Esse, in tre, coprono l'87% della variabilità delle componenti di primo livello. Rispetto agli indicatori originali si può ritenere che la variabilità spiegata sia dell'83% cioè quanto si poteva ottenere fattorizzando la matrice originaria S. Questo è coerente con l'idea che le componenti di secondo livello colgano, almeno in parte, le interazioni tra indicatori in macro-determinanti diverse trascurate dalle sotto-analisi.

Le componenti di secondo livello hanno però due debolezze: una di tipo tecnico ed una logica. Da un lato le componenti tralasciate in una sottoanalisi perché non rilevanti ai fini della variabilità spiegata di quella macro-determinante potrebbero invece avere legami molto forti con le componenti di altri blocchi. Il calcolo di queste ultime componenti risulterebbe perciò privato di una fonte potenzialmente rilevante di informazione. L'altra carenza riguarda la interpretabilità delle componenti di secondo livello: sono già note le difficoltà e la necessità di forti doti creative per comprendere il senso delle componenti di primo livello, tanto che, spesso, ci si deve contentare di descrizioni vaghe, di larga massima e soggettive. Il problema si dilata per le componenti di secondo livello, che sono labili fili che collegano fili già piuttosto evanescenti.

Al primo problema è difficile ovviare: la speranza è che i benefici siano maggiori dei costi connessi. Il secondo invece non è molto grave se si tiene conto che le componenti saranno utilizzate per la cluster analysis e che la caratterizzazione dei gruppi avverrà, come suggeriscono Friedman e Rubin (1967), nei termini degli indicatori originali e non sulla base degli indicatori artificiali generati dall'analisi delle compoenti principali.

3.3 La cluster analysis

Nella nostra ricerca siamo interessati a generare una tipologia di Comuni per gruppi distinti e non sovrapponibili che costituisca un quadro dei sistemi agricoli territoriali dell'Italia. Non si conosce a priori né il numero né la struttura dei gruppi; la stessa definizione di "gruppo" o di "cluster" è molto labile: le entità (i Comuni) formano un cluster quando si addensano (con

o senza vincolo di contiguità) intorno ad un polo per cui quelle interne al cluster sono più "vicine e simili" (o "meno lontane e dissimili") tra di loro di quanto non lo siano rispetto ad entità esterne. Le virgolettature sono d'obbligo in quanto non ci sono definizioni precise di termini quali cluster, distanza, similarità, che siano valide per tutte le applicazioni.

La cluster analysis, dopo un periodo di grande popolarità negli anni '60 e '70, ha conosciuto un declino soprattutto a causa della natura euristica della sua impostazione. R.M. Cormack (1971) inizia così una fondamentale rassegna sulla cluster analysis:

La disponibilità di packages per la classificazione automatica ha prodotto più spreco di tempo scientifico qualificato di qualsiasi altra "novità" statistica (ad eccezione forse della regressione multipla).

Il desiderio di definire una tipologia stringente ed esaustiva è spesso tanto forte da far dimenticare le debolezze teoriche delle procedure di cluster analysis nonché le precauzioni, sempre necessarie, nel valutarne i risultati. E' infatti possibile che i dati semplicemente non si prestino ad essere disaggregati, che non ci sia alcuna struttura di gruppo e tentare di imporla può portare a soluzioni poco significative (un esempio sono le sezioni di censimento in cui è suddivisa una città, che non necessariamente formano un quartiere). Potrebbe peraltro capitare che i dati siano un campione proveniente da una popolazione ben articolata in gruppi, ma che per effetto del caso il campione non rispecchi quella articolazione, ma un'altra, con più o meno gruppi, o con gruppi diversi. Anche in questo caso i risultati hanno scarsa affidabilità. Infine, l'interesse potrebbe essere non nel ricercare le unità che si prestano ad essere inserite in un gruppo, ma di altre che se ne stiano isolate e staccate dal resto (outliers) e non tutti i metodi sono in grado di evidenziarle.

E' però anche possibile che i gruppi proposti dalla cluster analysis risultino molto coesi e giustificabili teoricamente al punto da motivare la riduzione dell'insieme delle unità alle poche, una "tipica" per ogni cluster, realmente rappresentative. Il fine della

cluster analysis in questo caso sarebbe pienamente raggiunto. Ma è proprio questo il fine di tale analisi? Nella vastissima bibliografia sull'argomento si ritrova spesso la citazione di D.W. Goodall (riportata ad esempio in Cormack, 1971) che nel 1954 affermò:

"La tendenza a classificare si sviluppa sin dalla prima infanzia e permane come forma abituale di pensiero nell'età adulta".

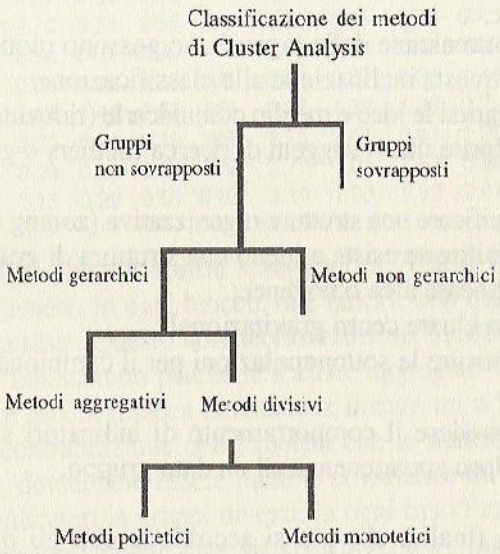
Vediamo alcune delle ragioni che possono motivare la persistenza di questa inclinazione alla classificazione:

a) Chiarirsi le idee e meglio comunicarle (riduzione dei dati);
b) Scoprire nuovi soggetti di ricerca (outliers o gruppi particolari);
c) Pianificare una struttura organizzativa (zoning territoriale);
d) Stabilire se esiste o meno una struttura di gruppo;
e) Delineare aree omogenee;
f) Individuare centri gravitazionali;
g) Costruire le sottopopolazioni per il campionamento stratificato;
h) Prevedere il comportamento di indicatori sulle entità in base alla loro appartenenza ad un dato gruppo.

Tra le finalità che più si accordano con gli obiettivi della nostra ricerca c'è certamente quella indicata al punto (d) perché siamo in una fase esplorativa e ci troviamo in uno stato di relativa ignoranza rispetto alla struttura di variabilità dei dati e all'esistenza di gruppi più o meno distinti di Comuni. Sono anche rilevanti gli obiettivi (b), (f) ed (h); è invece escluso il punto (c) in quanto le entità entrano nelle nostre elaborazioni senza riferimento alcuno alle loro coordinate geografiche. L'essenza è però, come si è già detto più volte, il punto (e), cioè l'obiettivo della nostra ricerca è la determinazione di una tipologia socio-economica dei Comuni che sia riconducibile ad un criterio ordinatore che ne spieghi le diversità.

La cluster analysis include numerose procedure che possono essere realizzate utilizzando svariati algoritmi. Data la evanescenza del concetto di cluster si può arrivare a soluzioni diverse

(peraltro, tutte giustificabili) anche agendo sullo stesso insieme di dati. Ad esempio, le carte di un mazzo francese si raggruppano per seme se si gioca a Bridge, ma si raggruppano per valore se si gioca a Ramino. Poiché i risultati sono influenzati sia dall'obiettivo dell'indagine che dal contesto applicativo occorre effettuare delle scelte che individuino la procedura più adatta.



Una prima valutazione deve riguardare la possibilità che l'appartenenza non sia esclusiva cioè che alcune entità possano ritrovarsi in più di un cluster (clumping; overlapping clusters). Una situazione di questo tipo potrebbe riscontrarsi nella classificazione dei pazienti secondo i sintomi mostrati, in quanto malattie diverse possono avere dei sintomi in comune. Un altro esempio è la classificazione delle parole: uno stesso termine può avere significati diversi e deve essere collocato in gruppi diversi. Tuttavia, visto l'obiettivo di zoning che ha la nostra ricerca, adotteremo il principio della appartenenza esclusiva delle entità ed in questa lavoro discuteremo solo di tecniche di classificazione esaustive (tutte le entità sono classificate) in clusters mutualmente esclusivi (ogni entità è posta in uno ed un solo cluster).

Una seconda scelta è necessaria tra l'adozione di procedure gerarchiche oppure non gerarchiche. Nel primo caso le entità si confrontano a coppie e si valuta se esse siano abbastanza simili per stare in uno stesso cluster oppure debbano stare in clusters diversi. Il metodo gerarchico cerca una sequenza di livelli che tracci un percorso dalla singola unità al complesso delle unità (o viceversa). Ad ogni livello si configura una struttura di gruppi ed è compito del ricercatore stabilire se essa sia significativa o meno rispetto alle finalità che si intendevano raggiungere. Nel caso dei metodi non gerarchici l'attenzione si sposta più sulla omogeneità interna dei gruppi e, attraverso un meccanismo iterativo, si tenta di ottimizzare, attraverso trasferimenti e scambi di entità tra gruppi, una qualche funzione obiettivo che misuri il grado di omogeneità all'interno dei singoli gruppi. Il risultato finale è una partizione ottimale (in base a certi parametri) delle entità in sottogruppi distinti.

Se si scelgono le procedure gerarchiche occorre anche scegliere in che direzione far muovere il collegamento tra l'unità e l'insieme delle unità. Si opta per le tecniche agglomerative se ad ogni passo si procede a raggruppare i clusters formatisi ai livelli precedenti in gruppi di ampiezza crescente fino a che il complesso delle entità formi esso stesso un unico grande gruppo.

Si può però optare per delle tecniche divisive o scissorie che, partendo dall'intero dataset procedano ad ogni passo con una bipartizione di uno dei clusters formatisi al livello precedenti, fino a definire una struttura ottima (in qualche senso definito) di raggruppamento. Le tecniche divisive, infine, possono essere monotetiche, ovvero la suddivisione passare, ad ogni cambio di livello, per uno solo degli indicatori, oppure politetiche, se la appartenenza o meno al cluster da bipartire è stabilita considerando più variabili.

3.3.1 Misure della distanza e della similarità.

I concetti di distanza e similarità hanno molta importanza per la cluster analysis ed è opportuno dedicare ad essi un po' di

attenzione. Diciamo subito però che limiteremo la discussione a concetti applicabili a variabili quantitative ed escluderemo perciò indici e misure basati sulle frequenze o sulla posizione d'ordine dei valori che si applicano in caso siano coinvolte variabili qualitative, nominali od ordinali.

Misure della distanza.

Indichiamo come al solito con X_i il vettore delle osservazioni sugli "m" indicatori rilevati sulla entità i-esima e sia $\Omega=\{X_i, i=1, 2, \ldots, n\}$ il dataset da analizzare. La distanza tra due entità qualsiasi è una funzione non negativa "d" definita su $(R^m x R^m)$ ed a valori in R^1 che gode delle seguenti, proprietà:

$d(X_i, X_j) \geq 0$	Non negatività
$d(X_i, X_j) = 0$ se e solo se $X_i = X_j$	Identità
$d(X_i, X_j) = d(X_j, X_i)$	Simmetria
$d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)$ per ogni "i,j,k"	Disuguaglianza triangolare

Tali proprietà derivano da quelle ormai familiari della distanza euclidea e, come questa, basate sull'idea che il percorso più breve tra due punti proceda in linea retta e la distanza tra quei punti sia la lunghezza del segmento che li unisce. Interpretando le entità come punti in uno spazio euclideo, le prime due proprietà affermano che la distanza è una quantità non negativa e che l'unica distanza nulla è quella di un punto da se stesso. La terza implica che data la distanza tra X_i ed X_j ne esiste un'altra, di pari lunghezza, procedendo all'indietro da X_j ad X_i. La quarta proprietà afferma che se scegliamo il percorso più breve da X_j ad X_k e poi il percorso più breve da X_k a X_i , avremo certamente fatto un cammino non inferiore di quello che avremmo percorso se fossimo andati direttamente da X_i a X_j. E' questa proprietà che rende "metrica" la funzione "d". Qui di seguito si riportano alcuni esempi di funzione di distanza correntemente in uso in varie procedure di clustering:

$$D_1 = \frac{1}{m}\left[\sum_{k=1}^{m}\left|\frac{X_{ik}-X_{jk}}{r_k}\right|^p\right]^{1/p} \qquad D_2 = \frac{1}{m}\sum_{k=1}^{m}\left|\frac{X_{ik}-X_{jk}}{X_{ik}+X_{jk}}\right|^p \qquad D_3 = \frac{1}{m}\sum_{k=1}^{m}\frac{|X_{ik}-X_{jk}|}{|X_{ik}|+|X_{jk}|}$$

$$D_4 = \frac{\left(\sum_{k=1}^{m} X_{ik}X_{jk}\right)^2}{\sum_{k=1}^{m} X_{ik}^2 + \sum_{k=1}^{m} X_{jk}^2} \qquad D_5 = \frac{1}{m}\sum_{k=1}^{m}\frac{|X_{ik}-X_{jk}|}{\max\{X_{ik},X_{jk}\}} \qquad D_6 = \frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{\min\{X_{ik},X_{jk}\}}{\max\{X_{ik},X_{jk}\}}\right)$$

In queste formule r_k è di solito, una misura di centralità o di dispersione dell'indicatore e "p" un intero maggiore o uguale ad uno. Poiché il risultato di ogni clustering è fortemente legato al modo in cui sono misurate le distanze, sarà sempre bene verificarne i risultati utilizzando definizioni alternative di misure della distanza.

Le $\{D_i\}$ sono tutte metriche. Consideriamone ora una molto usata e basata sul coefficiente di correlazione lineare:

$$r_{ij} = \frac{\sum_h (x_{ih}-\mu_i)(x_{jh}-\mu_j)}{\sigma_i \sigma_j}; \quad d_{ij} = 2(1-r_{ij});$$

questa misura non verifica la seconda proprietà (essere nulla esclusivamente in caso di identità tra i due punti) ed inoltre non necessariamente verifica la disuguaglianza triangolare e quindi non può considerarsi una vera e propria misura di distanza; piuttosto è una misura di similarità come quelle discusse nel prossimo sottoparagrafo.

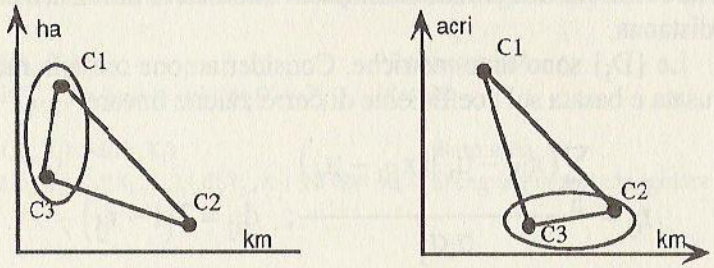
Le misure di distanza soffrono tutte del medesimo difetto: gli indicatori che più contribuiscono al valore di "d" sono quelli con unità di misura più grande: cioè un indicatore espresso in centimetri finisce col pesare di più che se fosse espresso in metri. Inoltre, una semplice alterazione della scala di misurazione non solo varia il valore numerico di "d", ma modifica anche la graduatoria delle distanze tra le entità.

Consideriamo ad esempio la seguente tabella relativa a tre comuni e due variabili: una misurata in chilometri e l'altra in ettari.

Comuni	V1(Km)	V2(ha)		C1	C2	C3
C1	5	5.00	C1		9.25	2.00
C2	8	5.50	C2	9.25		4.25
C3	6	6.00	C3	2.00	4.25	

Comuni	V1(Km)	V2(acri)		C1	C2	C3
C1	5	12.36	C1		10.53	7.11
C2	8	13.59	C2	10.53		5.53
C3	6	14.83	C3	7.11	5.53	

Nella parte destra della tabella è riportata la matrice delle distanze euclidee (D_1 con p=1 e r_k=1 per ogni indicatore). Come si vede dal grafico seguente



la gerarchia delle distanze è: $d_{12} > d_{23} > d_{13}$; se però V2 è espresso in acri (moltiplicando il secondo indicatore per 2.47) la gerarchia è: $d_{12} > d_{13} > d_{23}$: e quindi il "cluster" è formato da C2 e C3 e non più da C1 e C3.

La standardizzazione degli indicatori (già discussa nell'analisi delle componenti principali) eliminando gli effetti dell'unità di misura eliminerebbe questo problema (con la standardizzazione, tutti gli indicatori sono forzati ad avere la stessa media e la stessa varianza), ma ne porrebbe di nuovi: basti pensare alla conseguente attenuazione delle differenze nei gruppi che è invece il fondamento principale della cluster analysis. In questo senso apparirebbe meno restrittiva la normalizzazione dei dati descritta nel paragrafo 3.1.3. In realtà siamo di fronte ad un problema di circolarità: l'unico modo per risolvere il problema della unità di misura nella cluster analysis è quello di conoscere la struttura dei gruppi, che è invece proprio il problema da risolvere!

Ulteriori difficoltà scaturiscono dalla ponderazione degli indicatori nella formula della distanza. Uno schema di pesi molto flessibile è quello adottato nella D1, in cui si possono definire gli $\{r_k\}$ in modo da espandere o comprimere il ruolo della k-esima variabile nel definire il valore della distanza. Ma come scegliere gli r_k?

Una misura che non ricade fra le precedenti e che è abbastanza usata è la distanza di Mahalanobis:

$$M_{ij} = (X_i - X_j)^t S^{-1} (X_i - X_j)$$

dove "S" la matrice delle varianze-covarianze degli indicatori. La M_{ij} é invariante rispetto a trasformazioni del tipo:

$$Y_i = AX_i + b; \qquad i = 1, 2, \ldots, n$$

con "A" matrice non singolare. Se le variabili sono incorrelate la distanza di Mahalanobis coincide con la D1 con r_k pari alla devianza del k-esimo indicatore. Rispetto a M_{ij} si deve osservare che presuppone clusters che abbiano la stessa matrice di varianze-covarianze (un'ipotesi molto forte) e se questo non succede il suo uso è forse più dannoso che utile. Hartigan (1975) arriva ad ipotizzare che l'uso della distanza di Mahalanobis riduca la distinguibilità dei clusters più di quanto non faccia la standardizzazione degli indicatori.

Distanza ultrametrica.

Se l'ultima delle proprietà della funzione di distanza indicate in precedenza è sostituita dalla condizione seguente:

$$d(X_j, X_j) \le \max\{d(X_j, X_k); d(X_i, X_k)\}$$

allora "d" è una distanza ultrametrica ed ha particolare rilevanza nei metodi di cluster gerarchici. La condizione è più restrittiva

di quella già posta in quanto, essendo la distanza una funzione simmetrica a valori non negativi, si ha

$$\max\{d(X_j,X_k);d(X_i,X_k)\} \le d(X_j,X_k)+d(X_i,X_k)$$

per cui una distanza ultrametrica è anche metrica, ma non viceversa. Da notare che la distanza di Mahalanobis non necessariamente è ultrametrica.

La similarità.

La similarità riguarda la prossimità tra due entità quale risulta dalla rilevazione di "m" indicatori su di entrambe. Una misura della similarità deve soddisfare le seguenti condizioni:

$s(X_i, X_j) = S_{ij} \ge 0$	Non negatività
$s(X_i, X_j) = s(X_j, X_i)$	Simmetria
$s(X_i, X_j)$ = massima se i=j	

La disuguaglianza triangolare non è richiesta in quanto si preferisce privilegiare le proprietà ordinali delle misure piuttosto che i valori metrici. Quindi, S_{ij} misura il grado di somiglianza tra due entità e può assumere un insieme di valori che va dallo zero (entità diametralmente opposte) ad un massimo. Se si vuole usare un indice che varia tra zero ed uno si può porre:

$$S^*_{ij} = \frac{S_{ij}}{\max\{S_{ij}\}}$$

Se poi si ritiene utile un campo di variazione che includa anche dei valori negativi si può utilizzare l'espressione:

$$S^+_{ij} = 2^* \frac{S_{ij}}{\max\{S_{ij}\}} - 1$$

che varia tra meno uno ed uno. Esistono anche svariate misure della similarità che si rivolgono preferibilmente ad indicatori in

scala nominale od ordinale. Un esempio molto noto per indicatori su scala ordinale è l'indice di Spearman:

$$\rho_{ij} = \frac{6}{n(n-1)} \sum_h (g_{ih} - g_{jh}); \qquad -1 \le \rho_{ij} \le 1$$

in cui i valori degli indicatori sono sostituiti con le posizioni d'ordine da essi occupate nella graduatoria delle entità.

Anche per indicatori dicotomi, che possono cioè assumere solo due valori, diciamo 0 e 1, gli indici di similarità sono numerosi. Se indichiamo con a, b, c, d le combinazioni di

		entità "j"	
		1	0
entità "i"	1	a	b
	0	c	d

presenza/assenza dell'indicatore nelle due entità "i" e "j" possiamo costruire uno dei più noti coefficienti: *il simple matching coefficient:*

$$S^*_{ij} = \frac{a_{ij} + d_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}}$$

che è pari al numero di indicatori su cui le entità concordano – presente in entrambe o in entrambe assente – rapportato alla totalità dei confronti. Poiché non è ben chiaro il ruolo della contemporanea assenza degli indicatori ai fini della similarità (il fatto che il comune "i" ed il comune "j" non siano sul mare non li rende molto più simili) il d_{ij} è ignorato in molti altri indici. Citiamo ad esempio l'indice di Jaccard:

$$S_{ij} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

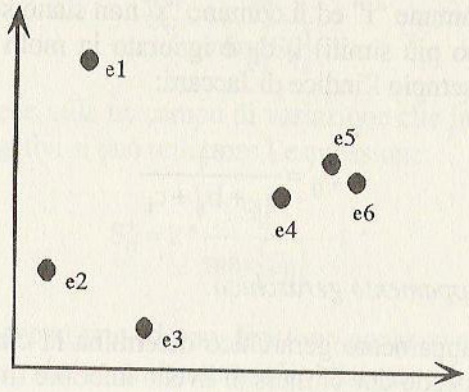
3.3.2 Raggruppamento gerarchico.

Il raggruppamento gerarchico determina la classificazione delle entità unendo due clusters di livello inferiore (metodi aggre-

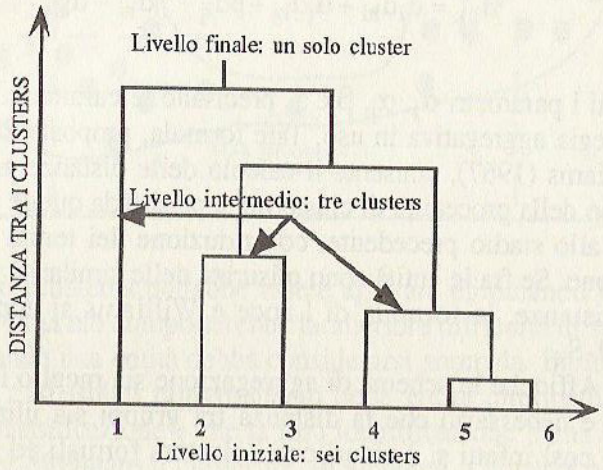
gativi) o bipartendo un cluster di livello superiore (metodi divisivi). Nel primo caso si parte dal massimo grado di specializzazione dei clusters ritenendo ogni X_i un'entità caratteristica e coincidente con uno specifico cluster. I clusters sono aggregati in base ad una qualche misura di similarità o distanza, in una sequenza concatenata e finita di passi ed i clusters del livello corrente derivano da quelli ottenuti al livello inferiore formando il cosiddetto *dendrogramma* (albero dei clusters) che rappresenta graficamente il risultato della classificazione. L'assegnazione ad un cluster è definitiva e le entità non possono essere riallocate in un altro cluster anche se questo potesse rivelarsi più proficuo ai fini della omogeneità della classificazione. Rimane poi solo da decidere qual'è il numero ottimo di clusters. Alternativamente, si potrebbe partire dal minimo livello di specializzazione ipotizzando che ci sia un unico grande cluster per tutte le entità. La sua suddivisione, in base ad un opportuno criterio, genera due clusters che possono a loro volta essere bipartiti se il numero ottimo di clusters è superiore a due. Poiché l'interesse della nostra ricerca è soprattutto per i metodi aggregativi, nella discussione che segue ci concentreremo solo su questi trascurando i metodi divisivi, cui è dedicato solo un sottoparagrafo.

