http://www.mauroennas.eu

Elementi di cluster analysis per la classificazione e il posizionamento nelle ricerche di marketing

Mauro Ennas

Allegati

Simulazioni SPSS

1 **Cluster_Analysis.sav**
OUTPUT_clustering_gerarchico_BAVERAGE.spv
OUTPUT_clustering_gerarchico_CENTROID.spv
OUTPUT_clustering_gerarchico_COMPLETE.spv
OUTPUT_clustering_gerarchico_MEDIAN.spv
OUTPUT_clustering_gerarchico_SINGLE.spv
OUTPUT_clustering_gerarchico_WARD.spv
OUTPUT_clustering_gerarchico_WAVERAGE.spv
OUTPUT_clustering_non_gerarchico_KMEANS.spv

Fogli di calcolo Excel

1 001_PW_PARMA_ENNAS_KM_k-means_due_passi.xls

Indice degli argomenti

Indice delle figure

Indice delle tabelle

Indice syntax e script

Cluster analysis

Ricerche di marketing

Il *marketing* è il ramo dell'economia che si occupa dello studio descrittivo del mercato e dell'analisi dell'interazione del mercato e dei suoi utilizzatori con l'impresa. Marketing significa letteralmente "piazzare sul mercato" e comprende quindi tutte le azioni aziendali riferibili al mercato destinate alla vendita di prodotti, avendo come fine il maggiore profitto.
Si distinguono quattro strategie di approccio al mercato da parte dell'impresa[1]:

1. orientamento alla produzione: caratterizzato da un eccesso di offerta rispetto alla domanda;
2. orientamento al prodotto: ci si concentra più su quest'ultimo che sul consumatore;
3. orientamento alle vendite: si cerca di vendere tutto ciò che si produce senza porre particolare attenzione alle esigenze del consumatore;
4. orientamento al *marketing*: l'approccio più recente, dove si parte dai bisogni del cliente per poi cercare di produrre un bene o un servizio che li soddisfi.

Quando si parla di strategia s'intende l'insieme delle iniziative che consentono all'impresa di primeggiare nel confronto competitivo. Partendo dal presupposto che il successo di una strategia si misura in base al raggiungimento o meno degli obiettivi preposti secondo un'analisi che considera come elementi fondanti:

- l'importanza degli obiettivi,
- la segmentazione del mercato,
- la scelta del vantaggio competitivo da raggiungere,
- l'analisi del posizionamento e l'applicazione del *marketing mix*.

Le ricerche di *marketing* permettono di ottimizzare gli investimenti al fine di massimizzare il ritorno dell'investimento, tramite l'analisi di possibili scenari orientata alle selezione di decisioni strategiche [].
Le fonti del vantaggio competitivo dell'azienda risiedono nel suo patrimonio di risorse e competenze ed in particolare quelle che godono dei seguenti attributi:

- scarsità,
- difendibilità,
- appropriabilità,
- economicità.

Queste caratteristiche convergono verso il consolidamento del vantaggio competitivo a medio e lungo termine.

La gestione della conoscenza aziendale (*company knowledge management*) è uno degli obiettivi cruciali in ambito competitivo. La conoscenza è un patrimonio generato con un grande dispendio di risorse materiali e temporali e, per questo motivo un bene intangibile di importanza strategica per l'azienda. Le ricerche di *marketing* si inseriscono in questo contesto aziendale, come solido armamentario, per la generazione di output strategici di primaria importanza, per la creazione di efficaci decisioni aziendali a partire dai dati provenienti dalle conoscenze consolidate dell'azienda stessa. Queste tecniche permettono di coniugare la

[1] Philip Kotler, Marketing management (2007).

visione soggettiva del *management* con i "dati di fatto" oggettivi, su base temporale storicizzata o su campionamenti sporadici mediante interviste/questionari. L'azienda orientata al mercato, con la sua cultura, le sue risorse e competenze, i suoi sistemi operativi e informativi, con il suo comportamento organizzativo consolidato, si pone come modello in competizione con altri modelli all'interno dell'arena competitiva rappresentata dal mercato di riferimento. Tale arena ha elevate dinamiche basate sul comportamento dei competitori diretti e degli utilizzatori/clienti del mercato stesso. Le modalità di acquisto e l'approccio psicologico al consumo caratterizzano in larga misura le decisioni strategiche di un'azienda al pari del comportamento strategico dei competitori che, nei moderni mercati, seguono comportamenti tipici da precursore o da inseguitore dell'innovazione. Per realizzare analisi di mercato nell'ambito più vasto delle ricerche di *marketing* è necessario avere un sistema informativo ricco di dati provenienti da precedenti attività di *marketing intelligence* o da sistemi di rilevazione aziendale su fonti informative interne, oppure da ricerche *ad hoc* o da fonti istituzionali esterne all'azienda [15].

I dati sottoposti ad analisi sono tra i più vari e possono riguardare aziende o *brand* in competizione, clienti, prodotti e servizi. Per quanto riguarda l'analisi dei dati dei clienti, ad esempio, generalmente si concentrano su alcuni aspetti cruciali quali:

- l'esistenza sul mercato di nuovi clienti o clienti potenziali,
- l'evoluzione dei clienti attuali,
- l'emergere di bisogni o comportamenti nuovi,
- l'entrata di nuovi concorrenti nel mercato o la minaccia di entrata di concorrenti potenziali,
- l'evoluzione dei concorrenti attuali.

Stabilito il *target* dell'analisi, è necessario raccogliere i dati che spesso devono essere ridotti e organizzati, e ciò è in stretta relazione con l'ambito interpretativo che s'intende adottare e con l'ambito analitico e decisionale di *marketing* dell'oggetto della specifica ricerca in atto.
Nelle seguenti pagine ci si concentra su alcune tecniche di classificazione utilizzate nelle ricerche di *marketing* allo scopo di evidenziare le tematiche fondamentali dei vari metodi con un approccio comparativo, ossia orientato alla ricerca di legami e dissonanze tra i metodi che permettano di focalizzare gli strumenti sugli specifici oggetti dell'indagine. In particolare si concentrerà l'attenzione sulla segmentazione del mercato e sul posizionamento dei prodotti.

Segmentazione

Col termine "segmentazione" del mercato s'intende l'attività di identificazione di "gruppi di clienti" cui è indirizzato un determinato prodotto o servizio. Molte aziende di successo decidono di focalizzare la propria attenzione su un determinato segmento di mercato: servire tale area significa soddisfare i bisogni di una particolare e ben determinata categoria di clienti. La segmentazione è dunque "il processo attraverso il quale le imprese suddividono la domanda in un insieme di clienti potenziali, in modo che gli individui che appartengono allo stesso insieme siano caratterizzati da funzioni della domanda il più possibile simili tra loro e, contemporaneamente, il più possibile diverse da quelle degli altri insiemi"[15].
Ciò significa specializzare le proprie strategie di *marketing*. Se, da un lato, è evidente che le politiche dei prezzi (focalizzate sulla riduzione dei costi di produzione) siano in grado da sole di produrre affetti consistenti sulla segmentazione del mercato è altrettanto vero che la complessità dei mercati moderni ha introdotto una notevole variabilità nei benefici richiesti e un'accresciuta disponibilità di informazione con un conseguente indebolimento della fideizzazione, ciò rende necessario l'utilizzo delle altre leve del *marketing* (prodotto, comunicazione, distribuzione e vendita) per ottenere una corretta e completa strategia.

I vantaggi derivanti da una corretta strategia di segmentazione sono i seguenti:

1. definizione del mercato in termini di bisogni del cliente;
2. capacità di percezione del mutamento dei bisogni (dinamiche di mercato);
3. valutazione della maggiore efficacia (punti di forza vs. debolezza) della propria offerta nei confronti dei competitori;

4. razionalizzazione nella definizione del portafoglio dei prodotti;
5. definizione e creazione di barriere all'ingresso di nuovi concorrenti;
6. misura *ex post*, più precisa, degli effetti su vendita e profitti di specifiche azioni di *marketing*.

In una ricerca di segmentazione, le fasi fondamentali sono le seguenti:

1. definizione dei criteri di segmentazione;
2. selezioni delle variabili utili per costruire e descrivere i segmenti;
3. scelta dell'approccio di segmentazione;
4. scelta della metodologia quantitativa più adatta;
5. valutazione dei risultati e scelta dei segmenti su cui concentrare le risorse aziendali.

Per quanto riguarda la scelta dell'approccio, si parla di:

1. segmentazione *a priori* quando le caratteristiche del segmento vengono definite sulla base di informazioni in possesso al *management* e legate all'esperienza operativa o a precedenti analisi;
2. segmentazione *a posteriori* quando le caratteristiche del segmento emergono dall'applicazione di opportune tecniche di analisi quantitativa non note in precedenza.

Queste tecniche *a posteriori* sono distinte a loro volta in segmentazione per omogeneità e per obiettivi.
Dal punto di vista applicativo, vi sono essenzialmente due modalità di segmentazione per omogeneità, la classica (combinazione di analisi fattoriale e *cluster analysis*) e la flessibile (combinazione di *conjoint analysis* e *cluster analysis*). Nella prima modalità si riducono le caratteristiche oggetto dell'analisi e si cerca di creare gruppi omogenei distinti rispetto a due o più delle caratteristiche in esame. La fase finale incrocia i risultati con caratteristiche socio- demografiche/anagrafiche per verificare l'effettiva raggiungibilità del target. La modalità flessibile è più complessa e richiede la valutazione di profili globali dell'offerta e la scomposizione dei giudizi globali in valutazioni di utilità dei livelli e degli attributi considerati: a questo fine si utilizza la *conjoint analysis*. Applicando la *cluster analysis* ai risultati della *conjoint analysis* si possono ottenere gruppi omogenei in relazione alle valutazioni di utilità implicitamente fornite. Anche in questo caso i risultati della *cluster analysis* vengono incrociati con dati socio-demografici/anagrafici. La flessibilità sta nella possibilità di analizzare profili innovativi dell'offerta (che non sono ancora presenti nel mercato) simulando variazioni di quote di preferenza in relazione alla modifica di una o più leve di *marketing*.
Per quanto riguarda la segmentazione per obiettivi, si suddivide il target in sub-popolazioni utilizzando una variabile "dipendente" conosciuta *a priori* (per esempio la redditività, la frequenza d'acquisto ….) e si valutano una serie di variabili "esplicative", per esempio le caratteristiche socio-demografiche, che influenzeranno in modo rilevante la variabile dipendente considerato. Tra le tecniche note ricordiamo, oltre all'analisi discriminante lineare (ADL) che verrà considerato nel prossimo capitolo, la metodologia CHAID [5], l'analisi logistica e le reti neurali [].

Posizionamento

Il posizionamento di un prodotto può essere visto come una decisione strettamente connesso a quella della selezione dei segmenti di mercato in cui l'impresa decide di competere e consiste nella misura della percezione che hanno i clienti di un prodotto o di una merce, relativamente alla posizione dei prodotti o delle marche concorrenti. Decidere di posizionare un prodotto in un certo segmento consiste nell'identificare le dimensioni su cui costruire tale percezione all'interno dei segmenti di un mercato di riferimento. Il posizionamento è legato strettamente alla comprensione delle motivazioni e delle aspettative dei clienti che costituiscono i segmenti ed è pertanto difficile assumere tale decisione in assenza di una chiara identificazione dei segmenti che compongono il mercato. Il posizionamento non è assoluto ma relativo alle posizioni dei concorrenti, di conseguenza, il posizionamento di un prodotto è strettamente legato oltre che alla percezione dei clienti anche alle decisioni strategiche dei concorrenti, che tendono a modificare il proprio comportamento in funzione dei *feed-back* del mercato.

L'impresa, nelle decisioni strategiche di posizionamento, analizza l'attuale posizionamento del proprio prodotto, linea o marca, nel mercato di riferimento e decide di consolidare la propria posizione oppure decide di riposizionarsi in base alla misura di percezione da parte dei clienti e delle aziende concorrenti. Costruire un

posizionamento di un prodotto significa sostanzialmente identificare delle dimensioni sulle quali differenziare il proprio prodotto da quello dei concorrenti. Le dimensioni tipiche della differenziazione sono:

- gli attributi del prodotto,
- i benefici ricercati dai clienti,
- le occasioni e le modalità di utilizzo,
- il posizionamento dei concorrenti.

La dimensione della differenziazione possono essere tangibili come le caratteristiche tecniche, il prezzo e la disponibilità di servizi di supporto o intangibili come la percezione del *brand*, il suo prestigio, in modo del tutto equivalente si possono costruire posizionamenti in base a specifici benefici richiesti da uno o più segmenti di clienti, o ancora da occasioni d'acquisto o utilizzazione. Il posizionamento può essere utilizzato per avvicinare la propria azienda ad aziende *leader* se si pensa di volere che il proprio prodotto possa essere percepito come simile ad un prodotto *leader* di mercato. Poiché il risultato che si vuole ottenere è una rappresentazione relative di come un certo prodotto viene percepito se confrontato con i prodotti dei concorrenti, l'output dell'analisi è una mappa delle percezioni e le procedure che la realizzano sono dette *perceptual mapping* [9]. In queste brevi note approfondiremo le procedure di analisi discriminante lineare (ADL) e il multidimensional scaling (MDS).

