TECNICHE DI ANALISI DEI DATI IN ECOLOGIA

Michele Scardi

Dipartimento di Biologia
Università di Roma “Tor Vergata”
Via della Ricerca Scientifica
00133 Roma

e-mail: mscardi@mclink.it
home page: http://www.mare-net.com/mscardi

Versione 1.5, gennaio 2009

Tavola dei contenuti.

1. Introduzione.

Gli insiemi di dati che vengono abitualmente prodotti nell'ambito delle attività di ricerca e/o monitoraggio svolte su ecosistemi marini o terrestri hanno la caratteristica di essere quasi sempre di tipo multivariato. E' molto raro, infatti, che nel corso di una campagna di campionamento si focalizzi l'attenzione su una sola variabile, anche nei casi in cui le operazioni di campo vengono svolte a fini estremamente specifici.

Le ragioni di ciò sono molteplici, ma certamente un ruolo primario è quello giocato dall'elevato costo delle operazioni di campo e dalla natura imperfetta e incompleta delle nostre effettive conoscenze ecologiche. Se il primo motivo spinge ad una acquisizione "a tappeto" di tutti i dati rilevabili su una singola stazione, il secondo è responsabile della natura tipicamente ridondante dei piani di campionamento per ciò che riguarda il numero di variabili di cui si prevede la misura. Infatti, poichè non sono note *a priori* le eventuali correlazioni fra di esse, non è possibile definire un filtro a monte delle operazioni di campo.

In generale un insieme tipico di dati ecologici può essere rappresentato in forma matriciale. Le righe della matrice corrispondono al vettore di tutte le misure previste per un campione, per una osservazione o per un oggetto. Al contrario, i vettori-colonna di questa stessa matrice conterranno l'insieme di tutti i valori relativi ad ogni singolo descrittore fra quelli previsti. Evidentemente è del tutto plausibile che si verifichi il caso opposto e che le righe corrispondano ai vettori-descrittore. In linea di massima, comunque, si tende ad organizzare i dati, per motivi pratici e, in qualche caso, anche computazionali, in modo da avere un numero di righe maggiore del numero delle colonne.

Ai fini della comprensione di quanto esposto nei capitoli che seguono, si tenga presente che si è preferito il termine *descrittore* a quello, più limitativo, di *variabile*. Analogamente, i termini *osservazione* ed *oggetto* sono stati preferiti ad altri più specifici, come *campione*, *prelievo*, *misura*, etc..

La maggior parte delle tecniche di analisi dei dati presentate in questo contesto hanno essenzialmente finalità descrittive e di sintesi dell'informazione. Solo in alcuni casi, infatti, è possibile ed utile, nel campo della ricerca ecologica, ricorrere ad una impostazione basata su test formali di ipotesi. La maggior difficoltà, in questo senso, sta nel fatto che i dati ecologici assai raramente possono soddisfare tutte le assunzioni necessarie a questo tipo di approccio.

D'altra parte, lo scopo dell'analisi dei dati in Ecologia è essenzialmente quello di fornire un supporto ad un percorso conoscitivo che si basa in larga misura sull'osservazione piuttosto che sulla sperimentazione in senso stretto: dunque, la possibilità di formulare delle inferenze informali è molto spesso più utile della possibilità di testare ipotesi formali.

Le tecniche di analisi che vengono presentate nei capitoli seguenti costituiscono un sottoinsieme rappresentativo di quello, più vasto, che raccoglie tutti gli strumenti dell'Ecologia Numerica. In molti casi l'esposizione fa riferimento a problemi correnti nel campo della ricerca ecologica, piuttosto che ad un eccesivo formalismo. Inoltre, si è preferito omettere la descrizione di tutte le possibili varianti delle singole tecniche, poichè la scelta dell'alternativa più corretta in funzione del problema da trattare costituisce un argomento di complessità superiore a quello compatibile con le finalità di queste pagine. Per lo stesso motivo, si è preferito non affrontare il problema della trasformazione dei dati.

Per quanto riguarda questi aspetti ed altri ancora fra quelli che non vengono trattati, si rimanda il lettore che desideri un approfondimento a testi specifici di maggior respiro (Davis, 1986; Legendre & Legendre, 1983, 1998; Pielou, 1984; etc.).

Infine, va sottolineato il fatto che queste pagine sono state assemblate raccogliendo ed adattando materiale prodotto in occasione di corsi e seminari dal 1986 ad oggi, senza però essere mai sottoposte ad una approfondita revisione. Al di là della possibilità di incontrare piccoli errori, ciò implica che lo spazio dedicato ai diversi argomenti non ne rispecchia necessariamente l'effettiva rilevanza.

2. Misure di distanza e di similarità.

2.1. Coefficienti di similarità.

2.1.1. Generalità.

I coefficienti di similarità forniscono una misura del grado di somiglianza fra osservazioni, campioni, oggetti o altre entità ed hanno valori che variano nell'intervallo compreso fra 0 ed 1. Tali valori limite corrispondono, rispettivamente, al caso di osservazioni del tutto disgiunte, prive di elementi comuni, ed al caso di osservazioni che soddisfano pienamente il criterio utilizzato per misurare la similarità (il che non implica che si tratti di osservazioni quantitativamente identiche fra loro).

Fra i molti coefficienti disponibili una importante distinzione è quella che deve essere fatta fra coefficienti simmetrici e coefficienti asimmetrici. All'interno di un vettore di misure relativo ad una osservazione può accadere che per uno o più descrittori siano stati rilevati dei valori nulli. E' evidente che in alcuni casi tali valori corrispondono ad un dato certo, almeno nei limiti dell'errore proprio dei metodi di campionamento e di determinazione (es. un certo inquinante è assente), mentre in altri casi lo zero indica piuttosto l'assenza di informazione (es. una certa specie non è stata rinvenuta in un certo campione). Nel primo caso la scelta dovrà cadere su un coefficiente simmetrico, ai fini del cui calcolo i dati nulli hanno il medesimo valore comparativo degli altri, mentre nel secondo caso dovranno essere utilizzati coefficienti asimmetrici, in modo tale da evitare di definire una elevata similarità sulla base di informazioni non certe (quale ad esempio, la simultanea assenza di un elevato numero di specie in due stazioni che hanno poche o nessuna specie in comune).

Nel seguito di questo capitolo vengono presentati alcuni coefficienti di similarità, scelti fra quelli il cui impiego in campo ecologico è più frequente. E' evidente che possono esistere dei casi specifici in cui un altro coefficiente, non compreso fra quelli descritti in questo contesto, potrebbe risultare più adatto ad affrontare una particolare

problematica, ma è bene sottolineare il fatto che la scelta di un coefficiente di similarità rappresenta comunque, in qualche misura, un passo arbitrario in una procedura di analisi. Proprio per questo motivo è consigliabile affinare le proprie esperienze su un insieme relativamente piccolo di coefficienti, piuttosto che spaziare su tutta la gamma di quelli noti senza una motivazione più che solida.

2.1.2. Coefficienti binari.

Ai fini della descrizione dei coefficienti binari è utile definire i quattro casi possibili nel confronto fra gli elementi corrispondenti di due vettori-osservazione. Tale definizione può essere rappresentata in forma schematica come segue:

		Osservazione *j* **1**	Osservazione *j* **0**
Osservazione *k*	**1**	*a*	*b*
	0	*c*	*d*

$p = a + b + c + d$

Dunque, con ***a*** si indica il numero di elementi in comune fra due vettori-osservazione, mentre con ***d*** si indica il numero di elementi nulli (assenti) in entrambi e con ***b*** e ***c*** il numero di elementi non nulli (presenti) esclusivamente nell'uno e nell'altro vettore. Con ***p***, infine, si identifica la somma dei quattro valori appena citati, cioè il numero totale di elementi (descrittori) dei vettori-osservazione.

Fra i coefficienti binari di tipo simmetrico più adatti ad un impiego in campo ecologico possono essere citati il coefficiente di concordanza semplice (Sokal & Michener, 1958) e due coefficienti da esso derivati. Il

coefficiente di concordanza semplice rappresenta il rapporto fra il numero di elementi che hanno il medesimo valore (e quindi concordanti) ed il numero totale di elementi:

$$S_{jk} = \frac{a+d}{p}$$

Poichè questo coefficiente non distingue fra casi di concordanza su valori 1 e su valori 0 (rispettivamente co-presenze e co-assenze), il criterio da utilizzare per la codifica binaria dell'informazione può essere considerato del tutto libero.

Il coefficiente proposto da Rogers & Tanimoto (1960) rappresenta una variante di quello di concordanza semplice poichè rispetto a quest'ultimo attribuisce un peso doppio alle discordanze:

$$S_{jk} = \frac{a+d}{a+2b+2c+d}$$

Una variazione sullo stesso tema, ma concettualmente opposta, è indicata da Sokal & Sneath (1963) ed attribuisce un peso doppio alle concordanze:

$$S_{jk} = \frac{2a+2d}{2a+b+c+2d}$$

Fra i coefficienti asimmetrici, il cui uso è da preferirsi quando si ha a che fare con liste di specie derivate da osservazioni di campo in cui la rappresentatività del campione non è del tutto certa, alcuni fra quelli più frequentemente utilizzati costituiscono la diretta trasposizione di quelli fin qui descritti al caso in cui lo zero si deve intendere come mancanza di informazione piuttosto che come assenza o come valore nullo di un descrittore.

Infatti, il coefficiente di Jaccard (1900, 1901, 1908) è simile a quello di concordanza semplice, ma non tiene conto delle assenze:

$$S_{jk} = \frac{a}{a+b+c}$$

e corrisponde quindi al rapporto fra concordanze e numero di elementi non nulli dei vettori-osservazione.

Il coefficiente di Sørensen (1948) è stato probabilmente il più utilizzato in Ecologia Marina ed è strettamente imparentato con il coefficiente simmetrico di Sokal & Sneath (1963) appena descritto:

$$S_{jk} = \frac{2a}{2a + b + c}$$

Si noti come, rispetto al coefficiente di Jaccard, il coefficiente di Sørensen attribuisce un peso doppio alle concordanze. Nel caso del confronto fra liste di specie, che rappresenta il tipico ambito di applicazione di queste misure di similarità, esso enfatizza il criterio di asimmetricità assegnando un peso doppio ai casi di co-presenza. Questi ultimi rappresentano, come è evidente, i soli casi certi di concordanza a causa della natura aleatoria del dato di assenza, che spesso è dovuto al sottodimensionamento del campione prelevato.

E' interessante rilevare che Sokal & Sneath (1963) propongono una versione asimmetrica anche del terzo dei coefficienti simmetrici precedentemente descritti, quello di Rogers & Tanimoto:

$$S_{jk} = \frac{a}{a + 2b + 2c}$$

Tuttavia, l'uso di questo coefficiente è poco interessante, per un motivo esattamente opposto a quello precedentemente esposto a proposito del coefficiente di Sørensen. Infatti, non sembra giustificata la scelta di un coefficiente asimmetrico se poi si attribuisce ai casi di discordanza (influenzati dalle assenze) un peso doppio rispetto ai casi di concordanza, che sono determinati con certezza.

2.1.3. Coefficienti semi-quantitativi e quantitativi.

I coefficienti di similarità basati su dati quantitativi veri e propri non sono, in realtà, molto numerosi, poichè nei casi in cui è necessario trattare questo tipo di dati molto spesso si preferisce l'uso di una misura

di distanza. Esistono, comunque, alcuni coefficienti sicuramente interessanti, i quali meritano una breve descrizione.

Il trattamento di dati di tipo semi-quantitativo (es. punteggi arbitrari) può essere affrontato nella maggior parte dei casi utilizzando i coefficienti che vengono descritti in questo paragrafo, mentre per ciò che riguarda insiemi di dati ai cui descrittori è applicata una codifica di tipo non ordinale (es. colore, forma, etc.) si deve considerare l'opportunità di tradurre l'informazione disponibile in forma binaria, utilizzando poi un coefficiente binario simmetrico. In alternativa, è possibile applicare il coefficiente di concordanza semplice, descritto nel paragafo precedente, ed inteso come rapporto fra numero di concordanze (uguale codifica di un descrittore in due osservazioni) e numero di descrittori.

Una interessante possibilità è quella offerta dal coefficiente di Gower (1971), che è formulato in modo tale da trattare ciascun descrittore di un insieme multivariato in maniera ottimale in rapporto alla sua natura. Questo coefficiente corrisponde alla media delle similarità calcolate individulamente per ogni descrittore disponibile in entrambe le osservazioni. Ciò è possibile grazie all'uso di una variabile ausiliaria, detta delta di Kronecker, che assume un valore unitario nel caso in cui i dati sono disponibili ed un valore nullo in caso contrario. E' evidente che questo coefficiente si presta assai bene al trattamento di insiemi di dati in cui uno o più valori risultano mancanti. La formulazione del coefficiente di Gower è la seguente:

$$S_{jk} = \frac{\sum_{i=1}^{p} w_i s_i}{\sum_{i=1}^{p} w_i}$$

dove w_i ed s_i sono rispettivamente il delta di Kronecker e la similarità relativi all'*i*-mo descrittore per le due osservazioni considerate.

La formulazione delle similarità per descrittore *s* può essere variata a piacimento in funzione della natura dei dati disponibili e del contesto da cui sono estratti, ma, in origine, l'Autore proponeva quanto segue:

- per i descrittori binari s_i=1 nei casi di concordanza e s_i=0 altrimenti, con il caso della concordanza da doppio zero che viene trattato in accordo con il significato dello zero (valore nullo o mancanza di informazione)
- per i descrittori semi-quantitativi ordinali e quantitativi si assume s_i=1-$|x_{ij}-x_{ik}|\ R_i^{-1}$

dove x_{ij} e x_{ik} sono i valori dell'*i*-mo descrittore nelle osservazioni *j* e *k* ed R_i è l'intervallo di variazione dell'*i*-mo descrittore nell'insieme di osservazioni disponibili o nella popolazione da cui sono estratte queste ultime.

Per ciò che riguarda i coefficienti di tipo asimmetrico va segnalata la possibilità di applicare, in forma modificata, coefficienti già descritti. Si consideri, ad esempio la possibilità di trattare insiemi di dati semi-quantitativi esprimendo la similarità come il rapporto fra il numero di descrittori in cui si osserva concordanza ed il numero totale di descrittori diminuito del numero di doppi zeri: la similarità che si ottiene, in caso di codifica binaria, è esattamente quella di Jaccard.

Il coefficiente di Steinhaus (Motyka, 1947) è legato da una analoga relazione al coefficente binario di Sørensen ed è noto, se moltiplicato per 100, anche come "similarità percentuale":

$$S_{jk} = \frac{2\sum_{i=1}^{p} \min(x_{ij}, x_{ik})}{\sum_{i=1}^{p} x_{ij} + x_{ik}}$$

Il complemento a uno del coefficiente di Steinhaus, ovvero la dissimilarità di Steinhaus coincide con la distanza di Bray-Curtis, che è molto più comune nelle applicazioni ecologiche.