Metodi aggregativi.

I metodi aggregativi partono dalla situazione di massima dispersione possibile in cui ogni unità forma un gruppo a sé.



Poi, attraverso una sequenza di fusioni in cui ogni volta si congiungono i due clusters meno distanti (gli altri clusters rimanendo invariati), si arriva alla situazione in cui tutte le entità si ritrovano in un solo cluster.



Per passare dal livello minimo a quello successivo si fondono i due clusters formati dalle entità "5" e "6"; al cluster così ottenuto si unisce quello formato dalla entità "4". Un ulteriore raggruppamento è ottenuto dall'unione dei clusters formati dalla "2" e dalla "3"per cui al livello intermedio esistono tre clusters {1}, {2,3}, {4,5,6}. Il livello finale è raggiunto prima fondendo i clusters {2,3} e {4,5,6} e poi unendo al cluster risultante l'entità "1" che, fino a questo livello, non era accostabile a nessun'altra.

Alla base di ogni metodo aggregativo c'è il calcolo delle distanze (o delle dissimilarità) per ciascuna delle n(n-1)/2 coppie di entità; le due entità con distanza minore si fondono per formare un nuovo gruppo. Negli stadi successivi andrà stabilita la distanza entità-cluster e cluster-cluster: sono queste scelte che caratterizzano le diverse strategie aggregative. Supponiamo che, ad un certo stadio, d_{ij} misuri la distanza tra il gruppo "C_i" contenente n_i entità ed il gruppo "C_j" che ne contiene n_j e che inoltre "C_i" e "C_j" debbano essere accorpati per formare il gruppo

"k" contenente $n_k=n_i + n_j$ entità. Per valutare adesso la distanza tra il nuovo gruppo ed un altro, diciamo il gruppo "h" con n_h entità, si può utilizzare la formula generale:

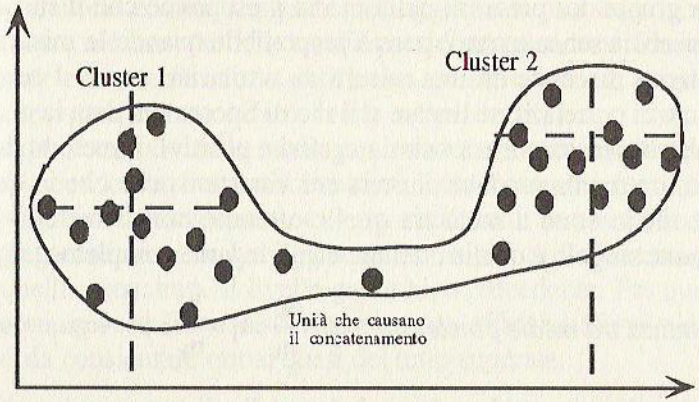
$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

in cui i parametri α_i, α_j, β e γ precisano le caratteristiche della strategia aggregativa in uso. Tale formula, proposta da Lance e Williams (1967), consente il calcolo delle distanze ad un dato stadio della procedura di clustering a partire da quelle già calcolate allo stadio precedente, con riduzione dei tempi di elaborazione. Se fra le entità sono misurate delle similarità S_{ij} invece di distanze, la formula di Lance e Williams si applicherà a $d_{ij}=1-S_{ij}$.

Affinché lo schema di aggregazione sia meglio interpretabile è necessario che la distanza tra gruppi sia ultrametrica: solo così infatti si è sicuri che i clusters formati ad un livello includano tutti i cluster già definiti al livello inferiore e sia possibile rappresentare perfettamente in un dendrogramma i vari livelli di agglomerazione. La formula di Lance e Williams definisce una distanza ultrametrica se i parametri verificano la disequazione: $\alpha_i + \alpha_j + \beta \geq 1$. Sono disponibili diverse specificazioni del legame generale ed ognuna porta ad una specifica tecnica gerarchica.

Distanza minima o legame singolo: $\alpha_i = \alpha_j = 0.5;\ \beta = 0;\ \gamma = -0.5$.

La distanza tra due gruppi è definita come la più piccola tra quelle riscontrabili tra coppie di entità di cui una appartenente al primo gruppo e l'altra al secondo. L'aggregazione dipende quindi dal legame tra due singole entità. Questa tecnica, pur avendo diversi pregi teorici, presenta la sgradevole caratteristica del "concatenamento" cioè di generare gruppi anche nel caso di due gruppi molto lontani, ma legati l'uno all'altro da un'esile "catena" di entità.



Questo problema potrebbe essere aggirato eliminando i casi "anomali", ma ciò comporterebbe la ulteriore difficoltà di specificare quando una entità debba considerarsi anomala. Da notare poi che i casi isolati rappresentano delle entità intermedie, di difficile classificazione e che la loro identificazione è una delle ragioni per effettuare il clustering. In aggiunta, i clusters hanno spesso forma allungata, col rischio che le entità agli estremi del cluster siano molto diverse, forse più diverse di quanto non lo siano rispetto ad entità in altri cluster. Il legame singolo è consigliato quando la contiguità delle entità è considerata rilevante.

Distanza massima o legame completo: $\alpha_i = \alpha_j = 0.5;\ \beta = 0;\ \gamma = 0.5$.

E' il logico contraltare della tecnica del legame singolo. Perché due gruppi siano aggregati è necessario che la distanza più grande riscontrabile tra due entità, una appartenente al primo gruppo e l'altra al secondo, sia inferiore alla distanza massima riscontrabile tra due entità appartenenti allo stesso gruppo. Il legame completo tende a formare gruppi molto compatti e sferici.

Distanza media $\alpha_i = \frac{n_i}{n_k};\alpha_j = \frac{n_j}{n_k};\ \beta = 0;\ \gamma = 0$.

La distanza tra due gruppi è data dalla media aritmetica delle distanze riscontrabili tra tutte le possibili coppie di entità dei

due gruppi. La presenza della media presuppone che il suo calcolo abbia senso e non è perciò proponibile quando la misura di distanza discende da una misura di similarità, come il coefficiente di correlazione lineare o il rho di Spearman, data la possibile compensazione tra valori negativi e positivi. Il metodo della distanza media produce clusters con caratteristiche che in qualche modo sono a metà tra quelle ottenute con il metodo del legame singolo e quelle ottenute con il legame completo.

Distanza tra medie ponderate: $\alpha_i = \frac{n_i}{n_k}; \alpha_j = \frac{n_j}{n_k}; \beta = -\alpha_i\alpha_j; \gamma = 0.$

La distanza tra due cluster è data dalla distanza tra i centroidi dei due clusters intendendo per centroide il vettore delle medie delle variabili, medie estese a tutte le unità incluse nel cluster. Il centroide di un cluster è utilizzato come rappresentazione del cluster. All'inizio, ogni entità coincide con il centroide del cluster costituito da se stessa. Non appena si ottiene un cluster dalla fusione di due clusters più piccoli, il centroide del nuovo cluster è dato da

$$\mu_h = \frac{n_i\mu_i + n_j\mu_j}{n_i + n_j}$$

Il metodo presuppone che i dati utilizzati siano tutti quantitativi cioè che abbia senso calcolare per essi la media aritmetica. Come è evidente dai valori dei parametri (la loro somma è inferiore all'unità) il metodo non è ultrametrico.

Distanza tra medie non ponderate: $\alpha_i = \alpha_j = 0.5; \beta = -0.25; \gamma = 0.$

E' molto simile al precedente tranne che il centroide rappresentativo di un nuovo cluster è dato dalla media aritmetica non ponderata dei valori delle entità dei clusters che si fondono per costituirlo:

$$\mu_h = \frac{\mu_i + \mu_j}{2}$$

Tale metodo è stato suggerito per aggirare il problema che si riscontra nel metodo delle medie ponderate quando si fonde un cluster contenente molte entità con un altro che ne contiene poche; quest'ultimo infatti, tende a "perdersi" nel cluster più grande. Il metodo delle medie non ponderate, come quello delle medie ponderate, non è ultrametrico, il che comporta (però raramente) delle conversioni, ovvero situazioni in cui i clusters aggregati ad un certo livello hanno una soglia di fusione minore di quella riscontrata al livello gerarchico precedente. Per queste ragioni le due ultime distanze non sono affidabilissime e sono anzi da considerare ormai quasi del tutto superate.

Metodo di Ward: $\alpha_i = \frac{n_i + n_h}{n_k + n_h}; \alpha_j = \frac{n_j + n_h}{n_k + n_h}; \beta = \frac{-n_h}{n_k + n_h}; \gamma = 0.$

L'idea di questa tecnica è di aggregare, ad ogni stadio, i due clusters per i quali è minore l'aumento della somma della variabilità all'interno di ciascun cluster. Al primo livello, quando tutte le unità formano un cluster di ampiezza uno, la varianza interna a ciascun cluster è zero. La fusione di due entità introduce una certa variabilità nel nuovo cluster. Il legame di Ward impone l'amalgama delle entità che danno il minor aumento nella variabilità. Quando i clusters sono dello stesso ordine di ampiezza e si presentano abbastanza compatti (cioè addensati intorno ad una moda significativa e con code sottili) il legame di Ward funziona al meglio ed i suoi risultati sono "robusti" rispetto a variazioni di numero, configurazione e separazione dei clusters. Molti autori sostengono che dovrebbe essere sempre scelto il legame di Ward a meno che non ci siano specifiche ragioni per usarne un altro, anche se esso, come è stato osservato, tende a favorire la formazione di clusters contenenti poche entità.

Metodo flessibile.

In questo tipo di legame rientrano moltissime strategie, tutte ultrametriche. Esse sono riconducibili alla formula generale dei

legami con i seguenti vincoli sui parametri: $\alpha_i + \alpha_j + \beta = 1$; 2) $\alpha_i = \alpha_j$; 3) $\beta < 1$; 4) $\gamma = 0$. In pratica però solo il metodo flessibile con $\beta = -0.25$ è menzionato nella letteratura.

I metodi gerarchici aggregativi sono utili quando le entità hanno una chiara tendenza a raggrupparsi e l'organizzazione in gruppi si presta naturalmente ad essere interpretata in termini di una struttura ad albero: è questo il caso di entità del mondo animale e vegetale. Se mancano queste condizioni i risultati possono facilmente portare fuori strada. C'è peraltro un uso molto efficiente di questi metodi ed è per la individuazione dei valori anomali (outliers) che infatti tendono a formare dei microclusters molto evidenti nel dendrogramma.

Un grande svantaggio di questi metodi è che spesso i raggruppamenti più interessanti si riscontrano ai livelli di fusione più elevati e che per ottenere questi bisogna necessariamente passare per tutti i livelli precedenti. Ne consegue che questi metodi sono lenti e dispersivi in quanto si spreca tempo e risorse di calcolo per ottenere risultati che spesso sono di scarso interesse. D'altra parte, le aggregazioni partono dal livello minimo di informazioni dove più frequenti sono gli errori di classificazione e poiché la sequenza gerarchica non consente di correggere errori fatti ai livelli precedenti qualche distorsione deve essere sempre preventivata.

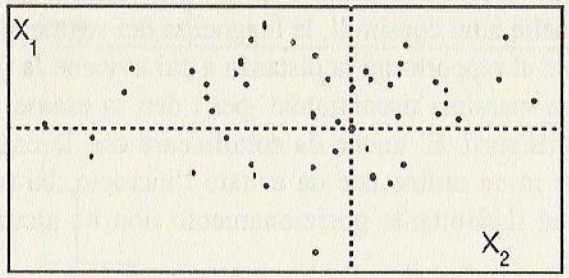
Nonostante gli indubbi vantaggi scientifici dati dal dendrogramma che mostra gli oggetti nel contesto di oggetti simili graduando inoltre la distanza, la clustering gerarchica non ha avuto la diffusione che merita a causa della rapida crescita delle dimensioni della matrice di distanze all'aumentare degli oggetti da considerare: per raggruppare i circa 8000 comuni italiani occorrerebbe calcolare, almeno una volta, un numero di distanze pari a circa 32,000,000 (anche calcolandone mille al secondo -tempo non disprezzabile- il solo calcolo della matrice delle distanze entità-entità richiederebbe circa 9 ore. Sibson (1973) ha proposto un algoritmo di calcolo per il legame singolo molto efficiente ed in grado di trattare datasets dell'ordine delle 10,000 entità.

Alcune simulazioni.

Al fine di dare un'idea dei problemi che si devono affrontare quando si sceglie di utilizzare i raggruppamenti gerarchici aggregativi, segnatamente il metodo della distanza minima (legame singolo) ed il metodo di Ward, cioè i due più diffusi, usando in entrambi i casi la metrica euclidea, si sono simulate tre situazioni.

A1) Campione proveniente da una popolazione senza struttura di gruppo.

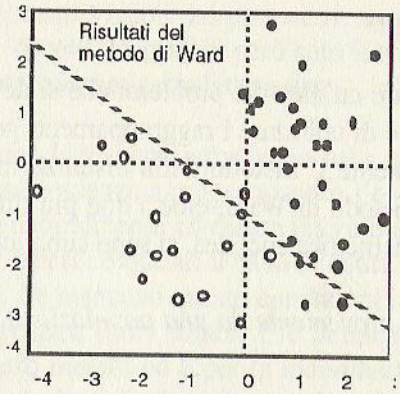
Si tratta 50 osservazioni generate dalla normale bivariata con medie nulle, varianze pari a due e covarianza zero.



Lo scattergram mostra che le entità si distribuiscono casualmente sul piano e la loro disposizione non suggerisce alcuna struttura di gruppo.

Tuttavia, il metodo di Ward, individua due clusters: uno di 31 entità ed un altro contenente le rimanenti 19, suddivise secondo linee puramente immaginarie. Ad esempio, bipartendo le unità nel primo e nel quarto quadrante da quelle del secondo e terzo (linea verticale passante per l'origine); ma anche la retta inclinata negativamente riportata nella figura potrebbe funzionare da demarcazione tra i due gruppi. Il grafico evidenzia queste ipotetiche bipartizioni.

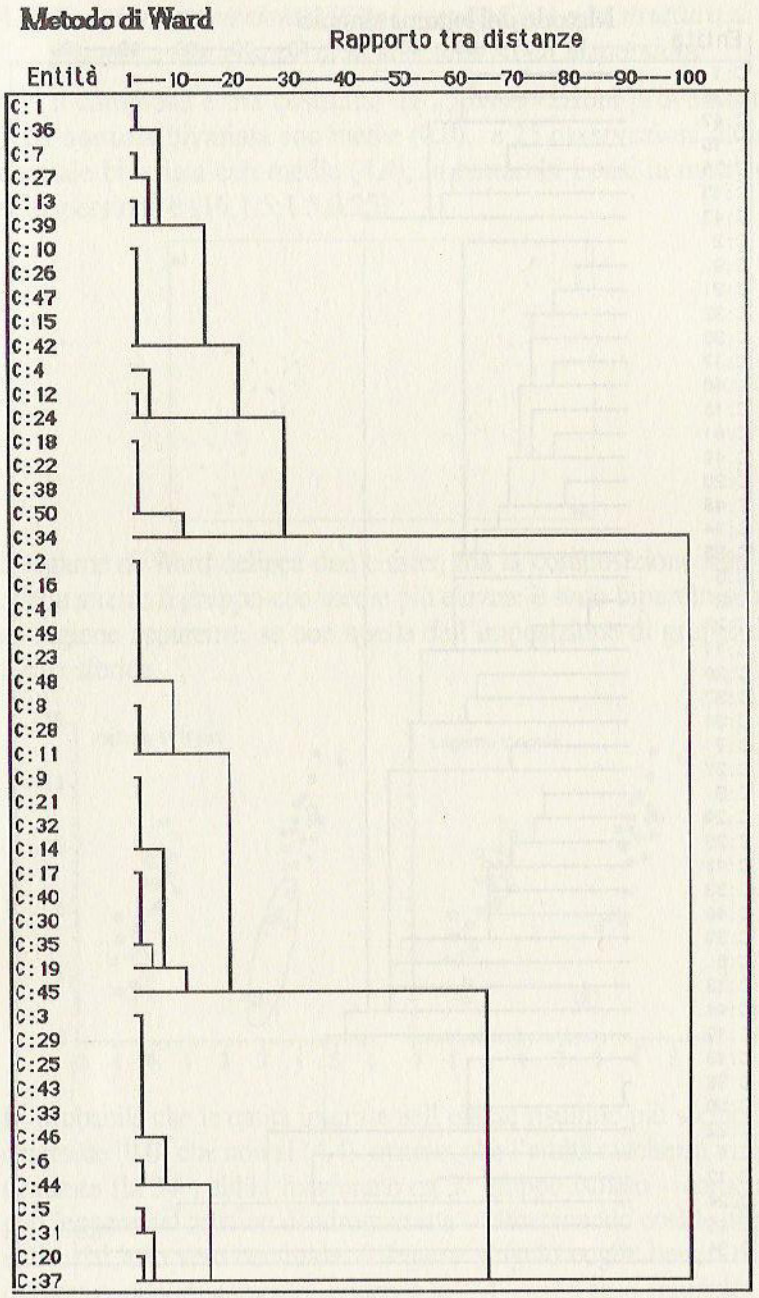
Il metodo di Ward ha forzatamente imposto una struttura di gruppo (sulla scelta del numero dei gruppi si veda più avanti il

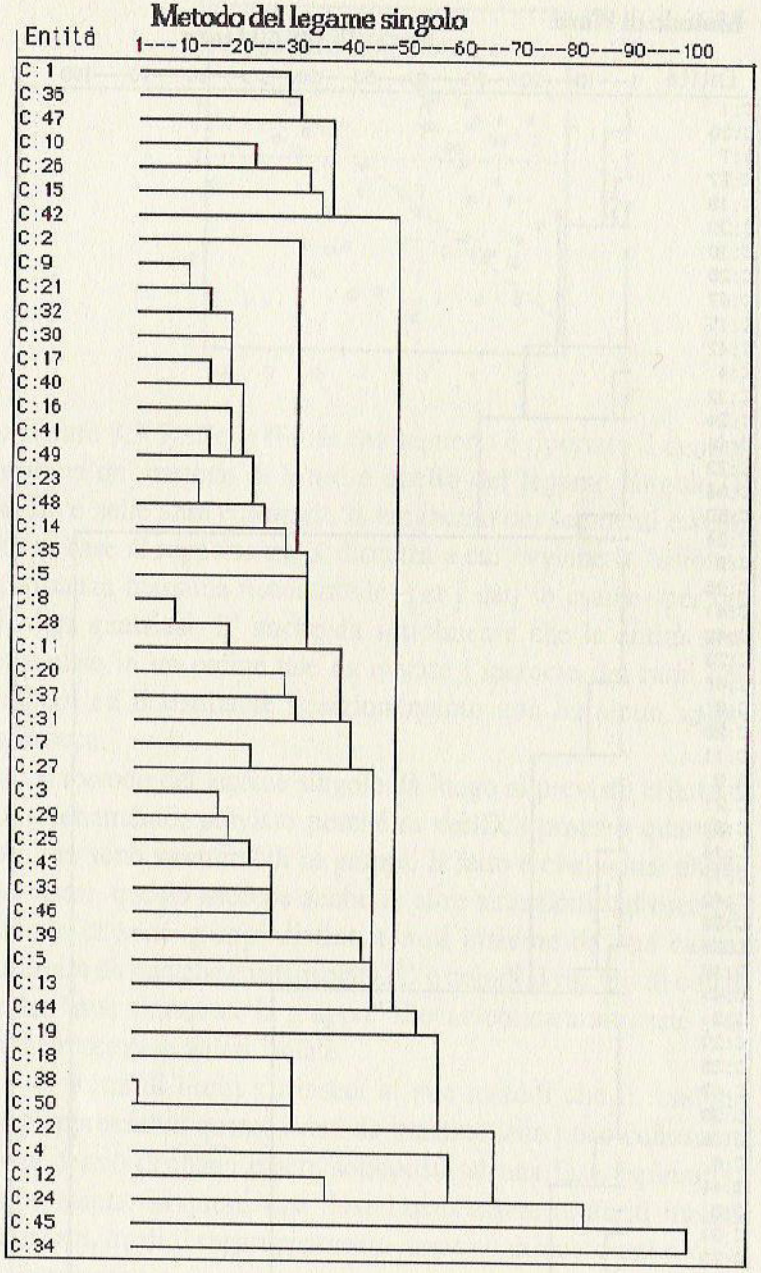


paragrafo 3.3.5). Nelle figure che seguono è riportato il dendrogramma del metodo di Ward e quello del legame singolo. In queste e nelle altre consimili, la lunghezza dei segmenti è stabilità in base al rapporto tra la distanza a cui avviene la fusione e la distanza massima riscontrabile -per i dati in esame- per due clusters qualsiasi. E' anche da sottolineare che le entità sono presentate in un ordine tale da evitare l'incrocio dei rami dell'albero ed il risultante posizionamento non ha alcun valore intrinseco.

Il metodo del legame singolo dà luogo al previsto effetto di concatenamento; previsto perché si verifica proprio quando i dati non sono strutturabili in gruppi. Il fatto è che, come abbiamo visto, questo succede anche in altre situazioni: ad esempio quando ci sono gruppi distinti tenuti insieme da una catena costituita da qualche entità isolata. C'è quindi il rischio di considerare "non strutturate in gruppo" rilevazioni caratterizzate solo dalla presenza di valori isolati.

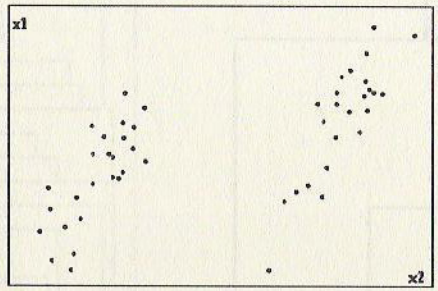
Si tratta di limiti intrinseci ai due metodi che li rendono poco apprezzabili quanto i dati da trattare siano poco conosciuti e che perciò debbano essere sottoposti ad una fase esplorativa molto spinta. In questi casi dovrebbero essere preferiti metodi più neutri, quali il raggruppamento iterativo discusso nel prossimo paragrafo.



