o di intervallo ma che sono frequentemente usate anche per trattare caratteri ordinali o caratteri nominali dicotomici.

2) Misure di *associazione* che risultano particolarmente indicate per risolvere problemi di raggruppamento relativi ad unità di osservazione caratterizzate da mutabili.

3) Misure di *correlazione* (r di Bravais-Pearson, τ di Kendall, ecc....), utilizzate soprattutto nei casi in cui si vuol procedere alla formazione di gruppi di variabili o mutabili ed il cui uso è invece da evitare nelle analisi dei gruppi di unità di osservazione.

4) Misure di similarità e diversità di tipo *probabilistico* che si differenziano dalle misure rientranti nelle categorie sopra elencate (che possono

[21] Indicazioni utili sul significato dell'entropia e su possibili usi nell'ambito dell'analisi dei gruppi si trovano in JARDINE e SIBSON (1971). Una esposizione sufficientemente dettagliata sulla metrica del χ^2 si trova in BENZÉCRI (1973).

[22] Una discussione ampia e dettagliata su questo punto si trova in CORMACK (1971), e SNEATH e SOKAL (1973).

[23] Fino ad oggi sono stati condotti pochissimi studi comparativi sulla bontà relativa dei vari indici e nessuno è riuscito a fornire una risposta valida di portata sufficientemente generale. Si tratta, per lo più, di studi svolti nell'ambito biologico ma, come si è già avuto modo di osservare, quello che risulta valido nel campo delle scienze biologiche non sempre rimane tale nel campo delle scienze non biologiche.

essere intese come tradizionali) in quanto non più basate sulla misura della similarità o diversità tra unità ma sulla *valutazione* della perdita o del guadagno di *informazione* (statistica) che comporta la fusione di due unità di osservazione.

6. CRITERI E ALGORITMI DI RAGGRUPPAMENTO

I criteri di analisi dei gruppi più comunemente usati — che sono stati proposti, nella generalità dei casi, direttamente in forma algoritmica — vengono ugualmente distinti in *gerarchici* e *non gerarchici* [24]. Nella prima classe rientrano quei criteri che prevedono la individuazione di n partizioni (cioè tante quante sono le unità di osservazione) ciascuna caratterizzata da un diverso numero m di gruppi, con m = 1, 2, ..., n; le n partizioni individuate vanno a costituire una struttura gerarchica di raggruppamento. Appartengono invece alla seconda classe tutti quei criteri che prevedono l'individuazione di una sola partizione [25] delle unità di osservazione in m gruppi, dove m può essere fissato a priori o derivare dal processo di raggruppamento stesso.

I criteri di tipo gerarchico vengono ulteriormente distinti in *divisivi* e *agglomerativi* [26]. Nell'ambito dei criteri di analisi dei gruppi di tipo divisivo, il processo di raggruppamento inizia considerando tutte le unità come appartenenti ad un solo gruppo ed opera una suddivisione in due gruppi secondo una regola predeterminata, ad esempio minimizzando una qualche misura della variabilità interna ai due gruppi. Il passo successivo è quello della suddivisione di uno dei due gruppi secondo la stessa regola e così via, fino a quando si perviene alla formazione di n gruppi ciascuno dei quali contiene una sola unità di osservazione.

Nell'ambito dei criteri di analisi dei gruppi di tipo agglomerativo, il processo di raggruppamento inizia considerando n gruppi ciascuno formato

[24] In questo saggio non verrà svolta un'analisi approfondita di tutte le dimensioni lungo le quali possono svilupparsi i diversi criteri e algoritmi di analisi dei gruppi, nè verrà tentata l'individuazione di una tipologia generale per un loro inquadramento. La discussione sarà necessariamente limitata e verterà sulle dimensioni che appaiono più importanti e significative. Il lettore interessato ad un approfondimento del tema può consultare, tra gli altri, i lavori di CORMACK (1971), BAILEY (1975), GOOD (1965, 1977).

[25] Si vedrà nelle pagine successive come anche in corrispondenza di particolari criteri di raggruppamento di tipo non gerarchico vengano individuate più partizioni; ma queste non vanno a costituire una struttura gerarchica di raggruppamento ed hanno natura strumentale ai fini della determinazione del numero ottimale m di gruppi rispetto al quale verrà individuata la partizione « *migliore* ».

[26] La distinzione dei criteri di analisi dei gruppi in divisivi e agglomerativi resta significativa anche nell'ambito dei criteri di tipo non gerarchico; ma in tale contesto la sua rilevanza è del tutto marginale e non verrà pertanto più ripresa in considerazione.

da una sola unità di osservazione ed opera l'aggregazione delle due unità più simili (o meno diverse). Il passo successivo è ancora costituito dalla fusione dei due gruppi più simili (tra gli n-1 gruppi esistenti); ciò può accadere combinando due gruppi formati da una sola unità oppure associando una terza unità di osservazione al gruppo di due unità formato al passo precedente. Il processo continua fino a quando si arriva alla formazione di un solo gruppo contenente tutte le unità di osservazione.

I risultati delle analisi dei gruppi di tipo gerarchico possono essere utilmente rappresentati con degli *alberi* (*dendrogrammi*). L'albero ottenuto costituisce la *struttura gerarchica* di raggruppamento, le unità, i gruppi di unità, i gruppi dei gruppi di unità rappresentano i *nodi* dell'albero. Nella fig. 1

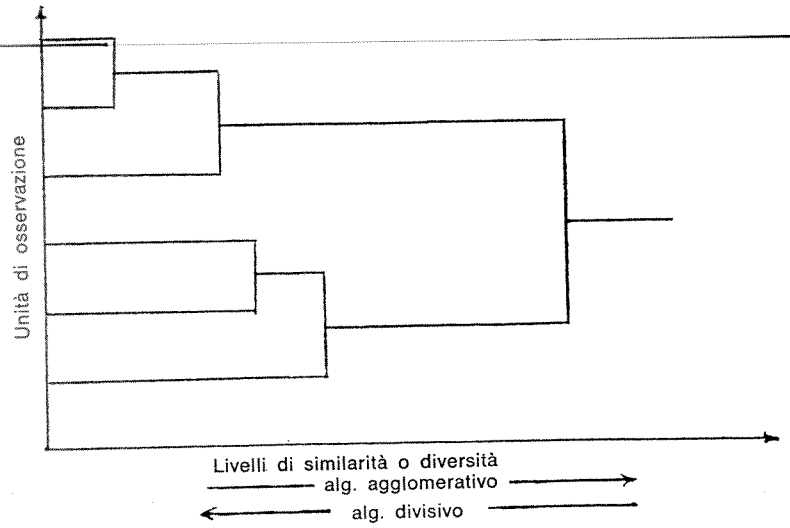


Fig. 1. Struttura gerarchica di raggruppamento (*albero* o *dendrogramma*).

viene riportato un albero ipotetico, dove sull'asse delle ascisse sono indicati i livelli di similarità o diversità rispetto ai quali due gruppi (che possono anche essere formati da singole unità) si fondono per dar luogo ad un nuovo gruppo.

L'albero può essere *tagliato* verticalmente a vari livelli; in corrispondenza di ciascun taglio si osserverà una partizione dell'insieme iniziale costituito dalle n unità di osservazione. Il problema del taglio dell'albero viene risolto, generalmente, in modo soggettivo ([27]) e dipende dalla natura del

([27]) Al fine di superare il carattere soggettivo connesso alla scelta del livello rispetto al quale operare il taglio dell'albero sono state proposte diverse procedure. Si tratta però di procedure che, nella generalità dei casi, hanno una validità limitata a situazioni molto

fenomeno che si sta studiando. Risulta spesso conveniente operare più tagli per procedere all'analisi simultanea di più partizioni in vista dell'eventuale soddisfacimento di fini differenziati che attraverso la ricerca stessa si vogliono perseguire. Può accadere infine che il ricercatore, per i propri fini, proceda ad una analisi completa dell'intera struttura gerarchica senza che questa comporti necessariamente un taglio dell'albero.

Si rileva immediatamente da quanto detto sopra che per i criteri di tipo gerarchico non si pone, almeno a priori, il problema della fissazione del numero m di gruppi che vanno a costituire la partizione; infatti, il numero m è una diretta conseguenza del taglio dell'albero.

Un ulteriore aspetto interessante da sottolineare riguardo ai criteri di tipo gerarchico è che essi non implicano necessariamente la massimizzazione o minimizzazione di una qualche funzione legata alla struttura gerarchica nella sua totalità ma si svolgono attraverso un processo di ottimizzazione a più stadi del cammino seguito per la formazione dei gruppi.

Tornando alla distinzione dei criteri gerarchici divisivi dai criteri gerarchici agglomerativi, si deve rilevare che il primo tipo di criteri, pur richiedendo una mole di calcoli notevolmente superiore a quella richiesta dai criteri del secondo tipo, non producono, generalmente, risultati più soddisfacenti. Nelle pagine successive l'esposizione verrà pertanto limitata ai soli criteri gerarchici di tipo agglomerativo; naturalmente, analoghe considerazioni possono essere svolte, *mutatis mutandis*, nei confronti dei criteri gerarchici di tipo divisivo.