Il coefficiente di Kulczynski (1928) ha una formulazione abbastanza simile e corrisponde alla media dei rapporti fra somma dei minimi e totale per le due osservazioni considerate:

$$S_{jk} = \frac{1}{2}\left(\frac{\sum_{i=1}^{p} \min(x_{ij}, x_{ik})}{\sum_{i=1}^{p} x_{ij}} + \frac{\sum_{i=1}^{p} \min(x_{ij}, x_{ik})}{\sum_{i=1}^{p} x_{ik}} \right)$$

Una ulteriore ed interessante variazione è quella rappresentata dal coefficiente di Rudjichka (Goodall, 1978), che, espresso senza essere trasformato in percentuale, ha la seguente formulazione:

$$S_{jk} = \frac{\sum_{i=1}^{p} \min(x_{ij}, x_{ik})}{\sum_{i=1}^{p} \max(x_{ij}, x_{ik})}$$

Il pregio di tale coefficiente sta nel fatto che il suo complemento all'unità, a differenza di quanto avviene per i due coefficienti descritti in precedenza, corrisponde ad una misura di distanza di tipo metrico.

Sia il coefficiente di Kulczynski, sia quello di Rudjichka, sono di tipo asimmetrico e si prestano a trattare dati quantitativi anche in forma non normalizzata.

2.2. Coefficienti di distanza.

2.2.1. Generalità.

I coefficienti di distanza forniscono una misura del grado di associazione fra due osservazioni, restituendo un valore nullo per osservazioni identiche ed un valore variabile da coefficiente a coefficiente per osservazioni totalmente differenti.

Le misure di similarità possono essere trasformate in distanza semplicemente prendendone il complemento a 1. In questo caso, tuttavia, al termine *distanza* si preferisce il termine *dissimilarità.* La distinzione non è di tipo esclusivamente formale, poichè molte misure di dissimilarità non godono delle proprietà metriche, le quali, se

soddisfatte, consentono di ordinare le osservazioni in uno spazio, per l'appunto, di tipo metrico.

Le proprietà che devono essere soddisfatte perchè un coefficiente di distanza o dissimilarità sia di tipo metrico sono le seguenti:

1. $D_{jk}=0$ se $j=k$;
2. $D_{jk}>0$ se $j\neq k$;
3. $D_{jk}=D_{kj}$;
4. $D_{jk}+D_{kh}\geq D_{jh}$ (assioma della diseguaglianza triangolare).

In generale è la quarta ed ultima proprietà quella che risulta discriminante ed il fatto che sia o meno soddisfatta distingue le misure metriche da quelle cosiddette semimetriche. In questo contesto, ai fini di una maggiore chiarezza, sarà utilizzato il termine di distanza solo per i coefficienti che soddisfano le proprietà metriche, mentre sarà comunque preferito il termine di dissimilarità per quelli che sono derivati da misure di similarità.

2.2.2. Distanze.

I coefficienti di distanza sono stati sviluppati per trattare dati di tipo quantitativo e, con poche eccezioni, trattano lo zero come una misura e non come una mancanza di informazione.

La più familiare fra le misure di distanza è certamente quella euclidea, che corrisponde esattamente a quella che si può calcolare o misurare nello spazio fra due oggetti fisici:

$$D_{jk} = \sqrt{\sum_{i=1}^{p} (x_{ij} - x_{ik})^2}$$

E' importante rilevare il fatto che il quadrato della distanza euclidea, che non di rado viene utilizzato al posto di quest'ultima, è una semimetrica.

E' evidente che la scala dei singoli descrittori è molto influente nel determinare una distanza euclidea fra due osservazioni. E' dunque

necessario riservare questa scelta ai casi in cui i descrittori sono dimensionalmente omogenei o a quelli in cui essi vengono centrati e standardizzati, al fine di eliminare l'effetto di eventuali differenze di scala.

Proprio al fine di ovviare a questo inconveniente Orloci (1967) propone di calcolare la distanza euclidea dopo aver normalizzato i vettori-osservazione in modo tale che la loro lunghezza sia unitaria. Questa distanza è detta "della corda" perchè la misura che si ottiene è proprio quella della corda che unisce due punti-osservazione all'interno di una ipersfera di raggio unitario. Questa distanza può anche essere calcolata direttamente dai dati non normalizzati utilizzando la seguente formulazione:

$$D_{jk} = \sqrt{2\left(1 - \frac{\sum_{i=1}^{p} x_{ij} x_{ik}}{\sqrt{\sum_{i=1}^{p} x_{ij}^2 \sum_{i=1}^{p} x_{ik}^2}}\right)}$$

La distanza della corda varia da 0, per due vettori identici per profilo, cioè proporzionali fra loro, a $p^{1/2}$, dove p è il numero dei descrittori.

Una soluzione molto flessibile è quella costuita dalla metrica di Minkowski:

$$D_{jk} = \sqrt[r]{\left(\sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|^r\right)}$$

dove r può essere assegnato in maniera teoricamente arbitraria. In realtà il caso r=2 corrisponde ad una distanza euclidea ed un valore di r maggiore di questo, in generale, non è desiderabile per non enfatizzare l'effetto della diversa scala dei descrittori.

Più interessanti sono i valori di r inferiori a questa soglia e, fra questi, un caso particolare è quello che si verifica per r=1. In questo caso la distanza che si ottiene è nota come metrica di Manhattan:

$$D_{jk} = \sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|$$

Il nome di questa misura di distanza è dovuto al fatto che essa misura la distanza fra due punti in un piano come la somma della distanza in ascissa e di quella in ordinata. Quest'ultima corrisponde al percorso più breve che unisce due punti muovendosi in una città le cui strade si incrociano ad angolo retto, come avviene, per l'appunto, a Manhattan.

La metrica di Manhattan presenta gli stessi problemi legati all'influenza della scala dei descrittori di cui si è detto a proposito della metrica euclidea. Una delle varianti che, laddove necessario, la correggono in questo senso è quella proposta da Lance & Williams (1966) con il nome di metrica di Canberra:

$$D_{ij} = \sum_{i=1}^{p} \frac{\left| x_{ij} - x_{ik} \right|}{\left(x_{ij} + x_{ik} \right)}$$

I doppi zeri, se presenti, devono essere esclusi dal calcolo per evitare problemi di indeterminazione. Pur senza normalizzare i dati, questa distanza assegna alla differenza fra i valori che un descrittore assume in due osservazioni un peso inversamente proporzionale alla somma dei valori stessi: dunque, la medesima differenza ha un peso maggiore se è osservata fra due valori piccoli. Uno degli inconvenienti di questa soluzione, comunque, è costituito dal fatto che, se uno dei due valori relativi ad un dato descrittore è uguale a zero, allora il contributo alla distanza totale sarà comunque pari a 1, cioè il massimo possibile. La metrica di Canberra, dunque, si presta meglio a trattare serie di dati in cui esista eterogeneità di scala fra i descrittori senza, però, che siano presenti molti valori nulli.

Una ulteriore variante della metrica di Manhattan è quella proposta da Czekanowski (1909) come "differenza media dei descrittori":

$$D_{jk} = \frac{1}{p} \sum_{i=1}^{p} \left| x_{ij} - x_{ik} \right|$$

Questa misura di distanza si presta all'esclusione dei casi in cui si osserva un doppio zero, laddove ciò sia necessario, ma risente comunque dell'eventuale eterogeneità di scala dei descrittori.

Infine, un coefficiente utilizzato una certa frequenza in applicazioni ecologiche è quello di Bray-Curtis (Bray & Curtis, 1957). Se ***s*** è il numero dei taxa presenti, esso si ottiene come:

$$D_{jk} = \frac{\sum_{i=1}^{s} \left| x_{ij} - x_{ik} \right|}{\sum_{i=1}^{s} \left(x_{ij} + x_{ik} \right)}$$

2.2.3. Dissimilarità metriche.

Come già accennato in precedenza, i coefficienti di similarità possono essere convertiti in misure di distanza o, più propriamente, di dissimilarità. Ciò si effettua semplicemente considerandone il complemento ad 1 (cioè: $D_{jk} = 1 - S_{jk}$).

Non tutte le dissimilarità, però, godono di proprietà metriche, poichè sono molte quelle per cui l'assioma della diseguaglianza triangolare non è verificato: in questo caso si usa la definizione di semimetrica o pseudometrica. Sono dissimilarità semimetriche, ad esempio, quelle derivate dai coefficienti di similarità di Sørensen, di Sokal & Sneath, di Steinhaus e di Kulczynski.

La dissimilarità derivata dal coefficiente di Rudjichka, al contrario, è di tipo metrico, così come quella derivata dal coefficiente di Jaccard, che è nota anche come distanza di Marczewski-Steinhaus (Orloci, 1978) e che può essere calcolata direttamente come segue:

$$D_{jk} = 1 - \frac{a}{a+b+c} = \frac{b+c}{a+b+c}$$

Anche la similarità di Gower, infine, può essere trasformata in una dissimilarità metrica, così come quella di Rogers & Tanimoto (sia nella forma simmetrica, sia in quella asimmetrica) e come l'indice di concordanza semplice.

Il principale vantaggio delle dissimilarità metriche è costituito dal fatto che esse si comportano esattamente come delle misure di distanza in uno spazio euclideo. Ciò rende più intuitiva la loro applicazione e rende possibile l'applicazione di alcune tecniche di analisi (es. Analisi delle Coordinate Principali, vedi §4.2.) che non possono essere applicate alle semimetriche.

2.3. Coefficienti di dipendenza.

Così come i coefficienti di similarità e di distanza descrivono le relazioni che esistono fra le osservazioni, i coefficienti di dipendenza sintetizzano quelle che esistono fra descrittori.

Esistono diversi tipi di coefficienti di dipendenza, fra i quali è possibile scegliere quello più adatto alla natura dei dati da trattare. Un caso particolare è quello delle relazioni fra specie animali o vegetali, che possono essere rappresentate mediante dei coefficienti di associazione.

A differenza delle misure di similarità e distanza, comunque, i coefficienti di dipendenza possono essere sottoposti a test statistici, sempre che la distribuzione dei descrittori studiati lo consenta. In generale, tali tests hanno come fine la verifica dell'ipotesi nulla di indipendenza fra i descrittori.

Per il trattamento di dati quantitativi i coefficienti di dipendenza di gran lunga più utilizzati sono certamente la covarianza e la correlazione di Pearson.

La covarianza fra due descrittori si può ottenere, sulla base di due vettori di *n* osservazioni, come:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

Si noti come il calcolo della covarianza richiede che sia disponibile un parametro statistico della distribuzione di frequenza dei descrittori,

cioè la media. E' evidente, inoltre, che nel caso particolare che si determina se $j=k$ la formula appena riportata restituisce la varianza di un descrittore stimata su n osservazioni. In altre parole, $s_{jj}=s_j^2$. Va sottolineato il fatto che la sommatoria degli scarti si divide per n anzichè per n-1 nel caso in cui la coverianza sia riferita ad una popolazione (in senso statistico) invece che ad un campione.

Il coefficiente di correlazione r di Pearson è strettamente legato alla covarianza ed esprime l'intensità della relazione lineare che lega due descrittori. Esso non è altro che una covarianza calcolata su dati standardizzati e può essere facilmente derivato, nel caso di dati non standardizzati, dalla covarianza e dalle varianze dei due descrittori:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_j^2 s_k^2}}$$

Ovviamente è anche possibile calcolare direttamente la correlazione r di Pearson fra due descrittori, partendo dai dati bruti**Error! Objects cannot be created from editing field codes.**:

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

Così come la covarianza, anche la correlazione r di Pearson è una misura parametrica di dipendenza, i cui parametri sono la media e la deviazione standard dei descrittori. Il coefficiente di correlazione r di Pearson varia da -1 a 1: questi limiti si ottengono per serie di dati esattamente proporzionali, rispettivamente in maniera inversa e diretta.

Il coefficiente di correlazione r di Pearson può essere sottoposto ad un test per verificare se esso differisce significativamente dallo zero. A questo fine si calcola la probabilità di ottenere un valore di r pari a quello osservato nel caso in cui i due descrittori siano totalmente

indipendenti fra loro e si considera significativa la correlazione se questa probabilità è sufficientemente piccola (es. P<0.05).

Per far ciò si utilizza il seguente rapporto, che è distribuito come un *t* di Student:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

La probabilità di ottenere un valore di *r* pari a quello osservato in assenza di correlazione lineare fra i descrittori è quella associata al valore di *t* ottenuto, con *n*-2 gradi di libertà.

Si tenga presente, comunque, che la non significatività della correlazione lineare non implica l'indipendenza dei descrittori, i quali possono essere legati da relazioni di ordine superiore.

Anche nel caso di descrittori semiquantitativi è possibile utilizzare dei coefficienti di dipendenza. In particolare, si presta molto bene a questo scopo il coefficiente di correlazione di rango *r'* (o ρ) di Spearman: questo coefficiente non-parametrico può essere applicato nel caso di relazioni di cui deve essere verificata la monotonicità, anche se di tipo non lineare. La "robustezza" della correlazione di rango in condizioni di non linearità delle relazioni fra descrittori, molto frequenti in Ecologia, è la caratteristica che rende particolarmente interessante l'applicazione di questo tipo di coefficiente.

Il coefficiente di correlazione *r'* di Spearman corrisponde esattamente ad un coefficiente di Pearson calcolato sui ranghi dei dati anzichè sui dati bruti. Esso può però essere ottenuto più direttamente come segue:

$$r'_{jk} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

dove *d* è la differenza fra il rango della *i*-ma osservazione per il descrittore *j* e quello per il descrittore *k*.

Se per entrambi i descrittori non esistono due o più osservazioni con il medesimo rango, allora il valore che si ottiene è identico a quello del coefficiente *r* di Pearson. Tuttavia, nel caso in cui l'informazione è di tipo semiquantitativo ed è codificata mediante un piccolo numero di punteggi è inevitabile che molte osservazioni abbiano lo stesso punteggio e quindi lo stesso rango. Ciò rende necessaria l'applicazione di una correzione che tenga conto del numero di casi assegnati per ciascun descrittore a ciascun rango. La formulazione del coefficiente *r'* di Spearman diventa allora:

$$r'_{jk} = \frac{2n^3 - 2n - \sum_{h=1}^{m}(q_{hj}^3 - q_{hj}) - \sum_{h=1}^{m}(q_{hk}^3 - q_{hk}) - 12\sum_{i=1}^{n} d_i^2}{2 \cdot \sqrt{\cdot n^3 - n - \sum_{h=1}^{m}(q_{hj}^3 - q_{hj})} \cdot \sqrt{n^3 - n - \sum_{h=1}^{m}(q_{hk}^3 - q_{hk})}}$$

dove, oltre a quanto descritto per la formulazione di base, *m* è il numero di ranghi e q_{hj} e q_{hk} sono il numero di osservazioni di rango *h* per il descrittore *j* e per quello *k*.

Per ciò che riguarda il test di significatività del coefficiente *r'* di Spearman è necessario fare riferimento a delle apposite tavole, poichè, malgrado le notevoli affinità con il coefficiente *r* di Pearson, non è possibile utilizzare il medesimo approccio. Infatti, la condizione di normalità della popolazione bivariata da cui sono estratti i campioni non è certamente soddisfatta nel caso di dati semiquantitativi.

Un caso particolare in cui è necessario disporre di un coefficiente di dipendenza è quello dello studio delle associazioni di specie. In questo caso i dati sono espressi tipicamente in forma binaria, poichè al centro dell'attenzione non sono i rapporti quantitativi, ma piuttosto la tendenza di più specie a ricorrere congiuntamente.

In questo contesto è possibile impiegare alcuni dei coefficienti di similarità asimmetrici già descritti a proposito dei dati binari. La scelta di coefficienti asimmetrici è motivata dal fatto che la co-assenza di specie non costituisce una informazione rilevante ai fini della definizione di eventuali associazioni.