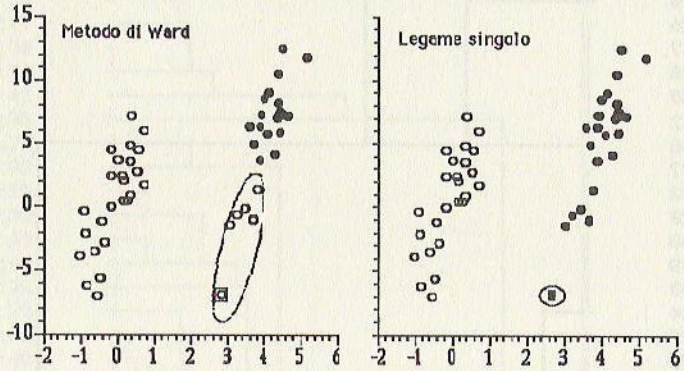


A2) Campione proveniente da una popolazione con struttura di gruppo e con gruppi di uguale matrice di dispersione

Il campione è ora costituito da 25 osservazioni provenienti dalla normale bivariata con medie (0,0), e 25 osservazioni dalla normale bivariata con medie (4,4); in entrambi i casi la matrice di dispersione è (16,1.5;1.5,0.25)

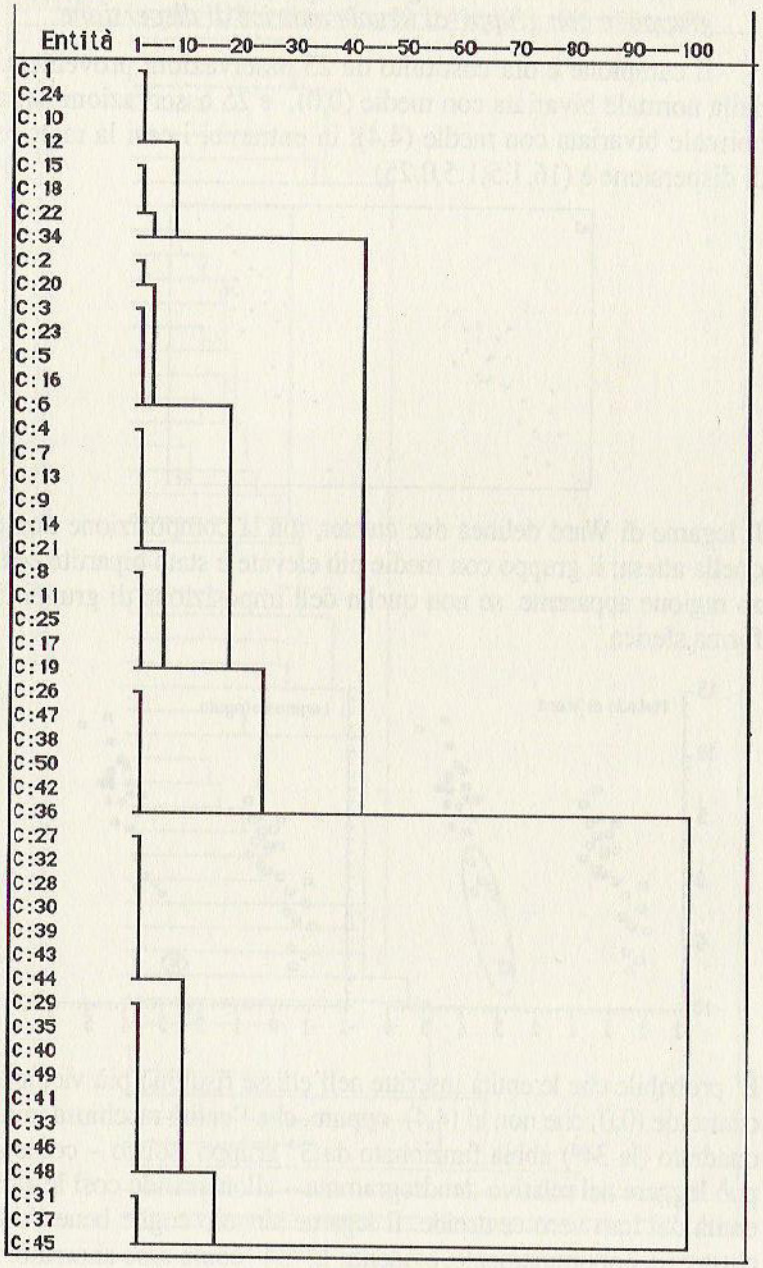


Il legame di Ward delinea due cluster, ma la composizione non è quella attesa: il gruppo con medie più elevate è stato bipartito senza ragione apparente, se non quella dell'imposizione di gruppi di forma sferica.

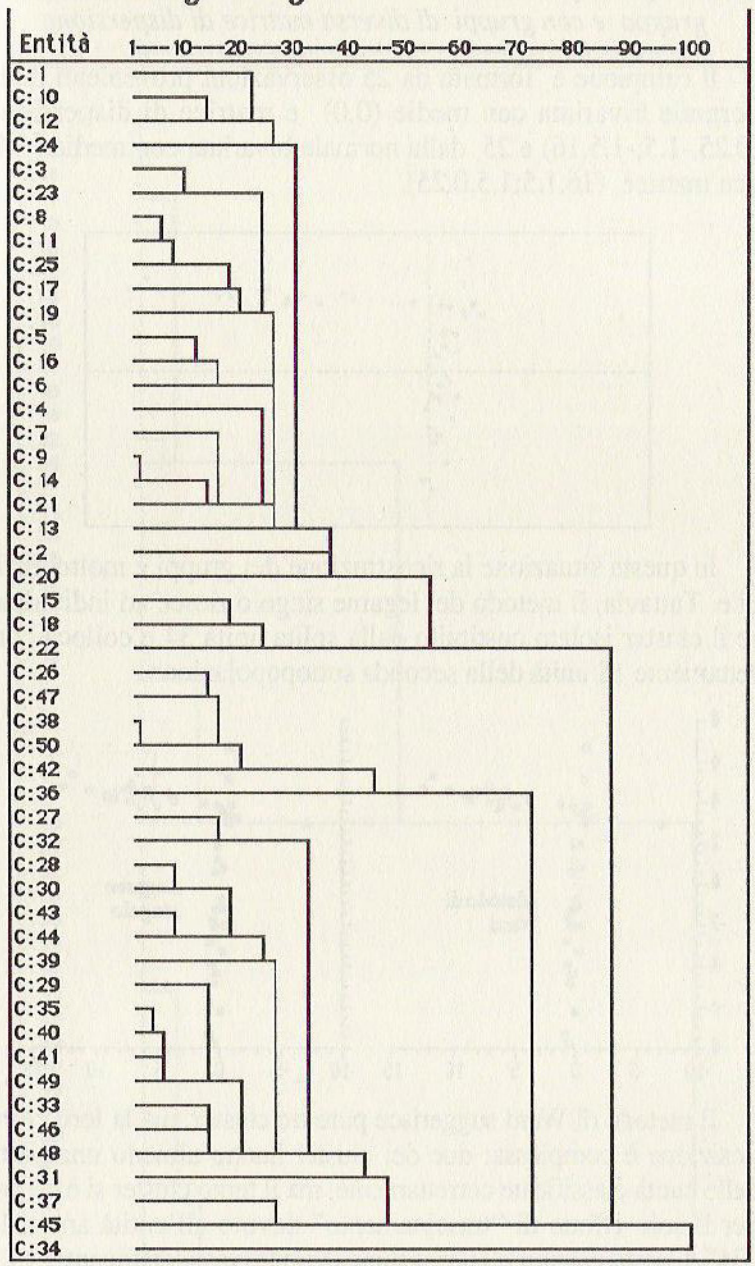


E' probabile che le entità inscritte nell'ellisse risultino più vicine al centroide (0,0) che non al (4,4) oppure, che l'entità racchiusa in un quadrato (la 34ª) abbia funzionato da 3° gruppo isolato – come si può leggere nel relativo dendrogramma – allontanando così le altre entità dal loro vero centroide. Il legame singolo coglie bene i due clusters e, opportunamente, evidenza la "34" come caso anomalo.

Metodo di Ward

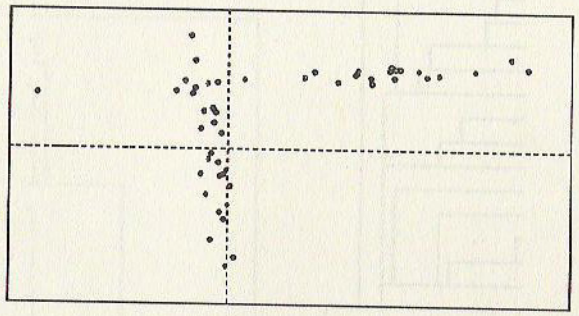


Metodo del legame singolo

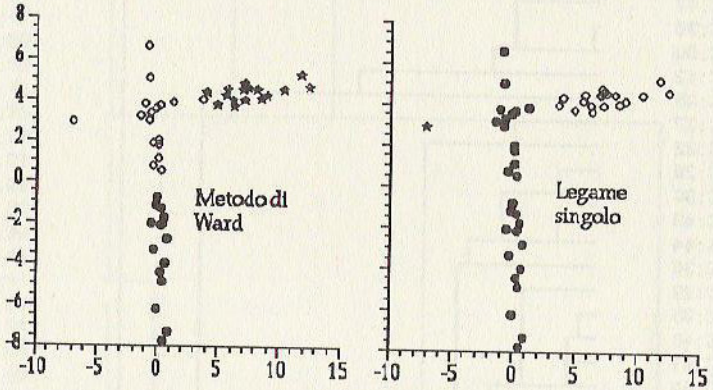


A3) Campione proveniente da una popolazione con struttura di gruppo e con gruppi di diversa matrice di dispersione

Il campione è formato da 25 osservazioni provenienti dalla normale bivariata con medie (0,0) e matrice di dispersione (0.25,-1.5;-1.5,16) e 25 dalla normale bivariata con medie (4,4) con matrice (16,1.5;1.5,0.25).

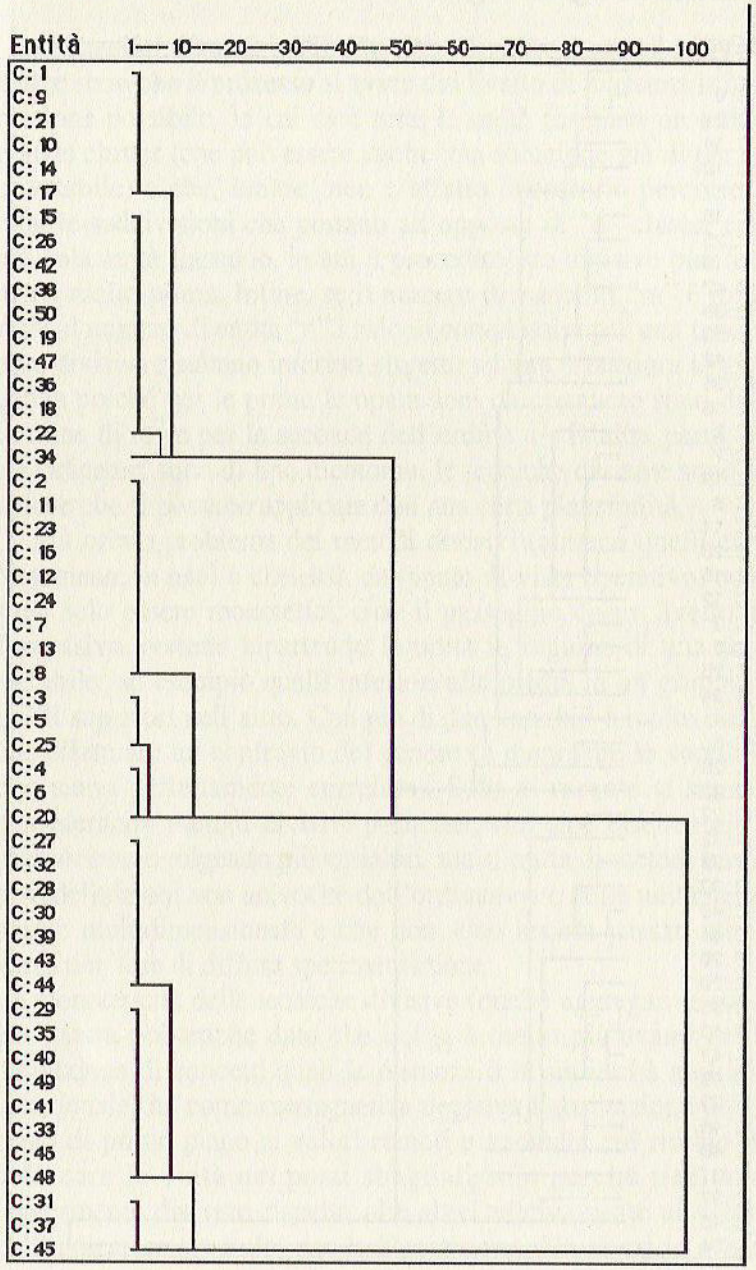


In questa situazione la ricostruzione dei gruppi è molto difficile. Tuttavia, il metodo del legame singolo riesce ad individuare il cluster isolato costituito dalla solita unità 34 e colloca correttamente 18 unità della seconda sottopopolazione.

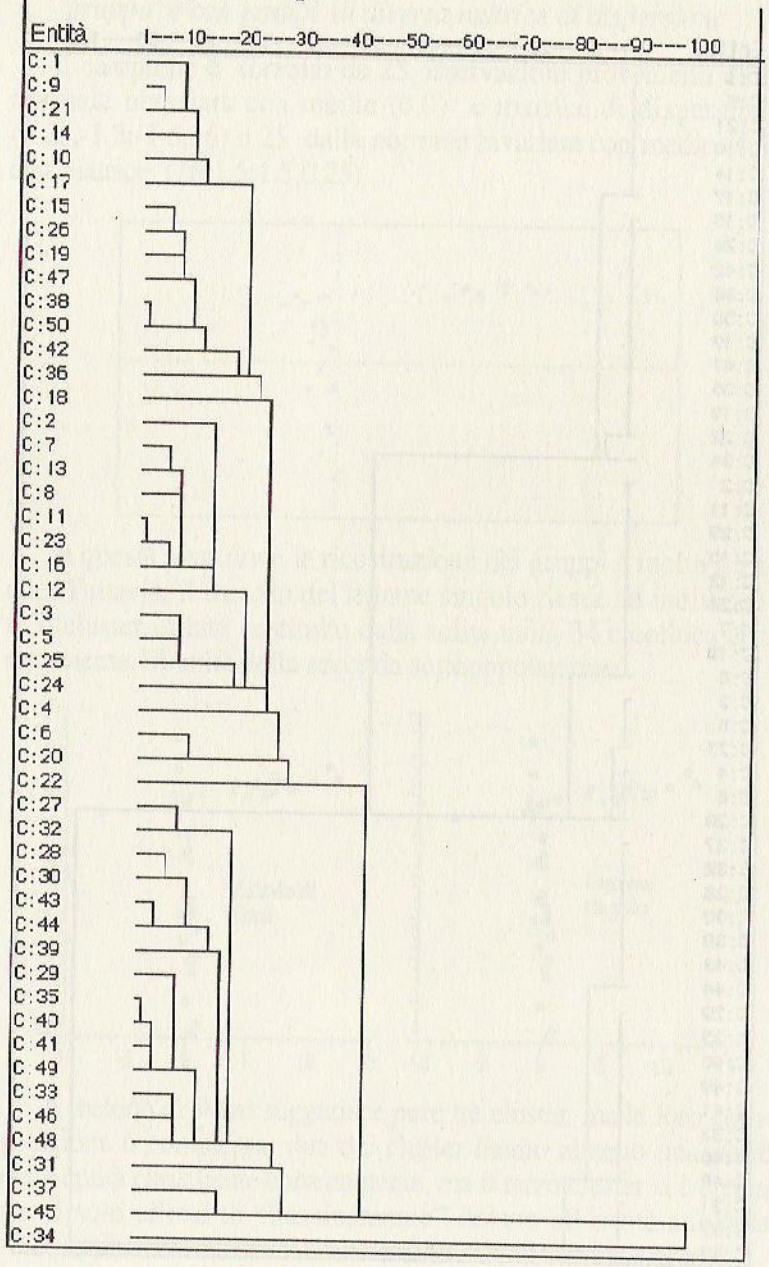


Il metodo di Ward suggerisce pure tre cluster, ma la loro composizione è complessa: due dei cluster hanno almeno una parte delle entità classificate correttamente, ma il terzo cluster si è creato per il solo effetto di "trascinamento" dovuto all'entità anomala "34". Questo costituisce un risultato del tutto insoddisfacente.

Metodo di Ward



Metodo del legame singolo



I metodi divisivi o scissori.

I maggiori vantaggi delle tecniche divisive su quelle aggregative sono che il processo si avvia dal livello di massima informazione possibile, in cui cioè tutte le unità formano un unico grande cluster (che può essere anche una soluzione già di per sé accettabile) e che, inoltre, non è affatto necessario percorrere tutte le suddivisioni che portano all'opposto di "n" cluster con una sola unità ciascuno; infatti il procedimento divisivo può fermarsi molto prima. Infine, se il numero di variabili "m" è inferiore al numero di entità "n" i calcoli complessivi per una procedura divisiva risultano inferiori rispetto ad una procedura aggregativa poiché per le prime le operazioni da compiere sono dell'ordine di m^2 e per le seconde dell'ordine n^2. D'altra parte, se gli indicatori sono di tipo dicotomo, le tecniche divisive sono le uniche che si possano applicare con una certa plausibilità.

Un primo problema dei metodi divisivi (almeno quelli correntemente in uso) è che essi, dal punto di vista operativo, possono solo essere monotetici, cioè il passaggio da un livello al successivo avviene bipartendo le unità in ragione di una sola variabile: ad esempio quelli inferiori alla media in un gruppo e quelli superiori nell'altro. Con più di due variabili è molto difficile effettuare un confronto del genere (a meno che le variabili non siano perfettamente correlate). Solo di recente si stanno considerando metodi divisivi politetici, che cioè effettuano la scissione coinvolgendo più variabili, ma si tratta di metodi basati su definizioni non univoche dell'ordinamento delle unità nello spazio multidimensionale e che non sono ancora passati attraverso una fase di diffusa sperimentazione.

La monoteticità delle tecniche divisive (quelle aggregative sono per natura politetiche dato che qui si è molto più avanti nella definizione di concetti quali la distanza o la similarità multidimensionale) ha come conseguenza negativa l'attribuzione di un ruolo di primo piano ai valori remoti o anomali, col rischio di collocare le unità nei posti sbagliati solo perché risultano leggermente discoste rispetto alle altre relativamente ai valori dell'indicatore prescelto per la bipartizione. Un secondo e ben

più importante problema è la mole di calcoli coinvolti con i metodi divisivi. Disponendo di "n" entità ci sono (2^{n-1} -1) possibili modi di suddividerle in due gruppi. Ad esempio, si potrebbe cercare la bipartizione dell'indicatore che rende massima la varianza tra i due gruppi

$$\max_{1\le i\le m}\left\{B_i = n_1(\bar{x}_{1i}-\bar{x}_i)^2 + n_2(\bar{x}_{2i}-\bar{x}_i)^2\right\}$$

dove $\bar{x}_1$ e $\bar{x}_2$ sono le medie dell'indicatore nei gruppi e $\bar{x}$ ne è la media totale. Dopo la prima suddivisione ci sono due gruppi che si possono suddividere e dopo la seconda suddivisione ce ne sono quattro. Il numero dei gruppi quindi cresce in ragione geometrica e ogni volta occorre esaminare un numero elevatissimo di possibili suddivisioni: se n=21 si dovranno valutare 1023 diverse bipartizioni e se n=41 se ne dovranno esaminare più di un miliardo.

3.3.3 Raggruppamento iterativo.

Le procedure che rientrano in questa categoria possono essere presentate secondo diversi approcci. Noi, per brevità, ne consideriamo solo alcuni tra quelli più accreditati: uno "metrico" incentrato sulla di distanza tra entità e clusters, un altro delle "normali distinte" che si basa su di un vettore di classificazione, ed infine l'approccio delle *"normal mixtures"*.

Approccio "metrico".

L'insieme di "n" entità $\Omega=\{X_a,\ a=1,2,\ \ldots,n\}$ deve essere ripartito in "k" sottoinsiemi (clusters) $C_1, C_2, \ldots, C_k$ ciascuno contenente almeno una entità e con nessuna entità in più di uno di essi. Se n_i indica il numero di entità nel cluster r-esimo si deve poi necessariamente avere: $n_1+n2+\ldots+n_k=n$ di modo che tutte le entità di "Ω" siano assegnate. Il cluster è l'insieme delle entità che sono più vicine al cluster:

$$C_j \subset \Omega \quad \text{con} \quad C_j=\left\{X_a\in\Omega \,\middle|\, \partial(X_a,C_j)=\min_{1\le r\le k}\{\partial(X_a,C_r)\};\ a=1,\ 2,\ \ldots,\ n\right\}$$

La valutazione della "∂" (distanza tra entità e cluster) su tutte le entità del dataset "Ω" dà luogo ad una partizione:

$$P_i=\{C_{1i},C_{2i},\ \ldots,\ C_{ki}\} \qquad C_{ir}\cap C_{is}=\varnothing; \qquad r\neq s$$

il problema del clustering nell'approccio metrico consiste nell'individuare la funzione di distanza "∂".

Per una data partizione P_i, definiamo il vettore delle medie e la matrice di varianze-covarianze osservate in ogni cluster:

$$\hat{\mu}_r=\frac{1}{n_r}\left[\sum_{X_a\in C_r}X_a\right]; \qquad \hat{\Sigma}_r=\frac{1}{n_r-1}\left[\sum_{X_a\in C_r}(X_a-\hat{\mu}_r)(X_a-\hat{\mu}_r)^t\right]; \qquad r=1,2,\ \ldots,\ k$$

A livello aggregato si ha l'identità: $T=\hat{\Sigma}+B$ dove T è la matrice delle devianze-codevianze totali, $\hat{\Sigma}$ è la somma delle matrici devianze-codevianze nei gruppi e B quella tra i gruppi:

$$\hat{\Sigma}=\sum_{r=1}^{k}(n_r-1)\hat{\Sigma}_r=\sum_{r=1}^{k}\sum_{a\in C_r}(X_a-\hat{\mu}_r)(X_a-\hat{\mu}_r)^t \qquad ; \qquad B=\sum_{r=1}^{k}n_r(\hat{\mu}_r-\hat{\mu})(\hat{\mu}_r-\hat{\mu})^t$$

con "$\hat{\mu}$" vettore delle medie calcolate su tutte le entità. In genere, la distanza tra entità e clusters si basa sulle matrici $\{\Sigma_r\}$ o meglio sulla loro stima in base ad una data partizione. Ecco degli esempi derivati da varie ipotesi circa la configurazione di tali matrici:

1. $d(X_a;C_r)=(X_a-\hat{\mu}_r)^t(X_a-\hat{\mu}_r)$;
2. $d(X_a;C_r)=(X_a-\hat{\mu}_r)^t\hat{\Sigma}^{-1}(X_a-\hat{\mu}_r)$;
3. $d(X_a;C_r)=(X_a-\hat{\mu}_r)^t\hat{\Sigma}_r^{-1}(X_a-\hat{\mu}_r)$;
4. $d(X_a;C_r)=(X_a-\hat{\mu}_r)^t\hat{\Sigma}_r^{-1}(X_a-\hat{\mu}_r)-n_rLn(n_r)$

Queste formule ipotizzano l'esistenza di un "centroide" o "polo" del cluster che costituisce il punto di massimo addensamento delle entità ovvero il punto verso cui queste confluirebbero se non ci fossero forze differenzianti (la variabilità del cluster) a tenerle distinte. Di solito, come mostrano le quattro formule presentate, il centroide è dato dal vettore delle medie calcolate su tutte (e solo) le unità incluse nel cluster, ma sono pos-

sibili altre definizioni (ad esempio il vettore delle mediane); talvolta, il polo del cluster non è localizzato in un particolare punto, magari artificiale come il vettore delle medie, ma attorno ad un certo numero di entità che si ritiene meglio caratterizzino quel cluster.

Nel corso della nostra trattazione assumeremo comunque che il polo del cluster coincida con il vettore delle medie delle entità del cluster. Qui però scatta la circolarità del problema della cluster analysis; infatti, come si vede dalle scelte di "∂", per definire la metrica è necessario conoscere la partizione P_i che, a sua volta, serve per determinare la metrica... Vedremo più avanti come i raggruppamenti iterativi superano i problemi della circolarità.

Approccio delle normali distinte.