6.1. *Criteri gerarchici di raggruppamento di tipo agglomerativo*

I processi di ottimizzazione che caratterizzano i *criteri gerarchici* di *tipo agglomerativo* operano, generalmente, su di una matrice (detta di similarità o diversità) ottenuta considerando tutti i valori degli indici di similarità o diversità calcolati su tutte le coppie di unità di osservazione. Se si sceglie, ad esempio, il quadrato della *distanza euclidea*

$$d_2^2(U_i, U_j) = d_2^2(X_i, X_j) = \sum_{h=1}^{k} (x_{ih} - x_{jh})^2 = d_{ij} \text{ per } i, j = 1, 2, \ldots, n$$

per esprimere il grado di diversità tra le coppie di unità di osservazione, ne risulterà una *matrice* (di diversità — *delle distanze*, in questo particolare caso —)

particolari e/o non superano completamente il fatto soggettivo; fornendo le tecniche stesse soltanto informazioni atte a facilitare ed indirizzare la scelta del ricercatore che rimane sostanzialmente soggettiva. Tra le varie procedure atte allo scopo, quelle che appaiono di portata più ampia sono state suggerite da Beale (1969) da Calinski e Harabasz (1971) e da Mojena (1977).

simmetrica di ordine n x n del tipo:

TAB. 1 - *Matrice delle distanze tra unità di osservazione.*

	U_1	U_2	U_3	.....	U_i	.....	U_j	.....	U_n
U_1	0	d_{12}	d_{13}	.....	d_{1i}	.....	d_{1j}	.....	d_{1n}
U_2		0	d_{23}	.....	d_{2i}	.....	d_{2j}	.....	d_{2n}
U_3			0	.....	d_{3i}	.....	d_{3j}	.....	d_{3n}
.					.		.		.
.					.		.		.
.					.		.		.
U_i					0	.....	d_{ij}	.....	d_{in}
.							.		.
.							.		.
U_j							0	.....	d_{jn}
.									.
.									.
.									.
U_n									0

Operando sui valori riportati nella matrice delle distanze, si procede alla formazione dei gruppi aggregando le due unità di osservazione più *vicine* attraverso la individuazione della distanza minima. Se si suppone, per semplicità e senza perdere in generalità, che le unità più vicine sono U_1 e U_2, si potrà calcolare la matrice delle distanze di ordine $(n-1) \times (n-1)$ che risulta dopo aver operato il primo passo del processo di raggruppamento (si veda Tab. 2). Per poter effettuare il passo successivo risulta necessario definire la distanza tra il gruppo formato da due unità ed i gruppi costituiti da una sola unità di osservazione, mentre al terzo passo del processo di raggruppamento si dovrà introdurre la distanza tra gruppi formati da più unità; infatti, dopo il secondo passo, degli $(n-2)$ gruppi formati, $(n-4)$ sono costituiti da una sola unità mentre i restanti due possono essere costituiti da due unità ciascuno.

Qualunque sia il tipo di distanza fra gruppi cui s'intende far ricorso, dopo che sono stati effettuati s passi del processo di raggruppamento, la matrice delle distanze assumerà la forma riportata nella Tab. 3 dove con $G_1, G_2, \ldots, G_{n-s}$, si sono voluti indicare i gruppi formati dopo s passi del processo e con $d(G_i, G_j)$ la misura della distanza tra i gruppi.

Si vede chiaramente come la definizione del criterio di ottimalità passi attraverso la specifica di un indice capace d'esprimere la similarità o diversità tra gruppi; ma, a prescindere dai problemi connessi alla scelta dell'indice

TAB. 2 - *Matrice delle distanze dopo l'effettuazione del primo passo del processo di raggruppamento.*

	$\{U_1\ U_2\}$	U_3	.....	U_i	.....	U_j	.....	U_n
$\{U_1\ U_2\}$	0	$d_{(12)3}$	.....	$d_{(12)i}$	.....	$d_{(12)j}$	.....	$d_{(12)n}$
U_3		0	.....	d_{3i}	.....	d_{3j}	.....	d_{3n}
.				.		.		.
.				.		.		.
.				.		.		.
U_i				0	.....	d_{ij}	.....	d_{in}
.						.		.
.						.		.
.						.		.
U_j						0	.....	d_{jn}
.								.
.								.
.								.
U_n								0

TAB. 3 - *Matrice delle distanze dopo l'effettuazione di s passi del processo di raggruppamento.*

	G_1	G_2	.....	G_i	.....	G_{n-s}
G_1	0	$d(G_1, G_2)$	.....	$d(G_1, G_i)$	.....	$d(G_1, G_{n-s})$
G_2		0	.....	$d(G_2, G_i)$	.....	$d(G_2, G_{n-s})$
.				.		.
.				.		.
.				.		.
G_i				0	.....	$d(G_i, G_{n-s})$
.						.
.						.
G_{n-s}						0

più opportuno, la procedura che caratterizza i criteri gerarchici di tipo agglomerativo si sviluppa secondo tre fasi distinte che possono essere riassunte, facendo riferimento alla rappresentazione tramite albero, nel modo che segue.

I fase: Attribuzione di ciascuna unità di osservazione ai *nodi terminali* dell'albero e calcolo della « *distanza* » tra tutte le coppie di nodi.

II fase: Aggiunta di un nuovo modo in corrispondenza della coppia di nodi più *vicini*. Il livello a cui viene posto il nuovo nodo è determinato dalla distanza tra i due nodi che lo hanno formato.

III fase: Considerazione del nuovo nodo come terminale e ripetizione della seconda parte della I fase, della II fase e della III fase fino a quando i nodi terminali si riducono ad uno soltanto.

In termini formali, e riprendendo le espressioni algebriche introdotte nel paragrafo 2, si avrà a che fare con una funzione del tipo

$$g\,[\pi\,(s)\,;\,X] = g\,[\pi_{\alpha_1}(s)\,,\,\pi_{\alpha_2}(s),\,\ldots,\,\pi_{\alpha_n}(s)\,;\;\;X_1,\;X_2,\,\ldots,\,X_n]$$

per

$$s = n-1\,,\;\;n-2,\;\ldots,\;1$$

(dove gli α_i ($i = 1, 2, \ldots, n$) possono assumere i valori interi da 1 a s) che dovrà essere massimizzata o minimizzata per ogni valore di s.

I criteri gerarchici di tipo agglomerativo che verranno considerati nelle pagine successive si distinguono tra di loro non per come si svolgono (la procedura comune è quella sopra tracciata) ma perché basati su misure di diversità tra gruppi aventi natura diversa.

Dei vari indici proposti per la misura della diversità tra gruppi [28] molti, tra quelli più comunemente usati, soddisfano la formula ricorrente [29].

$$d\,(G_k\,,\,G_{ij}) = \alpha_i\;d\,(G_k\,,\,G_i) + \alpha_j\;d\,(G_k\,,\,G_j) + \beta\;d\,(G_i\,,\,G_j) + \gamma\;|\,d\,(G_k\,,\,G_i) - d\,(G_k\,,\,G_j)\,|$$

dove: $d\,(G_k\,,\,G_{ij})$ rappresenta la diversità tra il gruppo G_k ed il gruppo G_{ij} ottenuto dalla fusione dei gruppi G_i e G_j; α_i, α_j, β e γ sono dei parametri la cui specificazione da luogo a indici differenti che servono a caratterizzare tutta una serie di criteri di raggruppamento di tipo gerarchico così come risulta dal prospetto riportato alla pagina seguente.

I criteri gerarchici di tipo agglomerativo indicati nella Tab. 4 presentano il notevole vantaggio, rispetto ai criteri che non soddisfano la formula ricorrente, di richiedere soltanto il calcolo delle diversità tra gruppi che devono essere conservate (in memoria dell'elaboratore) soltanto fino a quando due gruppi, qualsiasi, che si sono fusi ad un certo stadio del processo di raggruppamento, si fondano con un terzo gruppo.

Quanto affermato risulta facilmente confermato dalle considerazioni che seguono.

[28] La formula resta valida anche nel caso in cui si vogliano utilizzare indici di similarità $s\,(G_k\,,\,G_{ij})$ anziché indici di diversità, basterà in tal caso sostituire a $d\,(G_k\,,\,G_{ij})$ i valori $s\,(G_k\,,\,G_{ij}) = 1 - d\,(G_k\,,\,G_{ij})$. Occorre comunque tener presente che alcune proprietà, tra quelle che verranno in seguito illustrate facendo riferimento al quadrato della metrica euclidea, dei vari criteri di raggruppamento non risultano più soddisfatte quando si usano indici di similarità.

[29] La formula è stata introdotta da LANCE e WILLIAMS (1966, 1967) e successivamente estesa da WISHART (1969) e ANDERSON (1971).

TAB. 4 - *Criteri gerarchici di raggruppamento di tipo agglomerativo.*

Criterio di raggruppamento	Coefficienti della formula ricorrente			
	α_i	α_j	β	γ
1) Legame singolo	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
2) Legame completo	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
3) Centroide	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\frac{n_i \cdot n_j}{(n_i + n_j)^2}$	0
4) Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
5) Media tra gruppi	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
6) Media ponderata	$\frac{1}{2}$	$\frac{1}{2}$	0	0
7) Devianza minima	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$-\frac{n_k}{n_i + n_j + n_k}$	0
8) Flessibile	$\frac{1}{2}\,(1 - a)$	$\frac{1}{2}\,(1 - a)$	$a\;(< 1)$	0

Si assuma che i gruppi G_i e G_j siano composti, rispettivamente, da n_i e n_j unità, e che la misura della diversità tra questi due gruppi sia la minima, rispetto al criterio di raggruppamento prescelto, tra tutte le misure di diversità tra i gruppi che si sono già formati al generico s-esimo passo del processo di raggruppamento. In tali condizioni i due gruppi G_i e G_j si fonderanno per formare un nuovo gruppo $G_h = G_{ij}$ composto da $n_h = n_i + n_j$ unità di osservazione. Per calcolare la diversità tra il gruppo G_k ed il gruppo $G_h = G_{ij}$, attraverso la formula ricorrente, basterà conservare in memoria i valori $d\,(G_i\,,\,G_j)$, n_i e n_j; cosa questa immediatamente fattibile se alla matrice riportata nella Tab. 3 si aggiunge una riga dove vengono riportate le numerosità dei vari gruppi.