In particolare, possono essere considerati dei coefficienti di dipendenza fra specie sia il coefficiente di Jaccard (cfr. Reyssac & Roux, 1972), sia quello di Sørensen, che in questo caso viene indicato con il nome di indice di coincidenza (Dice, 1945).

Un coefficiente messo a punto espressamente per lo studio di associazioni di specie è quello proposto da Fager & McGowan (1963):

$$S_{jk} = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2 \cdot \sqrt{a+c}} \qquad (c \geq b)$$

Si noti come il secondo termine rappresenta una correzione per impedire che le specie rare risultino fortemente associate: esso, infatti, diminuisce il valore del coefficiente di una quantità tanto maggiore quanto più è rara la specie più frequente fra le due esaminate.

3. Tecniche di clustering.

3.1. Note introduttive.

Una delle esigenze più comuni nella ricerca ecologica (e non) è quella di raggruppare gli oggetti appartenenti ad un insieme dato in modo tale da definire dei sottoinsiemi il più possibile omogenei. Per raggiungere questo risultato, identificando una *partizione*, cioè una collezione d'oggetti tale che ogni oggetto appartenga ad un solo sottoinsieme o *classe*, è necessario disporre di una procedura o di un algoritmo adatti alla natura dell'informazione disponibile, del problema da affrontare e degli oggetti stessi.

Le procedure di tipo soggettivo, in quest'ambito, hanno un ruolo molto più importante di quanto non si pensi comunemente. Basti considerare il fatto che è su un approccio di questo tipo, per quanto codificato in un quadro tassonomico di riferimento, che è basata una delle attività fondamentali della ricerca ecologica, cioè la classificazione degli organismi animali e vegetali. Inoltre, prima che gli algoritmi di classificazione oggi disponibili venissero sviluppati, cioè fino a tutti gli anni '50, il modo più sofisticato di ottenere una partizione di un insieme di oggetti (o osservazioni) multivariati consisteva nel rappresentarli nello spazio dei loro descrittori o in quello definito da due o più assi principali (cfr. cap. 4), ricercando manualmente gli insiemi di punti più omogenei.

Come appena accennato, gli algoritmi di classificazione sono tutti abbastanza recenti, ma, nonostante ciò, essi costituiscono un insieme tanto ricco quanto diversificato. Gli algoritmi, in linea di massima, possono essere suddivisi in due grandi gruppi: quelli di tipo gerarchico, in cui si procede tipicamente per aggregazione successiva di oggetti, e quelli di tipo non gerarchico, in cui si procede per divisione dell'insieme di oggetti originale o per successivi aggiustamenti di una prima partizione.

Alcuni Autori prefiscono utilizzare il termine *clustering* per indicare i soli metodi non gerarchici, riservando il termine *classificazione* per quelli gerarchici. In questa sede, comunque, sarà utilizzato

esclusivamente il primo termine, poichè esso è largamente utilizzato e compreso, indipendentemente dal contesto applicativo. La trattazione sarà focalizzata sul clustering di oggetti (o osservazioni), ma è evidente che in alcuni casi può essere interessante e/o necessario ottenere piuttosto una partizione di un insieme di descrittori.

E' importante sottolineare il fatto che una partizione ottenuta mediante un algoritmo di clustering è a tutti gli effetti un descrittore aggiuntivo (e sintetico) dell'insieme di oggetti in esame. L'appartenenza ad un cluster, infatti, se codificata in maniera appropriata, può essere utilizzata come una variabile di sintesi per ulteriori elaborazioni dell'informazione disponibile.

Infine, anche se non sarà trattato in questo contesto, è interessante ricordare l'esistenza di un approccio del tutto particolare ai problemi di clustering, il quale pur essendo molto ben adattato ai problemi più disparati, non si è ancora ritagliato uno spazio significativo nell'ambito della ricerca ecologica. L'approccio in questione è quello basato sul concetto di *fuzzy sets,* secondo cui l'appartenenza di un oggetto ad una classe (cioè ad un *fuzzy set*) non viene espressa in forma binaria, ma piuttosto in forma probabilistica. E' evidente che questo tipo di logica è molto più vicina a quella che tutti noi utilizziamo nella vita di tutti i giorni, quando ci riferiamo a categorie i cui limiti sono difficilmente definibili in maniera univoca, poichè sfumano le une nelle altre senza soluzione di continuità.

3.2. Clustering gerarchico.

3.2.1. Generalità.

Gli algoritmi di clustering gerarchico utilizzano una matrice di similarità (o distanza) fra gli oggetti come base per l'aggregazione di questi ultimi. E' importante sottolineare il fatto che la scelta del coefficiente di similarità (o distanza) risulta in molti casi addirittura più determinante di quella dell'algoritmo di clustering ai fini del conseguimento dei risultati desiderati. Tale scelta, dunque, deve essere

preceduta da una accurata esplorazione dell'informazione disponibile e da una chiara identificazione del tipo di relazione fra gli oggetti che si intende rappresentare.

I risultati di una procedura di clustering gerarchico possono essere reppresentati in diversi modi, anche se in prevalenza si preferisce utilizzare un *dendrogramma*. I legami orizzontali in un dendrogramma vengono chiamati *nodi*, mentre le linee verticali sono dette *internodi*. La distanza di un nodo dalla base del dendrogramma è proporzionale alla similarità (o distanza) fra i due oggetti o gruppi di oggetti di cui il nodo rappresenta la fusione. La similarità (o distanza) è di solito riportata su una scala al lato del dendrogramma. La disposizione relativa degli oggetti alla base del dendrogramma è vincolata solo in parte dalla struttura di quest'ultimo e, entro questi limiti, gli oggetti possono essere liberamente riarrangiati.

In molti casi è utile anche visualizzare l'andamento progressivo delle similarità (o distanze) a cui via via avvengono le fusioni fra oggetti o gruppi di oggetti. Questa rappresentazione è fornita dal diagramma di aggregazione, grazie al quale è possibile individuare facilmente le discontiuità più rilevanti incontrate nella procedura di clustering. Tali discontinuità, in molti casi, possono corrispondere a partizioni "naturali" dell'insieme di oggetti analizzati e costituiscono un utile riferimento laddove sia necessario ripartire questi ultimi in un certo numero di classi (es. se si usa la partizione ottenuta come un nuovo descrittore sintetico dell'insieme degli oggetti).

3.2.2. Algoritmo del legame singolo.

L'algoritmo del legame singolo (o *nearest-neighbor*) è certamente il più semplice fra quelli disponibili e deve il suo nome al fatto che la fusione fra due oggetti o gruppi di oggetti può avvenire se la distanza fra due oggetti non appartenenti allo stesso gruppo è la più bassa fra quelle possibili.

La procedura operativa, supponendo di lavorare su una matrice di distanza, è la seguente:

1. si individua il valore minimo nella matrice (con esclusione, ovviamente della diagonale) e si fondono i due oggetti corrispondenti in un primo gruppo;
2. si individua il valore minimo residuo, cioè escludendo le distanze intra-gruppo, e si fondono i due oggetti che corrispondono a tale valore o i due gruppi a cui essi appartengono;
3. si procede fino a quando tutti gli oggetti sono assegnati ad un unico gruppo.

Come si può notare, la procedura di clustering è elementare e non richiede alcun calcolo aggiuntivo al di là di quello della matrice di similarità o distanza.

L'algoritmo del legame singolo, tuttavia, non è molto utilizzato, soprattutto per la sua tendenza al concatenamento degli oggetti, che rende sempre più facile l'aggregazione di nuovi elementi man mano che un gruppo diventa più numeroso. Ciò è dovuto al fatto che basta un solo legame, cioè una sola coppia di oggetti effettivamente simili fra loro, a far fondere due gruppi: è evidente quanto più è grande il numero di oggetti che appartengono ad un gruppo, tanto più è probabile che almeno uno di essi possa costiutire un "ponte" verso un altro oggetto o un altro gruppo di oggetti. In altre parole, si può immaginare che l'algoritmo del legame singolo provochi una contrazione dello spazio di riferimento intorno ai gruppi proporzionale alla loro dimensione.

3.2.3. Algoritmo del legame completo.

Una soluzione affine a quella appena descritta da un punto di vista procedurale, ma completamente opposta per ciò che riguarda le regole di fusione dei gruppi è quella che prevede l'uso dell'algoritmo del legame completo (o *farthest-neighbor*), proposto da Sørensen (1948)

In questo caso, infatti, si ammette la fusione di due gruppi di oggetti soltanto se tutte le distanze fra coppie di oggetti non appartenenti allo stesso gruppo sono inferiori alla soglia che permetterebbe la fusione di un'altra coppia di gruppi.

Ciò garantisce una notevole omogeneità intra-gruppo, favorendo la formazione di gruppi a cui appartiene un numero non troppo variabile di oggetti, poichè quanto più un gruppo è numeroso, tanto più è difficile che esso sia nel sua interezza sufficientemente simile ad un altro gruppo. In contrapposizione a quanto avviene per l'algoritmo del legame singolo, in questo caso si verifica una dilatazione dello spazio di riferimento intorno ai gruppi già formati che è proporzionale alla loro dimensione.

Le particolari caratteristiche dell'algoritmo del legame completo rendono quest'approccio particolarmente adatto ad applicazioni ecologiche, soprattutto quando si vogliono individuare le discontinuità più rilevanti in un insieme di dati.

Il rovescio della medaglia, peraltro comune ad altri algoritmi di cui si tratterà nel seguito di questo capitolo, è costituito dalla possibilità di incontrare casi particolari in cui la scelta dell'aggregazione non è definibile in maniera univoca. Pur senza scendere nel dettaglio, si tenga conto che queste situazioni possono essere risolte applicando nell'ordine alcune semplici regole (Sørensen, 1948). In particolare, sarà privilegiata l'aggregazione che: (a) genera il gruppo più numeroso; (b) accelera la diminuzione del numero di gruppi e (c) massimizza la similarità media intra-gruppo.

3.2.4. Algoritmi di legame intermedio.

Fra i criteri estremi utilizzati negli algoritmi del legame singolo e di quello completo esistono, evidentemente, delle possibilità intermedie. Una di queste è costituita dall'algoritmo del legame proporzionale, che prevede la fusione di due gruppi se una certa frazione, definita ***a priori***, delle distanze inter-gruppo è inferiore o uguale alla soglia necessaria per definire una nuova partizione (Sneath, 1966).

Nel caso in cui tale frazione è fissata al 50%, il cirterio adottato è esattamente a metà strada fra quello del legame singolo e quello del legame completo. Se l'impiego di questi ultimi provocava rispettivamente una dilatazione ed una contrazione dello spazio di riferimento intorno ai gruppi già formati, il criterio intermedio utilizzato

dall'algoritmo del legame proporzionale può garantire un accettabile grado di conservazione delle proprietà metriche dello spazio di riferimento.

L'algoritmo del legame proporzionale non è l'unico nella famiglia degli algoritmi di legame intermedio, dei quali Sneath (1966) descrive tre ulteriori forme.

3.2.5. Algortimi di legame medio.

Un'altra importante categoria di algoritmi di clustering è quella basata su misure di distanza (o similarità) media fra i gruppi.

Le varianti possibili sono quattro e derivano dalla combinazione di due scelte: il peso attribuito ai gruppi, che può essere uguale o proporzionale alla loro dimensione, e la procedura di calcolo della distanza inter-gruppo, che può essere basata sulla media delle distanze fra singoli oggetti o sulla distanza fra i centroidi dei gruppi. La tabella che segue fornisce un quadro d'insieme delle varie possibilità:

	Distanza fra gruppi definita come:	
	distanza media fra gli oggetti	*distanza fra i centroidi dei gruppi*
pesi uguali	clustering medio (UPGMA)	clustering centroide (UPGMC)
pesi proporzionali	clustering a pesi proporzionali (WPGMA)	clustering mediano (WPGMC)

Per ciascun algoritmo di clustering è indicata fra parentesi la denominazione utilizzata da Sneath & Sokal (1973).

Il clustering medio (*unweighted arithmetic average clustering*, Rohlf, 1963) utilizza come criterio per la fusione di due gruppi di oggetti

la media aritmetica delle distanze (o delle similarità) fra tutti gli oggetti dei due gruppi e ad ogni oggetto viene attribuito lo stesso peso.

Il clustering a pesi proporzionali (*weighted arithmetic average clustering*, Sokal & Michener, 1958) prevede l'assegnazione di un medesimo peso a ciascuno dei due gruppi che devono essere fusi: ciò implica che gli oggetti del gruppo più numeroso avranno un peso individuale minore di quello degli oggetti del gruppo meno numeroso. La distanza fra i gruppi si calcola poi come una somma ponderata di tutte le distanze inter-oggetto. Questo approccio è specificamente adattato al caso in cui si analizzano contemporaneamente diversi insiemi "naturali" di oggetti: in questo caso, infatti, se uno di tali insiemi contiene un numero di oggetti relativamente piccolo, il risultato della procedura di clustering potrebbe essere fortemente influenzato dall'insieme più numeroso.

Il clustering centroide (Sokal & Michener, 1958; *unweighted centroid*, Sneath & Sokal, 1973) è caratterizzato dal fatto che, dopo che due oggetti o gruppi di oggetti sono stati fusi, essi vengono rappresentati dal loro centroide. Ciò può essere ottenuto in diversi modi, ma in genere è possibile sostituire le righe e le colonne relative agli oggetti che appartengono ad un gruppo appena formato con un vettore di valori, uguale per tutti gli oggetti, che si ottiene utilizzando una media, meglio se ponderata (Gower, 1967) delle similarità relative ai singoli oggetti. Il clustering centroide può dare luogo, talvolta, a delle "inversioni" nella struttura del dendrogramma, cioè si può verificare il caso che un nodo di ordine gerarchico superiore corrisponda ad un livello di distanza (o di similarità) minore (maggiore) di quello relativo ad un nodo di ordine gerarchico inferiore. Il fatto che l'algoritmo non garantisce la monotonicità del diagramma di aggregazione e del dendrogramma rende talvolta difficile l'interpretazione dei risultati, che in ogni caso devono essere utilizzati con cautela per la definizione di partizioni vere e proprie.

Così come nel caso del clustering medio, se si considera un insieme di dati in cui le osservazioni relative ad uno o più ambienti (popolazioni) particolari predominano numericamente su quelle relative ad ambienti (popolazioni) meno rappresentati, può essere necessario

introdurre una correzione, basata sull'assegnazione di un peso uguale a ciscun gruppo, ogni volta che si effettua una fusione.

Questa soluzione prende il nome di clustering mediano (Gower, 1967, *weighted centroid*, Sneath & Sokal, 1973) e sta al clustering centroide esattamente come il clustering a pesi proporzionali sta al clustering medio.

3.3. Clustering non gerarchico.

Le procedure di clustering non gerarchico prevedono la ripartizione degli oggetti in un numero dato di gruppi, generalmente sulla base di un criterio di massimizzazione della omogeneità intra-gruppo.

A differenza delle procedure gerarchiche non è generalmente necessario disporre di una matrice di distanza o similarità fra gli oggetti: questa caratteristica è estremamente importante quando si devono analizzare grandi insiemi di dati.

Uno dei metodi di clustering non gerarchico più interessanti è quello noto come algoritmo delle Nubi Dinamiche (*Nuées Dynamiques*, Diday, 1971). Esso può essere sintetizzato come segue:

1. si assegna a caso ciascuno degli *n* oggetti ad uno degli *m* gruppi richiesti;
2. si calcolano le coordinate degli *m* centroidi dei gruppi appena formati nello spazio dei *p* descrittori considerati;
3. si riassegna ciascun oggetto al gruppo il cui centroide è più vicino;
4. se nessun oggetto ha cambiato gruppo, la partizione ottenuta è quella finale, altrimenti si torna al punto 2.