In questo caso si assume che $\Omega=\{X_a, a=1,2, \ldots, n\}$ sia un campione di osservazioni m-dimensionali indipendenti provenienti da una delle "k" popolazioni normali $\{N(\mu_r,\Sigma_r), r=1,2, \ldots, k\}$. Inoltre, si assume che alle entità sia associato un vettore di classificazione formato da "n" parametri politomi: $\gamma_a=r$ se $X_a \varepsilon C_r$. Il logaritmo della funzione di verosimiglianza del campione "Ω" è:

$$L(\gamma;\mu_1,\ldots,\mu_k;\Sigma_1,\ldots,\Sigma_k)=-\frac{1}{2}\sum_{r=1}^{k}\left[\sum_{\gamma_a=r}(X_a-\mu_r)^t\Sigma_r^{-1}(X_a-\mu_r)\right]-\frac{1}{2}\left[\sum_{r=1}^{k}n_r\mathrm{Log}(|\Sigma_r|)\right].$$

Nell'approccio delle normali distinte il problema del clustering consiste nella determinazione del vettore di classificazione γ e quindi nella stima delle

$$\{\hat{\mu}_r, r=1, \ldots, k\} \text{ e } \{\hat{\Sigma}_r, r=1, \ldots, k\}$$

Per un dato valore di g, le stime di massima verosimiglianza si ottengono secondo le formule standard:

$$\hat{\mu}_r(\gamma)=\frac{1}{n_r}\sum_{\hat{\gamma}_a=r}X_a; \qquad \hat{\Sigma}_r(\gamma)=\frac{1}{n_r}\sum_{\hat{\gamma}_a=r}[X_a-\hat{\mu}_r(\gamma)][X_a-\hat{\mu}_r(\gamma)]^t \qquad r=1,2, \ldots, k$$

Supponiamo per il momento che solo le medie siano incognite e sostituiamo la loro stima nella funzione L(.) che diventa:

$$L(\gamma;\Sigma_1,\ldots,\Sigma_k)=-\frac{1}{2}\sum_{r=1}^{k}\left[\sum_{\gamma_a=r}[X_a-\hat{\mu}_r(\gamma)]^t\Sigma_r^{-1}[X_a-\hat{\mu}_r(\gamma)]\right]-\frac{1}{2}\left[\sum_{r=1}^{k}n_r\mathrm{Log}(|\Sigma_r|)\right]$$
$$=-\frac{1}{2}\sum_{r=1}^{k}\left[\mathrm{Tr}\left(\hat{\Sigma}_r(\gamma)\Sigma_r^{-1}\right)\right]-\frac{1}{2}\left[\sum_{r=1}^{k}n_r\mathrm{Log}(|\Sigma_r|)\right]$$

Ciò che caratterizza questo approccio è l'ipotesi fatta sulle matrici di varianze-covarianze nei gruppi: $\{\Sigma_r\}$. Infatti, se si assume che gli indicatori siano incorrelati all'interno di ogni gruppo e con eguale varianza unitaria: $\Sigma_r=I$, la funzione di verosimiglianza diviene:

$$L(\gamma)=-\frac{1}{2}\sum_{r=1}^{k}\left\{\mathrm{Tr}\left[\hat{\Sigma}_r(\gamma)\right]\right\}=-\frac{1}{2}\mathrm{Tr}\left[\hat{\Sigma}(\gamma)\right]$$

cioè la traccia della matrice di devianze-codevianze totale. Quindi, la stima di massima verosimiglianza di γ è la partizione che minimizza la traccia di $\hat{\Sigma}$ (massimizzazione di L=minimizzazione di -L). Questo è analogo a determinare la partizione ottima in base alla distanza "1" dell'approccio metrico.

Se invece si assume che le matrici di varianze-covarianze nei gruppi siano diverse dalla matrice identità, ma tutte uguali: $\Sigma_1=\Sigma_2=\ldots=\Sigma_k=\Sigma$, la funzione di verosimiglianza diventa:

$$L(\gamma;\Sigma)=-\frac{1}{2}\sum_{r=1}^{k}\left\{\mathrm{Tr}\left[\hat{\Sigma}_r(\gamma)\Sigma^{-1}\right]\right\}-\frac{n}{2}\mathrm{Ln}(|\Sigma|).$$

Poiché Σ deve essere pure stimata, la L(.) diviene strettamente connessa al determinante della matrice di devianze-codevianze totale Σ ed il vettore di classificazione γ si determina in base alla regola

$$\max\left\{L(\gamma)\cong-\frac{n}{2}\mathrm{Ln}\left[\left|\hat{\Sigma}(\gamma)\right|\right]\right\}\Rightarrow\min\left\{\left|\hat{\Sigma}(\gamma)\right|\right\}$$

che corrisponde alla scelta della distanza "2" dell'approccio metrico.

Se si ritiene che le matrici Σ_r siano diverse e che in ogni gruppo ci siano almeno (m+1) entità, la funzione di verosimiglianza per il calcolo di γ si dovrà basare sul criterio:

$$\max\left\{L(\gamma) \cong -\frac{1}{2}\left[\sum_{r=1}^{k} n_r \mathrm{Log}\left|\hat{\Sigma}_r(\gamma)\right|\right]\right\}$$

la cui ottimizzazione porta agli stessi risultati dell'approccio metrico usando la distanza "3".

Anche nell'approccio delle normali distinte c'è circolarità. Infatti, la stima di γ non può essere disgiunta dalla stima di $\{\mu, \Sigma\}$ e le due classi di parametri vanno determinate alternativamente ed in modo iterativo: una prima approssimazione ad esempio di γ da cui ottenere la prima stima di medie e varianze-covarianze che a loro volta determinano una nuova stima $\hat{\gamma}$ con un più alto valore della funzione di verosimiglianza e così via. Naturalmente si può partire da una prima approssimazione di $\{\mu, \Sigma\}$ ed in base a queste determinare una stima iniziale di γ usata poi per correggere le stime dei parametri e così procedendo finché non si determini un massimo per la L(.)

Approccio delle normal mixtures.

L'ipotesi principale è che le entità: $\Omega=\{X_a, a=1,2, \ldots,n\}$ siano delle osservazioni indipendenti provenienti da un'unica distribuzione

$$f(X_a;\pi_1,\ldots,\pi_k;\mu_1,\ldots,\mu_k;\Sigma_1,\ldots,\Sigma_k)=\sum_{r=1}^{k}\pi_r N(\mu_r;\Sigma_r) \qquad \text{con } \pi_r \geq 0, \sum_{r=1}^{k}\pi_r = 1$$

dove π_r è un parametro esprimente la probabilità che una data X_a provenga dal cluster r-esimo C_r. Il cluster di appartenenza (incognito) della X_a è indicato con γ_a cioè $\gamma_a=r$ se $X_a \in C_r$. Il problema del clustsering, come nell'approccio precedente, è la determinazione del valore di γ che rende massima la funzione di verosimiglianza

$$L(X_a;\gamma;\pi_r;\mu_r;\Sigma_r)=\sum_{r=1}^{k} n_r \mathrm{Ln}(\pi_r)-\frac{1}{2}\sum_{r=1}^{k} n_r \mathrm{Ln}(|\Sigma_r|)-\frac{1}{2}\sum_{r=1}^{k}\sum_{\gamma_a=r}(X_a-\mu_r)^t\Sigma_r^{-1}(X_a-\mu_r)$$

Per un dato γ, le stime di massima verosimiglianza delle $\{\pi\}$, $\{\mu\}$ e $\{\Sigma\}$ si ottengono dalle formule:

$$\hat{\pi}_r(\gamma)=\frac{n_r}{n}; \quad \hat{\mu}_r(\gamma)=\frac{1}{n_r}\sum_{\hat{\gamma}_a=r} X_a; \quad \hat{\Sigma}_r(\gamma)=\frac{1}{n_r}\sum_{\hat{\gamma}_a=r}[X_a-\hat{\mu}_r(\gamma)][X_a-\hat{\mu}_r(\gamma)]^t \qquad r=1,2,\ldots,k$$

Supponendo γ noto e calcolando le stime delle medie possiamo studiare la L(.) secondo le tre usuali ipotesi sulle matrici di varianze-covarianze dei gruppi.

1. $\Sigma_r = I \quad r=1,\ldots,k \Rightarrow L(\gamma) \cong \sum_{r=1}^{k} n_r \mathrm{Ln}\left(\frac{n_r}{n}\right)-\frac{1}{2}\mathrm{Tr}\left[\hat{\Sigma}(\gamma)\right]$;
2. $\Sigma_r = \Sigma \quad r=1,\ldots,k \Rightarrow L(\gamma) \cong \sum_{r=1}^{k} n_r \mathrm{Ln}\left(\frac{n_r}{n}\right)-\frac{n}{2}\mathrm{Ln}\left|\hat{\Sigma}(\gamma)\right|$;
3. $\Sigma_r \neq \Sigma_s \quad r \neq s \quad \Rightarrow L(\gamma) \cong \sum_{r=1}^{k} n_r \mathrm{Ln}\left(\frac{n_r}{n}\right)-\frac{1}{2}\mathrm{Ln}\left[n_r\left|\hat{\Sigma}_r(\gamma)\right|\right]$;

Ciò che è cambiato rispetto alle normali distinte è la presenza del termine:

$$\sum_{r=1}^{k} n_r \mathrm{Ln}(n_r)$$

che permette di tenere conto della diversa ampiezza dei clusters (in termini di numero di entità incluse) nella determinazione del vettore di classificazione γ. Si può peraltro notare che l'ipotesi di varianze-covarianze diverse nei gruppi dell'approccio delle *normal mixtures* coincide con l'approccio metrico usando la distanza "4". La stima di γ come nell'approccio precedente, dovrà avvenire in modo iterativo alternandosi con le stime delle numerosità, delle medie e delle varianze-covarianze dei clusters.

L'approccio metrico, con la sua impostazione esplorativa, è applicabile in qualsiasi sistuazione abbia senso misure le distanze con la metrica euclidea. Gli altri due approcci hanno radici nell'ipotesi che il dataset Ω sia un campione casuale di entità,

scelto tra la popolazione di quelle potenzialmente trattabili. Non sempre tale impostazione è coerente con le applicazioni: in molte ricerche territoriali le entità esaminate sono l'intera popolazione (ad esempio i comuni di una regione) e sarebbe inopportuno proporre una interpretazione dei risultati del clustering che dipendano dalla distribuzione normale multivariata. Nel prosieguo, data la natura della ricerca cui questo studio è diretto, seguiremo sempre l'approccio metrico ed anche se faremo riferimento alla sua equivalenza con un criterio di ottimizzazione, sarà evitata ogni considerazione inferenziale.

I raggruppamenti iterativi si avviano da una partizione arbitraria, più o meno plausibile, in base alla quale determinare le caratteristiche dei gruppi che portino ad una partizione più corretta e che viene poi via via perfezionata in base a schemi prestabiliti. Le iterazioni terminano quando nessun miglioramento può più essere ottenuto spostando delle entità.

Gli elementi essenziali della realizzazione pratica di tutti gli approcci possono essere ricondotti ai punti seguenti:

a) Fissare il numero di clusters.
b) Scegliere la funzione di distanza (ovvero il criterio da ottimizzare).
c) Determinare una partizione iniziale.
d) Scegliere lo schema di riallocazione delle entità.
e) Valutare la partizione finale.

La scelta di "k" è in realtà un problema di post-ottimizzazione e cioè viene affrontato dopo che una data procedura è stata applicata con diversi valori del numero di clusters. Il raggruppamento iterativo presuppone che "k" sia già fissato, o meglio, lo considera un parametro della procedura, che può quindi cambiare da prova a prova, ma che è fisso in ogni data applicazione.

Scelta del criterio.

Il raggruppamento iterativo è imperniato sulla definizione di un criterio in base al quale orientare il calcolo di un vettore di classificazione; ciò equivale alla scelta di una metrica che per-

mette di decidere se una entità appartiene ad un cluster piuttosto che ad un altro e che risolve quindi operativamente il problema della definizione di cluster.

Nei vari approcci al raggruppamento iterativo sono stati proposti diversi criteri che possiamo ricondurre a tre principi essenziali:

1) ricerca di gruppi determinati dalla variabilità entro e/o tra i gruppi, ma non da legami tra gli indicatori (metodi di Edwards e Cavalli-Sforza);
2) ricerca di gruppi determinati anche dall'interazione tra gli indicatori, ma sempre a livello aggregato cioè per tutti i gruppi contemporaneamente (metodi di Friedman e Rubin);
3) ricerca di gruppi determinati dalla variabilità e dalle relazioni tra gli indicatori, ma a livello di singoli gruppi (metodi di Scott e Symon).

Metodi di Edwards e Cavalli-Sforza.

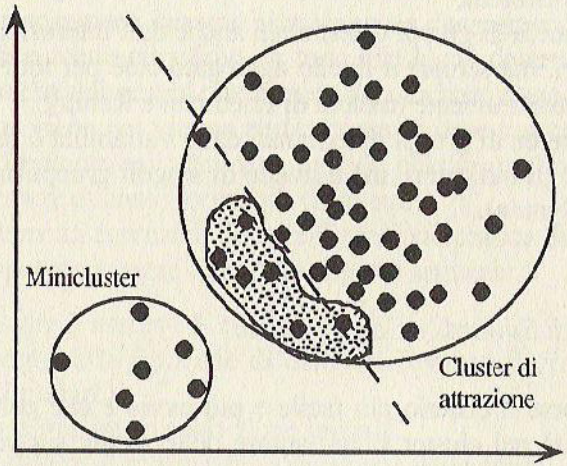
E' forse il criterio più facile e più ovvio e che colloca una data entità nel cluster il cui vettore delle medie sia più vicino alla entità stessa secondo la metrica euclidea

$$X_a \in C_j \Rightarrow (X_a - \mu_j)^t (X_a - \mu_j) = \min_{1 \le r \le k} (X_a - \mu_r)^t (X_a - \mu_r)$$

che, come si è visto, coincide con la minimizzazione della traccia di $\hat{\Sigma}$. Infatti, il presupposto di questo criterio è che le variabili risultino incorrelate in ogni cluster e che le varianze siano tutte uguali. Si ipotizza in pratica che la matrice di varianze-covarianze sia, in ogni cluster, uguale alla matrice identità: $\Sigma_i = I$ per $i=1,2,\ldots,k$ ipotizzando perciò clusters che formano delle ipersfere, di vario raggio, nello spazio m-dimensionale.

E' una ipotesi molto forte, raramente riscontrabile in pratica, ma che, sorprendentemente, non impedisce al criterio di funzionare bene anche in situazioni in cui gli indicatori presentino

significative relazioni lineari e/o siano eteroschedastici. La minimizzazione di $Tr(\hat{\Sigma})$ ha due difetti gravi: risente delle trasformazioni di scala (o, almeno, di quelle non ortogonali) e, concentrandosi esclusivamente sulla devianza, trascura le interrelazioni tra indicatori che potrebbero essere rilevanti per la descrizione di qualcuno o di tutti i clusters. Hand (1981) osserva che la presenza di valori anomali o di minigruppi disposti nelle vicinanze di un gruppo più grande può provocare degli inconvenienti del tipo di quelli rappresentati nella figura che segue.



Il mini gruppo rischia di essere assorbito da quello più grande oppure alcune entità del cluster maggiore (quelle della nube in grigio) sono trascinate in quello più piccolo; questo può succedere anche se i centroidi dei due clusters sono molto lontani.

Nonostante questi problemi, i metodi di Edwards e Cavalli-Sforza sono quelli più diffusi perché semplici e facilmente gestibili dal punto di vista computazionale. La versione più nota è sicuramente il metodo detto delle "k-medie" che ha una procedura di trasferimenti molto semplice ed efficace. Inoltre, il programma di calcolo realizzato da Hartigan e Wong (1979) lo rende competitivo con schemi più complessi e apparentemente più flessibili. Rimane però la sensibilità a variazioni proporzionali degli indicatori, cioè alla modifica della partizione ottimale

qualora uno o più indicatori siano moltiplicati per una costante. Ciò è di solito aggirato con la standardizzazione delle variabili anche se ci sono altri metodi per arrivare all'invarianza rispetto ai cambiamenti moltiplicativi (si veda il paragrafo 3.1.3).

Metodi di Friedman-Rubin.

Sotto questa etichetta si raccolgono diversi criteri, anche suggeriti da altri autori, ma con un comune denominatore: l'ipotesi che i dati analizzati siano stati ottenuti estraendo campioni di ampiezza indeterminata da "k" popolazioni normali con media diversa, ma comune matrice di varianze-covarianze, non necessariamente diagonale.

1) Minimizzazione di $|\hat{\Sigma}|$

Rispetto ai metodi Edwards-Cavalli Sforza, questo criterio tiene conto dell'esistenza di eventuali dipendenze lineari tra gli indicatori, ma assume che siano le stesse in tutti i gruppi. Ciò si riflette nella individuazione di clusters di numerosità simile e della stessa forma iperellissoidale che non sempre sono presenti nei dati. Vale poi la pena di ricordare che il criterio non può essere impiegato nei casi in cui m>n-k cioè quando il numero di indicatori è superiore al numero di entità ridotto del numero di clusters.

Il criterio $\min\{|\hat{\Sigma}|\}$ corrisponde, tenuto conto della invarianza di T nelle varie partizioni, alla massimizzazione di

$$\frac{|T|}{|\hat{\Sigma}|} = \left|I + \hat{\Sigma}^{-1}B\right| = \prod_{i=1}^{m}\left(1+\lambda_{(i)}\right) \quad \text{con} \quad \lambda_{(i)} = \text{autovalore ordinato di } \hat{\Sigma}^{-1}B$$

Se tra gli indicatori prescelti per il clustering sussistono forti relazioni lineari, ovvero qualcuno degli autovalori di $\hat{\Sigma}$ è prossimo allo zero, tale risulterà anche $|\hat{\Sigma}|$ ed il clustering basato su questo perderà molto della sua plausibilità. Il criterio infatti è so-

prattutto guidato dagli autovalori minori. Un'altra caratteristica negativa è che, se uno degli indicatori risulta nettamente partizionato, se cioè il suo istogramma presenta tante sottomode, la configurazione finale sarà dominata da tale variabile e saranno ignorate le altre. I metodi Friedman-Rubin, in fondo, cercano un raggruppamento ottimale delle entità e non un raggruppamento ottimale forzatamente basato su tutti gli indicatori. La minimizzazione di $|\hat{\Sigma}|$ ha però il vantaggio che la partizione ottimale da essa individuata non cambia se uno o più indicatori subiscono delle trasformazioni lineari, additive o moltiplicative che siano.

Diversi autori hanno osservato che questo criterio tende a favorire la formazione di clusters di eguale ampiezza. Un miglioramento è costituito dalla modifica suggerita da Symons (1981) che propone di minimizzare:

$$n\mathrm{Log}\left(\left|\hat{\Sigma}\right|\right) - 2\sum_{r=1}^{k} n_r \mathrm{Log}(n_r)$$

anche se rimane una leggera enfasi sui clusters più numerosi. Le prove da noi condotte non confermano alcun miglioramento per questa variante di min $|\hat{\Sigma}|$ che ci risulta comportarsi meglio sia in situazioni di clusters a numerosità uniforme che nei casi di coesistenza di clusters piccoli e clusters grandi. E' poi abbastanza riconosciuto che la qualità dei risultati ottenuti con i metodi Friedman-Rubin non è molto robusta rispetto alla eterogeneità delle matrici di varianze-covarianze nei gruppi. Anzi, se tali matrici sono molto diverse i raggruppamenti iterativi basati sull'ipotesi $\Sigma_1=\Sigma_2=\ldots=\Sigma_k$ porta a clusters del tutto artificiosi ed estranei alla realtà dei dati.

2) *Massimizzazione di* $\mathrm{Tr}(\hat{\Sigma}^{-1}B)$

Questo criterio tenta di riprendere la semplicità procedurale legata all'uso della traccia e nello stesso tempo di tener conto di eventuali relazioni lineari tra gli indicatori. In realtà, le elaborazioni sono ancora più difficili da gestire del criterio precedente e

non è poi dimostrato che i risultati siano significativamente diversi dal minimizzare la $\mathrm{Tr}(\hat{\Sigma})$. Il $\max\{\mathrm{Tr}(\hat{\Sigma}^{-1}B)\}$, ha il vantaggio di non risentire di trasformazioni lineari degli indicatori. E' però molto influenzabile attraverso gli autovalori più grandi di $\hat{\Sigma}^{-1}B$. Infatti, si ha

$$\mathrm{Tr}\left(\hat{\Sigma}^{-1}B\right) = \sum_{i=1}^{m} \lambda_{(i)} \quad \text{con} \quad \lambda_{(i)} = \text{autovalore ordinato di } \hat{\Sigma}^{-1}B$$

e se per caso i clusters della partizione iniziale sono posti lungo l'asse di massima variabilità e questa direzione è "sbagliata" rispetto alla partizione ottima, il metodo non potrà apportare correzioni di rotta, anzi l'errore verrà amplificato. D'altra parte, Spath (1985) è piuttosto perplesso sull'uso di questo criterio poiché non sono dimostrate certe proprietà ottimali. Le perplessità sono molto condivise in letteratura e tale metodo è stato del tutto abbandonato.

Metodi di Scott e Symons.

I metodi già descritti non riescono a governare la classificazione in gruppi quando le entità, oltre a differenziarsi per i livelli medi, sono anche molto diverse dal punto di vista delle interrelazioni tra indicatori e cioè nella situazione più realistica. Questo limite si supera se in qualche modo si coinvolge nel criterio di ottimizzazione la matrice di devianze-codevianze dei singoli clusters. In questo senso sono stati proposti diversi criteri, ma che fanno tutti riferimento ad una media generalizzata dei determinanti delle singole $\hat{\Sigma}_r$:

$$\left\{ \sum_{r=1}^{k} \frac{n_r}{n} \left|\hat{\Sigma}_r\right|^{1/p} \right\}^{p}$$

In particolare, per "p" tendente a zero, il metodo di Scott e Symons porta a minimizzare la media geometrica dei determinanti delle matrici di varianze-covarianze dei gruppi

$$\text{Min}\left\{\text{Ln}(\text{Mg})\right\} = \min\left\{\sum_{r=1}^{k} \frac{n_r}{n} \text{Ln}\left|\hat{\Sigma}_r\right|\right\}$$

che appare l'unico ad essere stato citato in letteratura insieme al criterio della media aritmetica ottenibile per p=1. Maronna e Jakovkis (1974) evidenziano che i raggruppamenti basati su questi criteri si fermano dopo pochi passi senza che necessariamente si determini un massimo locale. Segnalano inoltre effetti di "cannibalizzazione" da parte di gruppi che si dilatano a spese dei clusters loro vicini. Inoltre, tendono a formare comunque clusters con matrice di varianze-covarianze diversa, anche se non è così nei dati (ciò è particolarmente accentuato quando i centroidi dei clusters sono vicini). Symons (1981) propende per una leggera superiorità alla formula derivata dall'approccio delle normali distinte:

$$\sum_{r=1}^{k} n_r \left[\text{Ln}\left(\frac{\left|\hat{\Sigma}_r\right|}{n_r^2}\right)\right].$$

I metodi Scott-Symons richiedono però la stabilità delle varianze-covarianze nel corso delle varie allocazioni e riassegnazioni. In ogni passo, quindi, tutti i clusters debbono contenere un numero sufficiente di entità tale da consentire una stima accettabile dei parametri del clusters. In aggiunta, le entità debbono essere tali da garantire che $|\hat{\Sigma}|>0$ e questo anche se nella struttura "vera" dei dati i clusters hanno matrici di dispersione singolare. Infine, la partizione iniziale deve essere prossima a quella finale dato che per la applicazione di questi metodi sono necessari calcoli più laboriosi.

Il numero di partizioni.