Si è già avuto modo di affermare che i vari criteri di raggruppamento indicati nella Tab. 4 si differenziano fra loro per il diverso modo usato nella definizione della similarità o diversità tra i gruppi; pertanto le loro proprietà ed i loro limiti deriveranno direttamente dalle proprietà e dai limiti degli indici di similarità o diversità utilizzati. Ma non è questa la sede per svolgere una discussione approfondita sul tema [30]; qui basterà richiamare le carat-

[30] Il lettore interessato ad un approfondimento della conoscenza può utilmente riferirsi ai lavori di LETI (1961 e 1978).

teristiche principali dei vari criteri, sottolineandone soprattutto la rilevanza sul piano applicativo.

Bisogna innanzi tutto osservare che degli otto criteri elencati: i primi quattro sono basati sulla similarità o diversità tra *elementi particolari* o *indici caratteristici* dei gruppi posti a confronto; il quinto ed il sesto si fondano sul calcolo di un *valore medio* tra tutti i possibili confronti che possono essere istaurati tra le unità appartenenti ai due gruppi; il settimo criterio è basato sulla minimizzazione della devianza interna ai gruppi calcolata ad ogni stadio del processo di raggruppamento; l'ultimo criterio infine si rivela particolarmente interessante perché consente, attraverso l'attribuzione di particolari valori numerici al coefficiente β, di rendere lo spazio di definizione dei gruppi più o meno *dilatato* o *contratto* rispetto allo spazio originario di definizione delle unità di osservazione (31). La *distorsione* (dilatazione o contrazione) dello spazio originario di definizione delle unità ha come conseguenza più evidente la formazione di molti gruppi, relativamente poco numerosi (dilatazione), o quella della formazione di pochi gruppi composti da molte unità (contrazione). Questa seconda caratteristica si risolve, generalmente, nella formazione di gruppi di tipo non ellissoidale producendo il fenomeno del concatenamento tra unità (32).

Tra gli otto criteri, quelli del centroide, della mediana, della media tra gruppi e della media ponderata sono *spazio conservativi*, i criteri del legame completo e della minima varianza sono *spazio dilatativi* mentre il criterio del legame singolo è *spazio contrattivo*. Per quanto concerne il criterio flessibile, questo risulta essere tanto più *spazio contrattivo* quanto più β risulta prossimo ad 1 (33).

A prescindere dalle caratteristiche spazio conservative, dilatative o contrattive, tutti i criteri gerarchici considerati procedono, ad ogni stadio del processo di raggruppamento, alla fusione dei gruppi G_i e G_j che risultano più *simili* o meno *diversi* tra loro secondo la specifica accezione di *similarità* o *diversità* propria di ciascun singolo criterio. In particolare, se con

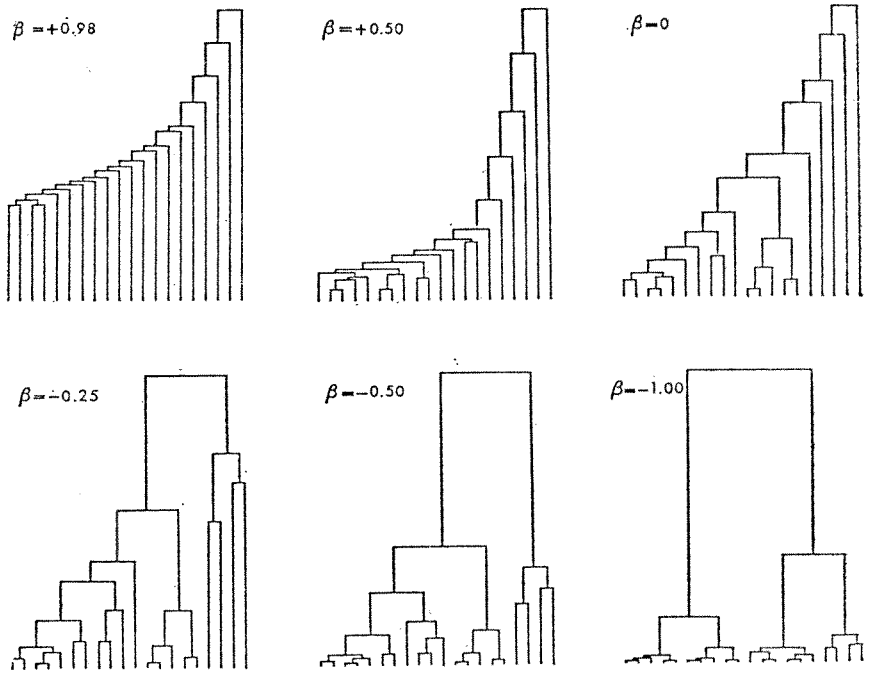
$$G_i = \{X_{i_1}, X_{i_2}, \ldots, X_{in_i}\} \text{ e } G_j = \{X_{j_1}, X_{j_2}, \ldots, X_{jn_j}\}$$

si indicano i due insiemi di misure relative ai due gruppi G_i e G_j composti, rispettivamente, da n_i ($n_i = 1, 2, \ldots$) e n_j ($n_j = 1, 2, \ldots$) unità di osservazione, i vari criteri risultano caratterizzati dagli indici di diversità (34) sotto illustrati.

Criterio del legame singolo. Secondo questo criterio, a ciascun stadio del processo di raggruppamento, i gruppi (che possono essere composti anche da una sola unità) che si fondono sono quelli che minimizzano la funzione (indice di diversità).

$$D_1(G_i, G_j) = \min_{\substack{r=1,2,\ldots n_i \\ s=1,2,\ldots n_j}} d(X_{ir}, X_{js})$$

da LANCE e WILLIAMS (1967). Dal grafico sono evidenti anche gli effetti prodotti dal variare dei valori numerici di β. La massima contrazione dello spazio è espressa dal primo albero.



I dati utilizzati si riferiscono a 20 unità di osservazione caratterizzate da 76 attributi binari. L'indice di diversità utilizzato è il quadrato della distanza euclidea.

(31) Una discussione approfondita su questo punto si trova in LANCE e WILLIAMS (1967).

(32) I criteri che soffrono dell'effetto di concatenamento vengono spesso criticati perché tendono a raggruppare anche unità molto diverse tra loro.

(33) Per illustrare gli effetti prodotti sulla strutture gerarchiche di raggruppamento (alberi) la contrazione e la dilatazione dello spazio, si riporta il seguente grafico ripreso

(34) Una esposizione perfettamente equivalente può essere fatta in riferimento agli indici di similarità, in tal caso si dovrebbe parlare però di massimizzazione del valore dell'indice anziché di minimizzazione.

I gruppi si fondono sulla base della diversità minima tra le singole unità appartenenti ai gruppi posti a confronto.

Criterio del legame completo. Questo criterio, come quello del legame singolo, è basato sul confronto tra singole unità, la funzione da minimizzare non è, però, più fondata sulla diversità minima ma su quella massima

$$D_2(G_i, G_j) = \max_{\substack{r=1,2,\ldots,n_i \\ s=1,2,\ldots,n_j}} d(X_{ir}, X_{is})$$

Criterio del centroide. Attraverso questo criterio vengano fusi i gruppi che minimizzano la funzione

$$D_3(G_i, G_j) = d(\bar{X}_i, \bar{X}_j)$$

dove $\bar{X}_i = \frac{1}{n_i}\sum_{r=1}^{n_i} X_{ir}$, $\bar{X}_j = \frac{1}{n_j}\sum_{s=1}^{n_j} X_{js}$

Il criterio del centroide non è, pertanto, basato sul confronto tra singole unità particolari (le più *vicine*, nel criterio del legame singolo; le più *lontane*, nel criterio del legame completo) ma tra indici caratteristici globali (vettori medi) calcolati considerando il contributo di tutte le unità appartenenti ai gruppi posti a confronto.

Criterio della mediana. Il principio ispiratore di questo criterio è analogo a quello del centroide ma, in questo caso, i *centroidi vengono calcolati* senza tener conto della dimensione dei gruppi che si fondono. In altri termini, nella determinazione dei centroidi, ciascun gruppo viene considerato come composto da una sola unità. La denominazione *criterio della mediana* deriva da fatto che i centroidi così calcolati cadono sempre nel punto intermedio, o punto mediano, della linea congiungente i centroidi dei due gruppi che si fondono. La funzione da minimizzare è quindi

$$D_4(G_i, G_j) = d(\bar{X}^*_i, \bar{X}^*_j)$$

cioè del tutto simile alla $D_3(G_i, G_j)$ con centroidi $\bar{X}_i^*$ e $\bar{X}_j^*$ calcolati diversamente ([35]).