Clustering

La *Cluster Analysis* (CLA) è l'insieme delle procedure [14] e delle metodologie utilizzate per ricavare, a partire da una popolazione di dati, una struttura di classificazione a gruppi. Dalla sua comparsa ad opera di *R. C. Tyron* nel 1939 [25] viene sperimentata, con estese applicazioni, a partire dagli anni '60. La prima esposizione sistematica risale al 1963 ad opera di *Sokal* e *Sneath* (*Principles of numerical taxonomy*) []. Le applicazioni delle tecniche numeriche associate alla *cluster analysis* sono molteplici e in campi multidisciplinari, dall'informatica, alla medicina e biologia, dall'archeologia al *marketing*: ogni qualvolta sia necessario classificare una grande mole di informazioni in gruppi distinguibili risulta uno strumento efficace e indispensabile; nelle ricerche di *marketing* riveste grande importanza per caratterizzare gli elementi fondamentali dei processi decisionali delle strategie commerciali: caratteristiche, bisogni e comportamenti degli acquirenti. Segmentando la clientela di un mercato si riesce a focalizzare le energie e gli investimenti su caratteristiche precise, riconoscibili e distinguibili. Posizionando aziende e marchi si riescono ad individuare le zone di mercato non ancora occupate rispetto ad alcune variabili caratteristiche del prodotto/servizio offerto sul mercato. Queste tecniche permettono di individuare all'interno di un insieme di oggetti, di qualsivoglia natura, sottoinsiemi, ossia *cluster*, che hanno la forte tendenza ad essere omogenei "in qualche senso" all'interno del gruppo di appartenenza. I criteri di similitudine sono stabiliti *a priori*. Il risultato dell'analisi dovrebbe, in linea con i criteri scelti, evidenziare una elevata omogeneità interna al gruppo (*intra-cluster*) ed un'alta eterogeneità tra gruppi distinti (*inter-cluster*).

Il punto di partenza di ogni analisi dei gruppi è la disponibilità di n dati di p variabili ciascuno. Tali dati sono rappresentati in forma di matrice $n \times p$. Le tecniche di CLA [14][26] sono tecniche di tipo esplorativo che non richiedono assunzioni *a priori* sui dati ma delle azioni e delle decisioni sia prima, durante e dopo l'analisi. In particolare è importante il criterio di scelta delle variabili, dei criteri di similarità (distanza), la scelta delle tecniche di aggregazione e di selezione del numero di gruppi da ottenere, nonché la valutazione della soluzione ottenuta e la scelta tra eventuali soluzioni alternative, tenendo presente che scelte diverse rappresentano risultati distinti e in qualche modo arbitrari (dipendenti fortemente dal criterio utilizzato per la selezione dei dati). Il fattore soggettivo accomuna tutti i procedimenti di analisi multivariata, caratteristica dei procedimenti di riduzione e semplificazione controllata delle informazioni.

Le fasi del processo generico di analisi dei *cluster* può essere sintetizzato nei seguenti passi:

1. scelta delle unità di osservazione,
2. scelta delle variabili e omogeneizzazione della scala di misura,
3. scelta della metrica di similarità o dissimilarità tra i dati,

4. scelta del numero di gruppi caratteristici,
5. scelta dell'algoritmo di classificazione (gerarchico, non-gerarchico),
6. interpretazione dei risultati ottenuti.

Similarità e dissimilarità

Per realizzare la classificazione è necessario individuare dei criteri di similarità e dissimilarità tra coppie; ciò si ottiene definendo degli indici che danno delle indicazioni preliminari, indispensabili per individuare le unità omogenee che caratterizzeranno i gruppi. Un indice di prossimità tra due generiche unità statistiche (dato di p componenti caratteristiche) è una funzione di due vettori riga x_i e x_j della stessa matrice dati.

$$IP_{ij} = f(x_i, x_j), \text{ i,j = 1, 2,... n.}$$

Due oggetti sono simili quando la loro dissimilarità è piccola ovvero quando la similarità è grande.
Se utilizziamo dati qualitativi gli indici di prossimità sono generalmente indici di similarità, se invece i dati sono quantitativi si utilizzeranno sia indici di similarità che di dissimilarità.
Esistono anche indici di prossimità che vengono utilizzati nel caso in cui le variabili siano miste, in parte qualitative e in parte quantitative (caso generale).

Tra le misure possibili di prossimità distinguiamo:

1. misure di prossimità per variabili categoriche (discrete)
2. misure di prossimità per variabili continue
3. misure di prossimità su variabili sia continue che discrete (insiemi misti)

Una tipologia molto comune è quella di dati con variabili binarie. In questo caso si possono realizzare tabelle di contingenza che evidenziano gli esiti nel caso di due soli oggetti esaminati, indicando quanti oggetti dell'analisi presentano lo stesso valore o valori diversi tra quelli della popolazione osservata.
La presenza contemporanea di una stessa caratteristica in due oggetti osservati posti a confronto o la co-assenza può essere associata ad un contenuto informativo tramite un indice o coefficiente di similarità.
Nel caso binario di confronto a coppie[2] avremo una tabella di contingenza del tipo:

		Oggetto i		
Oggetto j	**Esito**	**1**	**0**	**Totale**
	1	a	b	a+b
	0	c	d	c+d
	Totale	a+c	b+d	p=a+b+c+d

Tabella 1 - Tabelle delle contingenze nel caso del confronto a coppie con variabili binarie.

La scelta dei coefficienti di similarità è molto importante in quanto condiziona il risultato finale. Tra le possibili scelte vi sono quelle della Tabella 2 [].

	Misura	**Coefficiente**
S1	Coefficiente di corrispondente di Sokal e Michener	$S_{ij}^{SM} = \frac{a+d}{a+b+c+d}$
S2	Coefficiente di Jaccard	$S_{ij}^{J} = \frac{a}{a+b+c}$
S3	Coefficiente di Rogers e Tanimoto	$S_{ij}^{RT} = \frac{a+d}{a+2\cdot(b+c)+d}$

[2] I coefficienti a, b, c, d rappresentano le occorrenze delle configurazioni 11, 10, 01 e 00 nel caso binario, interpretabile come presenza o assenza di un determinato fattore/caratteristica.

S4	Coefficiente di Sokal e Sneath	$S_{ij}^{SS} = \frac{a}{a + 2 \cdot (b + c)}$
S5	Coefficiente di Gower e Legendre	$S_{ij}^{GL} = \frac{a + d}{a + \frac{1}{2}(b + c) + d}$
S6	Coefficiente di Dice	$S_{ij}^{D} = \frac{2 \cdot a}{2 \cdot a + b + c}$
S7	Coefficiente di Russel e Rao	$S_{ij}^{RR} = \frac{a}{a + b + c + d}$

Tabella 2 – Coefficienti di similarità.

Dati categorici con più livelli possono essere trattati allo stesso modo delle variabili binarie componendo ogni livello in una singola variabile binaria (presenza o assenza della caratteristica in esame). Un tale modo e sconveniente perché genera un numero elevato di corrispondenze negative. Un altro metodo è quello di costruire un parallelepipedo s_{ijk} con $k \in [1, p]$ (con p dimensione della variabile corrente), ponendo $s_{ijk} = 1$ se x_i e x_j hanno la stessa variabile k, in caso contrario si pone $s_{ijk} = 0$. Il valore di similarità si calcola semplicemente facendo la media su tutte le p variabili:

$$s_{ij} = \frac{1}{p} \cdot \sum_{k=1}^{p} s_{ijk}$$

Se le variabili considerate sono continue, allora la prossimità tra individui viene calcolata utilizzando misure di dissimilarità (distanze). La similarità s_{ij} (indice di similarità) può convertirsi nella dissimilarità calcolandone il complemento a uno, $d_{ij} = 1 - s_{ij}$ (indice di dissimilarità).

Distanze

La distanza tra due oggetti di dimensione p è una funzione d_{ij} con $i, j \in \Re^p$ che gode delle seguenti proprietà:

1. $d_{ij} \geq 0$, $\forall x, y \in \Re^p$ (non negatività)
2. $d_{ii} = 0 \Leftrightarrow x = y$ (identità)
3. $d_{ij} = d_{ji}$, $\forall x, y \in \Re^p$ (simmetria)
4. $d_{ij} \leq d_{ik} + d_{kj}$, $\forall x, y, z \in \Re^p$ (disuguaglianza triangolare)

Per costruire la matrice delle distanze si considerano i vettori riga della matrice dati e si calcola la distanza tra i due elementi per ogni coppia in modo da costruire una matrice di dissimilarità $\hat{D}_{n \times n} = \{d_{ij}\}$ con $d_{ij} = 0$ $\forall i = j$. Tale distanza è detta metrica se è valida per tutte le triplette $(i, j, l) \in \Re^p$ per tutte le coppie di oggetti $(i, j), (j, l), (l, i)$. Dalla disuguaglianza triangolare segue che la matrice $\hat{D}_{n \times n} = \{d_{ij}\}$ è simmetrica ossia $d_{ij} = d_{ji}$ $\forall i \neq j$. La generica matrice delle distanze è una matrice simmetrica con diagonale composta da elementi nulli (e quindi traccia uguale a zero).

	Misura	Distanza
D1	Distanza Euclidea	$d_{ij}^{E} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$
D2	Distanza Manhattan (Rettilinea o City Block)	$d_{ij}^{CB} = \sum_{k=1}^{p} \left\lvert x_{ik} - x_{jk} \right\rvert$
D3	Distanza Minkowski	$d_{ij}^{M} = \sqrt[r]{\sum_{k=1}^{p} \left\lvert x_{ik} - x_{jk} \right\rvert^{r}}\ ,\ r \geq 1$
D4	Distanza Camberra	$d_{ij}^{C} = \begin{cases} 0, & x_{ij} = x_{jk} \\ \sum_{k=1}^{p} \dfrac{\left\lvert x_{ik} - x_{jk} \right\rvert}{\left\lvert x_{ik} \right\rvert + \left\lvert x_{jk} \right\rvert}, & altrimenti \end{cases}$

Tabella 3 – Misure di distanza.

Un *dataset* di osservazioni multivariate può essere trasformato in una matrice di dissimilarità utilizzando una delle misure indicate nella Tabella 3; tali distanze possono anche essere pesate in modo non uniforme, con pesi w_p tali che ad esempio:

$$d_{ij}^{E} = \sqrt{\sum_{k=1}^{p} w_k \cdot \left(x_{ik} - x_{jk}\right)^2} \ .$$

Nel caso specifico di distanza Euclidea, la distanza d_{ij}^{E} corrisponde alla distanza tra due punti $\vec{x}_i = \left(x_{i1},...,x_{ip}\right)$ e $\vec{x}_j = \left(x_{j1},...,x_{jp}\right)$ in uno spazio a p dimensioni.
La distanza D3 generalizza la D1 e la D2. Per dati di tipo misto (contenenti variabili continue e categoriche) esistono diverse misure di similarità, qui consideriamo la sola misura di *Gower* (1971):

$$s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

con s_{ij} la similarità tra oggetti i e j, misurata sulla k-esima variabile, con w_{ijk} il speso corrispondente.

Definiremo i pesi nel modo seguente:

$$w_{ijk} = \begin{cases} 0, & \text{Se la k-esima variabile è mancante per uno o entrambi gli oggetti} \\ 0, & \text{Se, nel caso binario, si vogliono escludere corrispondenze negative} \\ 1, & \text{Altrimenti} \end{cases}$$

Il valore di s_{ijk} è valutato in modo diverso a seconda della natura delle variabili. Per variabili binarie o categoriche con più di due possibili valori $s_{ijk} = 1$ se i due oggetti hanno lo stesso valore della variabile k, in caso contrario $s_{ijk} = 0$; per variabili continue:

$$s_{ijk} = 1 - \frac{\left\lvert x_{ik} - x_{jk} \right\rvert}{R_k}$$

con R_k indichiamo il *range* della k-esima variabile (in pratica si utilizza la distanza *Manhattan* dopo avere scalato la k-esima variabile all'unità). Date n osservazioni di p componenti ciascuna si costruisce la matrice delle distanze:

$$\hat{D} = \begin{pmatrix} 0 & \dots & d_{1j} & \dots & \dots & d_{1n} \\ \dots & 0 & \dots & \dots & \dots & \dots \\ d_{i1} & \dots & 0 & \dots & \dots & d_{in} \\ \dots & \dots & \dots & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 & \dots \\ d_{n1} & \dots & d_{nj} & \dots & \dots & 0 \end{pmatrix},$$

calcolando le distanze con una delle espressioni di Tabella 3. Gli elementi della diagonale sono evidentemente nulli in quanto rappresentano la distanza di un punto da se stesso. Quando la distanza è Euclidea la matrice risulta simmetrica.