Uno degli inconvenienti di questo metodo sta nel fatto che la partizione finale non è determinata in maniera univoca: è infatti possibile che diverse configurazioni di partenza (cioè diverse ripartizioni casuali degli oggetti fra gli *m* gruppi) convergano verso stati finali leggermente differenti, soprattutto in assenza di una partizione

"naturale" degli oggetti. In questo caso è possibile iterare un certo numero di volte la procedura e mantenere la partizione per cui l'omogeneità intra-gruppo è massima. Questa soluzione non è così inefficiente come può sembrare in prima analisi, poichè questo algoritmo è estremamente rapido anche nel caso in cui si trattano insiemi di dati di grandi dimensioni.

Come già accennato, pur non potendo essere considerata come una procedura di clustering in senso stretto, la definizione euristica di sottoinsiemi di oggetti basata sull'uso di tecniche di ordinamento (vedi cap. 4) rappresenta una prassi consolidata e, con le dovute cautele, non priva aspetti interessanti. In generale, comunque, questa soluzione deve essere adottata a fini prettamente descrittivi, sfruttando soprattutto la possibilità di individuare in maniera immediata il descrittore o il complesso di descrittori che hanno il maggior peso nel determinare le differenze osservate fra i gruppi di oggetti.

3.4. Clustering vincolato.

In molti casi i risultati di una procedura di clustering dipendono in maniera determinante dalla scelta dell'algoritmo e da quella di una misura di distanza o similarità appropriata. Imporre dei vincoli ad un algoritmo di clustering implica la definizione di un modello *a priori* che guida il processo di aggregazione degli oggetti, limitando lo spettro delle partizioni valide ad un sottoinsieme di quelle possibili.

L'uso di tecniche di clustering vincolato (Legendre & Legendre, 1984; Legendre *et al.,* 1985; Legendre, 1987) si rivela di particolare utilità quando è necessario identificare le discontinuità più rilevanti in una serie spaziale o temporale. Questo approccio consente infatti di individuare i gruppi di campioni che presentano il massimo grado di omogeneità al loro interno, scegliendoli esclusivamente fra quelli che formano delle sequenze cronologicamente ordinate o spazialmente connesse.

Le tecniche di clustering vincolato, in breve, prevedono la fusione di due oggetti o gruppi di oggetti in un unico gruppo solo se essi sono

contigui nel tempo o nello spazio ed al tempo stesso sono soddisfatte le condizioni di fusione previste dall'algoritmo di clustering prescelto. In particolare, comunque, si deve sottolineare il fatto che gli algoritmi di clustering che meglio si adattano all'applicazione di vincoli sono quelli di tipo gerarchico.

Il concetto di contiguità, per poter essere applicato all'algoritmo di clustering, deve essere opportunamente formalizzato. Se per le serie monodimensionali (es. temporali) ciò non costituisce un problema, per ciò che riguarda le serie bi- o multidimensionali (es. un insieme di stazioni in un'area geografica) è necessario stabilire un criterio che definisca il concetto di contiguità. Tale criterio può anche essere di natura assolutamente soggettiva, ma esistono delle soluzioni che hanno il pregio di poter definire in maniera oggettiva ed univoca una matrice di connessione fra gli oggetti.

Una di queste soluzioni è rappresentata dalle reti di Gabriel (Gabriel & Sokal, 1969). In questo caso si considerano connessi due punti *A* e *B* se nessun altro punto cade all'interno del cerchio il cui diametro è il segmento che unisce i punti *A* e *B*. In altre parole, dati tre punti qualsiasi *A*, *B* e *C*, si connettono *A* e *B* se non esiste nessun punto C di coordinate (x_C,y_C) tale che sia

$$x_C^2 - x_C x_A - x_C x_B + x_A x_B + y_C^2 - y_C y_A - y_C y_B + y_A y_B < 0$$

Questa formula è derivata dall'equazione di una circonferenza avente come diametro il segmento che congiunge i due punti A e B, rispettivamente di coordinate (x_A,y_A) e (x_B,y_B). Ricordando che l'equazione di una circonferenza con centro nel punto di coordinate (x_0,y_0) e di raggio *r* è

$$\sqrt{\left(x - x_0\right)^2 + \left(y - y_0\right)^2} = r$$

e sostituendo x_0 e y_0 con le coordinate del punto medio di AB ed *r* con la metà della distanza fra A e B, l'equazione della circonferenza di diametro AB sarà

$$\sqrt{\left[x-\frac{x_A+x_B}{2}\right]^2+\left[y-\frac{y_A+y_B}{2}\right]^2}-\frac{\sqrt{\left(x_A-x_B\right)^2+\left(y_A-y_B\right)^2}}{2}=0$$

Sostituendo alla *x* ed alla *y* di questa equazione le coordinate di un punto qualsiasi, il primo termine dell'equazione avrà valore negativo se il punto è interno alla circonferenza, positivo se il punto è esterno alla circonferenza ed ovviamente nullo (da cui l'equazione stessa) se il punto giace sulla circonferenza. Sostituendo *x* e *y* rispettivamante con con x_C e y_C e semplificando opportunamente, si ottiene appunto la diseguaglianza di cui alla pagina precedente, che è verificata in tutti i casi in cui un punto C è interno alla circonferenza di diametro AB.

E' evidente che in alcuni casi può essere necessario imporre delle correzioni alla matrice di connessione. Ciò si verifica, ad esempio, quando due punti, considerati connessi nello spazio bidimensionale delle loro coordinate, sono funzionalmente disgiunti a causa della presenza di accidenti geografici (es. due stazioni in mare possono essere le più vicine fra loro in linea d'aria, pur essendo separate da una penisola).

4. Tecniche di ordinamento.

4.1. Analisi delle Componenti Principali.

L'Analisi delle Componenti Principali è la tecnica di ordinamento più semplice, nel senso che essa opera esclusivamente una rotazione rigida degli assi dello spazio multidimensionale dei dati in modo tale da orientarli in maniera coerente con i pattern di dispersione dei dati stessi. Ciò consente di rappresentare un insieme di dati in maniera più efficace anche in un numero ridotto di dimensioni, cioè in un sistema di assi ortogonali (le Componenti Principali) definiti come combinazioni lineari dei descrittori originali. Inoltre, è possibile ottenere anche una rappresentazione delle relazioni fra i descrittori stessi e fra questi ultimi e le Componenti Principali.

Come per la maggior parte delle tecniche di ordinamento, anche per l'Analisi delle Componenti Principali è necessaria l'estrazione di autovalori ed autovettori da una matrice. Nel caso specifico si tratta in genere di una matrice di covarianza o di correlazione.

La procedura di calcolo prevede che i dati siano organizzati in una matrice $\mathbf{X}_{nxp}$, dove *n* sono le osservazioni e *p* i descrittori. Gli elementi della matrice **X** dei dati bruti vengono quindi centrati sulle *p* colonne (cioè sui descrittori), in modo da ottenere una matrice **Y** di eguale dimensione:

$$y_{ij} = x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Moltiplicando la matrice **Y** per la sua trasposta **Y'** e dividendo il prodotto per il numero di osservazioni *n* si ottiene la matrice **S**, che è la matrice di covarianza dell'insieme dei dati originali contenuti nella matrice **X**:

$$\mathbf{S} = \frac{1}{n}\mathbf{Y'Y}$$

In realtà, va sottolineato il fatto che, pur essendo prassi abbastanza consolidata, la divisione per *n* del prodotto **Y'Y**, non è strettamente necessaria, poichè tale operazione non ha alcuna influenza sul risultato finale dell'analisi.

Si procede quindi ad estrarre gli autovalori λ_k (k=1,2,...,m) e gli autovettori u_{jk} (j=1,2,...p; k=1,2,...,m) della matrice **S**. Si noti che il numero *m* di autovalori ed autovettori da estrarre può essere fissato a piacere: in molti casi è sufficiente considerare i primi 2 o 3 autovalori, in ordine decrescente.

Le coordinate f_{ij} (o *scores*) delle osservazioni riferite al nuovo sistema di assi, cioè alla Componenti Principali, si calcolano moltiplicando la matrice dei dati centrati **Y** per la matrice **U** degli autovettori (o *loadings*).

Da un punto di vista pratico, la rappresentazione delle osservazioni nello spazio definito dalle Componenti Principali (*modello di ordinamento*) si può effettuare in una, due o tre dimensioni. Tuttavia, la rappresentazione di gran lunga più comune è quella che si ottiene nel piano definito da una coppia di Componenti Principali.

La qualità della rappresentazione ottenuta si può valutare sulla base degli autovalori estratti. La percentuale di varianza spiegata dalla prima Componente Principale è pari al rapporto fra il primo autovalore e la traccia della matrice **S** e così via.

Infine, è possibile proiettare anche i descrittori nello spazio delle Componenti Principali. Le coordinate g_{jk} (j=1,2,...p; k=1,2,...,m) dei descrittori si ottengono moltiplicando ciascun autovettore per la radice quadrata dell'autovalore corrispondente:

$$g_{jk} = u_{jk}\sqrt{\lambda_k}$$

La proiezione dei descrittori deve essere interpretata in maniera leggermente diversa da quella delle osservazioni. In quest'ultimo caso, infatti, è la distanza fra i punti che consente valutare la somiglianza delle osservazioni, mentre nel primo caso sono piuttosto gli angoli che formano i vettori che identificano i punti-descrittore nello spazio delle Componenti Principali a rappresentare le relazioni fra i descrittori stessi.

Le correlazioni fra Componenti Principali e descrittori originali possono essere calcolate semplicemente dividendo per la deviazione standard s_j le coordinate g_{jk} dei descrittori.

L'Analisi delle Componenti Principali richiede, per una corretta applicazione, che i descrittori siano di tipo quantitativo e che la loro distribuzione sia di tipo normale. Inoltre, si assume che essi siano legati da relazioni lineari e che la matrice dei dati non contenga un numero eccesivo di zeri. Nel caso in cui i descrittori non siano dimensionalmente omogenei, infine, è opportuno effettuare l'analisi su una matrice di correlazione: ciò si ottiene standardizzando i dati bruti o, ancor più semplicemente, dividendo ogni elemento di **Y** per la deviazione standard s_j del descrittore corrispondente.

4.2. Analisi delle Coordinate Principali.

Nel campo della ricerca ecologica non sempre gli insiemi di dati posseggono le proprietà necessarie ad una corretta applicazione dell'Analisi delle Componenti Principali. Si consideri il caso tipico di una lista di specie osservate in un certo numero di campioni: spesso l'informazione è espressa mediante una codifica binaria (presenza/assenza) ed anche nei casi in cui sono disponibili le abbondanze, queste ultime non sono certamente distribuite in modo normale. Inoltre, il numero di zeri, cioè di assenze di specie dai campioni esaminati, è molto spesso addirittura superiore al numero dei valori non nulli. In questi casi esistono numerose misure di similarità e/o di distanza che si prestano a rappresentare al meglio le relazioni fra gli oggetti (campioni), come ampiamente discusso nel capitolo 2.

Un ordinamento degli oggetti nello spazio definito da una qualsiasi matrice di distanza o di similarità, a condizione che essa goda di tutte le proprietà metriche, può essere ottenuto mediante l'Analisi delle Coordinate Principali (Gower, 1966). Tale tecnica di ordinamento ha la proprietà di preservare al meglio le distanze originali fra gli oggetti nello spazio ridotto definito dagli assi principali.

La matrice $\mathbf{D}_{nxn}$ delle distanze o similarità fra gli *n* oggetti viene dapprima trasformata nella matrice Δ:

$$\Delta = -\frac{1}{2}\mathbf{D}$$

La matrice **C** viene quindi ottenuta centrando la matrice Δ in modo tale che l'origine del sistema di assi che sarà definito si trovi nel centroide degli oggetti:

$$c_{ij} = \delta_{ij} - \frac{1}{n}\sum_{h=1}^{n}\delta_{ih} - \frac{1}{n}\sum_{k=1}^{n}\delta_{kj} - \frac{1}{n^2}\sum_{h=1}^{n}\sum_{k=1}^{n}\delta_{hk}$$

dove il secondo ed il terzo termine rappresentano le medie di riga e di colonna della matrice Δ (equivalenti nel caso di una matrice simmetrica) e l'ultimo termine rappresenta la media generale di questa stessa matrice.

Si calcolano quindi gli autovalori λ_j (j=1,2,...,m; $m \leq n$-1) e gli autovettori u_{ij} (i=1,2,...,n; j=1,2,...,m) della matrice **C**. Le Coordinate Principali f_{ij} degli oggetti si ottengono moltiplicando gli autovettori per la radice quadrata dell'autovalore corrispondente:

$$f_{ij} = \sqrt{\lambda_j} \cdot u_{ij}$$

Anche in questo caso la qualità dell'ordinamento ottenuto per ciascun asse principale può essere valutata sulla base del rapporto fra l'autovalore corrispondente e la somma degli autovalori estratti. Tuttavia, poichè è possibile che uno o più autovalori siano negativi, Cailliez & Pagès (1976) raccomandano di valutare globalmente la qualità di un ordinamento utilizzando il rapporto:

$$\frac{\sum_{i=1}^{q}\lambda_i + q\left|\lambda_{\min}\right|}{\sum_{i=1}^{n-1}\lambda_i + (n-1)\left|\lambda_{\min}\right|}$$

dove *q* è il numero di dimensioni in cui si è ottenuto l'ordinamento, *n* è il numero totale di dimensioni e λ_{min} è l'autovalore negativo di maggior valore assoluto.

4.3. Analisi Fattoriale delle Corrispondenze.

L'Analisi Fattoriale delle Corrispondenze, o semplicememte Analisi delle Corrispondenze, è una tecnica di ordinamento di grande interesse in ecologia (Benzecri ***et al.***, 1973). A differenza di altre tecniche, quali ad esempio i vari tipi di Analisi delle Componenti Principali, l'Analisi Fattoriale delle Corrispondenze consente di rappresentare simultaneamente i punti-variabile ed i punti-osservazione, con coordinate tali da rendere massima la correlazione fra i due insiemi per ogni fattore.

La dualità di questo tipo di analisi, tuttavia, non è il suo unico pregio. Una caratteristica di enorme interesse dell'Analisi Fattoriale delle Corrispondenze è l'equivalenza distribuzionale. In pratica, poichè ad essere analizzati sono sostanzialmente dei profili, il risultato globale dell'analisi non cambia se, ad esempio, le osservazioni relative a due entità tassonomiche la cui separazione è dubbia vengono cumulate o mantenute separate. Analogamente, se un'osservazione è replicata con risultati coerenti, può essere indifferentemente cumulata alla precedente o trattata come una nuova osservazione.

Tralasciando una trattazlone piu approfondita e centrata su aspetti piu strettamente formali, l'Analisi Fattoriale delle Corrispondenze può essere effettuata in tre fasi principali: calcolo di una matrice simmetrica di prodotti scalari, calcolo degli autovalori e degli autovettori di tale matrice ed infine calcolo delle coordinate e dei contributi assoluti (cioè dei contributi delle osservazioni e delle variabili agli assi fattoriali) e relativi (cioè degli assi fattoriali alla descrizione di osservazioni e variabili).