Criterio della media tra gruppi. L'applicazione di questo criterio fornisce dei risultati intermedi a quelli che si ottengono applicando i criteri del legame singolo e del legame completo in quanto che è basato sulla media delle misure di diversità tra tutte le unità appartenenti ai due gruppi posti a confronto. Questa particolarità del criterio della media tra gruppi lo rende simile al criterio del centroide ma da quest'ultimo si differenzia poiché, in questo caso, si operano medie di confronti e non confronti tra medie. La funzione da minimizzare è

$$D_5(G_i, G_j) = \frac{1}{n_i \cdot n_j}\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} d(X_{ir}, X_{js})$$

Criterio della media ponderata. Questo criterio svolge nei confronti del criterio della media tra gruppi la stessa funzione svolta dal criterio della mediana nei confronti del criterio del centroide, elimina cioè l'effetto della diversa dimensione dei gruppi nella formazione dei nuovi gruppi. La funzione da minimizzare è

$$D_6(G_i, G_j) = \sum_{r=1}^{n_i}\sum_{s=1}^{n_j} d(X_{ir}, X_{js})$$

Si tratta di una funzione analoga alla $D_5(G_i, G_j)$ con $n_i = n_j = 1$.

Criterio della devianza minima. Tra i criteri basati sulla minimizzazione della devianza totale interna ai gruppi il più noto, ed usato, è quello proposto da WARD (1963). Applicando tale criterio si procede alla formazione della struttura gerarchica di raggruppamento attraverso la minimizzazione della somma delle devianze interne ai gruppi relative a tutte le variabili considerate; in altri termini, ad ogni stadio del processo di raggruppamento vengono considerate tutte le possibili coppie di aggregazioni tra gruppi e si procede alla fusione dei due gruppi che comporta il minimo incremento della devianza totale interna ai gruppi. La funzione da minimizzare è

$$D_7(G_1, G_2, \ldots, G_m) = \sum_{j=1}^{m}\sum_{h=1}^{k}\sum_{i=1}^{n_h}(x_{ihj} - \bar{x}_{hj})^2 - \sum_{j=1}^{m-1}\sum_{h=1}^{k}\sum_{i=1}^{n_h}(x_{ihj} - \bar{x}_{hj})^2$$

dove: x_{ihj} rappresenta il valore assunto nella i-esima unità, appartenente al gruppo j-esimo, dalla h-esima variabile; $\bar{x}_{hj}$ è il valor medio dell'h-esima variabile nel j-esimo gruppo; m indica il numero dei gruppi formati dopo n — m passi del processo di raggruppamento. Come si può osservare, la fusione dei gruppi che comporta il minimo incremento di devianza totale interna ai gruppi, nel passaggio da una partizione strutturata in m gruppi ad una partizione strutturata in m — 1 gruppi, è quella cui corrisponde il minimo della devianza totale interna ai gruppi ad ogni singolo stadio del processo di raggruppamento.

Il criterio della devianza minima, praticamente ignorato dai biologi, si è dimostrato particolarmente efficace nel campo socio-economico; esso tende alla formazione di gruppi di tipo ellissoidale di uguale numerosità e, nella sua struttura, si avvicina molto al criterio del centroide quando quest'ultimo

([35]) Se ad un particolare stadio del processo si realizza la fusione dei gruppi G_i e G_j, il centroide del nuovo gruppo $G_{ij} = G_h$ risulta dalla relazione

$$\bar{X}_n = \frac{\bar{X}_i\, n_i + \bar{X}_j\, n_j}{n_i + n_j}$$

nel caso in cui s'intenda applicare il criterio del centroide, mentre è dato da

$$\bar{X}^*_h = \frac{\bar{X}^*_i + \bar{X}^*_j}{2}$$

per il criterio della mediana.

viene applicato facendo ricorso al quadrato della metrica euclidea quale indice base di diversità. Infatti, se si ipotizza la fusione dei gruppi G_p e G_q, dopo l'(n-m)-esimo passo del processo di raggruppamento, risulta facile dimostrare che

$$D_7(G_1, G_2, \ldots, G_m) = \frac{n_p \cdot n_q}{n_p + n_q} \sum_{h=1}^{k} (\bar{x}_{hp} - \bar{x}_{hq}))^2$$

dove n_p ed n_q rappresentano, rispettivamente, la numerosità del p-esimo e dell'h-esimo gruppo, mentre $\bar{x}_{hp}$ e $\bar{x}_{hq}$ sono le rispettive medie calcolate sull'h-esima variabile.

Il valore della funzione obbiettivo, quando si fa ricorso al metodo del centroide ed al quadrato della metrica euclidea, relativamente ai gruppi G_p e G_q, risulta dalla relazione

$$D_3(G_p, G_q) = d_2^2(\bar{X}_p, \bar{X}_q) = \sum_{h=1}^{k} (\bar{x}_{hp} - \bar{x}_{hq})^2$$

Risultato questo proporzionale a quello che si ottiene applicando la funzione D_7 (. , . , ... , .), con fattore di proporzionalità pari a

$$\frac{n_p \cdot n_q}{n_p + n_q}$$

Sostanzialmente, il criterio della devianza minima si risolve nella ricerca della *distanza minima ponderata* tra i centroidi dei gruppi ([36]).

Criterio flessibile. Su questo criterio si è già avuto modo di discutere; si tratta di un criterio del tutto particolare strettamente vincolato ai concetti di *conservazione*, *dilatazione* e *contrazione*, dello spazio originario di definizione delle unità di osservazione, introdotti da LANCE e WILLIAMS (1966).

I criteri gerarchici di tipo agglomerativo fin qui discussi sono tra quelli più comunemente usati nelle applicazioni. Un tale stato di fatto non sta però a significare che essi risultano i più appropriati in tutte le situazioni, si può anzi affermare che la loro larga diffusione è dovuta più a motivazioni di carattere contingente che a ragioni connesse alla loro superiorità, logica e metodologica, rispetto ad altri criteri quali, ad esempio, quelli di tipo non gerarchico. Si tratta, sostanzialmente, di criteri sviluppati nell'ambito della ricerca biologica, campo questo nel quale hanno trovato fertile terreno soprattutto perché si rifanno alla tradizionale struttura *tassonomica* di LINNEO.

La tendenza dei biologi, che per primi si sono interessati all'analisi dei gruppi, a sviluppare criteri tassonomici di tipo gerarchico, nel rispetto di una

([36]) Si deve osservare, peraltro, che l'introduzione della ponderazione, così come risulta dal criterio della devianza minima, elimina uno dei difetti tipici del criterio del centroide (e della mediana): « la produzione di *inversioni* »; elimina, cioè, la possibilità che la fusione dei gruppi a stadi più avanzati del processo di raggruppamento avvenga a livelli di *distanza* inferiori a quelli che hanno prodotto la fusione in stadi precedenti.

tradizione ormai consolidata, ha avuto dei riflessi (a dire il vero non sempre positivi) anche negli altri campi dell'indagine scientifica. Un numero non indifferente di ricercatori sono stati indotti all'applicazione, spesso acritica, di criteri il cui apparato concettuale, ancorché valido nell'ambito della problematica biologica, mal si adattava alle esigenze ed alle specifiche caratteristiche di fenomeni che con quelli biologici avevano poco o nulla in comune.

Le considerazioni sopra svolte non vogliono certo significare che i criteri gerarchici (di tipo agglomerativo) risultano utili soltanto nel campo biologico; ma, vogliono più semplicemente riaffermare il principio che essi conservano la loro validità, al di fuori di tale ambito disciplinare, soltanto nei casi in cui sia possibile riconoscere in natura strutture che siano di tipo gerarchico, e ciò non sempre accade quando si ha a che fare con fenomeni socio-economici.

Per quanto concerne le diverse possibilità di utilizzo dei vari criteri gerarchici di tipo agglomerativo considerati, si deve osservare che non esiste certamente un criterio che si rivela superiore agli altri in tutte le occasioni. Si deve anzi sottolineare che le discussioni più approfondite svolte sull'argomento conducono a delle conclusioni del tutto discordanti. Gli appartenenti alla cosiddetta scuola di Cambridge (soprattutto C. J. JARDINE e C. SIBSON) affermano che, tra tutti i criteri gerarchici di tipo agglomerativo proposti, soltanto il criterio del legame singolo rispetta una serie di condizioni analitiche che devono essere soddisfatte affinché la matrice delle similarità o diversità subisca la minima distorsione quando su di essa si procede alla costruzione di una struttura gerarchica. Per contro, gli appartenenti alla cosiddetta scuola australiana (soprattutto G. N. LANCE e W. T. WILLIAMS) obiettano che le condizioni analitiche e il concetto di distorsione introdotti dai fautori della scuola di Cambridge sono scarsamente significativi, in quanto impongono la scelta di un criterio (legame singolo) che dà luogo a dei risultati spesso del tutto inacettabili dal punto di vista della logica che ispira l'analisi dei gruppi che è tesa alla individuazione di *gruppi naturali*. Secondo gli australiani l'analisi dei gruppi deve essere usata, soprattutto, per semplificare e descrivere i dati di osservazione; devono essere pertanto delle *ragioni di ordine pratico* ad indirizzare il ricercatore nella scelta del criterio più appropriato allo specifico problema oggetto di studio ([37]). Le ragioni di ordine pratico portano spesso alla esclusione del criterio del legame singolo che dal punto di vista *analitico* risulta essere il più soddisfacente.