Clustering gerarchico

Quando si procede per partizioni successive a partire da un solo *cluster* iniziale contenente tutti i dati osservati oppure da un insieme di *cluster* pari al numero degli elementi osservati, uno per *cluster*, allora si parla di *cluster*ing gerarchico [31]. La classificazione procede dalla situazione iniziale per passi successivi e se non bloccata opportunamente, porta ad n *cluster* (uno per ogni punto) nel caso di inizio da un unico *cluster* contenente tutti i punti rappresentativi dei dati osservati e viceversa, ad un unico *cluster* nel caso di partenza delle iterazioni da n *cluster*. Gli algoritmi di *clustering gerarchico* procedono per partizioni (strategia *top-down*) nel primo caso e per fusioni (strategia *bottom-up*) nel secondo caso. Risulta fondamentale sapere quanti dovranno essere i *cluster* finali attesi per avere un criterio di "stop" dell'algoritmo.

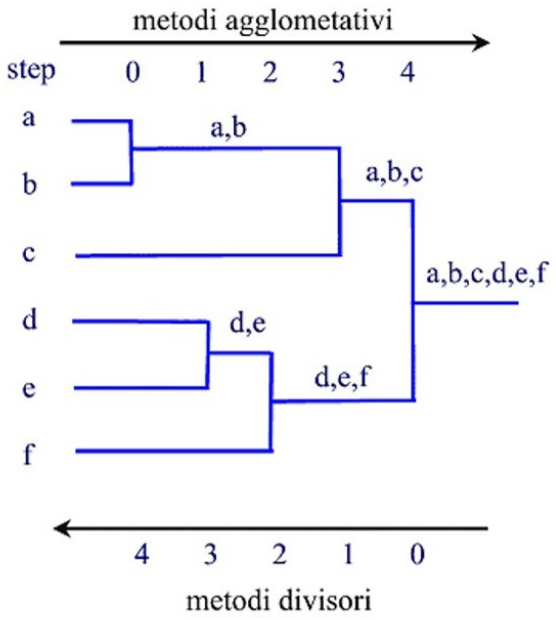

Figura 1 - Dendrogramma: clustering gerarchico.

I metodi per partizione (scissori o divisori) risultano più efficienti dal punto di vista computazionale, anche se hanno lo svantaggio di non potere permettere la correzione di eventuali errori di classificazione dei passi precedenti. La struttura ottenuta è la caratteristica struttura a dendrogramma (Figura 1) degli alberi di decisione.

La procedura operativa può essere schematizzata nei seguenti passi:

1. Inizializzazione: date n unità statistiche o osservazioni, ogni elemento rappresenta un gruppo di un elemento (si hanno n *cluster* iniziali) e vengono numerati da 1 a n.
2. Selezione: vengono calcolate le distanze e selezionati i *cluster* più vicini rispetto ad una misura di prossimità fissata (Tabella 3).
3. Aggiornamento: si aggiorna il numero di *cluster* (n-1) attraverso l'unione di due *cluster* a minima distanza tra loro; in corrispondenza si aggiorna la matrice delle distanze, sostituendo alle due righe che riferiscono la minima distanza, una colonna con le distanze aggiornate rispetto ai nuovi *cluster*, per tenere conto del nuovo gruppo.
4. Ripetizione: si eseguono nuovamente i passi (2) e (3) per n-1 volte.
5. Arresto: la procedura viene fermata quando tutti gli elementi vengono incorporati in un unico *cluster*.

In base a come vengono calcolate le distanze e a quali sono i dati di input si distinguono diversi metodi gerarchici di *cluster*ing.

Metodo del legame singolo (single linkage)

Questa tecnica è anche nota come "tecnica del confinante più vicino" (*nearest-neighbour technique*). Alla base di questo metodo c'è la definizione della similarità o distanza tra i *cluster*. Il grado di vicinanza tra due gruppi

viene stabilito prendendo in considerazione solo le informazioni relative ai due oggetti più vicini, ignorando quelle degli altri oggetti. La distanza tra due gruppi, diciamo A e B, è definita come la distanza minore rilevata tra la coppia di oggetti (i,j) con i∈A e j∈B, in altri termini si considera il minimo delle $n_A \times n_B$ distanze tra ciascuna delle unità A e ciascuna delle unità del gruppo B:

$$d_{AB} = \min_{i=A,\, j=B} d_{ij} .$$

Tale tecnica gode della particolare proprietà, delle sue partizioni, di essere invarianti rispetto a trasformazioni monotone delle variabili (Jardine, Sibson).

Metodo del legame completo (complete linkage)

Il metodo è anche noto come metodo "del confinante più lontano" (*farthest neighbour technique*) è l'opposto della tecnica del legame singolo, si considerano le similarità/distanze fra i gruppi più lontani (meno simili) come significative per la classificazione. La distanza tra due gruppi, diciamo A e B, è definita come la distanza maggiore rilevata tra la coppia di oggetti (i,j) con i ∈A e j∈B, in altri termini si considera il massimo delle $n_A \times n_B$ distanze tra ciascuna delle unità A e ciascuna delle unità del gruppo B:

$$d_{AB} = \max_{i=A,\, j=B} d_{ij} .$$

Questa tecnica tende ad identificare meglio gruppi relativamente compatti, composti da oggetti fortemente omogenei rispetto alle variabili impiegate.

Metodo del legame medio (average linkage)

Per determinare la distanza tra due gruppi A e B utilizzando questa tecnica, si prendono in considerazione tutte le distanze fra gli n_A oggetti membri del primo, rispetto tutti gli oggetti n_B del secondo. Con questa tecnica, la distanza fra due gruppi ci calcola in base alla media aritmetica fra le distanze (Sokal e Michener, 1958; McQuitty, 1964):

$$d_{AB} = \frac{1}{n_A \cdot n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij} .$$

I metodi seguenti richiedono la matrice dei dati, dalla quale si ricavano le distanze.

Metodo del centroide

La tecnica del centroidi fa riferimento ad una rappresentazione spaziale degli oggetti da classificare, infatti definisce per ogni gruppo un centroidi che ha per coordinate la media aritmetica di tutti gli oggetti appartenenti al gruppo. La distanza tra i gruppi coincide con la distanza dai rispettivi centroidi. La distanza tra i due gruppi A e B di numerosità rispettivamente n_A e n_B, è definita come la distanza tra i rispettivi centroidi (medie aritmetiche), $\bar{x}_A$ e $\bar{x}_B$:

$$d_{AB} = d(\bar{x}_A, \bar{x}_B) .$$

Realizzando la fusione tra i due gruppi A e B, il centroidi del nuovo gruppo sarà del tipo:

$$\bar{x}_{AB} = \frac{n_A \cdot \bar{x}_A + n_B \cdot \bar{x}_B}{n_A + n_B} .$$

Il metodo del centroidi e quello del legame medio presentano delle analogie: il metodo del legame medio considera la media delle distanze tra le unità di ciascuno dei suoi gruppi, mentre il metodo del centroidi calcola le medie di ciascun gruppo, e in seguito misura le distanze tra esse.

Metodo del Ward

Questa tecnica si propone di realizzare una classificazione gerarchica [31] tramite la minimizzazione della varianza delle variabili entro ciascun gruppo. La tecnica è iterativa e ad ogni passo vengono fusi i gruppi che presentano la minima variazione della varianza entro i gruppi (Ward, 1963). Questo metodo permette di generare dei gruppi composti da un numero di elementi comparabile. Il metodo è quello della minimizzazione di una funzione obiettivo che vuole realizzare la massima coesione interna a ciascun gruppo e la massima

separazione esterna tra gruppi diversi. La devianza totale delle p variabili viene scomposta in devianza nei gruppi e devianza fra i gruppi, e ad ogni passo della procedura gerarchica si aggregano tra loro i gruppi che comportano il minore incremento della devianza nei gruppi e il maggiore incremento della devianza tra gruppi in modo da ottenere la maggiore coesione interna possibile e la maggiore separazione esterna tra gruppi.

Tutte le tecniche gerarchiche esaminate finora dette tecniche gerarchico-agglomerative, possono essere viste come varianti di un unica tecnica generale (Lance e Williams, 1967) [13] che può essere espressa in forma compatta e ricorsiva nei termini seguenti:

1. si parte da una situazione con n *cluster* di un oggetto ciascuno;
2. si uniscono i due gruppi i e j che minimizzano la misura di dissimilarità d_{ij};
3. si ripete il basso (2) finché tutti gli oggetti non formano un solo gruppo.

	$\alpha(i)$	$\alpha(j)$	β	Γ
Legame singolo	$1/2$	$1/2$	0	$-1/2$
Legame completo	$1/2$	$1/2$	0	$1/2$
Legame medio	$n_i/(n_i+n_j)$	$n_j/(n_i+n_j)$	0	0
Ward	$(n_i+n_k)/(n_k+n_i+n_j)$	$(n_j+n_k)/(n_k+n_i+n_j)$	$-n_k/(n_k+n_i+n_j)$	0

Tabella 4 - Parametri per il calcolo delle misure di dissimilarità.

La misura di dissimilarità fra gruppi può essere calcolata ricorsivamente. All'inizio, al passo (1) le dissimilarità degli n gruppi coincidono con le dissimilarità tra gli n oggetti. Nei passi successivi la misura della dissimilarità fra il gruppo k-esimo e il gruppo ottenuto dalla fusione dei gruppi i-esimo e j-esimo (diciamo U_{ij}) si calcola sulla base della seguente espressione:

$$d_{k,ij} = \alpha(i)\cdot d_{ki} + \alpha(j)\cdot d_{kj} + \beta\cdot d_{ij} + \Gamma\cdot\left|d_{ki} - d_{kj}\right|,$$

nella quale i parametri $\alpha(i)$, $\alpha(j)$, β e Γ si possono determinare dipendentemente dalla tecnica adottata come illustrato in Tabella 4.

Esempio numerico: single linkage clustering

L'idea è quella di raggruppare i punti "primi vicini", ossia quelli più prossimi tra loro. Si procede in tre passi.

Passo 1

Sia data la matrice delle distanze euclidee (ricavata da dati standardizzati) in Tabella 5. La situazione iniziale consiste di n=10 *cluster* ognuno di un elemento. Vogliamo aggregare i *cluster* in modo tale da ottenere, al passo 2, n-1 *cluster* complessivi.

Passo 2

Se indichiamo con l ed m gli indici della riga e della colonna contenente il minimo, d(l,m) sarà la distanza minima: nel caso in esame d(5,6)=0.33. Poiché la matrice è simmetrica, la colonna m è uguale alla riga l: si aggregano le due righe l ed m, in pratica si eliminano le righe e le colonne l ed m e si sostituiscono con la colonna costituita dai valori minimi tra le due.

	P1	P2	P3	P14	P15	P16	P17	P18	P23	P24
Z1	1,48	1,16	0,11	-1,76	-1,08	-0,91	-0,74	-0,22	0,71	1,05
Z2	0,27	0,12	1,49	0,70	0,34	0,63	-0,38	0,27	-2,32	-1,97
	0,00	0,36	1,84	3,27	2,56	2,42	2,31	1,71	2,70	2,27
	0,36	0,00	1,72	2,97	2,24	2,12	1,96	1,39	2,48	2,09
	1,84	1,72	0,00	2,03	1,65	1,33	2,05	1,27	3,86	3,58
	3,27	2,97	2,03	0,00	0,77	0,86	1,49	1,60	3,90	3,87
	2,56	2,24	1,65	0,77	0,00	0,33	0,80	0,86	3,21	3,14
	2,42	2,12	1,33	0,86	**0,33**	0,00	1,02	0,77	3,36	3,25
	2,31	1,96	2,05	1,49	0,80	1,02	0,00	0,83	2,42	2,39
	1,71	1,39	1,27	1,60	0,86	0,77	0,83	0,00	2,75	2,57
	2,70	2,48	3,86	3,90	3,21	3,36	2,42	2,75	0,00	0,50
	2,27	2,09	3,58	3,87	3,14	3,25	2,39	2,57	0,50	0,00

Tabella 5 – Passo 1: distanze provenienti da dati standardizzati.

In pratica, si calcola:

$$d_{(l,m),k} = \min(d_{lk}, d_{mk}),\ \forall k \in [1, n-1].$$

Nel caso in esame avremo:

$$\begin{cases} d_{(16,15),1} = \min(d_{16,1}, d_{15,1}) = \min(2.42,2.56) = 2.42 \\ d_{(16,15),2} = \min(d_{16,2}, d_{15,2}) = \min(2.12,2.24) = 2.12 \\ d_{(16,15),3} = \min(d_{16,3}, d_{15,3}) = \min(1.33,1.65) = 1.33 \\ d_{(16,15),4} = \min(d_{16,4}, d_{15,4}) = \min(0.86,0.77) = 0.77 \\ \ldots \\ d_{(16,15),} = \min(d_{16,24}, d_{15,24}) = \min(3.25,3.14) = 3.14 \end{cases}$$

La matrice che si ottiene è quella in Tabella 6, nella quale la colonna (e la riga) indicate con (16,15) indicano la riga ottenuta dall'aggregazione delle righe l=16 ed m=15, secondo il minimo, elemento per elemento.

