Analisi di agglomerazione (cluster analysis)

Lo scopo dell'analisi di agglomerazione è di suddividere un campione multivariato, come quello rappresentato dalla tabella dei rilievi fitosociologici (matrice dei dati) in gruppi di composizione omogenea. Nel nostro caso le variabili sono le specie (righe della matrice) e i campioni sono i rilievi (colonne della matrice).

Esamineremo i due aspetti di base delle tecniche di agglomerazione:

le *misure* di similitudine e dissimilarità tra campioni

i *criteri* con cui calcolare le similitudini o le dissimilarità tra i gruppi.

1. Misure di similitudine e dissimilarità tra campioni multivariati

Le misure sono calcolate su coppie di campioni: *a* e *b* sono i numeri di specie dei campioni a confronto e *c* il numero di specie comuni; *n* è il numero totale dei campioni.

Di seguito, le tre funzioni più usate come misure relative a campioni di vegetazione

Misure di similitudine (le più usate):

Indice di *Jaccard:* $J = c/(a+b)$ ne deriva che $0 \leq J \leq 1$

Indice di *Sorensen* $S = 2c/(a+b)$ ne deriva che $0 \leq S \leq 1$

I campioni vengono considerati a coppie, per tutte le coppie possibili: *a* è il numero delle specie nel primo campione (A) e *b* il numero delle specie nel secondo campione (B); *c* è il numero delle specie in comune.

Misura di dissimilarità (la più usata)

Distanza Euclidea: $ED = \sqrt{\sum(p_i - q_i)^2}$

dove p_i e q_i sono le quantità della specie i-esima nei campioni a confronto, considerati a coppia, per tutte le specie e per tutte le coppie possibili di campioni. I valori di ED possono essere $0 \leq ED \leq \sqrt{n}$, dove n è il numero totale delle specie considerate nell'analisi.

Mediante queste misure vengono costruite matrici di similitudine o di dissimilarità tra campioni.

2. Criteri per l'agglomerazione

I criteri usati sono:

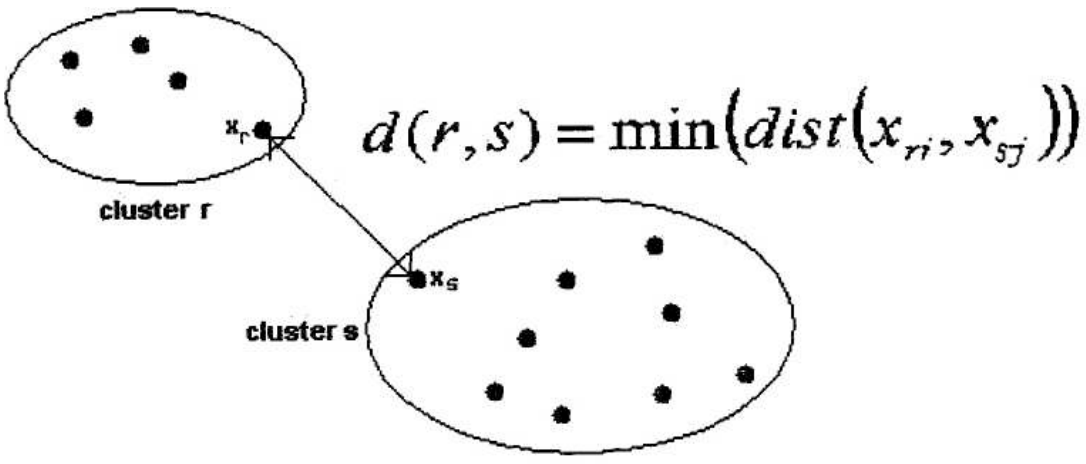
- metodo del legame singolo (del vicino più prossimo; *single linkage* o *nearest neighbor)*
- metodo del legame medio *(average linkage between groups)*
- metodo del legame completo (del vicino più lontano; *complete linkage; further neighbor)*

1.1 Metodo del legame singolo

Se abbiamo costruito una matrice di dissimilarità, usando la Distanza Euclidea, l'agglomerazione consiste nel selezionare inizialmente la distanza più piccola tra i campioni. Il primo gruppo che si forma sarà la coppia dei campioni a minore dissimilarità. Il secondo passo sarà quello di considerare la seconda più piccola distanza tra uno qualsiasi dei campioni del primo gruppo e quelli della matrice. Analogamente si procede con il terzo passo.

Se la procedura viene rappresentata graficamente in un dendrogramma,vengono visualizzate le relazioni di dissimilarità tra i gruppi e viene ricostruita la sequenza di agglomerazione.

Il metodo del legame singolo tende ad evidenziare catene tra i campioni (continuità) piuttosto che gruppi ben distinti (discontinuità). I gruppi che si formano per primi tendono ad accrescersi per aggiunta di campioni singoli.



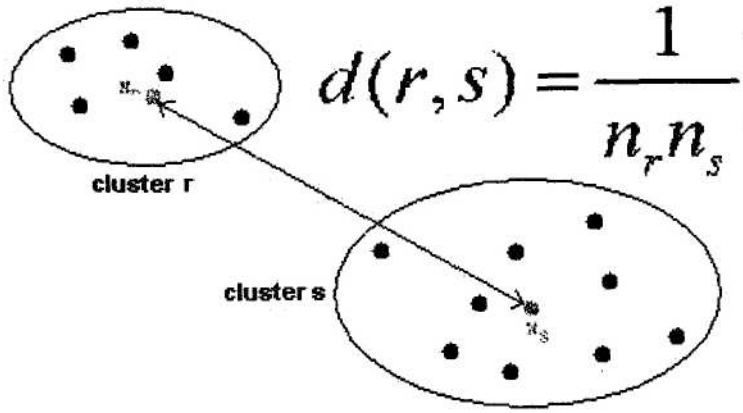
2.2 Metodo del legame medio

Un buon criterio di agglomerazione per evidenziare gruppi distinti è quello del legame medio.

Il primo passo è uguale a quello della procedura del legame singolo. Il secondo passo consiste invece nel sostituire ai due campioni del primo gruppo un campione "sintetico" comprendente le specie dei due campioni mediando i loro valori quantitativi. Questo campione sintetico è considerato nel ricalcolo di una nuova matrice (di dissimilarità, nel nostro caso). A questa nuova matrice di dissimilarità si applica la

procedura del primo passo: si sceglie una nuova coppia di campioni con la minore dissimilarità.

Si calcola anche in questo caso un campione sintetico, come nel passo precedente. Via via che si procede i campioni si riducono progressivamente (ad ogni passaggio i campioni si riducono di una unità) e i "campioni sintetici" vengono sempre considerati come "veri" campioni. Il dendrogramma mostrerà il risultato e la sequenza di agglomerazione.



$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist\left(x_{ri}, x_{sj}\right)$$

2.3 Metodo del legame completo

Se abbiamo costruito una matrice di dissimilarità, usando la Distanza Euclidea, l'agglomerazione consiste nel selezionare inizialmente la distanza più piccola tra i campioni. Il primo gruppo che si forma sarà la coppia dei campioni a minore dissimilarità. Successivamente, il livello di dissimilarità al quale un campione viene legato all'altro gruppo è però uguale alla maggiore dissimilarità tra il campione e uno qualsiasi dei campioni del gruppo. Si procede allo stesso modo negli altri passaggi.

Questo metodo è sconsigliabile quando è presente una grande quantità di "rumore" nel nostro set di campioni (molte specie con frequenze molto basse).