Le due scuole di pensiero, ispirate a due diverse filosofie — una teorica l'altra pragmatica —, non devono, a parere dello scrivente, essere contrap-

([37]) A questa filosofia sembra ispirarsi anche BENZÉCRI (1973). Questo autore, sulla scorta di molteplici esperienze, consiglia l'uso del criterio della devianza minima e, in seconda istanza, il criterio del legame medio, mentre afferma di aver ottenuto risultati significativi soltanto poche volte applicando il criterio del legame singolo.

poste ma vanno considerate come complementari tra loro. Entrambe devono essere tenute presenti dal ricercatore che intende procedere all'effettuazione di un'analisi dei gruppi poiché tutte e due sono *ragionevoli, significative* e tali da giocare un ruolo non irrilevante nella costruzione di quella *teoria generale* di cui si è più volte sottolineata la mancanza.

6.2. *Criteri non gerarchici di raggruppamento*

La procedura di base cui sono ispirati i criteri gerarchici di tipo agglomerativo inizia, generalmnete, con il calcolo delle similarità o diversità interindividuali e termina con la costruzione di un *dendrogramma* (*struttura gerarchica di raggruppamento*) che evidenzia i risultati dell'analisi, attraverso la *ottimizzazione* di una qualche *funzione obiettivo*.

Ad ogni specifico passo del processo di raggruppamento i vari criteri procedono alla fusione delle unità e dei gruppi più *simili* e le differenze tra i vari criteri risiedono nel diverso modo di definizione della similarità tra unità, tra unità e gruppi e tra gruppi. Si è visto infatti come, ad esempio, nel criterio della media tra gruppi la similarità (diversità) risulti determinata dalla media delle similarità (diversità) calcolate su tutte le coppie di unità di osservazione appartenenti ai due gruppi posti a confronto, mentre nel criterio del legame singolo la similarità (diversità) tra gruppi resta definita dalle due unità più simili (meno diverse).

Così come per i criteri gerarchici, anche per i criteri non gerarchici di raggruppamento la *attribuzione* delle unità di osservazione ai vari gruppi viene effettuata in modo da *ottimizzare* una *funzione obiettivo* predefinita e ben specificata. Questi ultimi criteri differiscono però dai criteri gerarchici sia perché prevedono, nella generalità dei casi, l'individuazione di una sola *partizione* delle unità di osservazione (non una struttura gerarchica formata da tante partizioni quante sono le unità), sia perchè consentono la rimozione e la riattribuzione delle unità stesse, rendendo così possibile la correzione di partizioni insoddisfacenti; ciò avviene attraverso l'applicazione di procedure di tipo iterativo.

Si è già avuto occasione di affermare che il modo migliore per individuare la *partizione ottima* è quello di considerare tutte le possibili partizioni dell'insieme originario delle unità di osservazione, ma si è anche detto della pratica impossibilità di attuazione di una tale procedura anche nei casi in cui si disponga di un elaboratore elettronico ad alta capacità di memoria ed elevata velocità di esecuzione. Vista l'impossibilità di analisi esaustive, che tengano conto di tutte le possibili partizioni, la ricerca dell'ottimo viene generalmente condotta in un ambito ristretto di partizioni.

Se con

$$\pi(r) = [P_{\alpha_1}, P_{\alpha_2}, \ldots, P_{\alpha_{i-1}}, P_{\alpha i}, P_{\alpha_i+1}, \ldots, P_{\alpha_n}; r] \quad (38)$$

(38) Si ricorda che gli α_i ($i = 1, 2, \ldots, n$) possono assumere i valori interi da 1

si rappresenta la partizione della unità che risulta alla r-esima replicazione del processo di raggruppamento, e con

$$\pi(r+1) = [P_{\alpha_1}, P_{\alpha_2}, \ldots, P_{\alpha_{i-1}}, P_{\alpha_j}, P_{\alpha i+1}, \ldots, P_{\alpha_n}; r]$$

la partizione delle unità di osservazione alla $(r+1)$-esima iterazione quando, rispetto alla iterazione precedente, si è realizzata una riattribuzione dell'unità i-esima che dal gruppo originaria α_i passa al gruppo α_j, si potrà affermare che i diversi criteri non gerarchici di raggruppamento differiscono tra loro in dipendenza della specifica della funzione $g[\pi(r); X]$ che s'intende ottimizzare o, più in particolare, della differenza tra i valori assunti dalla funzione $g(.,.)$ in due iterazioni successive

$$\Delta_{ijr}\, g(\pi; X) = g[\pi(r+1); X] - g[\pi(r); X]$$

dove i ($i = 1, 2, \ldots, n$) si riferisce alle unità, j ($j = 1, 2, \ldots, m$) si riferisce ai gruppi ed r ($r = 1, 2, \ldots$) alle iterazioni.

Ciascun processo non gerarchico di raggruppamento risulta pertanto definito dalle seguenti fasi:

I fase: Scelta di una partizione iniziale $\pi(0)$ caratterizzata generalmente, ma non necessariamente, da un numero m prefissato di gruppi.

II fase: Data la partizione $\pi(r)$, che risulta alla r-esima iterazione del processo, si calcolano i valori di $\Delta_{ijr}\, g(\pi; X)$ per $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$, dove m può essere fisso o variabile.

III fase: Per $i = 1, 2, \ldots, n$ si attribuisce l'i-esima unità al gruppo j* se

$$\Delta_{ij^*r}\, g(\pi; X) = \min_{j=1,2,\ldots,m} \Delta_{ijr}\, g(\pi; X)$$

od anche

$$\Delta_{ij^*r}\, g(\pi; X) = \max_{j=1,2,\ldots,m} \Delta_{ijr}\, g(\pi; X)$$

IV fase: Se $\pi(r+1) \neq \pi(r)$ si ritorna alla II fase altrimenti il processo si arresta.

Analizzando le varie fasi del processo di ottimizzazione, si deduce facilmente come i vari criteri di raggruppamento si possano differenziare tra loro almeno secondo tre diverse dimensioni; e cioè a seconda: a) della scelta della partizione iniziale $\pi(0)$; b) della specifica della funzione $\Delta_{ijr}\, g(\pi; X)$; c) che sia stato fissato a priori il numero m di gruppi, che non si modifica fino al termine del processo, oppure che tale numero risulti come conseguenza del processo stesso.

La prima dimensione lungo cui si svolgono e differiscono i vari criteri non gerarchici è la scelta della partizione iniziale $\pi(0) = [P_{\alpha_1}, P_{\alpha_2}, \ldots, P_{\alpha_i}, \ldots, P_{\alpha_n}; 0]$. Alcuni criteri prevedono la individuazione diretta di

ad m, m rappresenta il numero dei gruppi e $P_{\alpha i}$ indica il gruppo (subpopolazione) al quale è stata attribuita l'i-esima unità di osservazione.

una tale partizione, altri introducono un processo preliminare che conduce alla partizione stessa; il processo preliminare può consistere, ad esempio, nella scelta di un certo numero di *centroidi iniziali* (*seed points*) e la partizione $\pi(0)$ risulterà dopo l'attribuzione delle unità di osservazione al centroide più vicino.

In generale si deve osservare che per la scelta della partizione iniziale $\pi(0)$, operata in via diretta, o in via indiretta attraverso la introduzione di m centroidi, si possono seguire: *a*) procedure del tutto casuali; *b*) procedure dettate dalla natura e dalla conoscenza del fenomeno oggetto d'analisi; *c*) procedure di tipo meccanicistico, eventualmente suggerite dall'osservazione dei valori assunti dalle variabili che individuano ciascuna unità di osservazione o dai valori che esprimono le similarità o diversità interindividuali; *d*) procedure collegate a preventive analisi di raggruppamento di tipo gerarchico ([39]).

Relativamente alla seconda dimensione cui si è fatto sopra riferimento si deve osservare che gli elementi caratterizzanti le funzioni da *ottimizzare*, e che distinguono i diversi criteri non gerarchici di raggruppamento, hanno duplice natura: la prima si riferisce al tipo di *misura* introdotta per definire la *omogeneità* (*similarità* o *diversità*) tra unità e gruppi e all'interno dei gruppi, la seconda è invece relativa alla possibilità che certi elementi caratteristici di riferimento dei vari gruppi, ad esempio i centroidi, possano subire modifiche dopo lo spostamento di una singola unità o che tale modifica possa avvenire soltanto dopo il completamento di un intero ciclo iterativo.

Così come accade per i criteri gerarchici, anche per i criteri non gerarchici le varie funzioni si distinguono a seconda che nel processo di attribuzione e riattribuzione delle unità di osservazione ai vari gruppi queste vengano confrontate con: *a*) unità particolari dei gruppi (la più vicina — *legame singolo* — o la più lontana — *legame completo* —); *b*) indici sintetici caratteristici dei gruppi (*centroide*); *c*) tutte le unità, sintetizzando poi le varie misure attraverso una opportuna *media*. Il processo di attribuzione e riattribuzione può fondarsi, infine, sulla minimizzazione della devianza o della varianza interna ai gruppi ([40]).

([39]) Si osservi che i risultati ottenuti attraverso l'analisi di raggruppamento di tipo gerarchico possono fornire informazioni utilissime non solo ai fini della scelta della partizione iniziale, ma possono dare anche valide indicazioni per la scelta del numero di gruppi m da formare quando questo è richiesto dallo specifico criterio non gerarchico di raggruppamento.