	(16,15)	P1	P2	P3	P14	P17	P18	P23	P24
(16,15)	**0,00**								
P1	**2,42**	0,00							
P2	**2,12**	**0,36**	0,00						
P3	**1,33**	1,84	1,72	0,00					
P14	**0,77**	3,27	2,97	2,03	0,00				
P17	**0,80**	2,31	1,96	2,05	1,49	0,00			
P18	**0,77**	1,71	1,39	1,27	1,60	0,83	0,00		
P23	**3,21**	2,70	2,48	3,86	3,90	2,42	2,75	0,00	
P24	**3,14**	2,27	2,09	3,58	3,87	2,39	2,57	**0,50**	0,00

Tabella 6 – Passo 2: matrice derivata aggregando le righe contenenti il minimo assoluto.

Passo 3

Utilizzando la nuova matrice si ripete il passo 2 e si individua il minimo assoluto

$$d_{(l,m),k} = \min(d_{lk}, d_{mk}) = 0.36,$$

ripetendo l'aggregazione e la sostituzione come nel caso precedente.

	(2,1)	(16,15)	P3	P14	P17	P18	P23	P24
(2,1)	0,00							
(16,15)	**0,36**	0,00						
P3	**1,72**	1,33	0,00					
P14	**2,97**	0,77	2,03	0,00				
P17	**1,96**	0,80	2,05	1,49	0,00			
P18	**1,39**	0,77	1,27	1,60	0,83	0,00		
P23	**2,48**	3,21	3,86	3,90	2,42	2,75	0,00	
P24	**2,09**	3,14	3,58	3,87	2,39	2,57	0,50	0,00

Tabella 7 - Passo 3: matrice derivata aggregando le righe contenenti il minimo assoluto della matrice ottenuta al passo precedente.

Ci si ferma quando tutti gli elementi appartengono ad un unico *cluster*, oppure fissato un numero di *cluster* desiderabili , quando si raggiunge tale numero.

Clustering non-gerarchico

Nei metodi non gerarchici l'inizializzazione è definita a partire da centroidi scelti a caso tra i punti del *dataset* o all'esterno del *dataset*, fissato *a priori* il numero di *cluster* che si desidera popolare. Tra questi metodi quello più popolare è il cosiddetto metodo delle k-medie (*k-means method*), introdotto da MacQueen nel 1963 [14].

Si parte da un *dataset* di n osservazioni e si fissa il numero k di gruppi si raggruppano i dati in modo da avere massima omogeneità all'interno di ogni gruppo rispetto ad una metrica prefissata che rende distinguibile ogni gruppo da un altro. L'algoritmo può essere schematizzato in cinque passi:

1. si sceglie il numero k di *cluster* da formare;
2. si scelgono in modo casuale (spesso utilizzando generatori di numeri casuali) k valori del *dataset*, essi saranno i cosiddetti centroidi dei k *cluster*;
3. si utilizza la distanza euclidea per assegnare i restanti dati ai *cluster* attraverso la distanza dai centroidi rappresentativi dei *cluster*;
4. si utilizzano i dati così aggregati per calcolare, attraverso le medie delle coordinate dei punti appartenenti ad ogni *cluster*, le coordinate dei nuovi centroidi;
5. se le nuove medie sono uguali a quelle calcolate in precedenza il processo termina, in caso contrario si utilizzano tali medie come centroidi e si ripetono i passi dal (3) al (5) per determinare le distanze dai nuovi centroidi.

Esempio numerico: metodo K-means

Si consideri l'insieme di N=26 punti indicato in E si scelgano k=3 punti appartenenti al *dataset* (siano P5, P13 e P25). Scelta una metrica euclidea, si calcolano tutte le distanze dei punti del *dataset* dai tre punti scelti, come mostrato in Tabella 8, nelle colonne denominate rispettivamente D(P5), D(P13) e D(P25).

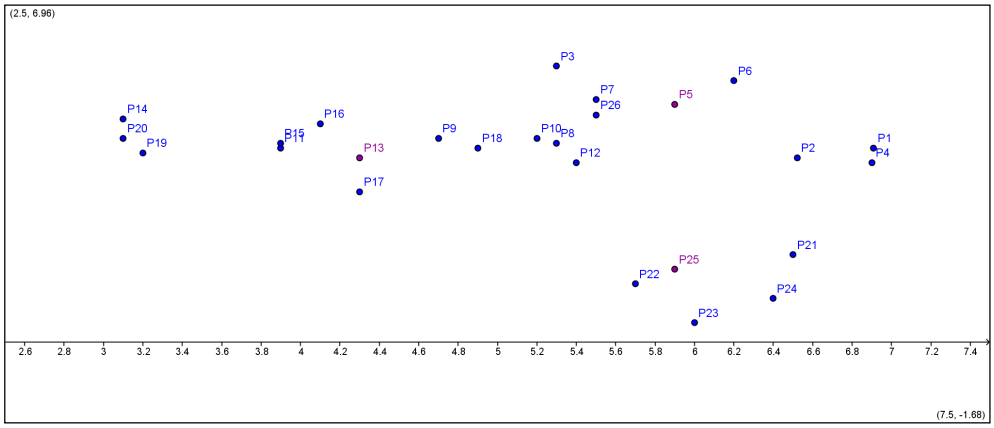

Figura 2 - Clustering K-means (passo 1: scelta dei centroidi di inizializzazione).

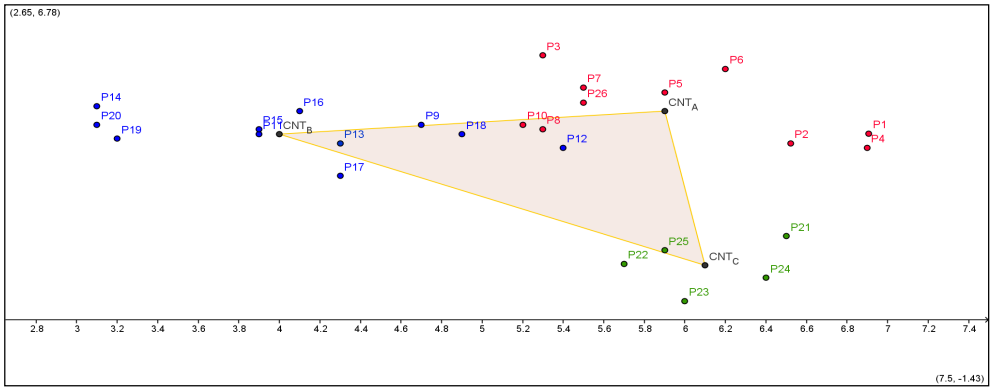

Figura 3 - Clustering K-means (passo 2: calcolo dei nuovi centroidi).

	Passo 1							Passo 2										
	X1	X2	D(P5)	D(P13)	D(P25)	MIN	Cluster	A		B		C		D(A)	D(B)	D(C)	MIN	Cluster
P1	6,91	4,00	1,35	2,61	2,70	1,35	A(P5)	6,91	4,00	0,00	0,00	0,00	0,00	**1,12**	2,82	2,95	1,12	A(CNT_A)
P2	6,52	3,80	1,26	2,22	2,38	1,26	A(P5)	6,52	3,80	0,00	0,00	0,00	0,00	**0,96**	2,45	2,67	0,96	A(CNT_A)
P3	5,30	5,70	1,00	2,14	4,24	1,00	A(P5)	5,30	5,70	0,00	0,00	0,00	0,00	**1,31**	2,08	4,60	1,31	A(CNT_A)
P4	6,90	3,70	1,56	2,60	2,42	1,56	A(P5)	6,90	3,70	0,00	0,00	0,00	0,00	**1,29**	2,83	2,65	1,29	A(CNT_A)
P5	**5,90**	**4,90**	0,00	1,94	3,40	0,00	A(P5)	5,90	4,90	0,00	0,00	0,00	0,00	**0,35**	2,02	3,74	0,35	A(CNT_A)
P6	6,21	5,41	0,59	2,49	3,92	0,59	A(P5)	6,21	5,41	0,00	0,00	0,00	0,00	**0,90**	2,54	4,24	0,90	A(CNT_A)
P7	5,50	5,01	0,42	1,70	3,53	0,42	A(P5)	5,50	5,01	0,00	0,00	0,00	0,00	**0,62**	1,73	3,89	0,62	A(CNT_A)
P8	5,31	4,10	1,00	1,04	2,67	1,00	A(P5)	5,31	4,10	0,00	0,00	0,00	0,00	**0,76**	1,22	3,04	0,76	A(CNT_A)
P9	4,71	4,20	1,39	0,57	2,96	0,57	B(P13)	0,00	0,00	4,71	4,20	0,00	0,00	7,47	0,65	3,34	0,65	B(CNT_B)
P10	5,21	4,20	0,99	0,99	2,79	0,99	A(P5)	5,21	4,20	0,00	0,00	0,00	0,00	**0,80**	1,14	3,16	0,80	A(CNT_A)
P11	3,90	4,00	2,19	0,45	3,20	0,45	B(P13)	0,00	0,00	3,90	4,00	0,00	0,00	7,47	**0,18**	3,59	0,18	B(CNT_B)
P12	5,41	3,70	1,30	1,11	2,26	1,11	B(P13)	0,00	0,00	5,41	3,70	0,00	0,00	7,47	**1,35**	2,63	1,35	B(CNT_B)
P13	**4,30**	**3,80**	1,94	0,00	2,80	0,00	B(P13)	0,00	0,00	4,30	3,80	0,00	0,00	7,47	**0,30**	3,19	0,30	B(CNT_B)
P14	3,10	4,61	2,82	1,44	4,18	1,44	B(P13)	0,00	0,00	3,10	4,61	0,00	0,00	7,47	**1,15**	4,56	1,15	B(CNT_B)
P15	3,90	4,10	2,15	0,50	3,28	0,50	B(P13)	0,00	0,00	3,90	4,10	0,00	0,00	7,47	**0,20**	3,67	0,20	B(CNT_B)
P16	4,10	4,51	1,84	0,73	3,50	0,73	B(P13)	0,00	0,00	4,10	4,51	0,00	0,00	7,47	**0,49**	3,89	0,49	B(CNT_B)
P17	4,30	3,10	2,41	0,70	2,26	0,70	B(P13)	0,00	0,00	4,30	3,10	0,00	0,00	7,47	**0,94**	2,64	0,94	B(CNT_B)
P18	4,91	4,00	1,34	0,63	2,69	0,63	B(P13)	0,00	0,00	4,91	4,00	0,00	0,00	7,47	**0,82**	3,08	0,82	B(CNT_B)
P19	3,20	3,90	2,88	1,11	3,61	1,11	B(P13)	0,00	0,00	3,20	3,90	0,00	0,00	7,47	**0,89**	3,98	0,89	B(CNT_B)
P20	3,10	4,20	2,89	1,27	3,89	1,27	B(P13)	0,00	0,00	3,10	4,20	0,00	0,00	7,47	**1,00**	4,27	1,00	B(CNT_B)
P21	6,50	1,81	3,15	2,97	0,67	0,67	C(P25)	0,00	0,00	0,00	0,00	6,50	1,81	7,47	3,27	**0,75**	0,75	C(CNT_C)
P22	5,70	1,23	3,67	2,92	0,33	0,33	C(P25)	0,00	0,00	0,00	0,00	5,70	1,23	7,47	3,21	**0,41**	0,41	C(CNT_C)
P23	6,00	0,40	4,50	3,80	1,10	1,10	C(P25)	0,00	0,00	0,00	0,00	6,00	0,40	7,47	4,09	**0,77**	0,77	C(CNT_C)
P24	6,40	0,90	4,03	3,58	0,78	0,78	C(P25)	0,00	0,00	0,00	0,00	6,40	0,90	7,47	3,88	**0,40**	0,40	C(CNT_C)
P25	**5,90**	**1,50**	3,40	2,80	0,00	0,00	C(P25)	0,00	0,00	0,00	0,00	5,90	1,50	7,47	3,10	**0,39**	0,39	C(CNT_C)
P26	5,50	4,68	0,46	1,48	3,21	0,46	A(P5)	5,50	4,68	0,00	0,00	0,00	0,00	**0,44**	1,56	3,56	0,44	A(CNT_A)
N								10	10	11	11	5	5					
media								**5,925**	**4,551**	**4,086**	**4,013**	**6,100**	**1,168**					
								CNT_A		CNT_B		CNT_C						
								Coordinate centroidi (step 2)										

Tabella 8 - Calcolo delle k medie con Excel.