La qualità della rappresentazione ottenuta nello spazio ridotto definito dagli assi fattoriali può essere stimata sulla base degli autovalori estratti, per quanto riguarda la qualità globale

dell'ordinamento ed il grado di strutturazione del sistema, e sulla base dei contributi relativi per quanto riguarda i singoli taxa e le singole stazioni.

La matrice dei dati $\mathbf{A}_{nxp}$ sarà organizzata in modo tale che risulti $p \leq n$, al fine di ottimizzare le procedure di calcolo. Ciò implica, nella maggior parte dei casi, che le osservazioni corrispondano alle righe ed i descrittori alle colonne, poichè le prime dovrebbero essere comunque più numerose dei secondi. Un caso tipico in cui ciò non si verifica, tuttavia, è quello, peraltro assai frequente, in cui si debbano trattare delle liste si specie osservate in un insieme di stazioni: in questo caso è del tutto normale che le specie (cioè i descrittori) siano molto più numerose delle stazioni (cioè delle osservazioni).

La matrice **A**, così organizzata, viene trasformata nella matrice **U**, in cui

$$u_{ij} = \frac{a_{ij}}{\sqrt{a_{i.}a_{.j}}} - \frac{\sqrt{a_{i.}a_{.j}}}{a_{..}}$$

La matrice **U** contiene dunque gli scarti degli elementi di **A** pesati sulla media geometrica delle somme marginali di riga e di colonna rispetto alla stessa media geometrica pesata sul totale generale.

La matrice dei prodotti scalari **S**, di rango *p*, si ottiene quindi moltiplicando la tale matrice per la sua trasposta **U'**

$$\mathbf{S} = \mathbf{U}'\mathbf{U}$$

Si calcolano quindi gli autovalori λ_j [j=1,2,...,m; $m \leq p-1$] e gli autovettori v_{jh} [j=1,2,...,p; h=1,2,...,m)] della matrice **S**. Si noti che, poichè non è strettamente necessario calcolare tutti gli autovalori e gli autovettori, spesso ci si limita ad estrarre solo i primi 2 o 3, i quali, peraltro, sono in generale largamente sufficienti ai fini dell'analisi.

Si calcolano quindi le coordinate delle osservazioni:

$$f_{ih} = \sum_{j=1}^{p} \frac{a_{ij} v_{jh}}{a_{i.} \sqrt{\frac{a_{.j}}{a_{..}}}}$$

per gli *h* assi fattoriali richiesti. Si passa poi alle coordinate delle variabili:

$$g_{jh} = \sum_{i=1}^{n} \frac{a_{ij} f_{ih}}{a_{.j} \sqrt{\lambda_h}}$$

Successivamente si calcolano i contributi assoluti all'*h*-mo fattore da parte della *i*-ma osservazione e della *j*-ma variabile:

$$ca(f_{ih}) = f_{ih}^2 \frac{a_{i.}}{a_{..} \lambda_h}$$

$$ca(g_{jh}) = g_{jh}^2 \frac{a_{.j}}{a_{..} \lambda_h}$$

Infine, si calcolano i contributi relativi dell'*h*-mo fattore, all'*i*-ma osservazione ed alla *j*-ma variabile

$$cr(f_{ih}) = \frac{f_{ih}^2}{\sum_{h=1}^{m} f_{ih}^2}$$

$$cr(g_{jh}) = \frac{g_{jh}^2}{\sum_{h=1}^{m} g_{jh}^2}$$

La significatività degli assi fattoriali può essere testata in maniera empirica in diversi modi. Il più semplice è quello che prevede il confronto della percentuale di varianza spiegata da ciascuno di essi con quella attesa in base al modello di Mac Arthur ("broken-stick").

E' inoltre possibile rappresentare altre osservazioni ad altre variabili nello spazio fattoriale così definito.

4.4. Analisi delle Correlazioni Canoniche.

Nell'ambito di uno studio ecologico è spesso necessario prendere in considerazione insiemi di variabili qualitativamente eterogenei. Ad esempio, è assai frequente il caso in cui si dispone della lista delle specie e delle misure dei principali parametri fisico-chimici relative ad un insieme di osservazioni distribuite nello spazio e/o nel tempo. Un insieme di dati organizzato in tal modo non può essere analizzato esaustivamente mediante le consuete tecniche di ordinamento, le quali, al di la dei problemi formali, non consentono di isolare i due sottoinsiemi di variabili e di valutarne globalmente il grado di correlazione.

L'Analisi delle Correlazioni Canoniche al contrario, ha come fine proprio l'esame di tali correlazioni. Per l'Analisi delle Correlazioni Canoniche la matrice dei dati può essere vista come l'insieme delle *n* osservazioni relative a due sottoinsiemi composti rispettivamente da *p* e da *q* variabili, con $p \leq q$. In altre parole, la *i*-ma osservazione può essere rappresentata da due vettori riga **x** ed **y**

$$\mathbf{x} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$$
$$\mathbf{y} = \begin{pmatrix} y_{i1} & y_{i2} & \cdots & y_{iq} \end{pmatrix}$$

in cui le **x** sono le misure del sottoinsieme di variabili meno numeroso e le **y** le rimanenti.

La matrice di covarianza **S** di rango *p+q* dell'insieme completo dei dati sarà quindi ripartibile in blocchi:

$$\mathbf{S} = \frac{1}{n}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}\begin{pmatrix} \mathbf{x}' & \mathbf{y}' \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}$$

In particolare, $\mathbf{S}_{11}$ è la matrice di rango *p* di covarianza delle variabili del sottoinsieme ***x***, così come $\mathbf{S}_{22}$ di rango *q* lo è del sottoinsieme ***y***. La $\mathbf{S}_{12}$ è una matrice *p*x*q* che contiene le covarianze fra i due sottoinsiemi di variabili. Poichè **S** è una matrice simmetrica, $\mathbf{S}_{21}$ è la trasposta di $\mathbf{S}_{12}$.

Lo scopo dell'Analisi delle Correlazioni Canoniche è trovare, partendo dalla matrice **S**, le *p* combinazioni lineari delle variabili **x** e le *p* combinazioni lineari delle variabili **y**

$$\begin{aligned} u_i &= a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ip}x_p \\ v_i &= b_{i1}y_1 + b_{i2}y_2 + \ldots + b_{iq}y_q \end{aligned} \qquad i = 1,2,\ldots,p$$

tali da soddisfare le seguenti condizioni:

1) tutte le u_i devono essere indipendenti fra loro;

2) tutte le v_i devono essere indipendenti fra loro;

3) le *p* coppie di combinazioni lineari devono essere tali da rendere massime le *p* correlazioni r_i fra le u_i e le v_i.

Le variabili *u* e *v* sono perciò dette variabili canoniche e le loro correlazioni *r* sono dette correlazioni canoniche.

Prescindendo in questa sede da una trattazione completa dal punto di vista formale, l'Analisi delle Correlazioni Canoniche può essere effettuata, sulla base della matrice di covarianza **S** ripartita in blocchi, calcolando innanzitutto gli autovalori delle due matrici ottenute dai prodotti

$$\begin{aligned} &\mathbf{S}_{22}^{-1}\,\mathbf{S}_{12}'\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12} \\ &\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{12}' \end{aligned}$$

Esistono al massimo *p* autovalori non nulli della prima matrice prodotto: tali autovalori sono uguali a quelli non nulli della seconda matrice prodotto.

I vettori dei coefficienti **a** e **b** si ottengono risolvendo i due sistemi, rispettivamente di *p* e *q* equazioni lineari

$$\begin{aligned} \left(\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{12}' - \lambda_i\mathbf{I}\right)\mathbf{a} = 0 \\ \left(\mathbf{S}_{22}^{-1}\,\mathbf{S}_{12}'\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12} - \lambda_i\mathbf{I}\right)\mathbf{b} = 0 \end{aligned}$$

per ogni λ_j (*i*=1,2,...,*p*). Per comodità, si pone

$$a_{i1} = 1 \qquad\qquad b_{i1} = 1$$

Si possono quindi ricavare le variabili canoniche mediante un prodotto fra vettori

$$\mathbf{u}_i = \mathbf{a}'\mathbf{x} \qquad \mathbf{v}_i = \mathbf{b}'\mathbf{y}$$

Per ciascuna coppia u_i e v_i la correlazione canonica sarà

$$r_i = \sqrt{\lambda_i}$$

Le variabili canoniche possono quindi essere impiegate per ulteriori analisi, come pure per un output grafico diretto, che rappresenta la correlazione fra i due sottoinsiemi di variabili eterogenee e l'ordinamento delle osservazioni in questo ambito.

Sulla base del primo autovalore estratto, e cioè della correlazione canonica piu alta, è possibile effettuare un test di indipendenza fra i due sottoinsiemi di variabili.

Va infine rilevato che, nel caso in cui le variabili originali presentino una sensibile eterogeneità di scala, puo essere conveniente effettuare l'analisi sui dati centrati e standardizzati: in tal modo la matrice **S** è in realtà una matrice di correlazione **R** e le variabili canoniche ottenute sono adimensionali. Questa soluzione consente, inoltre, di confrontare l'importanza delle variabili originali in base al valore dei coefficienti **a** e **b** delle variabili canoniche.

5. Analisi di serie spaziali e temporali.

5.1. Autocorrelazione.

Uno strumento di notevole utilità nello studio delle serie spaziali e temporali di dati è costituito dalle funzioni di autocorrelazione (Cliff & Ord, 1973 e 1981). Il concetto di autocorrelazione è legato alla possibilità di prevedere l'andamento di una variabile nel tempo o nello spazio sulla base dei valori misurati: una autocorrelazione positiva, ad esempio, implica una maggiore probabilità di osservare valori elevati della variabile considerata in prossimità di un punto in cui è stato effettivamente misurato un valore elevato.

La forma delle funzioni che legano l'autocorrelazione alla distanza fra coppie di punti (correlogrammi) consente di formulare delle inferenze sulla struttura spaziale (o temporale) della variabile studiata.

Una delle misure di autocorrelazione più utilizzate nel caso di serie spaziali di dati, soprattutto nel caso in cui le osservazioni non siano distribuite in maniera uniforme, è il coefficiente *I* di Moran (1950):

$$I(d) = \frac{\frac{1}{W}\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\frac{1}{p}\sum_{i=1}^{p}(y_i - \bar{y})^2} \quad \text{per } i \neq j$$

dove *d* è la distanza considerata, y_i è il valore della variabile *y* nell'*i*-mo punto della serie, w_{ij} è un delta di Kronecker, *W* è la somma dei delta di Kronecker per la distanza *d* e *p* è il numero di punti nella serie.

5.2. Test di Mantel.

Questo test, di recentissima introduzione in campo ecologico, è stato sviluppato in origine per lo studio della distribuzione spaziale dell'occorrenza di casi di tumori (Mantel, 1967). Esso consente di

ottenere una misura del grado di correlazione esistente fra due matrici di distanze (di cui una può essere di tipo geografico) o di similarità. L'ipotesi nulla che viene testata è quella di indipendenza fra le due matrici analizzate, mentre il livello di probabilità relativo al valore della statistica viene calcolato sulla base di una procedura iterativa.

La statistica Z di Mantel, che esprime il grado di correlazione fra la struttura delle due matrici, si calcola come la somma dei prodotti degli elementi corrispondenti delle due matrici di distanza, esclusi quelli sulla diagonale. Se gli elementi di ciascuna delle due matrici vengono preventivamente centrati e standardizzati, allora la statistica di Mantel (indicata in questo caso come R) risulta standardizzata ed assume lo stesso significato e lo stesso intervallo di variazione di un coefficiente di correlazione di Bravais-Pearson.

Il livello di probabilità associato al valore della statistica di Mantel si calcola sulla base di una procedura iterativa che prevede la permutazione casuale delle righe e delle colonne di una delle due matrici ed il ricalcolo della statistica di Mantel per un numero sufficientemente alto di volte. Il valore della statistica ottenuto per le matrici originali viene confrontato con la distribuzione empirica di quelli ottenuti ripetendo il calcolo su matrici permutate aleatoriamente: la percentuale delle iterazioni in cui si è ottenuto un valore inferiore a quello originale corrisponde al livello di probabilità di quest'ultimo. Dal punto di vista pratico si rigetterà l'ipotesi nulla di indipendenza fra le matrici se almeno il 95% o il 99% dei valori ottenuti per le matrici permutate è inferiore (o superiore) a quello originale.

Questo tipo di procedura consente, inoltre, di ottenere anche un'altra forma di standardizzazione della statistica di Mantel, che non richiede di intervenire sulle matrici originali. Questa standardizzazione, proposta da Hubert, si effettua riscalando il valore originale di Z rispetto al minimo ed al massimo ottenuti durante la procedura iterativa di permutazione delle matrici e ricalcolo di Z, che vengono assunti come estremi teorici della variazione della statistica (e cioè come -1 e 1, rispettivamente). Il significato di questa forma di standardizzazione è interessante soprattutto perchè essa viene effettuata in rapporto alla specifica natura delle matrici sottoposte al test: in altre parole anche

una correlazione debole, purchè sia realmente la migliore ottenibile sulla base dei dati originali, fa assumere a R un valore pari a 1 (che corrisponderà, evidentemente ad un livello di probabilità P(R) prossimo al 100%).

Dunque, questa forma di standardizzazione fornisce una misura del livello di correlazione relativa fra le matrici analizzate, mentre quella precedentemente illustrata fornisce una misura assoluta.

6. Interpolazione.

6.1. Note introduttive.

Una efficace rappresentazione grafica dei risultati è, indipendentemente dal contesto, un utile strumento di sintesi. In campo ecologico ciò è ancor più vero, in considerazione dell'eterogeneità delle variabili in gioco e della complessità delle relazioni che le legano.

La natura stessa della ricerca ecologica propone spesso situazioni in cui i dati quantitativi, semiquantitativi o addirittura qualitativi devono essere rappresentati in funzione della posizione delle stazioni di rilevamento. Assai spesso può essere utile, ad esempio, mappare la densità degli individui di una specie in una determinata area geografica, e magari confrontare il risultato con quello relativo ad una seconda specie o ad una variabile di altra natura.

L'esame ed il confronto di rappresentazioni di questo tipo possono consentire di evidenziare rapporti funzionali o di formulare nuove ipotesi di lavoro, ma in sostanza il risultato ultimo è una rielaborazione sintetica di dati disponibili sotto altra forma.

L'elaborazione dell'informazione disponibile in funzione della sua trasposizione cartografica è dunque il nocciolo del problema. Nel caso più semplice ciò si riduce alla determinazione delle coordinate di riferimento ed al semplice trasferimento sulla carta dei valori numerici in oggetto: il risultato è simile a quello che si ottiene riportando le quote dei rilievi principali su una carta geografica. Il problema si complica quando si desidera rappresentare una grandezza su tutta un'area, tracciando sulla mappa delle curve che congiungono i punti in cui essa assume lo stesso valore (isoplete).

La realizzazione di una mappa ad isoplete, infatti, presenta dei problemi di interpolazione, poiché è evidente che, per quanto numerose possano essere le osservazioni effettuate, il rilevamento dei dati non può comunque assumere il carattere di continuità che, invece, dovrà essere restituito dalla rappresentazione grafica. Si rende perciò

necessario formulare un'ipotesi sul comportamento della grandezza in esame fra due o più punti noti ed assumere la stessa come la migliore approssimazione possibile dei valori reali.