In realtà, il modo più sicuro di suddividere "n" unità in "k" clusters è di esaminare tutte le possibili partizioni. Questo però non è praticabile se non per poche entità che potrebbero facilmente ed efficacemente essere classificate con delle semplicissi-

me tecniche grafiche. Ad esempio, nel caso dell'approccio delle normali distinte o delle *normal mixtures* occorre valutare la funzione di verosimiglianza su un numero di partizioni pari a k^n che, per 20 entità da ripartire in 3 cluster, implica 3.5 miliardi di partizioni. Certo, con il vincolo che ogni cluster contenga almeno un elemento il loro numero si riduce diventando:

$$P_{k,n} = \frac{1}{k!}\sum_{j=1}^{k}\binom{k}{j}(-1)^{k-j}(j)^n$$

che però raggiunge sempre cifre elevate anche per valori modesti di "k" ed "n": $P_{20,3}$ = (580 082 158) ed un computer che riuscisse a valutare 1,000 partizioni al secondo (quindi, molto veloce) impiegherebbe quasi una settimana per trovare l'ottima partizione. Se poi "k" non è noto, ma deve essere determinato dai dati, il numero di partizioni da esaminare diventa proibitivo:

$$\sum_{k=1}^{\text{Max } k} P_{k,n}.$$

Ne consegue che la ricerca della partizione ottima, ammesso che ce ne sia una sola, non può che avvenire esaminando un numero limitato tra quelle possibili. E' questo uno dei limiti del metodo di raggruppamento iterativo: tranne che per le situazioni in cui sono coinvolte pochissime entità non si potrà essere sicuri che la soluzione trovata sia veramente quella ottima e che quella ottenuta non sia una delle tante soluzioni localmente valide, ma che risultano molto diverse da quella globalmente ottima.

La partizione iniziale.

Il raggruppamento iterativo non richiede la memorizzazione della matrice delle similarità o delle distanze e, potenzialmente, potrebbe trattare una cospicua mole di dati. Si è però visto che l'individuazione della partizione ottima richiederebbe calcoli innumerevoli; diventa allora essenziale avviare la procedura iterativa da un punto iniziale "buono" (nel senso

che esso, presumibilmente, non è troppo distante da quello ottimale).

Esistono diversi modi di definire la configurazione iniziale. Per comodità i più noti sono stati classificati in due tipologie, a seconda che si tratti di individuare una partizione iniziale o i poli dei clusters.

Scelta dei poli iniziali.

In questo caso si parte da poli predeterminati intorno ai quali si aggregano le entità in base alla loro prossimità a quei "semi". Ad esempio, se i poli iniziali sono i vettori s_i i=1, 2, ..., k lo schema di assegnazione che definisce la prima partizione si ottiene assegnando la X_j al cluster "h" se risulta:

$$d(X_j, s_h) \le d(X_j, s_i) \quad i = 1,2,\ldots,k$$

dove "d" è la misura di distanza o di dissimilarità scelta come meccanismo di costruzione dei clusters oppure derivata dal criterio da ottimizzare. Lo schema dovrà essere tale che si formino esattamente "k" clusters e che in ognuno ricada almeno una entità (in caso contrario uno o più semi debbono essere sostituiti). In che modo scegliere gli s_i? Queste alcune delle alternative:

1) I poli sono noti (in linea di massima) ovvero si ritiene che le unità si raggruppino intorno a centri che presentano certi valori degli indicatori.
2) I poli sono noti e coincidono con particolari entità tra le "n" osservate.
3) Per ognuno dei "k" gruppi è nota l'appartenenza di alcune entità ovvero i clusters si formeranno aggregandosi non intorno ad un singolo centro, ma ad una nube più o meno rarefatta di punti: Z_j, contenente q_j entità. I poli veri e propri sono dati dai baricentri degli insiemi Z_j.
4) I poli vengono generati scegliendo a caso "k" unità distinte in modo che ognuna di essa abbia la stessa probabilità di essere un

polo. Questa opzione è equivalente alla scelta delle prime (o delle ultime) "k" entità se il loro ordinamento non è rilevante.
5) La scelta "casuale" è ripetuta un certo numero di volte e sulle partizioni risultanti si valuta il criterio di ottimizzazione. Se il numero di ripetizioni è abbastanza elevato, diciamo dell'ordine di 4*(n/k) questo metodo fornisce poli abbastanza ragionevoli. Semplici ragionamenti probabilisti danno conforto a questa scelta di poli iniziali anche se la distribuzione degli "ottimi" nell'insieme delle partizioni può essere talmente degenere che anche la migliore scelta casuale risulta comunque scadente.
6) Per ciascuno degli "m" indicatori si determina il *range* $R_j = \max\{X_j\} - \min\{X_j\}$ j=1,2, ..., m scegliendo poi i poli iniziali secondo lo schema

$$\begin{cases} C_{1,j} = \min(x_j) + \dfrac{R_j}{2k} & i = 1 \\ C_{i,j} = C_{i-1,j} + \dfrac{R_j}{k} & i = 2,\ldots,k \end{cases} \qquad j = 1,2,\ldots,m$$

Questa scelta ha buone possibilità di funzionare se tutti gli indicatori hanno correlazioni positive (o negative, modificando opportunamente le formule) ed i gruppi si formano per valori medi tutti crescenti (o tutti decrescenti).
7) Si determina innanzitutto il vettore delle medie globali:

$$\mu = (\mu_1, \mu_2, \ldots, \mu_m); \quad \text{con} \quad \mu_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}; \qquad j = 1,2,\ldots,m$$

Il primo polo è l'entità che ha distanza massima da μ. il secondo è quella entità che ha distanza massima dal primo polo:

$$s_2 = X_h \quad \text{con} \quad d(X_h, s_2) = \max_{1 \le j \le n} \{ d(X_j, s_1) \}$$

Il polo successivo sarà quell'entità la cui distanza dal più vicino dei poli già determinati è maggiore

$s_p = X_h$ con X_h determinato da $\max_{1\le j\le n}\left\{\min_{1\le i\le p-1}\left[d(X_j, s_i)\right]\right\}$ per $p = 3,\ldots,k$

e così via fino al k-esimo "seme". Tale strategia dovrebbe essere in grado di cogliere sia un eventuale addensamento sul centroide globale che clusters nascenti intorno a valori remoti.

8) Qualora fosse praticabile (per dimensione della matrice dei dati e per i tempi di elaborazione) il raggruppamento iterativo potrebbe essere preceduto da un raggruppamento gerarchico dal quale trarre un'idea di quali siano i gruppi e le loro entità rappresentative.

Scelte della partizione iniziale.

Invece dei poli si può prefissare una partizione delle "n" entità in "k" gruppi da cui poi ricavare i poli come medie nei singoli clusters

$$\mu_i = \frac{1}{n_i}\sum_{j\in C_j} X_j$$

Come definire la partizione iniziale? Anche qui ci sono più possibilità:

1) E' nota una classificazione (magari ipotetica) delle entità nei clusters rispetto alla quale si vuole verificare lo scostamento della partizione finale.

2) Si assegnano sequenzialmente (n/k) unità ad ogni cluster: il primo blocco al primo cluster, il secondo blocco al secondo e così via; l'eventuale resto viene ripartito assegnando le unità residue una per cluster fino ad esaurimento.

3) I clusters vengono creati assegnando casualmente ciascuna delle "n" entità ai "k" clusters, generando perciò un numero casuale intero tra 1 e k per ognuna delle "n" entità. Questa assegnazione risulterebbe più efficace se si potesse stabilire una probabilità di appartenenza al cluster che sia vicina al peso del cluster nella partizione finale.

4) La configurazione iniziale è costruita assegnando un elemento ogni k allo stesso cluster; l'eventuale resto viene ripartito assegnando le unità residue una per cluster fino ad esaurimento.

Le scelte 2, 3, 4 formano clusters iniziali di numerosità uguale, o praticamente uguale, e sono perciò pessime scelte se questa parità di importanza tra clusters non è prevista per la partizione finale.

5) I clusters vengono creati con una procedura scissoria: ad esempio quella ricordata da Seber (1984, pp. 79-80). I primi due clusters si ottengono bipartendo le "n" entità secondo una linea di demarcazione che massimizzi la variabilità tra i due gruppi:

$$B_j = n_1(\mu_{j1} - \mu_j)^2 + n_2(\mu_{j2} - \mu_j)^2$$

dove n_i é il numero di entità nel cluster μ_{ji} è la media dell'indicatore j-esimo nel gruppo i-esimo e μ_j è la media di questo su tutte le entità. Il massimo di B_j si può determinare cercando l'indice "r_j" che massimizza

$$R_j = \max_{1\le r_j\le n} \frac{\left(\sum_{i=1}^{r_j} X_{(i)j}\right)^2}{r_j} + \frac{\left(\sum_{i=r_j+1}^{n} X_{(i)j}\right)^2}{n - r_j}$$

in cui $X_{(i)j}$ denota i valori ordinati in senso ascendente dell'indicatore j-esimo. Se "k" è maggiore di due, ciascuno dei due nuovi gruppi si candida ad essere diviso ulteriormente e, in questo senso, si sceglierà quello con l'R_j più grande e così procedendo fino ad arrivare ai "k" clusters. La scelta dell'indicatore "j" potrebbe cadere sulla prima componente principale che viene poi costantemente usata come riferimento in tutte le suddivisioni oppure, più correttamente, ad ogni suddivisione, si cercherà, tra gli "m" indicatori ed i "k-1" gruppi l'R_j più elevato.

6) La partizione iniziale deriva da un precedente raggruppamento gerarchico.

7) La partizione iniziale é formata a partire da una "nube" di entità di cui è nota l'appartenenza ad un dato cluster con una riassegnazione delle altre entità al cluster la cui "nube" è più vicina. Ad esempio:

$$X_h \in C_j \text{ se } j \text{ é } \min_{1\le i\le q_j} d(X_h, Z_{ij}) \le \min_{\substack{1\le i\le q_r \\ 1\le r\le k}} d(X_k, Z_{ir})$$

dove Z_{ir} è una delle entità che formano il polo del cluster r-esimo.

Nessuna delle procedure elencate (ed altre che per brevità non abbiamo riportato) è in grado di assicurare sempre il punto di partenza che porti poi certamente alla ottimizzazione globale del criterio di clustering. La strategia migliore è probabilmente quella di effettuare un alto numero di prove con vari metodi o combinazioni di metodi per decidere intorno a quale partizione si stabilizzi la soluzione finale ottima. In questo senso è determinante la scelta discussa nel prossimo paragrafo, cioè il criterio da ottimizzare.

Lo schema di riallocazione.

Tutte le procedure iterative sono centrate sullo schema di riallocazione delle entità tra i clusters. Il problema è molto semplice: in che modo e in che misura lo spostamento di una o più entità da un cluster ad un altro migliora il criterio scelto (implicitamente con l'approccio metrico ed esplicitamente con gli altri due approcci) come obiettivo da ottimizzare? La risposta dipende certo dal criterio, ma dipende anche dal modo in cui lo spostamento si realizza.

Occorre innanzitutto stabilire quante entità di un dato cluster coinvolgere in un trasferimento: di solito le entità sono considerate una alla volta, ma è una scelta di semplicità di calcolo che non è obbligata da considerazioni logiche o teoriche ed è perfettamente lecito impostare trasferimenti di un sottinsieme di entità da distribuire variamente negli altri clusters. Ammesso che il trasferimento riguardi una sola entità per volta si deve precisare il modo in cui si effettua il trasferimento.

Può ad esempio essere "combinatorio" per cui i poli dei clusters cedente e ricevente si aggiornano ad ogni spostamento, oppure può essere "non combinatorio" e cioè l'aggiornamento avviene solo dopo che siano state escusse tutte le entità. Nel primo caso si tiene subito conto della modifica intervenuta nei clusters e questi sono proposti alle altre entità già aggiornati nella loro composizione rendendo quindi più rapida la convergenza. Il passo combinatorio ha però delle controindicazioni: richiede molti più calcoli e, se si pensa ad applicazioni su larga scala, questo non è un aspetto trascurabile; in più, la soluzione finale risulta dipendente dall'ordine con cui le entità sono esaminate dato che la ridefinizione continua dei centroidi altera necessariamente le condizioni del confronto delle entità, particolarmente le ultime. Un passo non combinatorio fa in modo che il raggruppamento iterativo prescinda dalla sequenza con cui sono elaborati i dati. C'é però il rischio che i vari trasferimenti proposti, anche se singolarmente determinano un miglioramento, nell'insieme comportino un peggioramento del criterio finendo per far ripetere un infinito numero di volte la procedura iterativa.

Il calcolo della distanza dai poli può avvenire in modo "inclusivo" oppure "esclusivo" dell'entità considerata per la definizione dei centroidi dei clusters. Sembra corretto escluderla dal cluster di partenza ed includerla in quello di arrivo (passo completo) dato che il trasferimento dovrebbe basarsi sulla conoscenza completa di ciò che si riscontra *dopo* che esso è stato effettuato. Se così non fosse, si correrebbe il rischio che dopo un eventuale passo combinatorio, l'aggiornamento dei centroidi dei clusters contraddica il trasferimento e che l'entità si trovi più vicina al cluster da cui è stata rimossa. C'è però da tener conto che il modo inclusivo (e quindi il passo completo) implica elaborazioni più lunghe, anche se secondo schemi ben programmibili. Ad esempio, un trasferimento a passo completo in cui l'entità X_a dal cluster "u", cui attualmente appartiene, passa al cluster "e", ha l'effetto seguente sulla matrice di devianze-codevianze globali

$$\hat{\Sigma}' = \hat{\Sigma} - \alpha_u y_u y_u^t + \alpha_e y_e y_e^t \qquad \text{dove: } y_u = X_a - \mu_u; y_e = X_a - \mu_e; \quad \alpha_u = \frac{n_u}{n_{u-1}}; \alpha_e = \frac{n_e}{n_{e+1}}$$

Se il criterio è quello di Min{Tr($\hat{\Sigma}$)}, il trasferimento è effettuabile (cioè fa diminuire il criterio) se e solo se:

$$\alpha_e \|y_e\|^2 - \alpha_u \|y_u\|^2 < 0$$

Se invece il criterio è Min{$|\hat{\Sigma}|$}, il trasferimento è effettuabile se e solo se:

$$\left(1-\alpha_e y_e^t \hat{\Sigma}^{-1} y_e\right)\left(1+\alpha_u y_u^t W^{-1} y_u\right)+\alpha_e\alpha_u\left(y_e^t \hat{\Sigma}^{-1} y_u\right)^2 < 1$$

Nel caso si adotti la modifica proposta da Symons l'ammissibilità del trasferimento richiede che

$$\left(1-\alpha_e y_e^t \hat{\Sigma}^{-1} y_e\right)\left(1+\alpha_u y_u^t W^{-1} y_u\right)+\alpha_e\alpha_u\left(y_e^t \hat{\Sigma}^{-1} y_u\right)^2 < \left[\frac{(n_u-1)^{n_u-1}*(n_e+1)^{n_e+1}}{n_u^{n_u}*n_e^{n_e}}\right]^{\frac{2}{n}}$$

Per il criterio della minimizzazione della media geometrica delle matrici di devianze-codevianze nei gruppi, infine, il passo è ammissibile se

$$\mathrm{Ln}\left(\frac{|\hat{\Sigma}_u|}{|\hat{\Sigma}_e|}\right)-(n_u-1)*\mathrm{Ln}\left(1-\alpha_u y_u^t \hat{\Sigma}_u^{-1} y_u\right)-(n_e+1)*\mathrm{Ln}\left(1+\alpha_e y_e^t \hat{\Sigma}_e^{-1} y_e\right)>0$$

Lo spostamento della X_a dal cluster "u" potrebbe indurre migliorie nel criterio per diversi clusters di destinazione ed andrà spostata nel cluster in cui il miglioramento è maggiore. Da notare che se un cluster contiene una sola entità lo stesso cluster non potrà essere considerato come cluster di partenza per eventuali trasferimenti perché lascerebbe un cluster vuoto il che non è compatibile con un numero fisso di clusters. Se lo spostamento migliora il criterio allora l'entità X_a esce dal gruppo "u" ed entra nel gruppo "e" e, se il passo è di tipo combinatorio, occorrerà aggiornare le quantità coinvolte, segnatamente le matrici di devianze-codevianze dei gruppi "u" ed "e":

$$\hat{\Sigma}_u' = \hat{\Sigma}_u - \alpha_u y_u y_u^t \qquad \hat{\Sigma}_e' = \hat{\Sigma}_e + \alpha_e y_e y_e^t$$

Non tutti gli autori sono concordi sul fatto che i calcoli necessari per la ricerca del massimo effetto tra tutti i clusters dello spostamento della entità X_a meritino di essere effettuati. Alcuni sostengono che si debba considerare, come possibile destinazione, non tutti i clusters, ma uno o due e, segnatamente, quelli il cui centroide sia molto prossimo al centroide del cluster cui appartiene la X_a (questo non riduce affatto le elaborazioni come invece sembrerebbe ad una considerazione superficiale). Altri addirittura suggeriscono di spostare l'entità al primo cluster per cui si produca un miglioramento del criterio.

La valutazione dell'effetto dello spostamento deve essere ripetuta per tutte le "n" entità e solo dopo che tutte siano state considerate senza che si siano avuti spostamenti (a questo in effetti si arriva in pochi passi, raramente più di una dozzina) si è raggiunto un punto di arresto del raggruppamento iterativo. Che la partizione ottenuta sia poi un punto di ottimo globale oltreché locale non è garantito: anche questo è uno dei problemi irrisolti della cluster analysis.

Lo scambio.

Un'altra scelta da fare relativamente alla procedura di trasferimento delle entità è se si debba considerare solo il passaggio di una unità da un cluster ad un altro od anche lo scambio di unità tra due cluster (*swapping*). L'effetto dello scambio della unità X_a dal cluster "u" al cluster "e" e della unità X_b dal cluster "e" al cluster "u" è:

$$\hat{\Sigma}' = \hat{\Sigma} - (X_a-\mu_u)(X_a-\mu_u)^t + (X_b-\mu_u)(X_b-\mu_u)^t - (X_b-\mu_e)(X_b-\mu_e)^t + (X_a-\mu_e)(X_a-\mu_e)^t$$

che può essere utilizzato per valutare la plausibilità dello scambio in base ai criteri visti nel sottoparagrafo precedente.

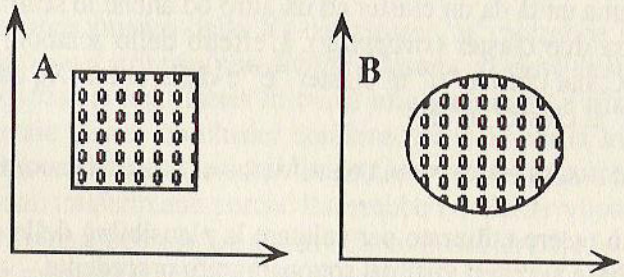
A differenza del trasferimento, lo scambio può avere luogo anche se uno dei clusters coinvolti include una sola entità: ovviamente solo uno dei clusters coinvolti può averne una sola, altrimenti l'effetto sarebbe nullo. Una procedura completa di riassegnazione dovrebbe prevedere sia le possibilità di tra-

sferimento che di scambio, magari da effettuare in fasi diverse: ad esempio, prima i trasferimenti e poi gli scambi, fino a che nessuno dei due meccanismi induca miglioramenti nel criterio adottato. Nulla vieterebbe di passare alla considerazione di trasferimenti e scambi relativi a blocchi di unità tra clusters, ma i tempi di esecuzione, anche se basati sul più semplice criterio della traccia, possono diventare impraticabili.

La partizione finale.

La cluster analysis iterativa è caratterizzata dall'incertezza tra località e globalità della soluzione ottima che si determina alla fine dei cicli di trasferimento e scambio. In effetti, a seconda della partizione iniziale cambia anche quella finale: le strategie di avvio, come si è detto in precedenza, rispondono ad istanze molto diverse e ciò non può che riflettersi nel punto di arresto della procedura

Uno dei problemi irrisolti della cluster analysis (Everitt, 1979; Mineo, 1986) è la mancanza di un termine di confronto per la soluzione trovata. Il fatto è che , se si attiva un algoritmo di cluster su dei dati del tutto privi di struttura quali quelli presentati nei grafici "A" e "B"



non è escluso che si arrivi ad una divisione in gruppi localmente "ottima" anche se nei dati non c'è alcuna struttura naturale in più gruppi. Se però i "k" gruppi ci sono resta da chiedersi come valutare la "qualità" della soluzione.

Confronto con una partizione standard.

Supponiamo che per le "n" entità esista una partizione "naturale" oppure ritenuta ottima per delle considerazioni teoriche od ancora ottenuta da un altro metodo e rispetto alla quale si vogliono misurare le eventuali discordanze:

$$P^* = \left\{C_1^*, C_2^*, \ldots, C_k^*\right\}; \quad \text{con} \quad C_i^* = \left\{X_{i1}^*, X_{i2}^*, \ldots, X_{in_i^*}^*\right\}$$

Supponiamo che la procedura di clustering abbia dato luogo alla partizione

$$P^\circ = \left\{C_1^\circ, C_2^\circ, \ldots, C_k^\circ\right\}; \quad \text{con} \quad C_i^\circ = \left\{X_{i1}^\circ, X_{i2}^\circ, \ldots, X_{in_i^\circ}^\circ\right\}$$

ed indichiamo con f_{ij} il numero di entità collocate simultaneamente nel cluster i-esimo della partizione standard e nel cluster j-esimo della partizione finale.

Per valutare la similarità tra le due partizioni, Rand (1971) ha proposto la seguente misura:

$$s\left(P^*, P^\circ\right) = \left\{1 - \left[\frac{n^*(n-1)}{4}\left(\sum_{i=1}^{k}\left(\sum_{j=1}^{k} f_{ij}\right)^2 + \sum_{j=1}^{k}\left(\sum_{i=1}^{k} f_{ij}\right)^2 - \sum_{i=1}^{k}\sum_{j=1}^{k} f_{ij}^2\right)\right]\right\}$$

La "s" è zero quando non c'è alcuna concordanza tra le due partizioni ed è uno quando esse coincidono elemento per elemento. Nelle altre situazioni assume valori intermedi tra zero ed uno e crescenti all'aumentare della similarità tra le due partizioni.

L'indice di Rand potrebbe essere usato per sottoporre a verifica l'ipotesi che nei dati non esista una struttura in gruppi contro l'alternativa che esistano "k" gruppi. In tal senso, seguendo il suggerimento di Aldenderfer e Blashfield (1984), si generano "n" entità pseudo-casuali da una distribuzione "m" variata che coincida con i dati reali almeno nel vettore delle medie globali e nella matrice globale delle varianze-covarianze ed applicare a questi dati simulati la stessa procedura di cluste-

ring applicata ai dati reali. Le due partizioni finali possono poi essere confrontate con l'indice di Rand: valori di $s(P^*,P^\circ)$ vicini all'unità, ad esempio superiori a 0.75, indicheranno partizioni in "k" gruppi non significativamente diverse da una suddivisione artificiosa in "k" sottoinsiemi di un blocco di dati unico. Per sicurezza, la prova dovrà essere ripetuta alcune volte usando pseudo-campioni diversi perché nulla vieta che il campione simulato da una distribuzione compatta non si disponga spontaneamente proprio in "k" clusters ben separati.

Il problema di comprendere almeno se in $X=\{X_a, a=1,2, \dots ,n\}$ c'è più di un cluster può anche porsi come una verifica di ipotesi su γ:

$$\begin{cases} H_0: \gamma_1 = \gamma_2 = \dots = \gamma_n \\ H_1: \gamma_r \neq \gamma_s \quad \text{per almeno un } r \neq s \end{cases}$$

che può essere basato sulla quantità:

$$\lambda = n\mathrm{Log}\left(\frac{|T|}{|\hat{\Sigma}(\hat{\gamma})|}\right)$$

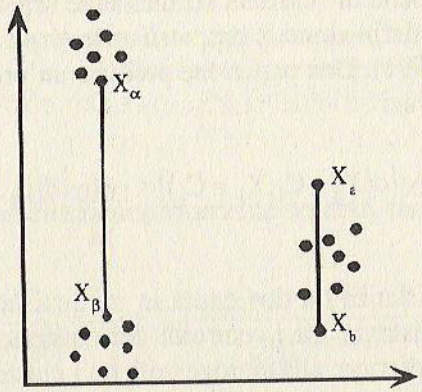
dove $\hat{\gamma}$ è il vettore di classificazione finale della procedura iterativa adottata. L'indice ricorda molto il test del rapporto di verosimiglianza anche se le usuali ipotesi che portano a questo tipo di statistica non sono applicabili nel raggruppamento iterativo (per il fatto che non si tratta di una partizione casuale, ma diretta alla ottimizzazione di un criterio) e ci si deve basare su valutazioni intuitive del suo comportamento. Riprenderemo questa discussione a proposito del problema della scelta del numero di clusters.