La distanza minima dal punto di inizializzazione (P5,P13,P25) deciderà l'appartenenza ad uno dei tre *cluster* (denominati A, B e C): se un punto del *dataset* ha distanza minima da P5 apparterrà al *cluster* A, se la distanza è minima da P13 a B altrimenti a C. La colonna MIN della Tabella 8 indica il minimo tra i tre valori delle distanze D(P5), D(P13) e D(P25), per ogni punto dell'insieme, mentre la colonna Cluster l'assegnamento al *cluster*, evidenziato meglio nelle colonne delle coordinate dei punti appartenenti ad A, B o C dl passo 2 sempre in Tabella 8. Queste ultime coordinate serviranno per calcolare le medie delle due coordinate dei punti di ogni *cluster*, sommando gli elementi vicini al punto rappresentativo iniziale e dividendo per la numerosità dei punti di ogni *cluster*. Le medie rappresentano le coordinate dei nuovi centroidi dei *cluster*, CNT_A=(5.925, 4.551), CNT_B=(4.086, 4.013) e CNT_C=(6.100, 1.168), come riportato in Figura 3.

Esempio grafico

Disegnando[3] gli assi del poligono (triangolo) che unisce i punti rappresentativi dei *cluster*, è possibile avere una prima idea qualitativa e visuale, della composizione dei *cluster*.

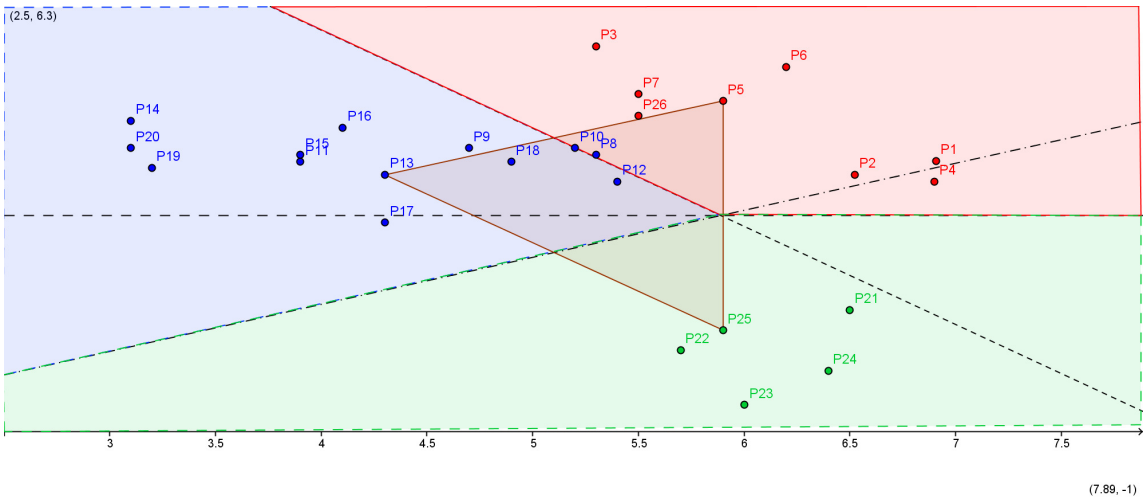

Figura 4 - Assi del poligono costruito sui centroidi di inizializzazione.

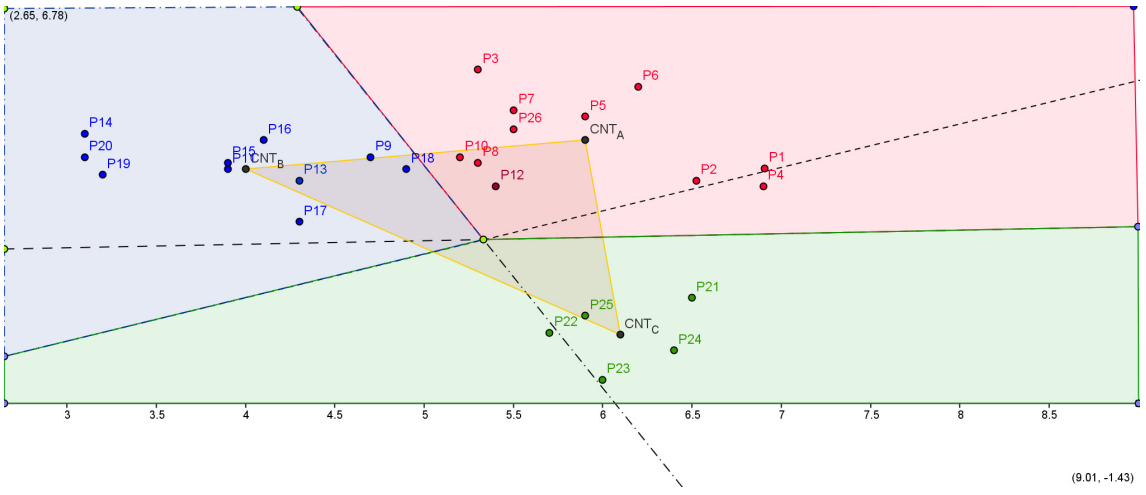

Figura 5 - Assi del poligono costruito sui centroidi calcolati con le medie al passo 2.

[3] I grafici sono stati realizzati col software GeoGebra: http://www.geogebra.org/cms/.

Analisi dei cluster con SPSS

Una buona analisi dei cluster dovrebbe avere le seguenti caratteristiche:

- **Efficienza**
  Utilizzare il minore numero di cluster possibile.
- **Efficacia**
  Evidenziare gruppi di dati d'importanza decisionale/commerciale.

Clustering gerarchico

```
CLUSTER
  /MATRIX IN('C:\DOCUME~1\user\IMPOST~1\Temp\spss1104\spssclus.tmp')
  /METHOD BAVERAGE
  /PRINT SCHEDULE CLUSTER(3)
  /PRINT DISTANCE
  /PLOT DENDROGRAM VICICLE
  /SAVE CLUSTER(3).
```

Syntax 1 - Clustering gerarchico (Between-groups linkage).

Il comando CLUSTER richiama la procedura di *cluster analysis* gerarchica seguito dalle variabili che vengono utilizzate per definire la soluzione di analisi dei cluster. Il comando /METHOD BAVERAGE (*between-groups linkage*) serve per specificare il metodo di classificazione che viene utilizzato, ovvero il "criterio di fusione" che viene utilizzato dal programma per calcolare la distanza tra i cluster ad ogni livello della gerarchia di partizioni definita.

SPSS /METHOD	Metodo
BAVERAGE	Between-groups linkage
WAVERAGE	Within-groups neighbour
SINGLE	Nearest-neighbour
COMPLETE	Furthest-neighbour
CENTROID	Centroid clustering
MEDIAN	Median clustering
WARD	Ward's clustering

Tabella 9 - Selezione del metodo di clustering.

L'opzione riportata è relativa al metodo di BAVERAGE, che costituisce il metodo di *default*. Il comando /PRINT SCHEDULE CLUSTER(3) DISTANCE serve per ottenere in output indici e informazioni che consentono di interpretare al meglio la soluzione. In particolare SCHEDULE (opzione di default) permette di ottenere in output il programma di agglomerazione. Le altre due opzioni invece consentono di ottenere in output l'appartenenza ai gruppi di tutti i casi nelle soluzioni anche a più gruppi (ove specificato), e la matrice delle distanze tra i casi. Specificando NONE invece nessuna di queste informazioni verrà inserita nell'output. Il comando /PLOT DENDROGRAM serve per ottenere in output la rappresentazione grafica del dendrogramma. Specificando invece VICICLE(min,max,inc) e HICICLE(min,max,inc) verranno prodotti i grafici a stalattite rispettivamente verticale e orizzontale, per tutti i cluster se non vengono riportati i valori tra parentesi, oppure per un intervallo di cluster che va da *min* a *max* con un incremento pari al valore *inc*. Anche in questo caso, specificando NONE invece nessuna di queste informazioni verrà inserita nell'output. Il comando /SAVE CLUSTER(3) serve per salvare nel file attivo l'appartenenza del soggetto ai gruppi specificati tra parentesi (nel caso di più gruppi, ad esempio da 2 a 5 avremo CLUSTER(2,5), verranno salvate le appartenenze per le soluzioni a 2, 3, 4 e 5 gruppi). Infine i comandi /MISSING che gestisce il trattamento dei casi con valori mancanti (le opzioni EXCLUDE, che è il valore di default ed esclude tutti i casi con almeno un valore mancante, ed INCLUDE che invece include i casi con valori mancanti); /MATRIX che consente di leggere un file in formato matriciale (IN) o di salvare un file in formato di matrice (OUT).

Com'è noto, nella *cluster analysis* gerarchica ad ogni passo si associano in un nuovo cluster gli oggetti (o cluster) più vicini. Se questo vale per tutti i metodi, cambia però il criterio rispetto al quale vengono calcolate le distanze tra i gruppi. Ogni metodo diverso infatti prevede un criterio diverso. Presentiamo di seguito i principali criteri di fusione utilizzati nella procedura della cluster analysis gerarchica di SPSS.

- Metodo del legame singolo
  La distanza tra due cluster è uguale alla distanza dei due individui nei due differenti cluster che risultano più vicini. Questo metodo viene definito tramite il comando "/METHOD SINGLE".

- Metodo del legame completo
  La distanza tra due cluster è uguale alla distanza dei due oggetti nei due differenti cluster che risultano più lontani. Questo metodo viene definito tramite il comando "/METHOD COMPLETE".

- Metodo del legame medio
  La distanza fra due cluster diversi corrisponde alla media aritmetica delle distanze definite su tutte le coppie di oggetti nei due cluster. In SPSS esistono due varianti di questo metodo.
  - Nel metodo del "legame medio fra i gruppi" la distanza tra due gruppi è uguale alla media delle distanze tra ogni coppia di elementi appartenenti a gruppi differenti. Questo metodo (*average linkage*) viene richiamato tramite la sintassi "/METHOD BAVERAGE", ed è il metodo di default di SPSS.
  - Nel metodo del "legame medio entro i gruppi" la distanza tra due gruppi è uguale alla media delle distanze tra ogni coppia di elementi, incluse le coppie di elementi che appartengono allo stesso gruppo. Questo metodo viene definito tramite il comando "/METHOD WAVERAGE".

- Metodo del Centroide
  La distanza fra due cluster è definita dalla distanza fra i rispettivi centroidi. Questo metodo viene definito tramite il comando "/METHOD CENTROID".

- Metodo della Mediana
  La distanza fra due cluster è definita dalla distanza fra le rispettive mediane. Questo metodo viene definito tramite il comando "/METHOD MEDIAN".

- Metodo di Ward
  In questo metodo il procedimento di associazione fra due cluster diversi è basato sulla minimizzazione della devianza entro i gruppi (ovvero, la massimizzazione delle distanze tra i centroidi dei gruppi). La devianza entro i gruppi è minima ($\rightarrow 0$) quando tutti i casi appartengono ad un gruppo unico ed è massima quando essi sono tutti separati. La coppia di cluster da aggregare in un certo passo è quella che determina un incremento minimo della varianza interna ai cluster. La distanza euclidea tra due oggetti viene calcolata con una funzione che considera sia la numerosità dei gruppi sia la distanza euclidea al quadrato tra i centroidi dei due gruppi [3]. Questo metodo viene definito tramite il comando "/METHOD WARD".

Anche nel caso degli indici di distanza la procedura per la *cluster analysis* gerarchica di SPSS prevede una serie di opzioni differenti, a seconda del livello di misurazione delle variabili prese in esame. Vediamone alcuni a titolo di esempio:

- Variabili su intervalli equivalenti
  Distanza Euclidea: viene richiesta con il comando "/MEASURE= EUCLID"
  Distanza Euclidea al quadrato: viene richiesta con il comando "/MEASURE= SEUCLID" (è l'opzione di default)

Distanza di Minkowski: viene richiesta con il comando "/MEASURE= MINKOWSKI(p)" dove p rappresenta l'esponente utilizzato per elevare alla p-esima potenza la differenza tra i punteggi degli oggetti, sulla la cui somma viene calcolata la radice p-esima.

- Variabili che rappresentano frequenze
  Chi-quadrato: viene richiesta con il comando "/MEASURE= CHISQ"
  Phi-quadrato: viene richiesta con il comando "/MEASURE= PH2"

- Variabili dicotomiche
  Indice di somiglianza è l'indice di Russell e Rao: "/MEASURE= RR"
  Coefficiente di concordanza semplice di Sokal e Michener: "/MEASURE= SM"
  Indice di distanza euclidea "/MEASURE= BEUCLID"
  Distanza euclidea al quadrato "/MEASURE= BSEUCLID".