([40]) La devianza totale interna ai gruppi non è certamente l'unico *indice statistico* significativo che può essere considerato nell'ambito dei processi di raggruppamento (sia di tipo non gerarchico che in quelli di tipo gerarchico). Almeno altri tre indici, sempre basati su sintesi della *variabilità totale* tra gruppi e nei gruppi, risultano particolarmente interessanti ai fini dell'analisi dei gruppi. Supponendo di aver operato una partizione in m gruppi dell'insieme delle osservazioni, la matrice T delle devianze e codevianze relative alle n osservazioni originarie può essere espressa dalla relazione

$$T = B + W$$

La terza dimensione lungo la quale si svolgono i processi di raggruppamento di tipo non gerarchico è relativa al numero dei gruppi che definiscono la partizione. Alcuni criteri prevedono la fissazione del numero m di gruppi a priori; in questo caso il ricercatore, nel fissare un tale numero, potrà avvalersi dei risultati di una analisi preliminare di tipo gerarchico, delle conoscenze specifiche sul fenomeno oggetto d'indagine. Qualora il ricercatore non disponga di informazioni *sufficienti*, la procedura da seguire può essere la ripetizione del processo iterativo di raggruppamento, ipotizzando diverse numerosità della partizione, e la scelta del numero più soddisfacente in funzione dei risultati ottenuti.

I criteri non gerarchici di raggruppamento che non prevedono la fissazione a priori del numero dei gruppi, lasciando che questo parametro sia libero di variare fino a fissarsi su di un certo valore in dipendenza del tipo di dati utilizzati, comportano necessariamente la fissazione a priori dei valori di altri indici o parametri di riferimento. Si tratta, nella generalità dei casi, della distanza minima tra gruppi (al di sotto della quale due gruppi vengono fusi) e della distanza massima tra ciascuna unità e il gruppo (al di sopra della quale l'unità non viene attribuita). S'intuisce facilmente come l'inserimento nel processo di raggruppamento dell'ultimo vincolo possa produrre un insieme di unità isolate che non trovano collocazione in nessun gruppo.

L'ultima osservazione fatta evidenzia una ulteriore dimensione dei processi non gerarchici di raggruppamento ed è quella relativa al carattere *esaustivo* o meno dell'analisi dei gruppi. Alcuni criteri, ad esempio, prevedono la possibilità che delle unità non vengano attribuite ai gruppi, altri prevedono invece la formazione, durante il processo iterativo di raggruppamento, di un particolare gruppo contenente le unità spurie (*outliers*) che non vengono attribuite agli altri gruppi perché superano un certo livello di distanza prefissato.

dove B rappresenta la matrice totale delle devianze e codevianze tra gruppi, W la matrice totale delle devianze e codevianze nei gruppi. Il criterio della minimizzazione della devianza totale interna ai gruppi si risolve nella minimizzazione della traccia della matrice W (tr (W)). Gli altri tre criteri si basano invece:

1) Sulla minimizzazione del rapporto tra i determinantidi W e di T

$$\lambda = \frac{|W|}{|T|}$$

Questo criterio è noto come il criterio λ di WILKS.

2) Sulla massimizzazione della traccia del prodotto tra la matrice inversa di W e B.

$$t = tr(W^{-1} B)$$

Questo criterio è noto come criterio della traccia di HOTELLING. Si osservi che la tr (W^{-1} B) è uguale alla somma degli autovalori λ_i calcolati risolvendo l'equazione caratteristica

$$|B - \lambda W| = 0.$$

3) Sulla massimizzazione del più elevato autovalore λ_1. Questo criterio è noto come criterio della radice massima di ROY.

Tra gli innumerevoli criteri non gerarchici di raggruppamento proposti, ne verranno esaminati tre soltanto: il criterio di FORGY, il criterio K-means convergente di MAC QUEEN ed il criterio ISODATA. I tre criteri appaiono, per le loro peculiari caratteristiche, sufficientemente rappresentativi di classi molto più vaste; a questi criteri base possono essere ricondotti, anche se a volte con notevoli varianti, molti altri criteri di raggruppamento qui non considerati.

Il criterio di Forgy

Il criterio di FORGY (1965) è basato sul confronto tra le varie unità di osservazione ed i centroidi dei gruppi relativi alla partizione iniziale ed alle partizioni che si vengono a determinare ad ogni stadio del processo iterativo di raggruppamento. Più in particolare, il processo si svolge secondo le seguenti fasi:

1) Individuazione di una partizione iniziale in modo diretto o indiretto operando una scelta, casuale o ragionata, di m centroidi. Il numero dei gruppi m è prefissato.

2) Attribuzione di ogni unità di osservazione al gruppo più vicino. La similarità o diversità viene calcolata in riferimento al centroide del gruppo. I centroidi non si modificano dopo l'attribuzione di ogni unità.

3) Calcolo dei nuovi centroidi dei gruppi.

4) Ripetizione delle fasi 2 e 3 fino a quando la somma delle distanze (se si utilizza un indice di distanza come generalmente accade) tra le unità ed i centroidi dei gruppi di appartenenza non si riduce più attraverso il trasferimento di unità da un gruppo ad un altro gruppo ([41]).

Ad ogni iterazione il criterio richiede il computo di nxm misure di diversità (o similarità) e ciascuna unità U_i, caratterizzata dal vettore $X_i = (x_{i1}, x_{i2}, \ldots, x_{2k})$, viene assegnata al gruppo il cui centroide $\overline{X}_j = (\overline{x}_{1j}, \overline{x}_{2j}, \ldots, \overline{x}_{hj}, \ldots, \overline{x}_{kj})$ [dove $\overline{x}_{hj}$, ad esclusione di quello relativo alla partizione preliminare $\pi(0)$ che può essere prefissato, deriva dalla uguaglianza

$$\overline{x}_{hj} = \frac{1}{n_j} \sum_{r=1}^{n_j} x_{rh}$$

e n_j è il numero delle unità che compongono il j-esimo gruppo] minimizza (o massimizza) l'espressione

$$d(X_i, \overline{X}_j) \quad \text{per } j = 1, 2, \ldots, m.$$

([41]) JANCEY (1966) propone una variante al criterio di FORGY suggerendo un calcolo diverso dei nuovi centroidi (fase 3). La variante dovrebbe, secondo l'autore, accelerare la velocità del processo di convergenza e facilitare il raggiungimento di minimi (o massimi) locali migliori. Prove pratiche effettuate smentiscono però questa congettura.

Di solito la convergenza viene raggiunta dopo un numero limitato di iterazioni, pertanto il ricercatore può condurre diverse analisi di raggruppamento in corrispondenza di diversi valori di m, ad un costo assai contenuto e comunque inferiore a quello cui andrebbe incontro attuando un'analisi di tipo gerarchico.

Il criterio K-Means convergente

Il criterio K-Means, proposto originariamente da MACQUEEN (1967) e successivamente sottoposto a modifiche e varianti da altri autori, non si discosta molto dal criterio di FORGY. La differenza fondamentale sta nel fatto che i centroidi dei gruppi non restano fissi lungo tutto un ciclo iterativo ma si modificano, e vengono aggiornati, dopo l'attribuzione di ogni unità di osservazione. Il processo di raggruppamento si svolge pertanto secondo le fasi seguenti.

1) Individuazione di una partizione iniziale. Il numero dei gruppi m è prefissato.

2) Confronto di ciascuna unità con i centroidi dei gruppi ed attribuzione dell'unità al gruppo che presenta il centroide più *vicino*. Se l'operazione comporta il trasferimento dell'unità da un gruppo ad un altro gruppo i centroidi dei due gruppi vengono ricalcolati in funzione della modifica in essi intervenuta.

3) Ripetizione della fase 2 del processo di raggruppamento fino a quando nessuna unità cambia il gruppo di appartenenza.

Il criterio K-means richiede una mole di calcoli superiore a quella richiesta dal criterio di FORGY, sia per il ricalcolo dei centroidi dopo ogni riattribuzione delle unità, sia per il più elevato numero di iterazioni necessarie per il raggiungimento della convergenza. I più elevati costi non sono comunque tali da sconsigliarne l'uso in favore del criterio di FORGY visto che, rispetto a quest'ultimo, presenta una certa maggiore ragionevolezza ([42]).

Il criterio ISODATA

Il criterio non gerarchico di raggruppamento denominato ISODATA (Iterative Self-Organizing Data Analysis Techniques) è stato proposto originariamente da BALL e HALL (1965) ed ha subito in seguito modifiche ed adattamenti da parte di diversi studiosi. Una delle versioni più utilizzate è quella sotto riassunta.

([42]) WISHART (1969) presenta una variante del criterio k-means che, attraverso l'assegnazione di valori numerici a certi parametri caratteristici di controllo (distanza massima tra centroidi, dimensione minima dei gruppi, numero massimo di iterazioni per ogni ciclo, numero minimo di gruppi), consente l'individuazione di una *partizione ottima* la cui numerosità non è fissata a priori ma è una risultante del processo.

1) Specificazione del valore numerico di alcuni parametri caratteristici di controllo quali: *a*) il numero minimo di elementi per ogni gruppo (THETAN); *b*) la distanza massima tra centroidi (THETAC); *c*) il numero massimo di iterazioni per cicli interni (NPARTS) ed esterni (ITERMAX); *d*) il numero massimo di fusioni tra gruppi (NCLST) i cui centroidi sono ad una distanza inferiore al limite prefissato THETAC; *e*) un numero di riferimento per la fusione o la divisione di gruppi (NWRDSD); *f*) un coefficiente di variabilità (THETAE) di riferimento per la divisione dei gruppi.