La qualità della partizione .

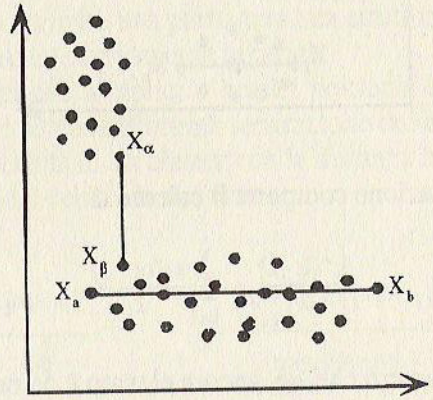
Una partizione è detta "ben strutturata" se:

$$\max_{1\le t\le k}\left\{d(X_a \in C_t, X_b \in C_t)\right\} \le \min_{\substack{1\le r,s\le k \\ r\neq s}}\left\{d(X_\alpha \in C_r, X_\beta \in C_s)\right\}$$

ciò implica che una entità appartenente un qualsiasi cluster dista dalle altre entità nello stesso cluster meno di quanto non disti da qualsiasi altra entità in un cluster diverso.



La definizione prescinde dal numero di clusters "k". Infatti, risulta ben strutturata anche la partizione con k=n, cioè quella in cui ogni entità forma un cluster separato.
Come mostra chiaramente il grafico qui sotto



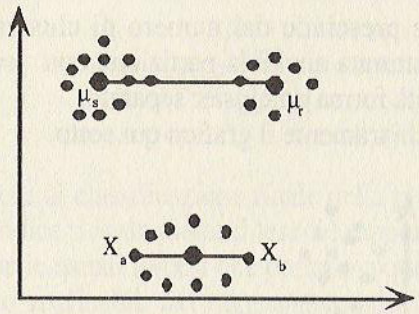
la definizione di partizione ben strutturata può risultare eccessivamente rigorosa ed esistono suddivisioni perfettamente accettabili che non rispecchiano la definizione di clusters che implica

la definizione. D'altra parte, la partizione ben strutturata è legata alle n*(n-1)/2 distanze entità-entità e quando "n" è elevato la sua verifica comporta calcoli piuttosto laboriosi.

Una diversa angolatura da cui guardare alla qualità di una partizione è quella di "clusters ottimamente separati" che si può derivare dalla definizione di dati *well-structured* fatta da Fisher e Van Ness (1971). Una partizione produce un'ottima separazione dei clusters se:

$$\max_{1\le i\le k}\left\{d(X_a \in C_i, X_b \in C_i)\right\} \le \min_{\substack{1\le j\le k \\ j\ne i}}\left\{d(\mu_i,\mu_j)\right\}$$

cioè tutte le distanze tra due entità in un qualsiasi cluster sono minori delle distanze tra i centroidi dei clusters: l'eterogeneità nei clusters è inferiore alla eterogeneità tra i clusters.



Questa formulazione comporta il calcolo di

$$\frac{k*(k-1)}{2} + \sum_{i=1}^{k}\frac{n_i(n_i-1)}{2}$$

distanze; un numero che, se ancora elevato è comunque inferiore alle distanze da calcolare per la verifica della partizione ben strutturata.

Perché le due definizioni precedenti diventino operative è necessario precisare la funzione di distanza "d"; per semplicità

ci limitiamo alle due metriche più usuali: quella euclidea e quella di Mahalanobis.

Una partizione risulterà ben strutturata, in base alle due metriche considerate, se

$$\rho = \frac{\max_{1\le i\le k}\left\{\gamma_a = i, \gamma_b = i \left\| X_a - X_b \right\|^2\right\}}{\min_{\substack{1\le r,s\le k \\ r\ne s}}\left\{\gamma_\alpha = r, \gamma_\beta = s \left\| X_\alpha - X_\beta \right\|^2\right\}} \le 1;\quad \rho^* = \frac{\max_{1\le i\le k}\left\{\gamma_a = i, \gamma_b = i (X_a - X_b)^t T^{-1}(X_a - X_b)\right\}}{\min_{\substack{1\le r,s\le k \\ r\ne s}}\left\{\gamma_\alpha = r, \gamma_\beta = s (X_\alpha - X_\beta)^t T^{-1}(X_\alpha - X_\beta)\right\}} \le 1$$

Risulterà con clusters ottimamente separati se

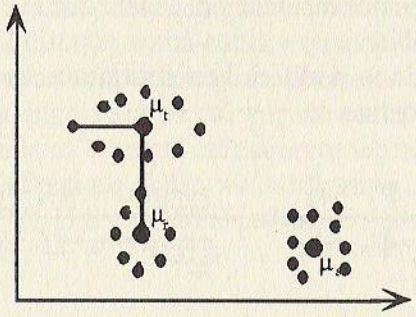
$$\tau = \frac{\max_{1\le i\le k}\left\{\gamma_a = i, \gamma_b = i \left\| X_a - X_b \right\|^2\right\}}{\min_{\substack{1\le r,s\le k \\ r\ne s}}\left\{\left\| \mu_r - \mu_s \right\|^2\right\}} \le 1;\quad \tau^* = \frac{\max_{1\le i\le k}\left\{\gamma_a = i, \gamma_b = i (X_a - X_b)^t T^{-1}(X_a - X_b)\right\}}{\min_{\substack{1\le r,s\le k \\ r\ne s}}\left\{(\mu_r - \mu_s)^t T^{-1}(\mu_r - \mu_s)\right\}} \le 1$$

Gli indici $\{\rho\}$ oltre a risultare eccessivamente laboriosi dal punto di vista dei calcoli (particolarmente quelli basati sulla Mahalanobis), potrebbero rivelarsi anche molto conservativi e indurre il rigetto addirittura della partizione dell'ottimo globale: i dati reali pur articolandosi in "k" gruppi non necessariamente sono disposti a formare una partizione ben strutturata o una partizione con clusters ottimamente separati.

Degli indici più semplici e pratici possono essere ricavati (rimanendo nel solco dell'ottima separazione dei clusters) misurando l'eterogeneità di un cluster con la distanza massima (nelle due metriche) dal centroide del cluster:

$$\tau_1 = \frac{\max_{1\le i\le k}\left\{\gamma_a = i \left| \left\| X_a - \mu_i \right\|\right.\right\}}{\min_{\substack{1\le j\le k \\ i\ne j}}\left\{\left\| \mu_i - \mu_j \right\|\right\}} \le 1;\qquad \tau_1^* = \frac{\max_{1\le i\le k}\left\{\gamma_a = i \left| (X_a - \mu_i)^t T^{-1}(X_a - \mu_i)\right.\right\}}{\min_{\substack{1\le j\le k \\ i\ne i}}\left\{(\mu_i - \mu_j)^t T^{-1}(\mu_i - \mu_j)\right\}} \le 1$$

Nessuna entità dista dal proprio centroide più di quanto questo non disti dai centroidi degli altri clusters.



Questa configurazione è meno severa e più semplice da verificare delle precedenti. Le distanze da calcolare sono soltanto

$$\frac{k*(k-1)}{2} + n$$

un numero molto ridotto rispetto alle altre.

Una terza caratterizzazione della partizione finale di un raggruppamento iterativo può essere ottenuta misurando l'eterogeneità del cluster con la distanza media tra due entità qualsiasi nello stesso cluster. Consideriamo *ammissibile* una partizione che verifichi la seguente condizione

$$\max_{1\le i\le k}\left\{\frac{\sum_{\gamma_a=i}\sum_{\gamma_b=i} d(X_a, X_b)}{n_i^2}\right\} \le \min_{\substack{1\le j\le k \\ i\ne j}}\left\{d(\mu_i,\mu_j)\right\}$$

Per semplificare le formule si utilizza la metrica euclidea in modo da sfruttare la relazione tra differenza quadratica media e varianza di un indicatore, cioè

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|^2}{n^2} = 2 * \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

dove $\bar{x}$ indica la media aritmetica dell'indicatore. Ne consegue che una partizione risulta ammissibile quando

$$\tau_2 = \frac{\max_{1\le i\le k}\left\{Tr(\hat{\Sigma}_i)\right\}}{\min_{\substack{1\le j\le k \\ i\ne j}}\left\{\left\|\mu_i-\mu_j\right\|^2\right\}} \le 1$$

se cioè, la distanza media tra due entità nello stesso cluster (uno qualsiasi) è sempre inferiore alla distanza tra i centroidi di due clusters qualsiasi.

Le forme forti di Diday.

Data l'incertezza dei concetti coinvolti, la soluzione finale di un raggruppamento iterativo non può essere espressa in termini di una sola partizione, ma dovrà basarsi su una molteplicità (diciamo in numero di "Q") di soluzioni locali, all'interno delle quali riconoscere delle "forme forti" (Diday, 1971). Costruiamo un indice di appartenenza al cluster C_j

$$e_{ijq} = \begin{cases} 1 & \text{se l'entità i - esima é posta nel cluster j - esimo nella partizione q - esima} \\ 0 & \text{altrimenti} \end{cases}$$

costruiamo anche un indice di permanenza nel cluster della entità i-esima con la somma degli indici di appartenenza

$$E_{ij} = \sum_{q=1}^{Q} e_{ijq}$$

Definiamo "stabile" un'entità posta sempre nello stesso cluster "j" cioè tale che $E_{ij}=Q$. Lo "zoccolo duro" di ogni cluster è costituito dalle sue entità stabili:

$$A_j = \left\{X_i \in C_j \middle| E_{ij} = Q\right\}; \qquad j = 1,2,\ldots,k$$

cioè A_j è formato da tutte quelle entità che nelle "Q" partizioni localmente ottime si è sempre trovato nello stesso cluster indipendentemente dalla partizione iniziale usata per avviare il rag-

gruppamento iterativo. La speranza è ovviamente che nessuno degli A_j sia vuoto (se questo dovesse succedere sarà segno che il numero di gruppi deve essere modificato). Poiché la pratica impone la scelta di una e una sola partizione come quella ottima, è opportuno riavviare la procedura iterativa di classificazione a partire dalle forme forti usando come partizione iniziale le Aj.

Definiamo "erratica" l'entità i-esima per la quale si abbia

$$\max_{1\le j\le k}\{E_{ij}\}=\left[\frac{Q}{k}\right]+1$$

cioè un'entità che è classificata ogni volta in un cluster diverso ovvero con ripetizioni obbligate solo dal fatto che $k\le Q$. Sia "D" l'insieme delle entità erratiche riscontrate tra le "Q" partizioni delle entità:

$$D=\left\{X_i;\ 1\le i\le n \middle| \max_{1\le j\le k}E_{ij}=\left[\frac{Q}{k}\right]+1\right\}$$

Una soluzione finale accettabile dovrebbe dare luogo ad Aj molto numerosi e ad un insieme D con pochissimi elementi. In letteratura non sono stati proposto indici della qualità della partizione basate sulle idee di Diday.

Alcune simulazioni.

I tre esempi utilizzati per valutare i metodi gerarchici aggregativi sono stati anche analizzati adoperando tre algoritmi che implementano le minimizzazioni di $Tr(\hat{\Sigma})$, $|\hat{\Sigma}|$ e della media geometrica dei determinanti delle matrici di varianze-covarianze "*within*": $Mg(|\hat{\Sigma}_i|)$. I tre algoritmi sono stati avviati assegnando casualmente le unità a uno dei due cluster. La configurazione iniziale è stata ovviamente la stessa per ogni procedura.

Nel primo caso, in mancanza di una vera e propria struttura in gruppi, i metodi considerati, come si era preannunciato nei

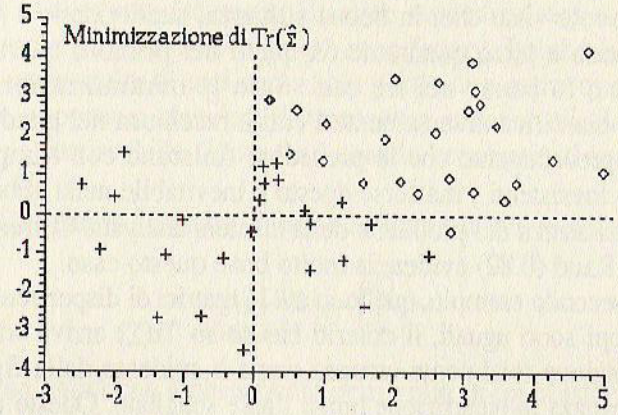
paragrafi precedenti, ne hanno inventata una, nemmeno tanto irragionevole visto che, in buona sostanza, suddividono i punti nel secondo e terzo quadrante da quelli del primo e quarto. Il risultato è lo stesso nei tre casi. Solo la minimizzazione di $Mg(|\hat{\Sigma}_i|)$ classifica diversamente l'entità racchiusa nel quadrato. Certo è preoccupante che le procedure finiscano con l'imporre strutture inesistenti , ma forse questo è inevitabile nella generale indeterminatezza del problema della clusters analysis. Comunque il test di Rand (0.82) evidenzia molto bene questo caso.

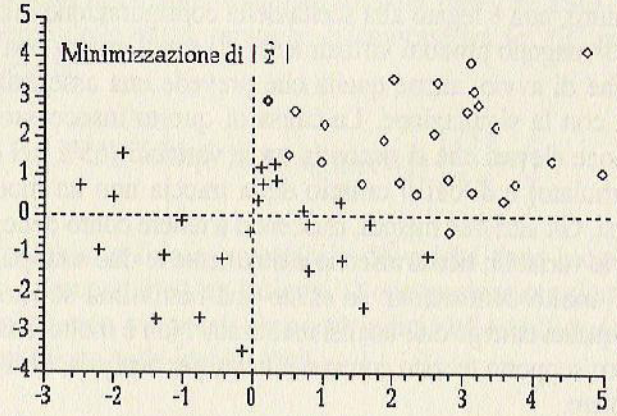
Nel secondo esempio, quello in cui le matrici di dispersione dei due gruppi sono uguali, il criterio basato su $Tr(\hat{\Sigma})$ arriva ad una classificazione totalmente erronea come è evidente dalla figura. Infatti realizza la bipartizione lungo l'asse sbagliato. Questo risultato, peraltro, non è legato alla scelta della configurazione iniziale, in quanto vengono prodotti virtualmente gli stessi gruppi con tutte le tecniche di avvio, anche quella che prevede una assegnazione allineata con la simulazione. La causa di questo insuccesso è la correlazione elevata che si riscontra tra le variabili (65% nel campione simulato) e di cui il criterio della traccia non ha modo di accorgersi. Gli altri due metodi, riuscendo a tenere conto delle relazioni tra le variabili, ricostruiscono esattamente le due sottopopolazioni. E' molto confortante: se esiste una nettissima struttura di gruppo, questa emerge dall'analisi realizzata. Non è molto, ma senza un vero supporto teorico come si è in cluster analysis, ci si deve accontentare.

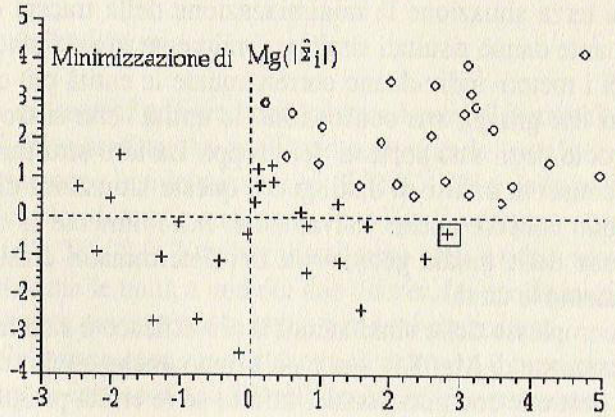
Nella terza situazione la minimizzazione della traccia e del determinante danno risultati simili e similmente insoddisfacenti. Entrambi i metodi individuano correttamente le entità più caratterizzanti dei gruppi, ma confondono le entità che si trovano sull'incrocio degli assi portanti dei gruppi. La loro strutturazione non consente infatti di distinguere queste situazioni che in effetti sono confuse anche visivamente. Mirabilmente la minimizzazione della media geometrica dei determinanti classifica correttamente le unità.

Nel complesso delle simulazioni il più efficace è risultato la minimizzazione di $Mg(|\hat{\Sigma}_i|)$, ma è purtroppo anche quello di più difficile gestione computazionale. Infatti, se le entità presenti in

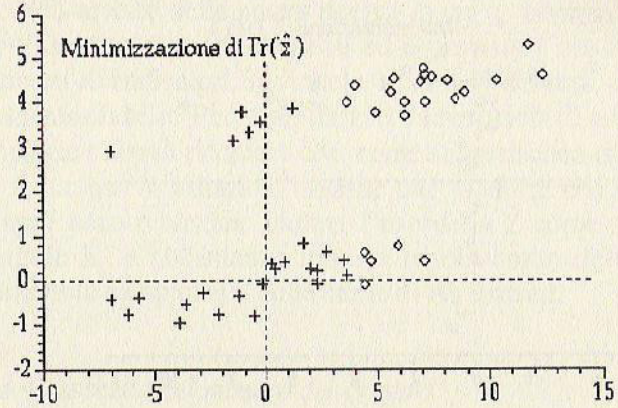
Prima simulazione

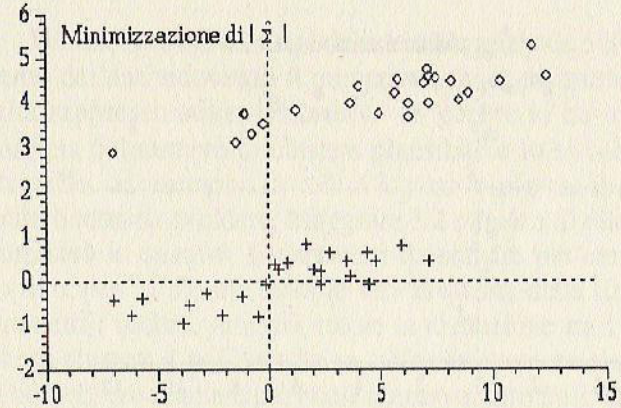


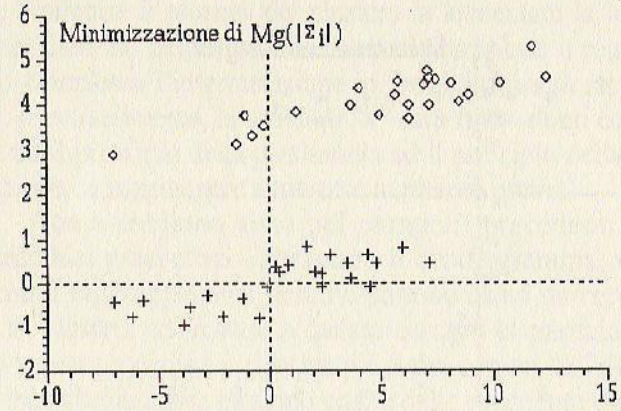




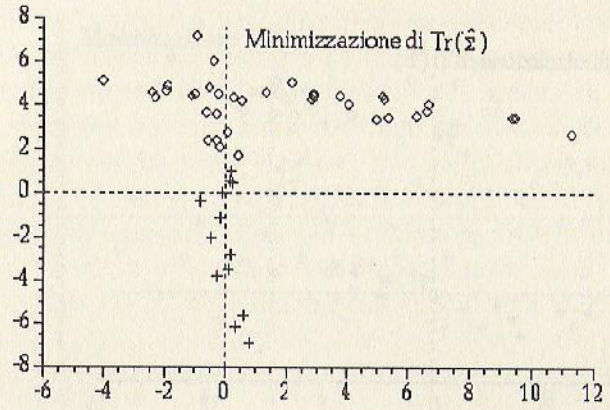
Seconda simulazione

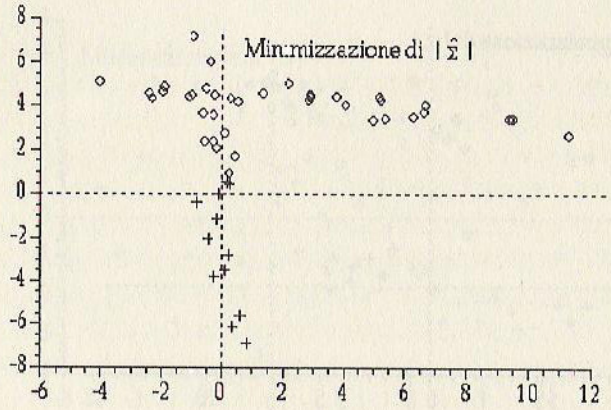


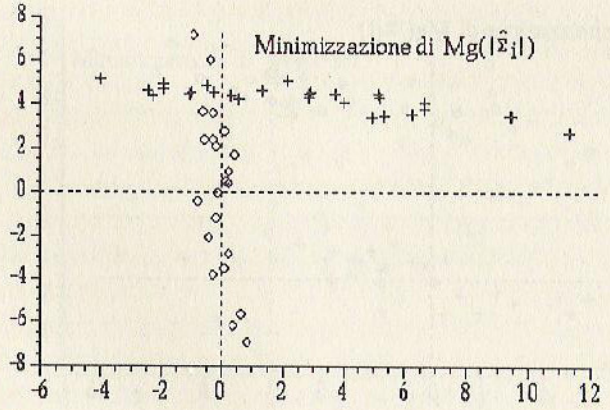




Terza simulazione







un dato cluster sono poche, la stima di $\hat{\Sigma}_i$ è piuttosto inefficiente. Nell'ambito della nostra ricerca in cui è scontata l'esistenza di clusters con pochi elementi ed è previsto l'uso di un buon numero di indicatori, il vincolo $n_i>(m+1)$ diventa un ostacolo insormontabile. Per attenuare il rammarico di non poterlo impiegare si può ricordare che, come suggeriscono diversi autori, il suo uso in situazioni inadatte può produrre una struttura in gruppi poco realistica; inoltre, l'uso della $\hat{\Sigma}$ come stima delle singole $\hat{\Sigma}_i$, è abbastanza robusta tenuto conto della estrema variabilità campionaria delle stime di tali matrici.

3.3.4 La scelta del numero dei gruppi.

Uno dei principali scopi della cluster analysis è il compattamento dei dati riducendo il numero di entità da trattare a poche unità rappresentative: i clusters. In genere si ha una idea di massima del numero di clusters plausibile e lo si colloca in un intervallo, ad esempio: $k_1 \leq k \leq k_2$, ma quale sia il suo valore esatto è tutto da decidere. Maggiore è il numero di clusters maggiore sarà la quantità di parametri da stimare per caratterizzarli (ovvero, per "n" fissato, minore sarà la qualità delle stime di quei parametri); inoltre, sarà più sottile la distinzione tra i profili dei diversi clusters (i poli), anche se, all'aumentare dei clusters, questi dovrebbero risultare più coesi intorno ai loro poli. D'altro canto, diminuire il numero dei clusters fa aumentare la loro dispersione interna, ne diminuisce la specializzazione e rende sempre più complessa l'interpretazione in termini di unità rappresentative. Inevitabilmente, la scelta di "k" sarà frutto di un compromesso tra il principio della parsimonia ed il principio della specializzazione da raggiungere attraverso numerose prove.