Dendrogramma

Il dendrogramma è una sintesi grafica della soluzione di *clustering* adottata. I casi vengono elencati a sinistra nell'asse verticale. L'asse orizzontale mostra le distanze tra i *cluster* quando vengono fusi insieme.
Il grafico può essere ispezionato in vari modi. Generalmente si cercano i gap tra i *cluster* fusi tra loro lungo l'asse orizzontale. Ad esempio, iniziando da destra verso sinistra, c'è un gap tra 10 e 25, che divide i 5 elementi del terzo gruppo dal resto degli elementi: questi ultimi sono suddivisi a loro volta in due *cluster* evidenziati dal gap tra 5 e 10, uno di 7 elementi e l'altro di 13 elementi. E' chiaro che procedendo ancora verso sinistra, il numero dei *cluster* aumenterà.

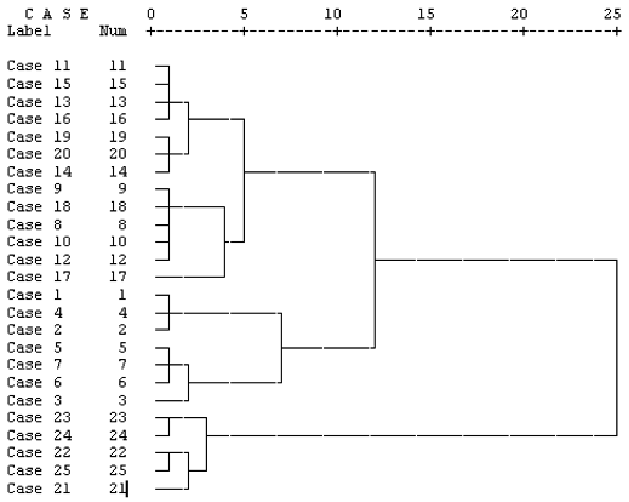

Figura 6 – Dendrogramma ottenuto utilizzando il metodo gerarchico Average Linkage (Between Groups).

Agglomerazione

La sequenza di agglomerazione è un sommario numerico della soluzione di clustering.
Al primo stadio i casi 11 e 15 vengono combinati tra loro perché hanno la minima distanza. Il cluster che li contiene apparirà allo stadio 9 come indicato nella colonna relativa allo "stadio successivo". Ispezionando la riga relativa allo stadio 9, notiamo che il cluster 11 viene accorpato col 13 al passo 13: ciò significa che il cluster risultante appare allo stadio 13. Se ci sono molti casi la tabella è difficile da esplorare ma può essere utilizzata insieme al dendrogramma in modo proficuo e alternativo. Una buona soluzione può essere determinata ancora una volta determinando i gap. In genere la soluzione che precede il gap è una buona soluzione.

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	11	15	1,000	0	0	9
2	8	10	1,000	0	0	10
3	9	18	,998	0	0	12
4	1	4	,997	0	0	7
5	19	20	,997	0	0	11
6	22	25	,996	0	0	16
7	1	2	,994	4	0	22
8	5	7	,994	0	0	15
9	11	13	,992	1	0	13
10	8	12	,992	2	0	12
11	14	19	,989	0	5	18
12	8	9	,989	10	3	20
13	11	16	,988	9	0	18
14	23	24	,986	0	0	19
15	5	6	,983	8	0	17
16	21	22	,975	0	6	19
17	3	5	,972	0	15	22
18	11	14	,964	13	11	21
19	21	23	,964	16	14	24
20	8	17	,943	12	0	21
21	8	11	,917	20	18	23
22	1	3	,882	7	17	23
23	1	8	,802	22	21	24
24	1	21	,549	23	19	0

Tabella 10 - Programma di agglomerazione.

Se la soluzione ottenuta con il metodo *between-groups linkage* (*average linkage*) non è soddisfacente perché la classificazione risulta debole (i gruppi si confondono tra loro) si può provare a riclassificati con un altro metodo come il *furthest neighbour* (linkage completo). La separazione tra i gruppi dovrebbe migliorare.

Cluster di appartenenza

Caso	3 cluster
1:Case 1	1
2:Case 2	1
3:Case 3	1
4:Case 4	1
5:Case 5	1
6:Case 6	1
7:Case 7	1
8:Case 8	2
9:Case 9	2
10:Case 10	2
11:Case 11	2
12:Case 12	2
13:Case 13	2
14:Case 14	2
15:Case 15	2
16:Case 16	2
17:Case 17	2
18:Case 18	2
19:Case 19	2
20:Case 20	2
21:Case 21	3
22:Case 22	3
23:Case 23	3
24:Case 24	3
25:Case 25	3

Cluster di appartenenza

Numero di caso	Cluster	Distanza
1	3	1,017
2	3	,901
3	3	1,360
4	3	1,208
5	3	,362
6	3	,839
7	3	,712
8	3	,898
9	2	,554
10	3	,940
11	2	,276
12	3	1,097
13	2	,257
14	2	1,221
15	2	,285
16	2	,482
17	2	,934
18	2	,726
19	2	,984
20	2	1,090
21	1	,754
22	1	,405
23	1	,775
24	1	,403
25	1	,387

Tabella 11 - Appartenenza ai cluster secondo il modello gerarchico (a sinistra) e non-gerarchico (k-means, a destra).

K-Means

L'analisi K-means è un metodo creato per assegnare i dati ad un numero di gruppi predefinito le cui caratteristiche non sono note a priori ma sono basate su un insieme di variabili. E' molto utilizzato per classificare estesi dataset, migliaia di dati.
I centri dei cluster iniziali possono essere assegnati arbitrariamente tramite un'apposita procedura che genera casualmente i centri, dato il loro numero. La procedura è quella esposta nel paragrafo "Clustering gerarchico" (pag. 23). La Tabella 16 - ANOVA. dell'analisi ANOVA indica quali variabili contribuiscono maggiormente alla soluzione di clustering. Le variabili con il valore del rapporto F elevato producono una maggiore separazione tra i cluster. I centri finali dei cluster vengono calcolati per caratterizzare la soluzione finale.

```
QUICK CLUSTER X1 X2
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(3) MXITER(10) CONVERGE(0)
  /METHOD=KMEANS(UPDATE)
  /SAVE CLUSTER DISTANCE
  /PRINT INITIAL ANOVA CLUSTER DISTAN.
```

Syntax 2 - SPSS K-Means.

Il comando QUICK CLUSTER richiama l'omonima procedura di *cluster analysis* non-gerarchica, seguito dalle variabili che vengono utilizzate per definire la soluzione di analisi dei cluster.
Il comando /MISSING=LISTWISE gestisce il trattamento dei casi con valori mancanti. Le opzioni sono le alternative LISTWISE e PAIRWISE, in liste o in coppie. Completa il quadro l'opzione INCLUDE che permette di includere in analisi tutti i casi con valori mancanti. Il comando /CRITERIA= CLUSTER(3) MXITER(10) CONVERGE(0) prevede una serie di opzioni che consentono di controllare il processo di analisi. Nel nostro esempio CLUSTER(3) specifica che verrà creata una partizione di 3 gruppi, MXITER(10) stabilisce il numero massimo di iterazioni previste per raggiungere la convergenza, CONVERGE(0) serve per determinare il cambiamento minimo nei centroidi dei cluster affinché il processo di convergenza possa considerarsi concluso. Un'ulteriore opzione, che non è riportata in queste linee di script, è NONINITIAL: se tale opzione viene specificata, il programma prenderà come centroidi iniziali per generare la partizione i primi n casi senza valori mancanti nel data file (dove n è il numero di gruppi in cui verranno suddivisi gli oggetti. L'opzione di *default* invece prevede che il programma esamini i dati una prima volta per scegliere come centroidi iniziali dei gruppi gli n soggetti che sono più distanti.

	Cluster		
	1	2	3
X1	6,000	3,203	5,300
X2	,400	3,904	5,700

Tabella 12 - Centri iniziali dei cluster.

Il comando /METHOD=KMEANS(UPDATE) consente di governare le operazione relative al ricalcalo dei centroidi dei gruppi dopo l'assegnazione dei soggetti ai gruppi. Se si lascia l'opzione di default NOUPDATE i centroidi sono ricalcolati dopo che tutti i casi sono stati assegnati, alla fine dell'iterazione. Se invece viene scelta l'opzione UPDATE i centroidi vengono ricalcolati dopo l'assegnazione di ogni caso ad un gruppo: si tratta dunque del metodo delle cosiddette "medie mobili", che prevede un aggiornamento dinamico dei centri dei gruppi. Con l'opzione CLASSIFY infine i casi vengono assegnati ai gruppi più vicini, non si effettua nessuna iterazione e i centroidi dei gruppi vengono ricalcolati quando tutti gli oggetti sono stati classificati. Il comando /SAVE CLUSTER DISTANCE prevede le due opzioni che consentono di salvare nel file attivo rispettivamente il numero del cluster in cui il soggetto è stato classificato e la distanza dal centroide del gruppo di appartenenza. Il comando /PRINT INITIAL ANOVA CLUSTER DISTAN prevede alcune opzioni che consentono di ottenere in output diverse informazioni utili per interpretare la soluzione. In particolare, l'opzione

INITIAL consente di avere i centroidi iniziali dei cluster, l'opzione ANOVA consente di esaminare la significatività statistica della differenza tra le medie delle variabili attraverso i gruppi, l'opzione CLUSTER consente di avere in output una tabella con specificati per ogni casi il gruppo cui appartiene e la distanza dal centroide, l'opzione DISTAN consente di avere in output una tabella con le distanze tra i centroidi dei cluster. Completa questo comando l'opzione "ID(nome della variabile)" che consente di utilizzare il valore della variabile specificata (solitamente si tratta di una variabile alfanumerica con un etichetta che identifica il caso) come identificatore supplementare oltre al numero del caso assegnato di default nel file dati. Come per i comandi analoghi della procedura gerarchica, anche in questa procedura, specificando NONE nessuna di queste informazioni verrà inserita nell'output.

Iterazione	Modifiche ai centri dei cluster		
	1	2	3
1	,646	,491	1,269
2	,108	,249	,202
3	,018	,231	,187
4	,003	,018	,021
5	,000	,001	,002
6	8,304E-5	,000	,000
7	1,384E-5	8,089E-6	2,852E-5
8	2,307E-6	6,222E-7	3,169E-6
9	3,845E-7	4,786E-8	3,522E-7
10	6,408E-8	3,682E-9	3,913E-8

a. Iterazioni interrotte perché è stato eseguito il numero massimo di iterazioni. Impossibile ottenere la convergenza tramite le iterazioni. La variazione massima assoluta delle coordinate per qualsiasi centro è 6,35E-008. L'iterazione corrente è 10. La distanza minima tra i centri iniziali è 2,761.

Tabella 13 - Cronologia delle iterazioni[a] (10 passi).

Ci sono infine tre ulteriori comandi che possono essere specificati nella procedura QUICK CLUSTER. Il comando "/INITIAL()" serve a specificare i valori dei centroidi iniziali; in particolare vanno inserire nella parentesi le medie di ciascuna variabile dal primo gruppo all'ultimo gruppo (nel nostro esempio dovremmo fornire 2*3= 6 differenti valori, dove 2 è il numero delle variabili, e 3 il numero dei gruppi). I valori delle medie possono essere letti anche da un file esterno utilizzando il comando "/FILE=nomefile": ovviamente il file deve essere in formato SPSS. Nell'esempio discusso nel testo non abbiamo utilizzato queste opzioni. Infine il comando /OUTFILE consente di salvare i valori finali dei centroidi in un file esterno in formato SPSS.

	Cluster		
	1	2	3
X1	6,100	3,954	5,915
X2	1,168	4,044	4,453

Tabella 14 - Centri finali dei cluster.

Cluster	1	2	3
1		3,588	3,290
2	3,588		2,003
3	3,290	2,003	

Tabella 15 - Distanze tra i centri dei cluster finali.

	Cluster		Errore		F	Sig.
	Media dei quadrati	df	Media dei quadrati	df		
X1	12,333	2	,377	22	32,671	,000
X2	19,392	2	,347	22	55,820	,000

I test F devono essere utilizzati solo per motivi descrittivi poiché i cluster sono stati scelti per ottimizzare le differenze tra i casi in diversi cluster. I livelli di significatività osservati non sono perciò corretti e non possono quindi essere interpretati come test dell'ipotesi che le medie dei cluster siano uguali.

Tabella 16 - ANOVA.

Cluster	1	5,000
	2	10,000
	3	10,000
	Validi	25,000
	Mancanti	,000

Tabella 17 - Numero di casi in ogni cluster.

Cenni di analisi fattoriale

L'analisi fattoriale [9] si pone l'obiettivo di riassumere l'informazione contenuta (*data reduction*) in una matrice di correlazione o di varianza/covarianza, cercando di individuare statisticamente le dimensioni latenti (*structure detection*) e non direttamente osservabili. Si può dire che se due variabili hanno una forte correlazione con uno stesso fattore, una parte non trascurabile della correlazione tra le due variabili si spiega col fatto che esse hanno quel fattore in comune.
Fornendo un principio di identificazione di questi fattori comuni, l'analisi fattoriale fornisce una descrizione in forma semplice, della complessa rete di relazioni esistente nell'ambito di un insieme di variabili associate. Questa descrizione consente di definire, all'interno della matrice di correlazione, un limitato numero di componenti indipendenti l'una dall'altra e identificate con i fattori: esse spiegano il massimo possibile di varianza delle variabili contenute nella matrice d'informazione originaria.
Data una matrice nxp, contenente p variabili rilevate su n unità osservate, si tratta di verificare in che misura ciascuna variabile costituisce una descrizione ridondante rispetto alle rimanenti p-1 e, quindi, se esiste la possibilità di raggiungere la stessa efficacia descrittiva con un numero minore di variabili non osservate (fattori).

