Appendice A: Statistica

Cluster analysis

La Cluster Analysis raggruppa un insieme di tecniche, proposte e sperimentate soprattutto dagli anni '60 in poi, anche se essa fa la sua prima comparsa nel 1939 quando Tryon pubblica una monografia proprio con il titolo di Cluster Analysis. La prima esposizione sistematica di tali tecniche risale al 1963 ad opera di Sokal e Sneath che in quell'anno pubblicano *Principles of numerical taxonomy*.

Con il termine *Cluster Analysis,* o analisi dei gruppi*,* si intendono le procedure che permettono di individuare, all'interno di un insieme di oggetti di qualsiasi natura, alcuni sottoinsiemi, i clusters appunto, tendenzialmente omogenei al loro interno, e sufficientemente distinti fra loro. Le tecniche di Cluster Analysis creano i gruppi in modo tale che ogni osservazione sia molto simile a tutte le altre che appartengono allo stesso gruppo, in funzione di alcuni criteri prestabiliti dal ricercatore. Alla fine del procedimento, i cluster finali dovrebbero esibire un'alta omogeneità interna (intra-cluster) ed un'alta eterogeneità esterna (inter-cluster). Quindi, se la classificazione ha successo, gli oggetti all'interno dei cluster saranno vicini tra loro, mentre gli oggetti che appartengono a differenti cluster saranno più lontani tra loro (Barbarito, 1999).

Interventi decisionali nelle diverse fasi della Cluster Analysis

PRIMA	Scelta delle variabili Criteri di similarità-distanza
DURANTE	Tecniche di aggregazione Numero dei gruppi da ottenere
DOPO	Valutazione della qualità della soluzione Scelta fra le diverse possibili soluzioni alternative

Ricorrere alla Cluster Analysis è particolarmente indicato quando si ha una grossa mole di informazioni che si desidera ridurre ad una quantità più facilmente trattabile. La classificazione ha infatti l'effetto di ridurre le dimensioni della matrice dei dati di partenza riducendo il numero delle righe della matrice, ossia dei casi. L'obiettivo quindi della Cluster Analysis è quello di raggruppare gli oggetti in gruppi che ad ogni passaggio diventano sempre più numerosi, usando alcune misure di similarità o di distanza. Il ricorso alla Cluster Analysis impone però una serie di decisioni da parte del ricercatore, prima, durante e dopo l'analisi.

Proprio questa componente di arbitrarietà, è stata fonte di notevoli critiche, ma nelle scienze sociali e non solo, questa componente di soggettività accomuna tutti i procedimenti di analisi multivariata dei dati. È infatti tipico dei procedimenti di conoscenza scientifica un processo di riduzione e di semplificazione controllata delle informazioni disponibili, per favorire la comprensione dei fenomeni (Massart *et al.*, 1988).

Le fasi del processo di analisi dei gruppi sono le seguenti:

- Scelta delle unità di osservazione;
- scelta delle variabili; omogeneizzazione delle scale di misura;
- scelta della misura di similarità o di diversità tra le unità statistiche (misure di somiglianza e di diversità);
- scelta del numero dei gruppi;
- scelta dell'algoritmo di classificazione: cluster analysis gerarchica; cluster analysis non gerarchica;
- interpretazione dei risultati ottenuti.