Come abbiamo visto nel paragrafi precedenti, i metodi gerarchici prospettano, attraverso il dendrogramma, varie soluzioni. I raggruppamenti iterativi partono da un numero prefissato di clusters ed in base a questo cercano la partizione ottima. La stessa procedura è ripetuta più volte con un "k" diverso. Per le tecniche iterative esistono anche delle procedure che includo-

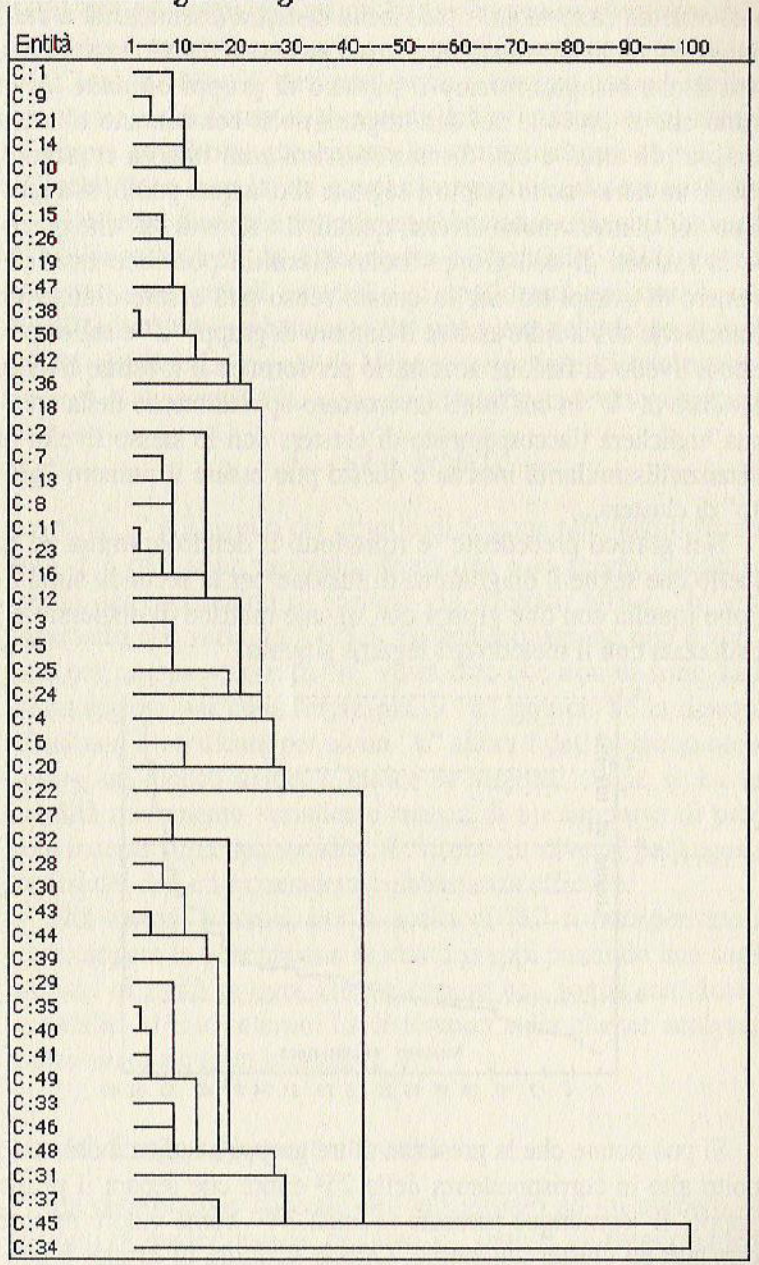
no automatismi che fondono o scindono due clusters nel tentativo di trovare la soluzione al problema del numero di gruppi (un esempio è la famosa tecnica ISODATA, adoperata per la classificazione dei dati da satellite), ma riteniamo si tratti di una "meccanizzazione" eccessiva della cluster analysis.

Se i dati presentassero un'articolazione spontanea in gruppi ben distinti e coesi la determinazione di "k" non comporterebbe alcuna difficoltà: sia il dendrogramma per i metodi gerarchici che la valutazione del criterio di ottimizzazione per i raggruppamenti iterativi non mancherebbero di evidenziare, anche con le più semplici tecniche grafiche, quale debba essere il corretto numero di clusters. D'altra parte, se per nessun valore di "k" si riscontrano "fratture" nel dendrogramma o miglioramenti ap-prezzabili nel criterio di ottimizzazione sarà segno che o i dati ruotano indifferenziati intorno ad un nucleo comune (distribuzione uniforme in una ipersfera) oppure che i dati sono un blocco indistinto privo di ogni inclinazione a formare dei gruppi (distribuzione uniforme in un ipercubo). In realtà, i dati si presentano come possono: i gruppi hanno forma molto irregolare, con centroidi ravvicinati ed entità a volte in numero talmente ridotto da portare a stime di medie e varianze ben poco affidabili; quando le estrazioni casuali da una popolazione in "k" gruppi non producano uin campione particolarmente scadente in cui uno o più clusters siano assenti. Stabilire "k" nelle situazioni vere è perciò una scelta più intuitiva che tecnica e potrebbe cambiare se cambia la persona che analizza i risultati. Esistono però molti indici che riducono la soggettività della scelta e permettono di guardare ai risultati del clustering con maggiore convinzione (non perdendo però mai di vista il fatto che la scelta di "k" non può avvenire solo su basi formali). Noi ci limiteremo a presentarne alcuni già noti ed altri di nostra proposta (per una loro rassegna si veda Vicari, 1990).

Metodi gerarchici.

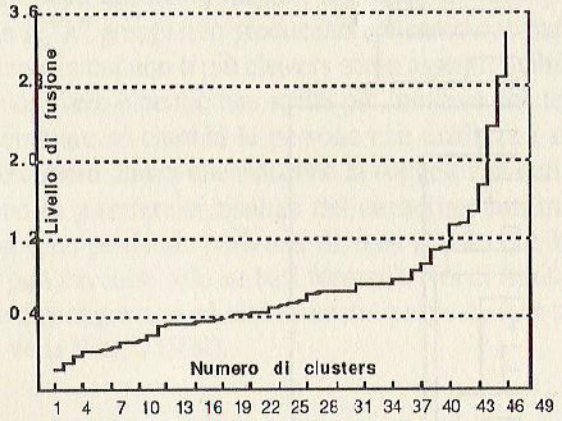
L'impiego delle tecniche gerarchiche è già di per sé una dichiarazione di incertezza, nel senso che i dati possono contenere ben più di una singola struttura di gruppo. Il problema è deci-

Metodo del legame singolo



dere a cuale livello del dendrogramma fermarsi. Ad occhio, se si intravede un salto molto ripido nella distanza/dissimilarità a cui i clusters di un livello si aggregano per passare al livello successivo, sarà lì che bisogna cercare il numero di gruppi ottimale. Man mano che si procede nel dendrogramma si considerano clusters sempre più ampi e con distanza/dissimilarità interna crescente, quindi un ramo molto ampio è segnale che, a quel punto, si aggregano due clusters molto diversi; quindi, il valore di "k" che precede la fusione di due gruppi molto dissimili potrebbe essere il numero di gruppi trovati. In questo senso può essere d'aiuto un grafico che abbia sulle ascisse il numero di gruppi "k" e sulle ordinate il livello di fusione necessario per formare il k-esimo livello. Il valore di "k" in cui inizia un marcato appiattimento nella spezzata, indicherà l'accorpamento di clusters con lo stesso livello di distanza/dissimilarità interna e questo può essere il numero "giusto" di clusters.

Nel grafico precedente è riprodotto il dendrogramma ed in quello che segue il diagramma di fusione per la seconda simulazione (quella con due gruppi con uguale matrice di dispersione) analizzata con il metodo del legame singolo.



Si può notare che la presenza di tre gruppi è indicata dal ramo molto alto in corrispondenza della 25ª entità che separa il primo gruppo. E' comunque presente la unità 34ª come valore remoto formante un cluster autonomo. Questa struttura di gruppi è con-

fermata dall'appiattimento riscontrabile dopo la fusione della 25ª entità nel primo gruppo ed un finale impennamento allorché si individua il valore remoto come cluster diverso dai precedenti.

L'analisi visiva di dati reali porta a risultati troppo variabili per essere soddisfacente: gli aspetti grafici indicativi del numero di gruppi potrebbero presentarsi in corrispondenza di più di un valore di "k" e la scelta tra questi diventerebbe del tutto arbitraria.

Un'indicazione più formale ed abbastanza efficace è la regola di Mojena (illustrata da Aldendenfer e Blashfield, 1984) che propongono una sorta di intervallo di confidenza unilaterale sulla significatività del cambiamento che interviene nel criterio di fusione dei clusters. La regola di Mojena suggerisce di aumentare il numero di cluster "k" fino a che risulti verificata la disequazione:

$$f_{k+1} > \bar{f} + h * \hat{\sigma}(f)$$

dove "f_{k+1}" è il livello del criterio di fusione allo stadio successivo, $\bar{f}$ e $\hat{\sigma}(f)$ sono la media aritmetica e lo scarto quadratico medio dei coefficienti di fusione già adottati; infine, "h" è un parametro che varia tra 1 e 3.5. Se la disequazione non è verificata per nessun valore di "h" vorrà dire che non ci sono sufficienti ragioni per considerare più di "k" gruppi. Se la disequazione non è verificata per alcun "k" allora i dati possono considerarsi un blocco unico. L'indice di Mojena, anche se ha una validità meramente euristica e manca di un supporto di prove convincenti sulla sua validità, è fornito in diversi packages di analisi dei dati ed è considerato abbastanza efficace.

Nei metodi scissori, per la scelta di "k" si può pensare ad interrompere la generazione di nuovi gruppi ponendo una soglia minima di unità in ogni cluster oppure una soglia massima di variabilità al loro interno. La letteratura sulla cluster analysis è molto avara a questo proposito.

Metodi di raggruppamento iterativi.

La disposizione naturale dei dati intorno ad un numero ben definito di nuclei distinti, diciamo k_0, induce un miglioramento

fisiologico della partizione man mano che il numero di clusters si avvicina, per eccesso o per difetto, a k_0; il che non potrà non riflettersi in una forte riduzione nella matrice di devianze-codevianze globali nei clusters. Il numero di clusters ottimale può essere determinato rilevando opportunamente questo miglioramento.

A tal fine possiamo impiegare una tecnica simile a quella usata per decidere il numero di componenti principali o il numero di clusters nei metodi gerarchici agglomerativi. Si può cioè rappresentare in un grafico $|\hat{\Sigma}|$ oppure $Tr(\hat{\Sigma})$ in funzione del numero dei gruppi; se nella risultante spezzata è possibile individuare la solita rapida caduta con successivo appiattimento, il punto in cui questo comincia a succedere dovrebbe indicare k_0. Sull'effettiva utilità di queste procedure però Everitt (1979) ha molti dubbi.

In generale, per i raggruppamenti iterativi, la determinazione di "k" avviene provando la tecnica di clustering prescelta con un numero variabile di clusters partendo da un "k" che superi almeno di tre o quattro il numero che si sospetta sia in effetti presente. Sulla partizione finale del raggruppamento iterativo si calcolano vari indici la cui concordanza su di un valore di "k" dovrebbe impedire scelte erronee. Esistono molti indici analitici che possono essere di ausilio a questo fine, ma qui presentiamo quelli che possono essere calcolati a valle della partizione ottima per "k" fissato e che comunque non richiedano elaborazioni troppo complesse (ad esempio, la misura di Rand è di gestione abbastanza difficile e preferiamo utilizzarla nella validazione dei risultati nel prossimo paragrafo).

Beale (1969) si è chiesto quale sia il numero di clusters "significativo" ed ha proposto la seguente pseudo-statistica F:

$$\beta_k = \frac{\dfrac{Tr\left[\hat{\Sigma}(k)\right] - Tr\left[\hat{\Sigma}(k+1)\right]}{Tr\left[\hat{\Sigma}(k+1)\right]}}{\dfrac{n-k}{n-k-1} * \left(\dfrac{k+1}{k}\right)^{\frac{2}{m}} - 1} > F_{m,m(n-k-1)}$$

se la disuguaglianza è verificata per un certo "k" significa che la rappresentazione in termini di "k" cluster è insoddisfacente

rispetto a quanto si può ottenere con un cluster in più. L'indice "β" opera al meglio quando i centroidi dei clusters sono ben separati e la distribuzione delle loro entità forma delle ipersfere intorno ai centroidi. Se, come sempre succede nelle situazioni reali, queste condizioni non sono assicurate, le indicazioni fornite da "β" andranno considerate con prudenza badando più al suo comportamento rispetto a "k" che a rigorosi confronti con le soglie della "F" di Fisher.

Hartigan (1975) ha proposto un indice basato sulla scomposizione della devianza globale che rilevi l'avvicinarsi al corretto numero di clusters:

$$H = Log\left(\frac{Tr(B)}{Tr\left(\hat{\Sigma}\right)}\right)$$

L'andamento di H dovrebbe subire un forte accrescimento in corrispondenza del valore ottimale di "k" per poi dar luogo a lievi incrementi dovuti alla sola crescita tecnica della dispersione tra gruppi ed alla diminuzione della dispersione nei gruppi conseguente all'aumento del numero di clusters.

Un indice classico che sonda la qualità della partizione è il λ che abbiamo già visto in precedenza:

$$\lambda = nLog\left(\frac{|T|}{|\hat{\Sigma}|}\right)$$

Friedman e Rubin (1967) sono ad esempio convinti che un'impennata del valore seguita da accrescimenti più moderati (dovuti alle stesse ragioni illustrate per la misura di Hartigan) indichi il corretto numero di gruppi.

Un altro criterio, proposto da Calinski e Harabasz (1974) è basato su quantità analoghe a quelle usate dall'indice di Hartigan, ma che coinvolge più direttamente il numero corrente di clusters:

$$CH = \frac{n-k}{k-1} * \frac{Tr(B)}{Tr\left(\hat{\Sigma}\right)}$$

Se CH aumenta monotonicamente i dati non sono interpretabili come una aggregazione di gruppi diversi, ma sono da considerarsi un blocco indistinto. La diminuzione uniforme di CH rispetto a "k" è segno che nei dati c'è una struttura di gruppo gerarchica. Infine, se CH aumenta fino ad un certo punto e poi diminuisce, il punto di massimo rivelerà il numero ottimale di clusters.

Marriott (1971) riflettendo su quanto succede nella distribuzione uniforme in cui i clusters tendono a riprodurre, in scala minore, la struttura complessiva dei dati, ha proposto di scegliere "k" in base alla quantità

$$M_1 = \min_{k_1 \le k \le k_2} \left\{ 2\mathrm{Log}(k) + \mathrm{Ln}\left|\hat{\Sigma}\right| \right\}$$

Se i dati formassero un blocco unico distribuito intorno ad un nucleo centrale il valore minimo si otterrebbe per k=1. Se invece i dati si distribuissero uniformemente su di un ipercubo, il valore di M_1 rimarrebbe pressoché costante al variare di k. Se poi si dovesse riscontrare:

$$M_2 = \frac{k^2 \left|\hat{\Sigma}\right|}{|T|} \ge 1$$

per ogni valore di "k " vorrà dire che non esiste una struttura di gruppo.

Le indicazioni su β_k, H, I, CH, M1, M_2 vanno seguite tenendo conto che esse sono basate sull'ipotesi di distribuzioni normali multivariate con la stessa matrice di dispersione e che perciò sono applicabili solo nella misura in cui tale ipotesi è plausibile.

Altre misure possono essere ricavate dalla definizione di partizione ben strutturata (ρ e ρ^*), partizione con clusters ottimamente separati (τ_1 e τ_1^*) e partizione ammissibile (τ_2). Ci si aspetta che, all'avvicinarsi alla partizione ottima, i valori degli indici siano inferiori all'unità. Il valore di "k" più grande per cui questo succede sarà il numero ottimo di clusters: k_0.

Simulazioni per lo studio degli indici sul numero di clusters.

Gli indici del precedente sottoparagrafo sono stati provati con un algoritmo orientato alla minimizzazione del determinante della matrice di devianze-codevianze globali nei gruppi, $\min\{\hat{\Sigma}\}$, con dei valori iniziali ottenuti con il metodo delle distanze massime dal vettore delle medie globali (questa combinazione di metodi è quella che, tra tutte, è risultata più produttiva e affidabile in tutte le prove effettuate).

Il dataset prevede 100 osservazioni su 5 indicatori per un totale di 500 informazioni ottenute con due tipologie: una struttura naturale in 4 gruppi ed un blocco unico. In ognuna delle due tipologie sono state distinte tre situazioni diverse per la matrice di varianze-covarianze globale nei gruppi.

Organizzazione naturale in 4 gruppi.

La numerosità dei gruppi ed i poli dei gruppi sono riportati nella tabella che segue

n_i	μ_{i1}	μ_{i2}	μ_{i3}	μ_{i4}	μ_{i5}
9	1	2	3	4	5
17	5	5	5	5	5
31	–5	–5	–5	–5	–5
43	–10	10	–10	10	–10

e sono state scelte per evidenziare il comportamento dei criteri sia di fronte alla contemporanea presenza di gruppi di peso molto diverso che di fronte a gruppi ben separati compresenti a gruppi ravvicinato quali il cluster "1" e il cluster "2".

Caso 1: $\Sigma_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ . & 1 & 0 & 0 & 0 \\ . & . & 1 & 0 & 0 \\ . & . & . & 1 & 0 \\ . & . & . & . & 1 \end{bmatrix}$ Caso 2: $\Sigma_i = \begin{bmatrix} 2 & 1 & 1.5 & 2 & 2.5 \\ . & 4 & 3 & 3.5 & 5 \\ . & . & 8 & 6 & 6.5 \\ . & . & . & 16 & 10 \\ . & . & . & . & 32 \end{bmatrix}$

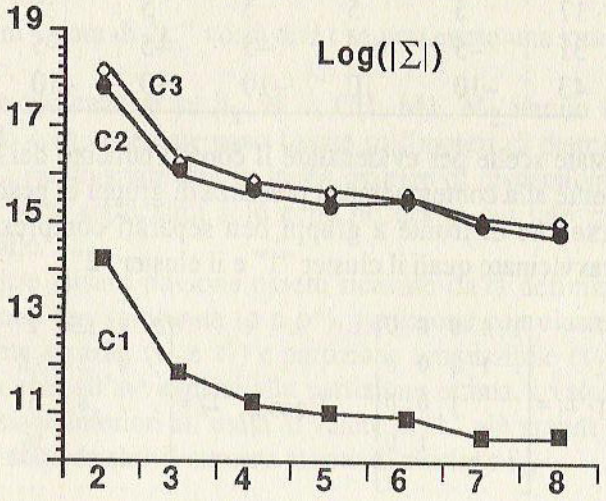
Caso 3: $\Sigma_i = \begin{bmatrix} 2 & a & a & a & a \\ . & 4 & a & a & a \\ . & . & 8 & a & a \\ . & . & . & 16 & a \\ . & . & . & . & 32 \end{bmatrix}$ con $\begin{cases} a=-1.5 \text{ se } i=1 \\ a=1.5 \text{ se } i=2 \\ a=2.0 \text{ se } i=3 \\ a=2.5 \text{ se } i=4 \end{cases}$

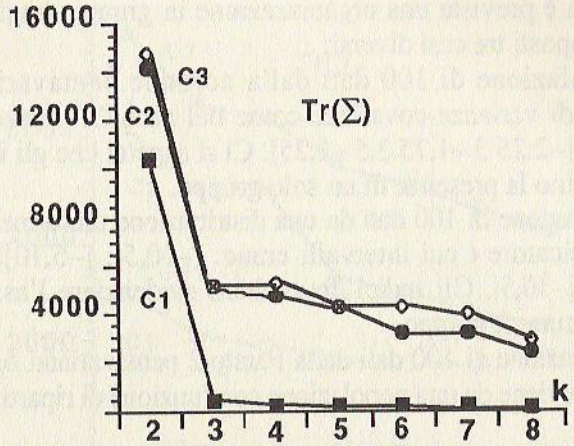
Nel primo caso si ha $\Sigma=I$, per cui i gruppi dovrebbero essere molto addensati intorno ai rispettivi centroidi e nessuna tecnica dovrebbe sbagliare nell'individuare il numero di gruppi.

Nel secondo caso, la dispersione relativamente elevata nei gruppi unita alla vicinanza del cluster "1" al cluster "2" potrebbe indurre qualche incertezza; nel terzo, il criterio di ottimizzazione è poco coerente con la filosofia costruttiva del dataset e ciò dovrebbe essere una prova piuttosto severa per tutti gli indici.

Vediamo innanzitutto come si comportano le rappresentazioni grafiche di due indicatori della qualità della partizione: il logaritmo del determinante della matrice di devianze-codevianze totali e la sua traccia, rispetto a "k".

I due grafici individuano il numero ottimale di clusters fra 3 e 4 nel caso 1. L'indicazione non è univoca in quanto il tratto della spezzata fra 3 e 4 può sia essere parte del tratto a caduta





ripida che il segmento di inizio dell'appiattimento. Negli altri due casi è indicato con relativa sicurezza che il numero ottimale di clusters è tre. C'è infatti il "gomito" per k=3 e poi la discesa regolare del criterio all'aumentare di k. Quindi le indicazioni dei due grafici porterebbero a sottostimare il numero di gruppi.

La situazione più limpida, tra quelle simulate per la prima tipologia è il caso 1. Fortunatamente, come si vede nella tabella presentata più avanti quasi tutti gli indici danno la corretta segnalazione di k_0=4 anche se non con la stessa leggibilità. Risulta troppo severo l'indice di partizione ben strutturata (lo stesso calcolato con la metrica di Mahalanobis, non riportato, è ancora più conservativo).

Nel caso 2, gli indici di Beale, Marriott, Calinsky e Harabasz, Hartigan e τ_1 e τ_1^* propongono tre gruppi. Solo il τ_2 è per k=4 mostrando di individuare la vera struttura dei dati. Nel caso 3, tutti gli indici puntano a k=3 anche se, nel progetto di simulazione era k=4. Questo cattivo esito è da attribuire alla non eccessiva robustezza del metodo del min $|\Sigma|$ alla violazione della condizione di uguaglianza tra le matrici di varianze-covarianze nei gruppi.

Assenza di una struttura multigruppo.

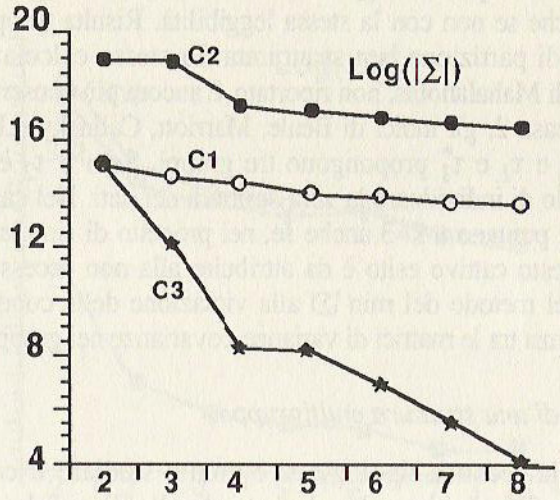
Con la seconda tipologia si è voluto studiare il comportamento degli indici quando, almeno a livello di popolazione teo-

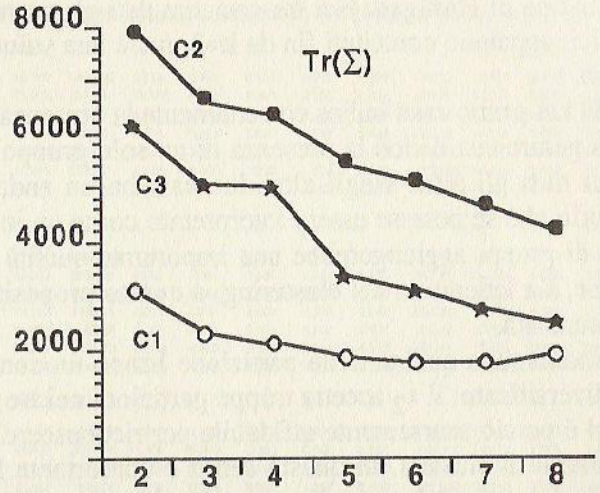
rica, non è prevista una organizzazione in gruppi. Anche qui si sono proposti tre casi diversi:
1) Simulazione di 100 dati dalla normale pentavariata con matrice di varianze-covariaze come nel caso "2" precedente e con μ = [–2.25 3 –1.75 3.5 –1.25]. Ci si aspetta che gli indicatori segnalino la presenza di un solo gruppo.
2) Simulazione di 100 dati da una distribuzione uniforme per ciascun indicatore i cui intervalli erano: [–10,5]; [–5,10]; [10,5]; [–5,10]; [–10,5]. Gli indici dovrebbero evidenziare l'assenza di una struttura di gruppo
3) Simulazione di 100 dati dalla Pareto/2 pentavariata: ogni indicatore proviene da una popolazione con funzione di ripartizone:

$$F(X_i) = 1 - \left[\frac{1}{2} + X_i\right]^{-3}$$

anche in questo caso ci si aspetta che gli indici suggeriscano k=1.