Le dimensioni latenti possono essere determinate in vari modi grazie alle svariate tecniche di estrazione dei fattori di cui l'analisi dei fattori si avvale. Tra le più utilizzate ricordiamo: l'analisi delle componenti principali e l'analisi fattoriale canonica[4].

Analisi delle componenti principali

Il metodo delle componenti principali [1] si propone di sostituire le p variabili date con un certo numero di variabili (tra loro non interdipendenti), ottenute come trasformazione lineare delle variabili originarie, riducendo così il numero di variabili necessarie a descrivere un certo ambito. Si tratta cioè di ricercare una serie di trasformate della matrice originaria dette componenti principali, che spieghino quanto più possibile la varianza delle variabili originarie ed inoltre che siano tra loro ortogonali. È possibile estrarre tante componenti quante sono le variabili originarie, quando però lo scopo è quello di conseguire un'economia nella descrizione, in termini quantitativi di un certo fenomeno: il risultato fornito dall'applicazione del metodo è tanto più utile quanto minore è il numero di componenti prese in considerazione. In genere il processo viene arrestato non appena la parte di varianza delle p variabili estratte dalle prime q componenti è sufficientemente grande. Un test comunemente usato per la scelta del numero di componenti da considerare, che utilizza la matrice della varianza e covarianza tra le variabili standardizzate, è il *Test di Bartlett* (1950).

Analisi fattoriale canonica

Il principio che guida questa analisi è quello di trovare una soluzione fattoriale nella quale la correlazione tra il *set* di ipotetici fattori e il *set* di variabili sia massima. Il metodo parte considerando due serie di variabili x_1 e x_2, la prima contiene p variabili osservate e la seconda contiene invece q variabili ortogonali incognite, le cui trasformate, opportunamente ridotte in forma standardizzata (z_1 e z_2), costituiscono le colonne della matrice dei fattori da determinare. L'analisi fattoriale canonica si scosta poco dalla precedente, ma essa opera sulla matrice di correlazione parziale invece che sulla matrice di correlazione totale delle variabili. In presenza di un ristretto ventaglio di variabili osservate sulle osservazioni, essa consente di legare più nitidamente i fattori latenti ad esse.

La rotazione dei fattori

Il problema della rotazione si pone perché le variabili possono venire rappresentate in modo simile da diversi fattori, ossia esistono più soluzioni conformi con l'obiettivo di individuazione di poche dimensioni fondamentali di un certo fenomeno mediante un elevato numero di variabili quantitative.

[4] Queste tecniche, per le caratteristiche dei loro algoritmi sono orientate all'analisi della varianza.

Individuare i fattori è legata al grado di correlazione tra il fattore e le variabili osservate. Nei casi pratici, spesso, i fattori estratti nella fase iniziale dell'analisi risultano correlati con un numero elevato di variabili rendendo poco chiara la corrispondenza tra fattori e loro significato. La rotazione si concretizza nella riduzione dei pesi dei fattori e dell'incremento, sia positivo che negativo, dei valori dei pesi fattoriali che erano preponderanti in prima istanza. La matrice delle saturazioni non presenta un'unica soluzione e, attraverso la sua trasformazione matematica, si possono ottenere infinite matrici dello stesso ordine. In una soluzione non ruotata, infatti, ogni variabile è spiegata da due o più fattori comuni, mentre in una soluzione ruotata ogni variabile è spiegata da un singolo fattore comune.

Attualmente vengono utilizzati diversi metodi di rotazione che possono essere suddivisi essenzialmente in due gruppi: quelli che producono "rotazioni ortogonali dei fattori" e quelli che invece prediligono "rotazioni oblique". Il vincolo dell'ortogonalità non sempre viene preferito, in quanto si sostiene che le dimensioni fondamentali modellizzate come indipendenti, possono essere in realtà tra loro interdipendenti, e sono stati messi a punto una serie di metodi di rotazione obliqua nei quali gli assi, presi a due a due, sono lasciati liberi di disporsi in modo da formare un angolo o maggiore o minore di 90 gradi.

La pluralità di queste tecniche di rotazione dei fattori provoca una indeterminatezza nella soluzione fattoriale, poiché non è possibile stabilire quale delle rotazioni sia migliore in assoluto; e questo non solo per la scelta tra rotazione obliqua e rotazione ortogonale, ma anche all'interno dei due tipi di rotazione. Un criterio spesso utilizzato con successo è quello di confrontare tra loro i risultati di diverse applicazioni e scegliere quella che meglio si adatta ai risultati osservati. Una soluzione fattoriale è determinata se i fattori comuni che si adattano al modello sono unici. La condizione di indeterminatezza implica che insiemi contraddittori di punteggi fattoriali risultano ugualmente plausibili e che la scelta di una soluzione piuttosto che di un'altra è arbitraria.

Nell'analisi fattoriale l'indeterminatezza si verifica a due livelli:
- nell'accettazione della soluzione che soddisfa il modello in senso statistico;
- nella ricerca di una soluzione più facilmente interpretabile di quella ottenuta col primo approccio.

Nell'ambito delle analisi esplorative condotte per acquisire informazioni sulla struttura latente dei dati osservati, o altre collegate a quelle presenti nel modello, disporre di più interpretazioni mutuamente consistenti deve considerarsi vantaggioso. Una altro problema è la determinazione del numero di fattori. Il rapporto tra numero di fattori q e il numero di variabili osservate p permette di misurare la determinatezza dei fattori comuni e specifici in termini matematici (Shonemann e Wang, 1972).

Vale in generale che:
- all'aumentare del numero di variabili studiate è necessario incrementare il numero di fattori al fine di ottenere una buona interpolazione del modello di analisi;
- all'aumentare del numero dei fattori estratti, cresce il rischio di indeterminatezza della soluzione.

Nella ricerca empirica di tipo esplorativo è ammissibile l'adozione di uno dei seguenti criteri:

- Varianza
  Se la selezione delle variabili non è casuale ma, come si verifica spesso, si inseriscono per prime quelle ritenute più appropriate (e quindi più informative: maggiore varianza spiegata) per il modello che il ricercatore ha in mente, è ragionevole supporre che la comunanza delle variabili aggiunte sia proporzionalmente inferiore a quella delle variabili introdotte in precedenza e che il numero dei fattori richiesti per raggiungere la stessa frazione di varianza spiegata sia relativamente più elevato.

- Autovalori
  Si raccomanda di considerare solo i fattori di una matrice di correlazione ai quali è associato un autovalore maggiore o al più uguale a 1. Il numero di tali fattori dovrebbe variare tra 1/6 e 1/3 del numero di variabili. In genere si considerano significativi fattori che globalmente contribuiscono almeno al 75% della varianza spiegata, anche se si tollerano valori di poco inferiori a questa soglia.

- Scree-plot
  La rappresentazione grafica degli autovalori (in ordinata) e dell'ordine di estrazione dei fattori (in ascissa) dà un'immagine dell'importanza relativa dei primi autovalori nella sequenza ricavata. Si escludono quei i fattori i cui valori appartengono alla spezzata che corre quasi parallela all'asse delle ascisse, in quanto hanno tasso di variazione prossima allo zero (derivata nulla). E' possibile che questo metodo consigli l'introduzione di più fattori rispetto a quanto suggerito dagli altri indicatori.

- Comunanze
  Sostituire i valori sulla diagonale principale della matrice di correlazione con le comunanze determina l'estrazione di un numero di fattori inferiore al rango della matrice.

Mapping multidimensionali delle percezioni

La misura della percezione di un prodotto da parte dei clienti permette di realizzare il posizionamento di tale prodotto rispetto ad altri prodotti concorrenti. In pratica, la rappresentazione grafica delle percezioni dei clienti, nell'ambito di segmenti del mercato di riferimento, costituisce il posizionamento del prodotto. La percezione di un prodotto racchiude in se componenti derivanti dalle motivazioni all'acquisto e dalle aspettative su un determinato prodotto in termini relativi ad altri prodotti dello stesso tipo. Questa dipendenza relazionale tra prodotti di differenti produttori in competizione tra loro è l'elemento di maggiore rilievo nella rappresentazione dei prodotti in termini relativi e fa si che il posizionamento sia influenzato non soltanto dalle azioni del produttore del prodotto in esame ma anche dalle decisioni dei concorrenti che, mutando le loro strategie, conferiscono al problema una elevata dinamicità e una caratteristica complessità d'analisi. Le tecniche di analisi multivariata utili alla costruzione di mappe di percezione che verranno analizzate in queste brevi note[5] sono l'analisi discriminante lineare e il *multidimensional scaling*: in particolare, l'ampia disponibilità di *software* per l'analisi discriminante lineare (ADL) rende tale procedura metrica (che utilizza in *input* valutazioni misurate almeno a livello di intervallo) di ampio utilizzo per il *mapping* multidimensionale.

[5] In realtà, tra le tecniche *attribute based* come l'ADL si annovera anche l'analisi delle corrispondenze, che non verrà trattata in queste note.

Glossario

Analisi dei fattori
E' una tecnica di statistica multivariata che si propone di individuare le dimensioni fondamentali di un fenomeno descritto da un insieme di n variabili quantitative.

Analisi statistica multivariata
Con statistica multivariata s'intende quella parte della statistica in cui l'oggetto dell'analisi è per sua natura formato da almeno due componenti, il che è spesso il caso nell'ambito di scienze quali la medicina, psicologia, sociologia, ecologia e biologia. Fanno parte della statistica multivariata metodi quali:

- analisi della correlazione canonica e analisi delle componenti principali
- analisi fattoriale
- analisi delle corrispondenze
- analisi dei cluster
- analisi discriminante
- analisi di regressione multidimensionale

Cluster
Con il termine *cluster* si intende un gruppo di unità simili o vicine tra loro, dal punto di vista della posizione o della composizione.

Correlazione
Per correlazione si intende una relazione tra due variabili casuali tale che a ciascun valore della prima variabile corrisponda con una certa regolarità un valore della seconda. Non si tratta necessariamente di un rapporto di causa ed effetto ma semplicemente della tendenza di una variabile a variare in funzione di un'altra. Talvolta le variazioni di una variabile dipendono dalle variazioni dell'altra (relazione tra la statura dei padri e quella dei figlio ad esempio), talvolta sono comuni (relazioni tra la statura e il peso di un individuo); talvolta sono reciprocamente dipendenti (relazione tra prezzo e domanda di una merce: il prezzo influisce sulla domanda e la domanda influisce sul prezzo). Il grado di correlazione fra due variabili viene espresso mediante i cosiddetti indici di correlazione. Questi assumono valori compresi tra meno uno (quando le variabili considerate sono inversamente correlate) e l'unità (quando vi sia correlazione assoluta cioè quando alla variazione di una variabile corrisponde una variazione rigidamente dipendente dall'altra), ovviamente un indice di correlazione pari a zero indica un'assenza di correlazione e quindi le variabili sono indipendenti l'una dall'altra. I coefficienti di correlazione sono derivati dagli indici di correlazione tenendo presenti le grandezze degli scostamenti dalla media. In particolare, il coefficiente di correlazione di Pearson è calcolato come rapporto tra la covarianza delle due variabili ed il prodotto delle loro deviazioni standard [7].

Covarianza/Varianza
E' un indice che misura la "*contemporaneità*" della variazione (in termini lineari) di due variabili casuali. Essa può assumere sia valori positivi che negativi. Nel caso di valori positivi indica che al crescere di una caratteristica statisticamente cresce anche l'altra, nel caso di valori negativi accade il contrario. Nella statistica inferenziale, quando due variabili sono tra di loro indipendenti, allora la loro covarianza è nulla (l'inverso non è necessariamente verificato). Si utilizza spesso la notazione:

$$\operatorname{cov}(x, y) = \sigma_{xy}$$

$$\sigma_{xy} = \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

essendo μ_x e μ_y rispettivamente la media aritmetica di x e y.

In caso di ponderazione,

$$\sigma_{xy} = \sum_{j=1}^{k} f_j (x_j - \mu_x)(y_j - \mu_y)$$

È un operatore simmetrico, cioè

$$\operatorname{cov}(x, y) = \operatorname{cov}(y, x)$$

La covarianza può essere scomposta in due termini, diventando

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_j y_j - \mu_x \mu_y$$

ovvero la media dei prodotti meno il prodotto delle medie.
Quando y=x, allora la covarianza si trasforma in varianza:

$$\sigma_{xx} = \operatorname{cov}(x, x) = \operatorname{var}(x) = \sigma_x^2 .$$

Deviazione standard *(Standard Deviation)*
Indice di dispersione della popolazione o del campione. Detto anche *Scarto Quadratico Medio*.