L'interpolazione può essere effettuata empiricamente, cioè sulla base del buon senso e dello spirito di osservazione, o mediante l'impiego di strumenti matematici. La prima soluzione è indubbiamente la più diffusa, a tutt'oggi, in un contesto di tipo ecologico. La soggettività delle scelte che la guidano ne è insieme il maggior pregio ed il peggior difetto, poiché, se è vero che si possono ottenere rappresentazioni sintetiche ed efficaci come nessun algoritmo potrebbe produrre, è vero che è molto difficile non lasciarsi sfuggire qualcosa (magari certi dettagli che non sembrano proprio al loro posto...).

Le tecniche non soggettive di interpolazione sono senza dubbio quelle che possono consentire di estrarre il massimo dell'informazione dai propri dati, ma, al tempo stesso, impongono l'uso di metodi rigorosi già a partire dal rilevamento degli stessi, poiché è del tutto evidente che l'elaborazione dei dati affetti da fonti di errore non controllabili resta un semplice esercizio formale. E' importante, tuttavia, rimarcare come non sia l'errore di misura in sè, quanto piuttosto la mancanza di informazioni sulla natura dello stesso, ad inficiare i risultati.

L'ipotesi che ogni misura effettuata, ogni dato rilevato sia soltanto una delle possibili manifestazioni di una variabile aleatoria ed il completo trattamento dell'errore sono alla base della tecnica di interpolazione nota come kriging.

6.2. Le tecniche di interpolazione

La mappatura di qualsiasi tipo di variabile, dunque, richiede due cose: un certo numero di misure, effettuate in punti identificati da un sistema di coordinate, e una tecnica di interpolazione, la quale consenta di "ricostruire", cioè di stimare in maniera non soggettiva, i valori assunti dalla grandezza in oggetto negli intervalli compresi fra i punti noti.

Le tecniche di interpolazione possono essere di due tipi: deterministiche o stocastiche. Quest'ultimo è il caso del kriging (Matheron, 1969 e 1970).

Si può affermare che le tecniche deterministiche stimano, sulla base delle osservazioni effettuate, una funzione o una combinazione lineare di funzioni che descrive l'andamento medio di una grandezza, senza però riprodurne i valori nei punti noti (es.: metodo dei minimi quadrati), o che assume i valori esatti nei punti noti, fornendo stime poco attendibili nelle regioni comprese fra questi (es.: interpolazione polinomiale). Fra le tecniche deterministiche impiegate per la mappatura di grandezze di interesse ecologico si può ricordare la trend-surface analisys: anch'essa, come il kriging, è stata inizialmente sviluppata in campo geologico.

Il campo ottimale di applicazione delle tecniche di interpolazione deterministiche, oltre alla descrizione di fenomeni mediante funzioni dal preciso significato fisico o biologico, è probabilmente quello della definizione di trends mono-, bi- o pluridimensionali sulla base di osservazioni regolarmente distribuite nello spazio, possibilmente con errore nullo.

La disponibilità di dati di questo tipo, tuttavia, rappresenta l'eccezione, piuttosto che la regola, in campo ecologico. Di solito, infatti, è molto difficile poter effettuare tutte le misure di cui si vorrebbe poter disporre ed ancor più difficile, nel caso di grandezze biologiche, è che esse non siano affette da errore nè distribuite irregolarmente. Una tecnica di interpolazione stocastica quale il kriging, però, può essere efficace anche in queste condizioni. Ogni osservazione, infatti, viene considerata come una singola realizzazione di una variabile aleatoria di cui sia noto (o ipotizzato) il valore medio in ogni punto, cioè il trend, e le cui proprietà statistiche siano definite da una funzione detta variogramma. Sulla base delle osservazioni disponibili vengono poi stimati tutti i valori desiderati, mentre quelli noti sono ricostruiti esattamente, a meno che non si sia introdotta nel modello di interpolazione una stima dell'errore strumentale o di campionamento.

La caratteristica più interessante del kriging, tuttavia, sta nella possibilità di disporre, per ogni valore ricostruito, di una stima

dell'affidabilità della ricostruzione. Ciò consente, ad esempio, di definire per quest'ultima un intervallo fiduciale od ancora di individuare le aree in cui è necessario aumentare la densità dei rilevamenti.

E' necessario, tuttavia, rilevare come anche nell'uso del kriging esista un rovescio della medaglia. In primo luogo si deve sottolineare la necessità di disporre di strutture di calcolo alquanto più potenti di quelle richieste dall'uso di tecniche più convenzionali, soprattutto in funzione dei non brevi tempi di elaborazione. In secondo luogo va considerato il fatto che la definizione del modello di interpolazione, contrariamente ad altre tecniche, non avviene in maniera del tutto univoca: infatti, pur esistendo dei criteri guida, il risultato è affidato in qualche misura all'abilità del modellista ed alla sua esperienza nel campo specifico di applicazione.

6.3. Il kriging: teoria.

Una grandezza da ricostruire su di un'area geografica può essere considerata, in genere, come una variabile aleatoria di cui è nota una singola misura od una stima z per un certo numero di punti di rilevamento. Sull'intera area in esame la grandezza può dunque essere rappresentata da una funzione aleatoria $Z(x)$, di cui sono noti i valori osservati $z(x)$ in un insieme di n stazioni. Nonostante il kriging sia una tecnica nata e sviluppata in un contesto geologico, il concetto di variabile aleatoria può essere facilmente esteso ad applicazioni di tipo ecologico ed in particolare al caso delle stime (di abbondanza, di intensità, di factor-scores, etc.) ottenute mediante campionamento di un sottoinsieme di nodi (stazioni) di un reticolo sovraimposto all'area geografica in esame.

La densità di una popolazione, ad esempio, pur non essendo teoricamente una variabile aleatoria, poiché il numero degli individui in un qualsiasi istante ed in una qualsiasi area è comunque determinato, può essere considerata come tale, e quindi descritta nello spazio da una funzione aleatoria, se si tiene conto del fatto che le misure di densità si stimano campionando piccole superfici o piccoli volumi,

considerati rappresentativi della stazione in esame. Se il piano ed il metodo di campionamento sono corretti, dunque, la densità può essere certamente trattata come una variabile aleatoria.

La funzione *Z(x)* si può considerare come la somma di un valore atteso *t(x)*, che descrive un trend, e di uno scarto *e(x)*, il quale è tanto minore quanto più efficacemente il trend descrive il comportamento della variabile aleatoria *z*:

$$Z(x) = t(x) + e(x)$$

Il trend *t(x)* è generalmente espresso come una combinazione lineare di funzioni $f_i(x)$

$$t(x) = a_0 + \sum_{i=1}^{p} a_i f_i(x)$$

Un caso particolare è quello in cui *p=0*, cioè il trend è costante.

Una stima $z'(x_0)$ della variabile da interpolare nel punto x_0 si ottiene mediante una combinazione lineare di valori noti $z(x_i)$ nei punti x_i:

$$z'(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i)$$

Le stime *z'* sono delle realizzazioni della funzione aleatoria *Z'(x)*, che quindi si può esprimere come

$$Z'(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

Il problema consiste, dunque, nel determinare i coefficienti λ_i. Al fine di evitare che l'errore

$$E(x_0) = Z(x_0) - Z'(x_0)$$

sia sistematico e di ottimizzare la stima si impongono i due vincoli

$$Media\left[E(x_0)\right] = 0$$
$$Varianza\left[E(x_0)\right] \rightarrow \min$$

I coefficienti λ_i si ottengono risolvendo il sistema

$$\sum_{j=1}^{n} \lambda_j \gamma(d_{ij}) + \mu_1 + \mu_2 t(x_i) = \gamma(d_{i0}) \qquad i = 1,2,...,n$$
$$\sum_{j=1}^{n} \lambda_j = 1$$
$$\sum_{j=1}^{n} \lambda_j t(x_j) = t(x_0)$$

dove il primo blocco e l'ultima equazione consentono di soddisfare rispettivamente il primo ed il secondo dei vincoli appena imposti, minimizzando la varianza dell'errore ed imponendo a quest'ultimo una media nulla, mentre la penultima equazione è necessaria per ottenere coefficienti dimensionalmente indipendenti dalla varianza della funzione aleatoria *Z(x)*. I moltiplicatori lagrangiani μ_1 e μ_2 sono relativi ai vincoli imposti mediante le due ultime equazioni. Nel caso in cui il trend ipotizzato *t(x)* sia costante, si elimina dal sistema l'ultima equazione, in quanto proporzionale alla penultima. Oltre a determinare i coefficienti λ, la soluzione del sistema consente di stimare la varianza dell'errore di $Z'(x_0)$:

$$Var\left[E(x_0)\right] = \sum_{j=1}^{n} l_j g(d_{j0}) + \mu_1 + \mu_2 t(x_0)$$

La funzione *g(d)*, che è già comparsa nel sistema la cui soluzione fornisce i coefficienti λ_i, è il variogramma della variabile da ricostruire, cioè è un'espressione della variazione della differenza tra i valori assunti dalla variabile stessa nei punti x_i e x_j in funzione della loro distanza *d*. Una stima di tale funzione, detta variogramma empirico (o semi-variogramma, poichè in effetti essa è moltiplicata per ½), si ottiene sulla base delle osservazioni disponibili:

$$g(d) = \frac{1}{2n(d)} \sum_{i=1}^{n} \sum_{j=1}^{n} w(d, x_i, x_j) \cdot \left[z(x_i) - z(x_j)\right]^2$$

dove *n(d)* è il numero delle coppie di punti la cui distanza è *d* e $w(d,x_i,x_j)$ è una funzione che assume valore 1 se la distanza fra x_i e x_j è pari a *d* e valore 0 altrimenti.

Il variogramma empirico, essendo stimato su un numero finito e spesso limitato di osservazioni (e quindi di distanze), presenta di solito notevoli irregolarità, oltre a non essere, evidentemente, una funzione continua della distanza d fra due punti qualsiasi. Poiché ai fini pratici la funzione *g(d)* deve essere definita per qualsiasi valore di *d*, è necessario che essa sia una funzione continua. Inoltre, poiché *g(d)* è una varianza, non deve poter assumere valori negativi.

Si determina, dunque, sulla base del variogramma empirico, un variogramma teorico che soddisfi tali condizioni. Fra i variogrammi teorici più flessibili e più largamente impiegati possono essere segnalati i seguenti:

$$g(d) = \sigma_0^2 + ad^b$$

$$g(d) = \sigma_0^2 + a\left[1 - e^{-\frac{d}{b}}\right]$$

$$g(d) = \begin{cases} \sigma_0^2 + a\left(\frac{3d}{2b} - \frac{d^3}{2b^3}\right) & \text{per } b \le d \\ \sigma_0^2 + a & \text{per } b > d \end{cases}$$

I parametri *a* e *b* vengono stimati mediante una qualsiasi tecnica di fitting (es.: minimi quadrati) a partire dal variogramma empirico ed in modo tale da ottenere il miglior adattamento possibile ad esso del variogramma teorico, soprattutto per bassi valori della *d*, laddove, cioè, la stima empirica è effettuata su di un maggior numero di osservazioni. La varianza locale σ^2_0 può essere assunta con un valore non nullo se esistono dati relativi alla variabilità intrinseca della grandezza in esame o su quella legata al campionamento o alla misurazione.

E' importante sottolineare il fatto che i valori dei parametri sono comunque suscettibili di successivi aggiustamenti, effettuati sulla base di test di validazione. Tali tests si rendono necessari poiché sia il tipo di variogramma prescelto, sia la prima stima dei suoi parametri possono non essere ottimali dal punto di vista della bontà dell'interpolazione: in

altre parole, il primo variogramma ipotizzato rappresenta il punto di partenza per una serie di aggiustamenti successivi effettuati mediante tests di validazione. Il ruolo di ciò è tanto più importante quanto più irregolare è il comportamento della grandezza da ricostruire e quanto minore è il numero di osservazioni.

I tests di validazione dei variogrammi teorici si effettuano sulle osservazioni disponibili, di ognuna delle quali viene stimato il valore sulla base delle rimanenti, in modo da ottenere, per ciascuna di esse, un errore $e(x_i)$:

$$e(x_i) = z(x_i) - z'(x_i)$$

Le condizioni che devono essere soddisfatte per considerare accettabile dal punto di vista formale un variogramma teorico sono legate alla distribuzione dell'errore $e(x_i)$. In particolare è necessario che esso abbia media nulla, cioè che non sia sistematico, e che assuma valori coerenti, in media, in rapporto alla deviazione standard s_i stimata sulle rimanenti n-1 osservazioni. Cioè:

$$\frac{1}{n}\sum_{i=1}^{n} e(x_i) \approx 0$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{e(x_i)}{s_i}\right]^2 \approx 1$$

In fase di validazione si considera ottimale il variogramma che soddisfa tali condizioni minimizzando l'errore quadratico medio e che, quindi, garantisce sia la migliore interpolazione complessiva, sia la più uniforme distribuzione dell'errore su tutti i punti stimati.

Quanto esposto sin qui rappresenta la base teorica del kriging come tecnica d'interpolazione. Dal punto di vista operativo, però, essa può essere applicata tanto globalmente quanto localmente. Nel primo caso tutte le n osservazioni concorrono a ciascuna stima, mentre nel secondo vengono considerate solo quelle comprese entro una circonferenza di raggio dato centrata nel punto da interpolare. Nel caso del kriging locale è necessario anche, se i punti noti non sono distribuiti

in maniera regolare, definire il numero minimo di punti che devono essere compresi in tale circonferenza.

Laddove questo numero non sia raggiunto, il raggio viene aumentato opportunamente. La definizione del raggio e del numero minimo di punti noti in esso compresi può richiedere ulteriori tests di validazione.

6.4. Il kriging: note applicative.

Il problema di maggior rilievo nell'applicazione del kriging è senza dubbio quella della risoluzione del sistema di equazioni lineari, che richiede, nel caso del kriging globale, l'inversione di una matrice di dimensioni *n* x *n* per ogni punto da stimare.

Questo ostacolo può essere risolto mediante l'uso di un algoritmo di kriging locale, considerando ai fini dell'interpolazione solo un certo numero di punti noti, scelti fra i più vicini al punto da stimare. In particolare, questa soluzione si rivela vantaggiosa, sia dal punto di vista dei tempi di calcolo, sia da quello della bontà dell'interpolazione, quando l'andamento del variogramma appare regolare solo per piccole distanze.

Un aspetto di non trascurabile importanza nell'ambito del kriging, come di altre tecniche di interpolazione, è la definizione di un sistema di coordinate per la localizzazione dei punti noti e di quelli stimati. Le applicazioni presentate qui sono basate su sistemi di coordinate arbitrarie intere, con origine nell'angolo inferiore sinistro dell'area in esame. L'interpolazione, in questo caso, è stata effettuata ai nodi di un reticolo a maglia quadrata, ma è comunque possibile intensificare o diradare a piacere la densità dei punti da stimare, come pure effettuare delle stime relative ad un punto qualsiasi.

La presentazione grafica dei risultati può essere fornita sotto forma di mappe ad isoplete o di proiezioni assonometriche di superfici. La prima tecnica, che può essere vantaggiosamente integrata dall'uso del colore, si presta meglio a mappature anche complesse mentre la

seconda può essere utilizzata in funzione della sua migliore resa dal punto di vista della sintesi descrittiva.

La ricostruzione di valori di densità può fornire, in alcuni casi, valori negativi: poiché di ogni valore è possibile determinare l'intervallo fiduciale, una stima negativa è da considerarsi pari ad una densità nulla con un livello di probabilità uguale a quello necessario a comprendere lo zero nell'intervallo fiduciale.