I grafici segnalano giustamente l'assenza della struttura multigruppo senza però distinguere tra k=1 e k=0 cioè tra la presenza di un solo forte nucleo centrale che attrae tutte le entità e la distribuzione uniforme sull'ipercubo.





Nel terzo caso (i valori dei criteri sono stati opportunamente scalati) c'è il sospetto di una articolazione in 4 o 5 gruppi (i grafici qui sono discordi). E' vero che la "frattura" nel grafico non sembra esagerata, ma se fossimo alla ricerca di una struttura di quattro o cinque clusters sarebbe stata certamente considerata interessante. In realtà, i cinque clusters sono in parte confermati dall'indice di Beale che solo qui, nei tre casi considerati della seconda tipologia, assume un valore non troppo banale (4.280).

Nei casi "1" e "2" c'è da notare il comportamento molto aderente alle previsioni teoriche del secondo indice di Marriott che, rimanendo costantemente sopra l'unità, indica l'assenza della struttura di gruppo in tutti i tre casi considerati. E' meno sicuro nel caso "3" anche se i valori sono piuttosto sospetti, soprattutto se confrontati con quelli assunti nei casi della precedente tipologia. In questo senso è più nitido il comportamento dell'indice M1 che aumenta sistematicamente con l'aumentare di "k" ponendo il minimo a k≤1.

Il λ di Friedman e Rubin è poco chiaro per definizione. Per comprenderlo si dovrebbe infatti badare ai suoi incrementi ed al loro eventuale saturarsi. Analogo comportamento dovrebbe poi

avere l'indice di Hartigan. Nei tre casi simulati gli scarti, per i due indici, appaiono contenuti fin da k=2, ma è una valutazione arbitraria.

Il CH nel primo caso indica correttamente la presenza di una struttura gerarchica e cioé la presenza di un solo gruppo contenitore di tutti gli altri. Negli altri due casi ha un andamento oscillatorio che se potesse essere interpretato come un indice di assenza di gruppi aggiungerebbe una importante qualità a questo indice. La letteratura sul clustering, a questo proposito, non si è pronunciata.

Gli indici sulla qualità della partizione hanno un comportamento diversificato: il τ_2 accetta troppe partizioni nei tre casi di prova ed è perciò scarsamente affidabile per riconoscere il corretto numero di clusters. In questo senso è confortante l'andamento di τ_1 e τ_1^* (il loro andamento è molto ben abbinato) secondo i quali nessuna delle partizioni finali proposte nei tre casi della seconda tipologia può considerarsi come una ottima separazione dei clusters.

In conclusione si può dire che se nei dati c'è una suddivisione in gruppi molto netta questa verrà evidenziata da tutti gli indici come verrà evidenziata, ad esempio dall'indice CH, l'assenza di ogni raggruppamento o la presenza di un gruppo unico. Dalle nostre prove sembra più semplice riuscire a negare la presenza nei dati di una organizzazione multigruppo rispetto alla individuazione del numero di gruppi che la costituiscono. Le situazioni più difficili sono i casi intermedi (quelli reali, quasi sempre) in cui però l'uso congiunto delle informazioni dei vari indici e le conoscenze sui dati dovrebbero portare ad una scelta molto tranquilla del numero ottimale di clusters.

Risultati delle simulazioni sul numero dei clusters presenti nei dati.

Σ=I	k	Log|Σ|	Tr(Σ)	b	t_1	τ_1^*	t_2	r	I	M_1	M_2	CH	H
	2	14.773	9893.62	2.714	0.467	0.421	0.278	1.209	640.57	15.559	0.007	88.560	-0.101
	3	11.809	647.79	75.862	0.067	0.057	0.028	0.416	376.89	14.007	0.001	362.048	3.535
	4	11.146	439.47	3.547	0.358	0.339	0.133	1.013	943.21	13.919	0.001	339.845	3.735
	5	10.973	412.53	0.623	3.464	3.945	1.252	3.148	960.51	14.192	0.002	1060.892	3.799
	6	10.847	393.03	0.570	3.370	3.381	1.236	3.149	973.15	14.430	0.002	882.369	3.849
	7	10.456	363.36	1.068	3.006	3.036	1.430	2.993	1012.25	14.348	0.002	788.156	3.929
	8	10.519	351.30	0.518	5.669	5.312	2.362	4.042	1005.97	14.677	0.003	691.689	3.964

Σ=cost.	k	Log|Σ|	Tr(Σ)	b	t_1	τ_1^*	t_2	r	I	M_1	M_2	CH	H
	2	17.741	14351.39	1.957	0.632	0.817	0.352	1.706	336.62	19.727	0.138	93.865	-0.428
	3	16.029	5334.32	8.982	0.391	0.330	0.144	1.133	507.81	18.226	0.056	167.025	1.237
	4	15.606	4949.48	0.582	3.598	2.662	0.877	3.598	550.13	18.378	0.065	121.260	1.333
	5	15.293	4440.59	1.093	4.751	4.807	1.676	3.348	581.38	18.512	0.075	103.033	1.468
	6	15.437	3405.26	3.491	3.431	3.194	0.952	2.977	567.04	19.020	0.124	112.072	1.785
	7	14.901	3421.50	-0.063	9.859	8.237	2.234	4.823	620.61	18.793	0.099	91.387	1.780
	8	14.752	2606.18	4.716	2.540	2.086	0.732	3.349	635.47	18.911	0.111	106.400	2.091

BU-Nor.	k	Log|Σ|	Tr(Σ)	b	t_1	τ_1^*	t_2	r	I	M_1	M_2	CH	H
	2	15.129	3123.58	1.917	1.328	3.201	0.421	2.031	118.763	16.515	1.220	62.557	-0.449
	3	14.676	2321.66	1.335	1.968	2.353	0.603	2.543	164.060	16.373	1.744	58.433	0.186
	4	14.418	2154.96	0.579	2.300	3.198	0.767	2.518	189.854	17.190	2.397	43.992	0.318
	5	14.091	1889.76	1.338	2.340	5.362	1.296	2.953	222.510	17.310	2.701	40.565	0.535
	6	13.989	1792.90	0.620	4.160	4.656	1.615	3.460	232.762	17.572	3.511	34.861	0.618
	7	13.716	1800.16	-0.054	6.412	3.454	2.736	4.219	260.063	17.606	3.637	28.563	0.611
	8	13.628	1920.24	-0.943	5.579	5.789	2.103	3.633	263.051	21.883	4.342	21.883	0.510

BU-Uni.	k	Log|Σ|	Tr(Σ)	b	t_1	τ_1^*	t_2	r	I	M_1	M_2	CH	H
	2	18.900	7885.35	0.484	3.262	3.273	1.584	3.189	69.047	19.476	2.005	15.782	-0.823
	3	18.781	6610.45	1.025	1.910	1.956	0.912	2.417	111.029	19.868	2.965	18.671	-0.955
	4	17.202	6338.48	0.321	3.057	2.958	1.429	3.057	157.876	19.974	3.300	14.220	-0.811
	5	16.978	5424.77	1.606	2.622	2.754	1.246	2.273	180.234	20.197	4.121	16.332	-0.374
	6	16.751	5075.42	0.790	1.957	1.951	0.856	2.529	202.930	20.334	4.728	15.112	-0.218
	7	16.524	4618.37	1.317	2.526	2.611	1.021	2.530	225.631	20.416	5.132	15.222	-0.018
	8	16.347	4194.77	1.525	2.278	2.191	0.941	2.768	243.336	20.506	5.616	15.542	0.164

BU-Par.	k	Log|Σ|	Tr(Σ)	b	t_1	τ_1^*	t_2	r	I	M_1	M_2	CH	H
	2	2.980	19.925	2.767	1.884	1.834	0.379	1.695	197.232	4.365	0.557	90.304	-0.012
	3	2.406	16.660	1.041	1.803	1.828	0.488	1.692	254.261	4.606	0.708	62.952	0.261
	4	2.034	16.033	0.293	6.906	7.949	1.407	3.314	291.634	4.807	0.866	44.413	0.323
	5	1.609	11.066	4.280	4.079	4.490	2.456	3.482	334.114	4.329	0.385	58.421	0.900
	6	1.369	10.039	1.175	4.613	5.058	2.156	3.476	358.223	4.952	1.001	52.891	1.034
	7	1.083	9.223	1.179	3.965	3.536	1.680	3.178	386.619	4.975	1.024	46.843	1.143
	8	0.801	8.193	1.895	2.235	2.700	1.303	2.724	414.551	4.960	1.009	46.273	1.301

Replicabilità della soluzione.

I risultati di ogni clustering dovrebbero essere valutati alla luce di due considerazioni:

1) Ogni metodo può suggerire l'esistenza di una struttura di gruppi anche in presenza di dati del tutto privi di un raggruppamento naturale;

2) Metodi diversi producono soluzioni differenti (talvolta molto differenti).

Nei paragrafi precedenti si è visto come ciascun metodo abbia delle "tendenze intrinseche": il legame singolo tende a trovare clusters di forma allungata e concatenati; il $\min\{Tr(\hat{\Sigma})\}$ è influenzato dai valori anomali; il $\min\{|\hat{\Sigma}|\}$ poggia molto sull'ipotesi di matrici di varianze-covarianze uguali nei gruppi. C'è quindi il rischio concreto che la soluzione trovata non sia ottima, o meglio che possa essere determinata dalle caratteristiche del metodo più che derivare dalla struttura effettiva dei dati considerati. Gli indici di qualità della partizione finale riescono a contenere l'arbitrarietà delle valutazioni, ma il giudizio rimane, nel complesso, soggettivo.

Se si disponesse di dati sufficienti si potrebbero ripartire casualmente le entità in due blocchi: uno operativo ed uno di controllo e provare su ciascuno il metodo prescelto: se i risultati convergessero allora l'esito potrebbe essere accettato con tranquillità. In genere, però, la disponibilità di dati non è mai tale da consentire questo tipo di verifica o per i costi di raccolta dei dati o perché i dati stessi sono già tutti quelli che si potevano acquisire. D'altra parte, la replicabilità della soluzione per blocchi casuali è solo una verifica di tipo confermativo: il fatto che si ripeta è confortante, ma la mancanza di ripetizione non necessariamente invalida una delle soluzioni trovate; ogni partizione è un ottimo condizionale e, nell'ambito delle condizioni poste, è la migliore soluzione possibile. Esiti basati su condizioni diverse non sono sempre comparabili.

Fisher e Van Ness (1971) hanno introdotto il concetto di ammissibilità frazionale che evidenzia più l'aspetto geometrico dei clusters che la densità delle entità al loro interno. L'ammissibilità frazionale è vigente a livello di entità se, duplicando sequenzialmente e singolarmente ogni entità e rifacendo per intero la clustering, l'entità duplicata è classificata sempre nello stesso cluster. Esiste a livello di cluster se, duplicando sequenzialmente e singolarmente ogni cluster e ripetendo ogni volta la procedura, i clusters non duplicati rimangono invariati. Esiste a livello di partizione se, duplicando l'intero dataset, nessuna entità cambia di cluster. Solo quest'ultimo tipo di ammissibilità frazionale sembra suscettibile di una verifica pratica non troppo

complessa e relativamente rapida: basterà usare l'indice di Rand sulle partizioni che si ricavano dai due blocchi ed un valore di questo indice molto prossimo all'unità ci convincerà sulla stabilità della soluzione. In realtà, nelle prove da noi effettuate, questo tipo di validazione non è risultato di alcuna utilità.

L'unico modo di utilizzare l'idea di semplicità del test di Rand e quella della stabilità della soluzione trovata è il confronto con delle distribuzioni multivariate ipotetiche in cui non sussistano, almeno a livello potenziale, strutture multigruppo. Dalla discussione fatta nei paragrafi 3.3 e 3.4 potremo ritenere abbastanza stabile una partizione finale multigruppo che sia dissimile:

1) da una normale multivariata con medie globali e matrice di varianze-covarianze globali calcolate sul dataset sottoposto alla clustering;
2) da una uniforme multivariata sull'ipercubo definito dai valori estremi degli indicatori.

Valori bassi dell'indice di Rand (ad esempio inferiori a 0.6) indicheranno che la partizione ottenuta è almeno più informativa rispetto ad un cluster unico o rispetto ad blocco indistinto.

3.3.5 Cluster analysis e analisi delle componenti principali.

La metodologia adottata in molte ricerche diretta alla individuazione di aree omogenee prevede di solito l'uso dell'analisi delle componenti principali (o dell'analisi fattoriale) per ridurre la dimensionalità del problema ed una successiva classificazione (gerarchica o iterativa) delle entità a partire non dagli indicatori originali, ma dai punteggi fattoriali. Questo però non è necessariamente l'iter più naturale ed esistono schemi alternativi che è utile richiamare.

Cluster analysis sulle componenti principali o sugli indicatori.

L'efficacia (e la gestibilità in concreto) di tutte le tecniche di classificazione è fortemente legata al numero di indicatori con

cui si descrivono le entità. All'aumentare di tale numero non solo crescono le esigenze di memoria, i tempi di elaborazione e di precisione nei calcoli, ma diminuiscono le possibilità di comprendere i risultati della classificazione, di individuare gli indicatori che determinano il numero e la configurazione dei gruppi e di spiegare il perché certe combinazioni di valori stiano al centro di tali gruppi. D'altra parte, la stessa impostazione dell'analisi dei SATI in fondo parte dell'ipotesi che, a priori, siano note alcune macrodeterminanti (demografia, struttura produttiva, disponibilità dei servizi, etc.) che differenziano o possono differenziare le entità comunali, ma che non si sappia quali esattamente siano gli indicatori che esprimono tali macrodeterminanti e quanti ne abbisognano per rappresentarle esaurientemente.

Il modo di procedere che ci sembra più produttivo è quello di attivare una preventiva analisi delle componenti principali ed avviare poi la clustering sui "punteggi" ottenuti a partire dalle prime più significative componenti. Su questo è opportuno riportare le considerazioni suggerite da Marriott (1971):

1) Il fatto che le componenti siano ortogonali non semplifica la cluster analysis: né quella gerarchica né quella iterativa (il fatto che T sia ora una matrice diagonale non implica che lo sia anche $\hat{\Sigma}$.

2) Ciò che conta nella clustering non sono i punteggi fattoriali con cui sono espresse le componenti principali, ma la variabilità globale da esse riprodotta. Ad esempio, la clustering basata su $\min\{|\hat{\Sigma}|\}$ è indipendente dal tipo di trasformazione lineare effettuata sui dati, ma non lo è l'analisi delle componenti principali nel senso che ogni diversa trasformazione produce componenti molto diverse, ma lo spazio da esse coperto sarà virtualmente lo stesso. Questo in pratica ci tranquillizza rispetto al problema di quale formula adoperare per lo *scaling* dei dati, anche se continuiamo a preferire la normalizzazione che mantiene le differenze tra medie e varianze covarianze tra gli indicatori. E' utile peraltro richiamare che, secondo Everitt (1979), l'ordine logico di analisi delle componenti principali e cluster analysis dovrebbe essere ribaltato: classificare in gruppi le entità e poi effettuare l'analisi delle componenti principali separatamente per ogni gruppo. Questo eviterebbe l'effettuazione dell'analisi su dati composti molto eterogeneamente ed in cui la correlazione globale non potrebbe che essere la sintesi delle correlazioni nei gruppi (il dilemma scompare se tutti i gruppi hanno la stessa matrice di varianze-covarianze, ma questo non è il caso più frequente). Lo stesso Everitt ammette però che quando il numero di indicatori è elevato (diciamo 40 o 50) si può usare l'analisi delle componenti per ridurre gli indicatori da coinvolgere poi nella procedura di classificazione.

Si pone però un problema: fino a che punto si può esser sicuri che il clustering sulle componenti principali produca la stessa struttura (o una struttura molto simile) di gruppi rispetto a quella ottenibile usando tutti gli indicatori? I risultati sono controversi. Se la percentuale di variabilità spiegata dalle componenti trattenute è molto elevata e se i gruppi sono molto caratterizzati non ci si aspettano differenze. Non è più così se i gruppi non sono ben individuabili. Ad ogni buon conto, e con la sola pretesa di illustrare il problema abbiamo effettuato una prova su di un collettivo di 50 entità e 10 variabili così costituito: 25 osservazioni sono state generate da una normale multivariata con medie nulle e matrice di varianze-covarianze debolmente decomponibile discussa nel paragrafo 3.2.6; altre 25 entità sono state ottenute da un'altra normale multivariata con medie $\mu_i=i/4$, i=1,2, ..., 10 e con matrice varianze-covarianze utilizzata nello stesso paragrafo. I risultati della classificazioni sono esposti nella tabella seguente:

Collocazione	Clustering su tutti gli indicatori (100%)		Clustering sulle prime due C.P. (95%)	
vera	1	2	1	2
1	[24]	1	[24]	1
2	7	[18]	0	[25]
	31	19	24	26

Come si vede, la definizione del secondo gruppo riesce bene con le prime due componenti ed è confusa se si usano tutti gli

indicatori. Nulla di conclusivo, si capisce, ma è un segno che non ci si deve preoccupare troppo di procedere secondo una logica non "alla Everitt".

Clustering delle variabili.

Una procedura potenzialmente molto utile è la classificazione gerarchica (con una distanza ultrametrica) applicata non più alle entità, ma agli indicatori. L'idea è quella di suddividere le variabili in gruppi o clusters al solito omogenei al loro interno ma diversi da quelli esterni. Se tra le variabili esiste una netta suddivisione in gruppi netta questa si rifletterà nell'analisi delle componenti principali che potrà essere effettuata separatamente per ogni cluster di variabili (si veda il paragrafo 3.2.6) per scegliere uno o pochi *leading indicators* per ogni macrodeterminante.

La clustering delle variabili richiede che sia ben definita la misura di distanza o di similarità in base alla quale giudicare la diversità tra indicatori. La misura che viene subito in mente è il coefficiente di correlazione lineare o, comunque, un indice basato su di esso. In alternativa si possono considerare delle distanze normalizzate (il rinvio è al paragrafo 3.3.1).

L'utilizzazione pratica di questa procedura può però andare incontro a delle difficoltà perché ora non si agisce sulla matrice dei dati, ma sulla sua trasposta, in cui cioè le unità territoriali sono gli "indicatori" in base ai quali si raggruppano le "entità" ovvero gli indicatori. Poiché le entità possono essere molto numerose, al punto da superare le capacità di calcolo disponibili, può rendersi necessario ripetere l'operazione diverse volte per campioni casuali di entità oppure applicarla una sola volta, ma per un campione ragionato delle entità scelte secondo criteri opportuni.

Riferimenti bibliografici

Aldenderfer M.S. Blashfield R.K. (1984): *Cluster Analysis*. Sage University Paper, London.

Anania G. M. Bonetti e G. Cannata (1984): *L'agricoltura in un sistema integrato: una proposta metodologica per l'analisi spaziale delle emergenze di marginalità socio-economica a livello comunale*. CNR/IPRA, Quaderni Metodologici n. 2.

Anania G. e F. Gaudio (1988): *La periferia emergente. Analisi spaziale delle caratteristiche dei sistemi socio-economici territoriali in Basilicata e Calabria*. CNR/IPRA.

Beale E.M. (1969): *Euclidean Cluster Analysis*. Bulletin of International Statistical Institute. Vol. 43, pp. 92-94.

Calinski T. Harabasz J. (1974): *A Dendrite Method for Cluster Analysis*. Communications in Statistics. Vol. 3, pp. 1-27.

Cannata G. (1989): *I sistemi territoriali agricoli italiani*, Franco Angeli, Milano.

Chatfield C. Collins A.J. (1980): *Introduction to Multivariate Analyis*. Chapman and Hall, London.

Cormack R.M. (1971): *A Review of Classification*. Journal of the Royal Statistical Society (series A). Vol. 134, pp. 321-367.

Diday E. (1971): *Une nuovelle méthode en classification automatique et reconnaissance des formes: la méthode des nueés dynamiques*. Revue de Statistique Appliquée. vol. 19, n.2, 19-33.

Everitt B.S. (1979): *Unresolved Problems in Cluster Analysis*. Biometrics. Vol. 35, pp. 169-181.

Fisher L. Van Ness J.W. (1971): *Admissible Clustering Procedures*. Biometrika, vol.58, n.1, pp. 91-104.

Friedman H.P. Rubin J. (1967): *On Some Invariant Criterion for Grouping Data*. Journal of the American Statistical Association. Vol. 62, pp. 1159-1178.

Gantmacher F.R. (1974): *The Theory of Matrices. Vol. II*. Chelsea Publishing Company, New York.

Hand D.J. (1981): *Discrimination and Classification*. John Wiley & Sons. New York.

Hartigan J.A. (1975): *Clustering Algorithms*. John Wiley & Sons. New York.

Hartigan J.A. Wong M.A. (1979): *Algorithm AS136. A K-means Clustering Algorithm*. Applied Statistics. Vol. 28, N.1, pp. 100-108.

Jolliffe I.T. (1986): *Principal Component Analysis*. Springer & Verlag. New York.

Lance G.N. Williams W.T. (1967): *A General Theory of Classificatory Sorting Strategies. Part I. Hierarchical systems.* Computer Journal, Vol. 9, 373-380.

Maronna R. Jacovkis M.P. (1974): *Multivariate Clustering Procedures with Variable Metrics.* Biometrics, Vol.30, pp. 499-505.

Marriott F.H.C. (1971): *Practical Problems in a Method of Cluster Analysis.* Biometrics, vol 27, pp. 501-514.

Mineo A. (1986): *Problemi e metodi di classificazione.* Atti della 33a riunione scientifica della SIS. Cacucci editore, Bari.

Morrison D.F. (1967): *Multivariate Statistical Methods.* McGraw Hill, New York.

Narayanaswamy C.R. Raghavarao D. (1991): *Principal Component Analysis of Large Dispersion Matrices.* Applied Statistics, vol. 40, N.2, pp. 309-316.

Rand W.M. (1971): *Objective Criteria for the Evaluation of Clustering Methods.* Journal of the American Statistical Association. Vol. 66, n. 336, pp. 846-850.

Rizzi A. (1985): *Analisi dei dati.* La Nuova Italia Scientifica. Roma.

Seber G.A.F. (1984): *Multivariate Observations.* John Wiley & Sons. New York.

Sibson R. (1973): *SLINK: An Optimally Efficient Algorithm for the Single-link Cluster Method.* Computer Journal. Vol. 16, pp. 30-34.

Spath H. (1985): *Cluster Analysis and Dissection.* Ellis Horwood.Chichester.

Symons M.J. (1981): *Clustering Criteria and Multivariate Normal Mixtures,* Biometrics, vol.37, pp. 35-43.

Vicari D. (1990): *Indici per la scelta del numero dei gruppi.* Metron, 47, 1-4, pp; 473-492

Wilkinson L. (1987). *Systat: The System for Statistics.* SYSTAT, Inc. Evanston, Il.