Incertezza standard
Incertezza del risultato di una misurazione espressa. Nelle statistiche quantitative essa è uguale alla Deviazione Standard *(Standard Deviation)*.

Incertezza standard composta
L'incertezza tipo del risultato di una misurazione si ha quando il risultato è ottenuto mediante i valori di un certo numero di grandezze; essa è uguale alla radice quadrata positiva di una somma di termini, che sono le varianze o le covarianze di quelle grandezze, pensate secondo la variazione del risultato della misurazione al variare di esse.

Intervallo di confidenza
Intervallo di valori costruito con una procedura statistica che garantisce che, su 100 intervalli realizzati con tale procedura, il 95% di questi conterrà il valore medio incognito del campione.

Mappa/Mappatura delle percezioni (Perceptual mapping)
E' una tecnica grafica utilizzata nel marketing per tentare di visualizzare la percezione che hanno i potenziali clienti di un determinato prodotto in relazione ad altri prodotti di riferimento.

Marketing
E' un ramo dell'economia che si occupa dello studio descrittivo del mercato e dell'analisi dell'interazione del mercato, degli utilizzatori con l'impresa. Il termine prende origine dall'inglese market, cui viene aggiunta la desinenza del gerundio per indicare la partecipazione attiva, cioè l'azione sul mercato stesso. Marketing significa letteralmente "piazzare sul mercato" e comprende quindi tutte le azioni aziendali riferibili al mercato destinate al piazzamento di prodotti, considerando come finalità il maggiore profitto e come causalità la possibilità di avere prodotti capaci di realizzare tale operazione [15].

Marketing analitico
Consiste in un insieme di tecniche e metodologie volte ad analizzare con metodi quantitativi, il mercato nella sua accezione più larga (dei clienti finali, o degli intermediari, ecc.) per mappare i desideri del cliente, oppure i suoi comportamenti (segmentazione), e per conoscere gli ambiti di mercato già eventualmente occupati dai rivali diretti e indiretti (posizionamento).

Marketing strategico
Si basa sull'analisi dei bisogni degli individui e delle organizzazioni. Questo primo aspetto del processo di marketing riguarda anzitutto l'individuazione, all'interno del mercato di riferimento, dei prodotti-mercato e dei segmenti già esistenti o potenziali. Di questi il marketing strategico misura l'attrattività in termini quantitativi,

qualitativi (con riferimento all'accessibilità al mercato) e dinamici (con riferimento alla durata economica che è rappresentata dal ciclo di vita del prodotto). Tali operazioni consentono di scegliere una strategia di sviluppo che colga le opportunità esistenti sul mercato (rappresentate sostanzialmente da bisogni insoddisfatti) e che, tenendo conto delle risorse e competenze dell'impresa, offrano alla stessa un potenziale di crescita e di redditività attraverso l'acquisizione ed il mantenimento di un vantaggio competitivo [12][15].

Marketing operativo
E' la parte applicativa dell'intero processo di marketing, a monte del quale ci sono le fasi di marketing analitico e marketing strategico. La componente operativa (o tattica) del marketing ha il compito di realizzare concretamente le strategie definite nelle fasi precedenti. le caratteristiche: orientamento all'azione, opportunità esistenti, ambiente stabile, comportamento reattivo, orizzonte a breve termine, responsabilità della funzione di marketing [12][15].

Marketing Mix
Indica la combinazione (*mix*) di variabili controllabili (leve decisionali) di marketing che le imprese impiegano per raggiungere i propri obiettivi. Le variabili che tradizionalmente si includono nel *marketing mix* sono le 4P teorizzate da Jerome McCarthy e riprese in seguito da molti altri: *Product* (Prodotto), *Price* (Prezzo), *Place* (Distribuzione), *Promotion* (Comunicazione) [12].

Percezione
Il complesso processo elettrochimico che connette i livelli sensoriali di un organismo attraverso il sistema nervoso e che opera la sintesi dei dati sensoriali in forme dotate di significato.

Posizionamento
Il posizionamento di un prodotto può essere visto come una decisione strettamente connesso a quella della selezione dei segmenti di mercato in cui l'impresa decide di competere. Il posizionamento consiste nella misura della percezione che hanno i clienti di un prodotto o di una merce, relativamente alla posizione dei prodotti o delle marche concorrenti.

Regressione lineare
La regressione formalizza e risolve il problema di una relazione funzionale tra variabili misurate sulla base di dati campionari estratti da un'ipotetica popolazione infinita. Più formalmente, in statistica la regressione lineare rappresenta un metodo di stima del valore atteso condizionato di una variabile dipendente, dati i valori di altre variabili indipendenti.

Segmentazione
Col termine "segmentazione" del mercato s'intende l'attività di identificazione di "gruppi di clienti" cui è indirizzato un determinato prodotto o servizio. La segmentazione è "il processo attraverso il quale le imprese suddividono la domanda in un insieme di clienti potenziali, in modo che gli individui che appartengono allo stesso insieme siano caratterizzati da funzioni della domanda il più possibile simili tra loro e, contemporaneamente, il più possibile diverse da quelle degli altri insiemi" [15].

Bibliografia

1. Abdi Hervvé e Williams Lynne J., Principal Component Analysis - Wiley Interdisciplinary Reviews: Computational Statistics (2010) - http://www.utdallas.edu/~herve/abdi-wireCS-PCA2010-inpress.pdf
2. Abe Shigeo – Pattern Classification – Springer (2001)- http://www.amazon.com/Pattern-Classification-Shigeo-Abe/dp/1852333529/ref=sr_1_1?ie=UTF8&s=books&qid=1262625717&sr=1-1
3. Barbaranelli, C. - Analisi dei dati con SPSS Vol. II. Milano: LED (2003) - http://www.ibs.it/code/9788879163156/barbaranelli-claudio/analisi-dei-dati-con.html
4. Bracalente, Mulas, Cossignani - Statistica aziendale - McGraw Hill (2009) - http://www.ibs.it/code/9788838664960/bracalente-mulas-cossignani/statistica-aziendale.html
5. Brasini Sergio, Freo Marzia, Tassinari Franco, Tassinari Giorgio - Statistica aziendale e analisi di mercato - Il Mulino, Bologna (2002) - http://www.ibs.it/code/9788815088765/zzz1k1456/statistica-aziendale-e-analisi.html
6. De Finetti B. – Sul significato soggettivo della probabilità (1931) http://www.brunodefinetti.it/Opere/Sul%20significato%20soggettivo%20della%20probabilit%E0.pdf
7. Field Andy – Discovering Statistics using SPSS for Windows – SAGE Publication (2000) - http://www.ibs.it/book/9781412977524/field-andy/discovering-statistics-using.html
8. Hartigan J. A. – Clustering Algorithm – New York Wiley (1975) - http://www.amazon.com/Clustering-Algorithms-Probability-Mathematical-Statistics/dp/047135645X
9. Hauser J. R., Koppelman F. S. - Alternative perceptual mapping technique – Journal of Marketing Research (1979) - http://web.mit.edu/hauser/www/Papers/Alternative_Perceptual_Mapping_Techniques.pdf
10. Howard Martin, Sappiamo cosa vuoi, Minimum Fax 2005, http://www.ibs.it/code/9788875210687/howard-martin/sappiamo-cosa-vuoi-chi.html
11. Jiawei Han – Data Mining, concepts and techniques – Morgan Kaufmann (2001).- http://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/1558609016/ref=ntt_at_ep_dpt_1
12. Kotler Philip, Marketing management (2007) - http://www.ibs.it/code/9788871922935/kotler-philip/*marketing*-management.html
13. Lance G. N., Williams W. T. – A general theory of classification sorting strategies, hierarchical systems – Computer Journal (1967) - http://users.informatik.uni-halle.de/~huebenth/lance67.pdf
14. MacQueen J. – Some Methods for classification and *analysis* of multivariate observations – University of California, Los Angeles (1967) – http//projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.bsmsp/1200512992
15. Molteni Luca, Gabriele Triolo - Ricerche di *marketing* - McGraw Hill (2003). - http://www.ibs.it/code/9788838663925/molteni-luca/ricerche-di-*marketing*.html
16. Piccolo Domenico - Statistica per le decisioni - Il Mulino, Bologna (2004) - http://www.ibs.it/code/9788815097705/piccolo-domenico/statistica-per-le-decisioni.html
17. R (The) Development Core Team - R: A Language and Environment for Statistical Computing - Copyright (©) 1999–2009 R Foundation for Statistical Computing, ISBN 3-900051-07-0 - http://cran.r-project.org/doc/manuals/refman.pdf
18. Raghavan Vijay V. Tandhe M.Y.L. Ip – Techniques for measuring stability of clustering: a comparative study (1982) - http://elvis.slis.indiana.edu/irpub/SIGIR/1982/pdf16.pdf
19. Randy Julian – Lecture slides – Lilly Research Laboratories http://miner.chem.purdue.edu/Lectures/
20. Ramsey F. P. - Truth and probability (1926) - http://fitelson.org/probability/ramsey.pdf
21. Ruminati Rino, Psicologia economica, a cura di Ruminati Rino Enrico Rubatelli e Maurizio Mistri, Carocci 2008, http://www.ibs.it/code/9788843044290/zzz1k1456/psicologia-economica.html
22. Russel S. J., Norvig P. – Intelligenza artificiale – Pearson Education Italia (2005) - http://www.ibs.it/code/9788871922287/russell-stuart-j/intelligenza-artificiale-approccio.html

23. Sharma Anil K., Sheikh Sohail - Classification and Clustering: Using Neural Networks (1994) - http://pubs.acs.org/doi/pdf/10.1021/ci00021a019
24. SPSS for Windows Documentation
    http://support.spss.com/ProductsExt/SPSS/Documentation/SPSSforWindows/index.html
    http://support.spss.com/ProductsExt/SPSS/Documentation/SPSSforWindows/SPSS 16.0 Algorithms.pdf
    http://support.spss.com/ProductsExt/SPSS/Documentation/SPSSforWindows/SPSS Conjoint 16.0.pdf
25. Tryon R. C. – Cluster Analysis – New York Mc Graw Hill (1939)
26. Tryon R., D. Bailey – Cluster Analysis – New York Mc Graw Hill (1983). http://www.amazon.com/Cluster-Analysis-Robert-Tryon/dp/0226813126/ref=sr_1_3?ie=UTF8&s=books&qid=1262628707&sr=1-3
27. Universität Hamburg – SPSS Algorithms –
    http://www1.uni-hamburg.de/RRZ/Software/SPSS/Algorith.120/
    http://www1.uni-hamburg.de/RRZ/Software/SPSS/Algorith.115/proxscal.pdf
    http://www1.uni-hamburg.de/RRZ/Software/SPSS/Algorith.115/alscal.pdf
28. Ulrich K. T., Eppinger S. D., Filippini R. – Progettazione e sviluppo del prodotto _ McGraw-Hill (1995) http://www.ibs.it/code/9788838663970/ulrich-karl-t-eppinger/progettazione-e-sviluppo-di.html
29. Vigneau E., Qannari E. M., Punter P. H., Knoops S. - Segmentation of a panel of consumers using clustering of variables around latent directions of preference – ENITIAA/INRA, Unite´ de Sensome´trie et de Chimiome´trie, la Ge´raudie`re, BP 82225, 44322 Nantes Cedex, France – Food Quality and Preference (2001) - http://www.dict.uh.cu/Bib_Dig_Food/elsevier/fqp/FQP12/FQP12_359.pdf
30. Vigneau E., Qannari E. M. - Segmentation of consumers taking account of external data. A clustering of variables approach - ENITIAA/INRA, Unite´ de Sensome´trie et de Chimiome´trie, la Ge´raudie`re, BP 82225, 44322 Nantes Cedex, France – Food Quality and Preference (2002) – http://www.typic.org/free/publications/consumer/Vigneau_qannari_FQP2.pdf
31. Ward J. – Hierarchical grouping to optimize an objective function – Journal of the American Statistical Association (1963) - http://iv.slis.indiana.edu/sw/data/ward.pdf
32. Wyne S. Desarbo, Rajdeep Grewal, and Crystal J. Scott - A Clusterwise Bilinear Multidimensional Scaling Methodology for Simultaneous Segmentation and Positioning Analyses – Journal of Marketing Research, June 2008 – http://www.personal.psu.edu/rug2/DeSarbo%20et%20al%202008%20JMR.pdf
33. Written Ian H., Eibe Frank – Data Mining, practical machine learning tools – Morgan Kaufmann (2000) http://www.amazon.com/Data-Mining-Techniques-Implementations-Management/dp/1558605525/ref=sr_1_1?ie=UTF8&s=books&qid=1262625852&sr=1-1

Indice dei nomi