2) Specificazione di m centroidi iniziali.

3) Attribuzione delle unità attraverso l'applicazione del criterio di FORGY fino al raggiungimento della convergenza o del numero prefissato NPARTS.

4) Eliminazione dei gruppi che contengono un numero di unità inferiore a THETAN.

5) Fusione o divisione dei gruppi residui a seconda che essi siano, rispettivamente, in numero superiore al doppio di NWRDSD o inferiore alla metà di NWRDSD.

6) Calcolo dei centroidi dei gruppi e riattribuzione delle unità secondo la procedura indicata al punto 3.

7) Ripetizione delle fasi 4, 5 e 6 fino alla convergenza o del numero prefissato ITERMAX.

Da rilevare che nel processo di fusione o divisione dei gruppi si procede al calcolo delle distanze fra unità e centroidi, tra centroidi, e della variabilità interna ai gruppi; i valori ottenuti, insieme ai parametri prefissati THETAN, THETAC, NCLST e THETAE, sono gli elementi che condizionano e determinano i risultati dell'analisi di raggruppamento.

Il criterio ISODATA richiede, per la sua applicazione, la fissazione di tutta una serie di valori alla cui specifica si può arrivare soltanto per tentativi. Questo fatto, insieme a problemi relativi alla convergenza del processo, ha suscitato non poche perplessità in merito alla *bontà* ed alla *oggettività* dei risultati cui l'applicazione del criterio di raggruppamento dà luogo; si tratta però, in genere di perplessità fondate su valutazioni aprioristiche piuttosto che su analisi specifiche.

Un'ultima annotazione da fare in merito al criterio ISODATA riguarda la sua natura. Il criterio pur essendo sostanzialmente non gerarchico tende, attraverso le fusioni e le aggregazioni dei gruppi, a produrre delle strutture di gruppo di tipo gerarchico.

Come già sottolineato, nelle pagine precedenti, in questa sede ci siamo limitati a considerare soltanto tre criteri non gerarchici di raggruppamento; la scelta è stata dettata dalla necessità di presentare i criteri di più largo impiego e che fossero nello stesso tempo basati su una logica sufficientemente chiara ed estensibile ad una classe molto più vasta di criteri. D'altra parte sarebbe risultato praticamente impossibile, e poco utile, elencare tutti i criteri esistenti alcuni dei quali costituiscono delle semplici varianti dei criteri discussi ([43]).

Sui meriti relativi dei criteri non gerarchici di raggruppamento, rispetto ai criteri gerarchici, si è già avuto occasione di discutere nelle pagine precedenti. La loro superiorità, se di superiorità si può parlare, risiede soprattutto nella possibilità offerta dai criteri non gerarchici di correggere, attraverso iterazioni successive, partizioni iniziali insoddisfacenti; cosa questa che non è invece possibile nel processo di costruzione delle strutture gerarchiche. Accanto a questo vantaggio, si devono però considerare almeno due aspetti negativi insiti nelle procedure non gerarchiche: il primo è relativo alla necessità d'introdurre, a priori, degli elementi (quali ad esempio il numero dei gruppi, il numero massimo di unità per gruppo, la distanza minima tra gruppi, ecc.) alla cui definizione il ricercatore giunge sovente ulizzando informazioni scarsamente attendibili; il secondo riguarda invece il possibile condizionamento della *partizione iniziale* sui risultati finali dell'analisi.

Si deve comunque osservare che il problema connesso alla attribuzione di valori numerici ai parametri segnaletici di riferimento può essere risolto in modo abbastanza soddisfacente per *tentativi*; cosa questa possibile dato il costo relativamente poco elevato che molte procedure non gerarchiche comportano. Per quanto concerne poi il possibile condizionamento della partizione iniziale sui risultati dell'analisi, esso risulta generalmente poco rilevante e può, in ogni caso, essere valutato ed eventualmente eliminato attraverso reiterazioni del processo di raggruppamento basate su partizioni iniziali diverse.

Per quanto riguarda, infine, il problema della scelta della procedura più appropriata, nell'ambito dei criteri non gerarchici, si deve semplicemente osservare che fino ad oggi non sono stati prodotti studi capaci di fornire una risposta significativa e di portata sufficientemente vasta ([44]).

([43]) Risulta tuttavia opportuno segnalare al lettore alcuni autori che hanno prodotto lavori contenenti proposte di criteri non gerarchici di raggruppamento di un certo interesse, o per l'originalità del contenuto o perché, pur trattando problemi specifici, i suggerimenti forniti si rivelano utili in una gamma piuttosto vasta di situazioni. Gli autori sono FRIEDMAN e RUBIN (1967), WALLACE e BOULTON (1968), LERMAN (1970), DIDAY (1971), BADALONI e RIZZI (1972). KENDALL (1972), MINEO (1974), SADOCCHI (1977), LIS e SAMBIN (1978).

([44]) Riguardo agli studi sulle tecniche di valutazione dei risultati delle analisi di raggruppamento e agli studi comparativi tra i vari criteri (gerarchici e non gerarchici), oltre ai già citati lavori di JARDINE e SIBSON e di LANCE e WILLIAMS, si devono segnalare quelli di SOKAL e ROHLF (1962), RAND (1971), LING (1973), BAKER e UBERT (1975), BLASHFIELD e ALDENDERFER (1977), GNANADESIKAN, KETTENRING e LANDWEHR (1977). In particolare questi ultimi tre autori discutono alcune tecniche di valutazione dei risultati ottenuti da analisi di raggruppamento considerando specifiche caratteristiche dei gruppi quali la stabilità e la sensitività rispetto a lievi mutamenti dell'insieme originario dei dati di osservazione (omissione di variabili e/o di unità).

ABSTRACT

The main objective of cluster analysis is to identify possible « category structures » present in the observation units of a particular « statistical collective », the study of which has provided the values assumed by a number of variables. This paper will look into the logical and methodological principles that underlie this type of analysis.

First, an attempt is made to clarify the overall problem and to place it in its proper framework: and then the various phases of the classification process are discussed. Particular attention is given to those aspects which seem particularly crucial to a correct application of the various alternative grouping criteria that are proposed. Finally, the more important hierarchical and non-hierarchical clustering criteria are analyzed with the purpose of singling out their advantages as well as their limits of applicability.

BIBLIOGRAFIA

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

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York.

ANDERSON, A. J. B. (1971). Numeric examination of multivariate soil samples. *Math. Geol.* Vol. 3.

BADALONI M. e RIZZI A. (1972). Contributi alla cluster analysis. *Metron*, Vol. XXX.

BAILEY, K. D. (1975). Cluster analysis, sta in Heise D. ed., *Sociological Methodology*. Jossey-Bass, San Francisco.

BAKER, F. B. e HUBERT, L. J. (1975). Measuring the power of hierarchical cluster analysis, *J. Amer. Statist. Ass*, Vol. 70.

BALL, G. H. e HALL, D. J (1965). ISODATA, *A Novel Method of Data Analysis and Pattern Classification*, AD699616. Stanford Res. Inst., Menlo Park, California.

BEALE, E. M. L. (1969). Euclidean cluster analysis. *Bull. I.S.I.*, Vol. 43.

BELLACICCO, A. (1976). *Metodologia e Tecnica della Classificazione Matematica*. La Goliardica, Roma.

BENEDETTI, C. (1959). Di alcuni criteri di decisione statistica basati sulla soluzione di problemi di minima distanza. *Metron*, Vol. XIX.

BENZÉCRI, J. P. (1973). *L'Analyse des Donnes: I La Taxinomie*. Dunod, Paris.

BIGGERI, L. (1977). Some applications of cluster analysis in sampling design. *Bull. I.S.I.*, Vol. 47.

BLASHFIELD, R. K. e ALDENDERFER, M. S. (1977). *A Consumer Report on Cluster Analysis Software: 1 Cluster Analysis Methods and Their Literature*. The Pennsylvania State University.

BIJNEN, E. J. (1972). *Cluster Analysis: Survey and Evaluation of Techniques*. Tilburg University Press, Groningen.

CACOULLOS, T. ed. (1972). *Discriminant Analysis and Applications*. Academic Press, New York.

CALINSKI, T. e HARABASZ, J. (1971). A dentrite method for cluster analysis. *Biometrics*, Vol. 27.

CASTELLANO, V. (1965). *Istituzioni di Statistica*, Ilardi, Roma.

CHIANDOTTO, B., GHILARDI, G. e LEONI, R. (1978). *Un Modello Statistico-Matematico di Localizzazione Industriale*. Federazione Regionale fra le Associazioni Industriali della Toscana, Firenze.

COOLEY, W. W. e LOHNES, P. R. (1971). *Multivariate data Analysis*. John Wiley, New York.

CORMACK, R. M. (1971). A review of classification. *J. R. Statist. Soc.*, Vol. 134.

DAGNELIE, P. (1966). A propos des differentes méthodes de classification. *Revue de Statist. Appl.*, Vol. 14.

DICE, L. R. (1945). Measures of the amount of ecological association between species. *Ecology*, Vol. 26.

DIDAY, E. (1971). *The Dinamic Cluster Method*. Fascicule IRIA, Rocquencourt 78 France.

DIDAY, E. e SIMON, J. C. (1976). Clustering analysis, sta in Fu, K. S. ed., *Digital Pattern Recognition*. Springer-Varlag, Berlin.