Infine, è interessante rilevare come la caratteristica più interessante di questa tecnica di interpolazione sia proprio la possibilità di disporre di una stima dell'errore di interpolazione, utile sia per definire degli intervalli fiduciali intorno ai valori interpolati, sia per riconoscere le aree dove effettuare nuove osservazioni o dove riorganizzare il piano di campionamento.

7. Diversità.

7.1. L'indice di Shannon.

Una delle maniere più utilizzate per sintetizzare l'informazione contenuta nella struttura di una comunità animale o vegetale è senza dubbio rappresentato dagli indici di diversità. Il più noto fra tutti è quello di Shannon-Weaver (Shannon, 1948):

$$H = -\sum_{i=1}^{p} \frac{n_i}{N} \log_2 \frac{n_i}{N}$$

dove p è il numero di specie, n_i è il numero di individui che appartengono alla i-ma specie ed N è il numero totale di individui di tutte le specie presenti nel campione.

Questa espressione rappresenta la quantità media di informazione per individuo, secondo un criterio per cui ogni individuo di una specie, una volta identificato, ha un contenuto di informazione tanto più rilevante quanto più la specie è rara. Si può facilmente dimostrare che il massimo valore di H si ottiene quando tutte le specie hanno la medesima frequenza, mentre il minimo si osserva quando tutte le specie sono rappresentate da un solo individuo, tranne una a cui appartengono tutti i rimanenti individui:

$$H_{\max} = \log_2 p$$
$$H_{\min} = \log_2 N - \frac{N - p + 1}{N} \log_2 (N - p + 1)$$

Sulla base di queste misure è possibile definire alcuni indici derivati da H, che hanno il vantaggio di rapportare il valore di quest'ultimo al suo valore massimo o all'intervallo di variazione possibile dati N e p.

Questi indici sono stati definiti in maniera diversa da vari Autori ed è perciò essenziale associare al loro nome il corretto riferimento. Ad esempio, per quanto riguarda la ***evenness*** (talvolta tradotta come

regolarità o equitabilità), vengono correntemente utilizzate le seguenti formulazioni:

$$R = \frac{H}{H_{\max}} \qquad \text{Pielou (1966)}$$

$$R = \frac{H_{\max} - H}{H_{\max} - H_{\min}} \qquad \text{Patten (1962)}$$

E' interessante sottolineare il fatto che esiste, teoricamente, la possibilità di stimare l'errore standard di queste misure e, quindi, i loro intervalli fiduciari (Pielou, 1975), ma questa prassi è assai poco comune. A questo fine è necessario assumere che il numero di specie effettivamente presenti in natura possa essere stimato sulla base di quelle identificate nel campione: tuttavia, tale condizione non è sempre del tutto verosimile.

7.2. Diagrammi rango-frequenza e modello di Zipf-Mandelbrot.

Poichè lo studio della diversità ha implicazioni di notevole interesse teorico e pratico, è certamente desiderabile esprimere questa proprietà in una maniera più articolata ed informativa di quella consentita da un semplice indice numerico.

La via più immediata per raggiungere questo scopo è quella di rappresentare la distribuzione degli individui fra le specie in forma grafica, ad esempio mediante un istogramma. Oltre che alle distribuzioni empiriche è possibile fare riferimento anche a diversi modelli teorici (es. log-normale).

Questa soluzione, però, non sempre fornisce dei risultati realmente utili, soprattutto a causa del fatto che le distribuzioni empiriche tendono a diventare irregolari e poco informative quando il numero di specie è piccolo o quando le abbondanze di ciscuna di specie sono modeste.

Per aggirare questo tipo di difficoltà è però possibile rappresentare lo stesso tipo di informazione plottando la frequenza di ciascuna specie contro il relativo rango, meglio se su scala logaritmica.

Il diagramma rango-frequenza che si ottiene in questo modo ha per costruzione un andamento decrescente in maniera monotona: tuttavia, è il profilo di tale andamento che sintetizza l'informazione pertinente la struttura della comunità.

È interessante notare, in particolare, che la forma della curva così ottenuta esprime precise caratteristiche strutturali dei popolamenti, pur essendo (teoricamente) invariante, a differenza degli indici di diversità, rispetto a diverse proprietà del campione analizzato (numero di individui, numero di specie, ambiente consiederato, etc.).

In particolare, la forma del profilo, soprattutto nella sua parte iniziale, consenta di inferire le proprietà globali della comunità studiata, almeno per grandi linee. Un profilo tendenzialmente concavo indica uno stato di stress (es. da inquinamento) o una condizione estremamente giovanile di una comunità (es. prime fasi della colonizzazione di un substrato artificiale), mentre un profilo nettamente convesso è associato ad una struttura più stabile e matura della comunità, in cui le interazioni sono più complesse.

Questo tipo di diagramma è stato utilizzato soltanto di recente in ecologia, ma la sua applicazione in altri campi (sociologia, econometria, linguistica, etc.) è ben consolidata.

Infatti, è in campo linguistico che è stato sviluppato il modello teorico di base per le curve rango-frequenza, il quale è noto come modello di Zipf, dall'Autore che ne ha presentato l'applicazione (Zipf, 1949-1965). La sua formulazione è la seguente:

$$f_r = f_1 \cdot r^{-\gamma}$$

dove f_r è la frequenza dell'item (della specie, nel nostro caso) di rango r e γ rappresenta la pendenza della retta che corrisponde al modello in un sistema log-log.

In tempi più recenti Mandelbrot (1953, 1982), noto per aver organizzato e divulgato la teoria dei frattali, ha proposto una formulazione generalizzata del modello di Zipf, che è stata prontamente adottata anche nel campo dell'ecologia:

$$f_r = f_0 \cdot (r + \beta)^{-\gamma}$$

dove f_r è la frequenza relativa dell'item di rango r, β e γ sono parametri ed f_0 è calcolato in modo tale che la somma delle frequenze di tutti gli items sia pari a 1.

I parametri di questo modello, β e γ, hanno un significato di notevole interesse in rapporto alla struttura di una comuntià. Il parametro γ rappresenta la pendenza dell'asintoto obliquo del modello, mentre il parametro β descrive la deviazione dall'asintoto per gli items (specie) più frequenti.

Anche prescindendo dalla stima dei parametri del modello di Mandelbrot, comunque, l'uso dei diagrammi rango-frequenza può essere molto utile. Ad esempio, si pensi alla possibilità di confrontare i profili ottenuti per stazioni diverse o per la medesima stazione in momenti diversi: una sostanziale coerenza fra di essi indica un assetto omogeneo nello spazio o stabile nel tempo delle comunità studiate, mentre, al contrario, una variazione nei profili osservati è certamente indice di eterogeneità/variabilità.

8. Bibliografia.

Benzécri J.P. *et al*., 1973. *L'Analyse des Données*. 2 vols, Dunod, Paris, France.

Bray R.J. & Curtis J.T., 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.*, 27: 325-349.

Cailliez F. & Pagès J.-P., 1976. *Introduction à l'analyse des données*. Société de Mathématiques appliquées et de Sciences humaines, Paris, xxii+616 pp.

Cliff A.D. & Ord J.K., 1973. *Spatial autocorrelation*. Pion Limited, London, 178 pp.

Cliff A.D. & Ord J.K., 1981. *Spatial processes: models and applications*. Pion Limited, London, 266 pp.

Czekanowski J., 1909. Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenz-Blatt deutsch. Ges. Anthropol. Ethnol. Urgesch*., **40**: 44-47.

Davis J.C., 1986. *Statistics and data analysis in Geology*, 2nd ed., J. Wiley & Sons, New York, 646 pp.

Dice L.R., 1945. Measures of the amount of ecological association between species. *Ecology*, **26**: 297-302.

Diday E., 1971. Une nouvelle méthode en classification automatique et reconnaissance des formes: les nuées dynamiques. *Rev. Stat. appl.*, **19**: 19-35.

Fager E.W. & McGowan J.A., 1963. Zooplankton species groups in the North Pacific. *Science* (Wash. D.C.), **140**: 453-460.

Field J.G., Green R.H., de L. Andrade F. A., Fresi E., Gros P., McArdle B.H., Scardi M. & Wartenberg D., 1987. Numerical ecology: developments for studying the benthos. In: *Developments in Numerical Ecology*, Legendre P. & Legendre L. eds., NATO ASI Series, vol. G14, Springer-Verlag, Berlin Heidelberg: 485-494.

Gabriel K.R. & Sokal R.R., 1969. A new statistical approach to geographic variation analysis. *Syst. Zool.*, **18**: 259-278.

Goodall D.W., 1978. Sample similarity and species correlation. In: *Ordination of plant communities* (R.H: Whittaker, Ed.), W. Junk, The Hague: 99-149.

Gower J.C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**: 325-338.

Gower J.C., 1967. A comparison of some methods of cluster analysis. *Biometrics*, **23**: 623-637.

Gower J.C., 1971. A general coefficient of similarity and some of its properties. *Biometrics*, **27**: 857-871.

Jaccard P., 1900. Contribution au problème de l'immigration post-glaciaire de la flore alpine. *Bull. Soc. vaudoise Sci. nat.*, **36**: 87-130.

Jaccard P., 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. vaudoise Sci. nat.*, **37**: 547-579.

Jaccard P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. vaudoise Sci. nat.*, **44**: 223-270.

Kulczynski S., 1928. Die Pflanzenassoziationen der Pieninen. *Bull. int. Acad. polonaise Sci. et Lettres. Classe Sci. math. et nat.*, *Ser. B*, suppl. II (1927): 57-203.

Lance G.N. & Williams W.T., 1966. Computer programs for classification. *Proc. ANCCAC Conference*, Canberra, May 1966, Paper 12/3.

Legendre L. & Legendre P., 1983. *Numerical ecology*. Elsevier, Amsterdam, 419 pp.

Legendre P. & Legendre L., 1998. *Numerical ecology*. 2nd English edition. Elsevier Science BV, Amsterdam. xv + 853 pp.

Legendre P. & Legendre V., 1984. Postglacial dispersal of freshwater fishes in the Québec peninsula. *Can. J. Fish. Aquat. Sci.*, **41**: 1781-1802.

Legendre P., 1987. Constrained clustering. In: *Developments in Numerical Ecology*, Legendre P. & Legendre L. eds., NATO ASI Series, vol. G14, Springer-Verlag, Berlin Heidelberg: 289-307

Legendre P., Dallot S. & Legendre L., 1985. Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. *Am. Nat.*, **125**: 257-288.

Mantel N., 1967. The detection of desease clustering and a generalized regression approach. *Cancer Res.*, **27**: 209-220.

Matheron G., 1969. Le krigeage universel. *Cah. Cent. Morphol. Math.*, **1**: 1-83.

Matheron G., 1970. La théorie des variables regionalisées et ses applications. *Cah. Cent. Morphol. Math.*, **5**: 1-212.

Moran P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika*, **37**: 17-23.

Motyka J., 1947. O zadaniach i metodach badan geobotanicznych. Sur le buts et le méthodes des recherches géobotaniques. *Ann. Univ. Mariae Curie-Sklodowska Sect. C*, *Suppl. I*, viii+168 pp.

Orloci L., 1967. An agglomerative method for classification of plant communities. *J. Ecol.*, **55**: 193-205.

Orloci L., 1978. *Multivariate analysis in vegetation research*. *2nd ed*., W. Junk, The Hague, ix+451 pp.

Patten B.C., 1962. Species diversity in net phytoplantkon of Raritan Bay. *J. mar. Res.*, **20**:57-75.

Pielou E.C., 1966. The measurement of diversity in different types of biological collections. *J. theor. Biol.*, **13**: 131-144.

Pielou E.C., 1975. *Ecological diversity*. John Wiley & Sons, New York, viii+165 pp.

Pielou E.C., 1984. *The interpretation of ecological data*. John Wiley & Sons, New York, viii+263 pp.

Reyssac J. & Roux M., 1972. Communautés planctoniques dans les eaux de Côte d'Ivoire. Groupes d'espèces associeées. *Mar. Biol.* (Berl.), **13**: 14-33.

Rogers D.J. & Tanimoto T.T. - 1960. A computer program for classifying plants. *Science* (Wash. D.C.), **132**: 1115-1118.

Rohlf F.J., 1963. Classification of *Aedes* by numerical taxonomic methods (Diptera: Culicidae). *Ann. entomol. Soc. Am.*, **56**: 798-804.

Shannon C.E., 1948. A mathematical theory of communications. *Bell System technical Journal*, **27**: 379-423, 623-656.

Sneath P.H.A., 1966. A comparison of different clustering methods as applied to randomly-spaced points. *Classification Soc: Bull.*, **1**: 2-18.

Sneath P.H.A. & Sokal R.R., 1973. *Numerical taxonomy - The principles and practice of numerical classification*. W.H. Freeman, San Francisco, xv+573 pp.

Sokal R.R. & Michener C.D., 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, ***38***: 1409-1438.

Sokal R.R. & Sneath P.H.A., 1963. *Principles of numerical taxonomy*. W.H. Freeman, San Francisco, xvi+359 pp.

Sørensen T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr.*, **5**: 1-34.

APPENDICE

In questa Appendice sono raccolte, in ordine sparso, alcune note su alcuni dei (tanti) metodi non presentati nel testo precedente. La bibliografia è essenziale e riportata per ciascuna nota.

Tests su proporzioni.

Se intendiamo come "proporzione" il rapporto fra il numero di oggetti in un dato sottoinsieme (es. *m*) e quello degli oggetti nell'insieme a cui esso appartiene (es. *n*), cioè $p=m/n$, con $0 \leq \pi \leq 1$, allora è possibile verificare se tale proporzione si discosta in maniera significativa dallo zero, cioè se il numero di casi in cui si è osservato un determinato carattere o evento può essere considerato sufficientemente alto rispetto al numero di osservazioni disponibili da farlo ritenere non casuale. In particolare, si tratta di testare l'ipotesi nulla H_0: la proporzione *p* è uguale a 0.

Per effettuare questo test è necessario che il numero di osservazioni disponibili (cioè *n*) sia sufficientemente grande. In linea di massima, il test può essere applicato se $n \geq 200$. La ragione di questo limite sta nel fatto che la statistica che verrà calcolata è distribuita come un *t* di Student soltanto per campioni di grandi dimensioni.

Nel caso in cui questa condizione sia verificata, allora:

$$t_{GL=1} = \frac{p}{s_p} \qquad \text{dove } s_p = \sqrt{\frac{p(1-p)}{n}}$$

In altre parole, se la statistica ha un valore superiore a quello del *t* di Student per il livello di significatività prescelto, si può rigettare l'ipotesi nulla di identità con lo zero e quindi concludere che la proporzione osservata differisce significativamente da una proporzione nulla, cioè dall'assenza del carattere o dell'evento considerati nell'insieme di osservazioni di riferimento. I valori di riferimento per il *t* di Student sono 12.706 e 63.656, rispettivamente per livelli di probabilità P=0.05 e P=0.01.

Un caso analogo, ma leggermente più complesso, è quello per cui si vogliono confrontare fra loro due proporzioni, p_1 e p_2. In effetti, quello appena descritto è un caso particolare di questo confronto più generale, cioè il caso per cui $p_2=0$.

Per poter testare l'ipotesi nulla H_0: $p_1=p_2$, è necessario innanzitutto che n_1 e n_2, cioè il numero totale di osservazioni su cui sono stati

determinate p_1 e p_2, sia sufficientemente grande, quindi almeno $n_1 \geq 200$ ed $n_2 \geq 200$.