DURAN, B. S. e ODELL, P. J. (1974). *Cluster Analysis: A Survey*. Springer-Verlag, Berlin.

FORGY, E. W. (1965). Cluster analysis of multivariate data: Effeciency Versus interpretability of classifications. *Biometrics*, Vol. 21.

FRIEDMAN, H. P. e RUBIN, J. (1967). On some invariant criteria for grouping data. *J. Amer. Statist. Ass.*, Vol. 62.

FU, K. S. (1977). Linguistic approach to pattern recognition, sta in Van Ryzin J. ed., *Classification and Clustering*. Academic Press, New York.

FUKUNAGA, K. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, New York.

GINI, C. (1951). Lezioni di Statistica. Eredi Virgilio Veschi, Roma.

GNANADESIKAN, B., KETTENRING, J. R. e LANDWEHR, J. M. (1977). Interpreting and assessing the results of cluster analysis. *Bull. I.S.I.*, Vol. 47.

GOOD, I. J. (1965). Categorization of classification, sta in *Mathematics and Computer Science in Biology and Medicine*. HMSO and Medical Reseach Council, London.

GOOD, I. J. (1977). The botryology of botryology, sta in Van Ryzin J. ed., *Classification and Clustering*. Academic Press, New York.

GOODMAN, L. A. e KRUSKAL, W. H. (1954). Measures of association for cross classifications. *J. Amer. Statist. Ass.*, Vol. 49.

GOODMAN, L. A. e KRUSKAL, W. H. (1959). Measures of association for cross classifications: II Further discussion and references. *J. Amer. Statist. Ass.*, Vol. 54.

GOWER J. C. (1972). Measurements of taxonomic distance and their analysis, sta in WEINER J. S. e HUZINGA J. eds., *The Assessment of Population Affinities*. Claredom Press, Oxford.

GREEN, P. E., FRANK, R. E. e ROBINSON, P. J. (1967). Cluster analysis in test market. *Management Science*, Vol. 13.

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

HARTIGAN, J. A. (1977). Distribution problems in clustering, sta in Van Ryzin J. ed., *Classification and Clustering*. Academic Press, New York.

HERZEL, A. (1974). Un criterio di quantificazione. Aspetti statistici. *Metron*, Vol. XXXII.

HUBERT, J. e BAKER, F. B. (1977). An empirical comparison of baseline models for goodness-of-fit in r-diameter hierarchical clustering, sta in Van Ryzin J. ed., *Classification and Clustering*. Academic Press, New York.

JACCARD, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vand. Sci. Nat.*, Vol. 44.

JANCEY, R. C. (1966). Multidimensional group analysis. *Austr. J. Botany*, Vol. 14.

JARDINE, N. e SIBSON, R. (1971). *Mathematical Taxonomy*. John Wiley, New York.

JOHNSON, R. L. e WALL, D. D. (1969). Cluster analysis of semantic differential data. *Educ. Psychol. Measur.*, Vol. 29.

KENDALL, M. G. (1966). Discrimination and classification, sta in Krishnaiah P. R. ed., *Multivariate Analysis*. Academic Press, New York.

KENDALL, M. G. (1972). The basic problems of cluster analysis, sta in Cacoullos T. ed., *Discriminant Analysis and Applications*. Academic Press, New York.

KRUSKAL, J. (1977). The relationship between multidimensional scaling and clustering, sta in Van Ryzin J. Ed., *Classification and Clustering*. Academic Press, New York.

LANCE, G. N. e WILLIAMS, W. T. (1966a). Computer programs for hierarchical polythetic classification (similarity analyses). *Comp. J.*, Vol. 9.

LANCE, G. N. e WILLIAMS, W. T. (1966b). A generalized sorting strategy for computer classifications. *Nature*, Vol. 212.

LANCE, G. N. e WILLIAMS, W. T. (1967a). A general theory of classificatory sorting strategies: I. Hierarchical systems. *Comp. J.*, Vol. 9.

LANCE, G. N. e WILLIAMS, W. T. (1967b). A general theory of classificatory sorting strategies: II Clustering Systems. *Comp. J.*, Vol. 10.

LERMAN, I. C. (1970). *Les Bases de la Classification Automatique*. Gauthier-Villars, Paris.

LETI, G. (1961). Nuovi tipi di distanze fra insiemi di punti e loro applicazioni alla statistica. *Metron*, Vol. XXI.

LETI, G. (1978). *Distanze e Indici Statistici* (*stesura provvisoria*). Istituto di Statistica — Facoltà di Scienze Statistiche, Roma —.

LING, R. F. (1972). On the theory and construction of k-cluster. *Comp. J.*, Vol. 15.

LING, R. F. (1973). A computed generated aid for cluster analysis, *Comm. ACM*, Vol. 16.

LIS, A. e SAMBIN, M. (1977). *Analisi dei Clusters*. Cleup. Padova.

LIS, A. e SAMBIN, M. (1978). Verso il non parametrico: dalla analisi fattoriale all'analisi dei clusters. Lavoro presentato al Seminario di aggiornamento su « *Due temi di analisi statistica multivariata* », Bressanone.

LUNETTA, G. (1973). *Variabilità a più Dimensioni e Analisi dei Gruppi*. Istituto di Statistica, Catania.

MAC QUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th. Berkeley Symp.*, Vol. 1.

MATULA, D. W. (1977). Graph theoretic techniques for cluster analysis algorithms, sta in Van Ryzin J. ed., *Classification and Clustering*. Academic Press, New York.

MINEO, A. (1974). *Un nuovo metodo per l'analisi dei gruppi*. Facoltà di Economia e Commercio, Palermo.

MOJENA, R. (1977). Hierarchical grouping methods and stopping rules. An evaluation. *Comp. J.*, Vol. 20.

MORRISON, D. F. (1967). *Multivariate Statistical Methods*. Mc Graw-Hill, New York.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Ass.*, Vol. 66.

RAO, C. R. (1977). Cluster analysis applied to a study of race mixture in human populations, sta in Van Ryzin J. ed. *Classification and Clustering*. Academic Press, New York.

RIZZI, A. (1978). *Analisi dei Gruppi*. La Goliardica, Roma —.

ROGERS, D. J. e TANIMOTO, T. (1960). A computer program for classifying plants. *Science*, Vol. 132.

RUSSEL, P. F. e RAO, T. R. (1940). On habitat and association of species of anopheline larval in South-Eastern Madras. *J. Malor Inst. India*, Vol. 3.

SADOCCHI, S. (1977). Un metodo di cluster analysis non stratificato. *Rivista di Statistica Applicata*, Vol. 10.

SALVEMINI, T. (1960). Lezioni sull'analisi discriminatoria, sta in *Lezioni di Metodologia Statistica per Ricercatori*, Vol. 3. Facoltà di Scienze Statistiche, Roma.

SNEATH, P. H. A. (1957). The application of computers to taxonomy. *J. Gen. Microb.*, Vol. 17.

SNEATH, P. H. A. e SOKAL, R. R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

SOKAL, R. R. (1977). Clustering and classification background and current directions, sta in Van Ryzin J. ed. *Classification and Clustering*. Academic Press, New York.

SOKAL, R. R. e MICHENER, C. D. (1958). A statistical method for evaluating systematic relationship. *Un. Kansas Sci. Bull.*, Vol. 38.

SOKAL, R. R. e ROHLF, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, Vol. 11.

SOKAL, R. R. e SNEATH, P. H. A. (1963). *Principles of Numerical Taxonomy*. Freeman London.

SOLOMON, H. (1970). *Numerical Taxonomy*. Department of Statistics, Stanford University. Tech. Report n. 167.

SOLOMON, H. (1977). Data dependent clustering techniques, sta in Van Ryzin J.ed., *Classification and Clustering*. Academic Press, New York.

VAN RYZIN J. ed. (1977). *Classification and Clustering*. Academic Press, New York.

VITALI, O. (1972). Su un particolare aspetto della problematica dell'assetto del territorio: La classificazione dei comuni italiani secondo il grando di urbanità e ruralità e l'impiego dell'analisi discriminante. *Atti della XXVII riunione della S.I.S.*, Palermo.

TRYON, R. C. e BAILEY, D. E. (1970). *Cluster Analysis*. Mc Graw-Hill, New York.

WALLACE, C. S. e BOULTON, D. M. (1968). An information measure for classification. *Comp. J.*, Vol. 11.

WISHART, D. (1969). An algorithm for hierarchical classifications. *Biometrics*, Vol. 25.

ZADEH, L. A. (1977). Fuzzy sets and their application to pattern classification and clustering analysis, stà in Van Ryzin J. ed, *Classification and Clustering*. Acedemic Press, New York.

ZANI, S. (1975). Sulle proprietà degli indici di distanza nell'analisi classificatoria *Studi e Ricerche* della Facoltà di Economia e Commercio, Padova.

ZANI, S. (1977). L'analisi classificatoria: Contributi metodologici ed impiego per l'individuazione di aree omogenee, sta in *Aree di Sviluppo Socio-Economico e Comprensori in Emilia-Romagna*. Istituto di Statistica, Facoltà di Economia e Commercio, Padova.

ZANI, S. e SICURI, R. (1978). Stratificazione dei comuni della Emilia-Romagna in base alle caratteristiche dell'offerta di lavoro. *A.I.R.O.* Sogesta, Urbino.