Nel caso in cui questa condizione sia verificata, allora:

$$t_{GL=\infty} = \frac{|p_1 - p_2|}{s_p}$$

dove, in questo caso, sarà

$$s_p = \sqrt{p^*(1-p^*)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \quad \text{con} \quad p^* = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Anche in questo caso, se la statistica ha un valore superiore a quello del *t* di Student per il livello di significatività prescelto, si può rigettare l'ipotesi nulla di identità fra le due proporzioni e quindi concludere che esse differiscono significativamente fra loro. I valori di riferimento per il *t* di Student per infiniti gradi di libertà sono 1.960 e 2.576, rispettivamente per livelli di probabilità P=0.05 e P=0.01.

MRPP

L'MRPP (Multi-Response Permutation Procedure) è una procedura non parametrica che serve ha verificare l'ipotesi di identità fra due o più gruppi di oggetti definiti ***a priori***. Per esempio, si può confrontare la composizione del popolamento di aree esposte ad un trattamento o ad un impatto e di aree indenni per verificare l'ipotesi di mancanza di effetto da parte del trattamento. Questa procedura presenta delle affinità, fra le tecniche parametriche, con l'analisi discriminante. Tuttavia, l'MRPP presenta il vantaggio di non richiedere che siano soddisfatte particolari condizioni (come ad esempio la multinormalità e l'omogeneità delle varianze), che sono raramente soddisfatte dai dati ecologici.

Il metodo è stato introdotto da Mileke et al. (1976), ma una buona presentazione del metodo stesso si può trovare in appendice in Biondini *et al.* (1985). Ulteriori dettagli possono essere reperiti in Mielke (1984) e

in Berry *et al.* (1983). Infine, si può far riferimento a Zimmerman *et al.* (1985) per un esempio di applicazione ecologica del metodo.

Se Δ è la distanza media intragruppo osservata fra i gruppi definiti *a priori*, pesata per la dimensione dei diversi gruppi, è possibile calcolare il valore atteso e la varianza della medesima statistica nel caso di una disposizione casuale degli oggetti nei diversi gruppi mediante opportune permutazioni degli oggetti. La statistica T che viene testata descrive la separazione fra gruppi ed definita come la differenza fra il Δ osservato e quello atteso, divisa per la radice quadrata della varianza dei Δ:

$$T = \frac{\Delta_{osservato} - \Delta_{atteso}}{\sqrt{\sigma^2_{\Delta}}}$$

La statistica osservata viene testata in rapporto alla distribuzione empirica dei Δ e viene stimata la probabilità di osservare un Δ uguale o maggiore di quello osservato.

La statistica R ha invece il ruolo di descrivere l'omogeneità dei gruppi in rapporto ad una attesa di casualità. Essa si ottiene come:

$$R = 1 - \frac{\Delta_{osservato}}{\Delta_{atteso}}$$

Quando tutti gli oggetti sono uguali all'interno dei gruppi, allora il Δ osservato è pari a 0 ed R è pari a 1, cioè è massimo. Se l'eterogeneità nei gruppi è uguale all'attesa in condizioni di casualità, allora R tende ad essere nullo. Infine, se l'accordo fra gruppi è inferiore a quanto atteso sotto l'ipotesi di casualità, allora R è negativo.

In generale, il test viene eseguito utilizzando la distanza euclidea ed applicando ad ogni gruppo un peso proporzionale alla sua dimensione. Altre soluzioni sono teoricamente possibili, anche se meno frequentemente applicate.

E' importante sottolineare il fatto che esiste la possibilità di effettuare questo test anche in maniera esatta, senza calcolare la statistica *T*, ma piuttosto confrontando il Δ osservato con la distribuzione dei Δ ottenuti per tutte le combinazioni possibili dei diversi oggetti nei diversi gruppi. E' evidente, tuttavia, che il numero di

combinazioni in gioco è elevatissimo anche per insiemi di dati di modeste dimensioni e che il metodo non è praticamente applicabile a problemi reali. In particolare, le possibili combinazioni di n oggetti ripartiti in p gruppi contenenti rispettivamente m_1, m_2, …, m_n oggetti sono:

$$N = \frac{n!}{m_p! \prod_{i=1}^{p-1} m_i!}$$

Considerando un caso pratico, con 30 oggetti ripartiti in 3 gruppi di 10 oggetti, le combinazioni possibili sono 5.550.996.791.340, cioè più di 5000 miliardi ed il tempo di calcolo necessario su un PC di buone prestazioni sarebbe nell'ordine dei 2-3 anni!

Berry, K. J., K. L. Kvamme, and P. W. Mielke, Jr. 1983. Improvements in the permutation test for the spatial analysis of the distribution of artifacts into classes. American Antiquity, 48: 547-553.

Biondini, M. E., C. D. Bonham, and E. F. Redente. 1985. Secondary successional patterns in a sagebrush (Artemisia tridentata) community as they relate to soil disturbance and soil biological activity. Vegetatio, 60: 25-36.

Mielke, P.W., K.J. Berry, and E.S. Johnson. 1976. Communications in Statistics: Theory and Methods, 5:1409-1424.

Mielke, P. W., Jr. 1984. Meteorological applications of permutation techniques based on distance functions. Pages 813-830. In P. R. Krishnaiah and P. K. Sen, eds., Handbook of Statistics, Vol. 4. Elsevier Science Publishers.

Zimmerman, G. M., H. Goetz, and P. W. Mielke, Jr. 1985. Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology, 66: 606-611.

Indicator species analysis.

Al fine di valutare con esattezza quali siano gli elementi faunistici o floristici più caratterizzanti di una serie di osservazioni ripartite in gruppi o classi definite ***a priori***, si può fare ricorso ad una tecnica di recente introduzione, nota come *Indicator Species Analysis* (Dufrene & Legendre, 1997).

Questa consente di valutare il valore delle singole specie come indicatrici di particolari condizioni ambientali o, comunque, di condizioni che caratterizzano un gruppo di osservazioni. La procedura di calcolo è estremamente semplice e combina le informazioni relative alla densità delle specie con quelle relative alla costanza della loro presenza in un gruppo di osservazioni, restituendo per ogni gruppo di osservazioni un valore indicatore (*indicator value* o INDVAL) per ogni specie analizzata.

Se A è una matrice in cui le n righe rappresentano altrettanti campioni e le s colonne altrettante specie e se i campioni sono ripartiti fra g gruppi di numerosità n_k, allora a_{ijk} è l'abbondanza della j-ma specie nell'i-mo campione che appartiene al gruppo k.

Si può quindi calcolare l'abbondanza media x_{kj} della specie j nel gruppo di campioni k:

$$x_{kj} = \sum_{i=1}^{n_k} \frac{a_{ijk}}{n_k}$$

e da questa l'abbondanza relativa RA_{kj} della specie j nel gruppo di campioni k:

$$RA_{kj} = \frac{x_{kj}}{\sum_{k=1}^{g} x_{kj}}$$

Successivamente si deve calcolare la frequenza media RF_{kj} della presenza di una specie j nel gruppo di campioni k. Per ottenere tale frequenza si trasforma innanzitutto la matrice A in una matrice binaria B e quindi si calcola:

$$RF_{kj} = \sum_{i=1}^{n_k} \frac{b_{ijk}}{n_k}$$

Combinando abbondanze relative (RA) e frequenze medie (RF) si ottiene quindi il valore indicatore (IV), che può essere poi espresso come una percentuale:

$$IV_{kj} = RA_{kj} \cdot RF_{kj} \cdot 100$$

Il valore indicatore più elevato riscontrato per ciascuna specie fra i diversi gruppi di campioni è quindi considerato come il valore indicatore generale per la specie in esame.

La significatività dei valori indicatori viene poi testata utilizzando una tecnica di Monte Carlo, cioè confrontando i valori indicatori osservati per ciascuna specie con una distribuzione empirica di riferimento generata riassegnando casualmente i singoli campioni ai diversi gruppi e ricalcolando i valori indicatori per un numero molto elevato di volte (es. 1000 volte). In particolare, saranno considerati significativi i valori indicatori per i quali il valore osservato supera almeno il 95% di quelli ottenuti per permutazione aleatoria dell'insieme dei dati.

Dufrene, M. and P. Legendre, 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs, 67(3): 345-366.

Analisi Canonica delle Corrispondenze.

L'Analisi Canonica delle Corrispondenze (CCA; ter Braak 1986, 1994) è una delle tecniche di ordinamento più largamente utilizzate ed è affine per molti aspetti all'Analisi delle Corrispondenze. Rispetto a quest'ultima, tuttavia, essa è vincolata da una regressione multipla su un secondo insieme di variabili, usualmente rappresentate da descrittori ambientali. Se è vero che ciò può facilitare l'interpretazione dei risultati, è anche vero che molta cautela va posta nella selezione delle variabili ambientali, che devono essere assolutamente pertinenti rispetto all'insieme di dati in esame. L'approccio basato sull' Analisi Canonica delle Corrispondenze è quindi differente da quello, per certi versi più

conservativo, che prevede un ordinamento dei dati relativi al popolamento e, in un secondo momento, un'analisi indipendente delle correlazioni fra i risultati di quest'ultimo ed i descrittori ambientali. Una discussione delle diverse filosofie che sono alla base dell'uso dell'Analisi Canonica delle Corrispondenze e delle altre tecniche di ordinamento è stata presentata da Okland (1996).

In pratica, per l'Analisi Canonica delle Corrispondenze sono necessarie due matrici di dati: la prima è organizzata in modo tale che le righe corrispondano alle stazioni (o ai campioni) e le colonne alle specie; la seconda contiene lo stesso numero di righe (stazioni) e tante colonne quante sono le variabili ambientali (o comunque quelle su cui vincolare l'ordinamento). E' necessario, in particolare, che queste ultime siano meno numerose delle righe delle due matrici (cioè delle stazioni o dei campioni analizzati).

L'Analisi Canonica delle Corrispondenze può operare in maniera ottimale se sono soddisfatte due condizioni: le risposte delle specie ai gradienti ambientali devono essere unimodali e le variabili ambientali utlilizzate devono essere quelle che effettivamente determinano la struttura dei cenoclini.

Okland, R. H., 1996. Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? Journal of Vegetation Science, 7:289-292.

Ter Braak, C. J. F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology, 67:1167-1179.

Ter Braak, C. J. F., 1994. Canonical community ordination. Part I: Basic theory and linear methods. Ecoscience, 1:127-140.

Il test U di Mann-Whitney.

Il test U di Mann-Whitney è un'alternativa non-parametrica al test t di Student sulle differenze fra medie. Il test U assume che la variabile in considerazione sia stata misurata su una base almeno ordinale (come rango, quindi), ma può ovviamente essere applicato anche a dati quantitativi se si ritiene che le condizioni necessarie affinchè il test t di Student possa essere applicato non siano soddisfatte (es. nel caso in cui la distribuzione dei dati non sia normale). L'interpretazione del test è essenzialmente identica a quella di un test t di Student, con la sola eccezione che la statistica U è calcolato a partire da somme di ranghi piuttosto che da valori medi. Il test U è generalmente assai potente (cioè sensibile), anche se è di tipo non-parametrico. In alcuni casi particolari, infatti, può addirittura avere una capacità di rigettare l'ipotesi nulla superiore a quella del test t.

Con campioni di dimensione maggiore di 20 la distribuzione della statistica U si avvicina rapidamente ad una distribuzione normale (cfr. Siegel, 1956). Di conseguenza, la statistica U (se corretta per i casi di pareggio fra ranghi), può essere associata ad un valore di probabilità derivato da una distribuzione normale.

Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Il test di Kolmogorov-Smirnov

Il test di Kolmogorov-Smirnov verifica l'ipotesi che due campioni siano stati estratti dalla medesima popolazione. A differenza del test parametrico t di Student o della sua controparte non-parametrica, il test U di Mann-Whitney, che considerano le tendenze centrali dei due campioni sottoposti al test (rispettivamente come medie o come somme dei ranghi), il test di Kolmogorov-Smirnov è anche sensibile alle differenze in termini di "forma" delle distribuizioni dei due campioni. In altri termini, esso è sensibile, oltre che alle differenze in termini di valori

medi, anche a quelle relative alla dispersione, all'asimmetria ed ad altre caratteristiche dei campioni (es. bimodalità).

Multidimensional Scaling Non-metrico

Il Multidimensional Scaling Non-metrico (NMDS) è una tecnica di ordinamento sostanzialmente differente dalla maggior parte di quelle solitamente utilizzate. Infatti, esso non è basato su una procedura che preveda l'estrazione di autovalori ed autovettori da una matrice di distanze, similarità, correlazione, etc. Al contrario, questo metodo è basato su un algoritmo iterativo che prevede un aggiustamento progressivo della posizione dei punti nel piano o nello spazio più o meno complesso in cui si desidera ottenere l'ordinamento.

Come nel caso del Multidimensional Scaling Metrico (o Analisi delle Coordinate Principali), ciò che viene minimizzato è lo scarto fra la struttura delle distanze nello spazio originale e quella ottenuta nello spazio ridotto dell'ordinamento (tale scarto è anche detta *stress*). Ciò viene ottenuto mediante un algoritmo basato sulla regressione monotona delle distanze nello spazio dell'ordinamento su quelle originali (Kruskal, 1964).

Come tutti i metodi iterativi, il Multidimensional Scaling Non-metrico non necessariamente fornisce la stessa soluzione se eseguito una seconda volta sul medesimo insieme di dati, poichè l'algoritmo impiegato può essere attratto verso un minimo locale di stress. Tuttavia, esso ha il vantaggio di poter essere applicato a qualsiasi matrice di distanza o di dissimilarità, anche qualora quest'ultima non goda di proprietà metriche.

Kruskal, J. B., 1964. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115-129.

ANOSIM (ANalysis Of SIMilarities)

Il test ANOSIM (acronimo per ANalysis Of SIMilarities) è una procedura non-parametrica che consente di verificare se le differenze fra due o più gruppi di osservazioni multivariate sono significative o meno (Clarke, 1993). Il test può essere effettuato su una qualunque misura di distanza o di dissimilarità fra le osservazioni da analizzare ed è concettualmente affine al Multidimensional Scaling Non-Metrico (NMDS) per il fatto che utilizza il rango delle distanze (o dissimilarità) piuttosto che i loro valori effettivi.

Come tutti i test parametrici e non-parametrici che testano le differenze fra gruppi, è opportuno operare in condizioni di eteroscedasticità il più possibile ridotta, ovvero evitando di confrontare gruppi la cui variabilità interna sia troppo diversa. Purtroppo, poiché non esistono procedure che consentano di stimare la soglia critica di eteroscedasticità accettabile per determinato insieme di dati e per una determinata misura di distanza o dissimilarità, la valutazione della corretta applicazione del test non può che essere affidata ad una valutazione euristica, confidando nella robustezza del test rispetto a deviazioni dalle condizioni di applicazione ideali.

Ai fini del calcolo della statistica R, su cui è basato il test, si devono trasformare i singoli elementi della matrice di distanza o dissimilarità nei rispettivi ranghi, calcolando poi il rango medio delle distanze o dissimilarità intra-gruppo (r_w) e di quelle inter-gruppo (r_b). Se N è la dimensione della matrice analizzata, la statistica R si ottiene poi come:

$$R = \frac{r_b - r_w}{N(N-1)/4}$$

Valori positive di R indicano che le distanze fra gruppi sono maggiori di quelle all'interno dei gruppi. Per testare la significatività della statistica R si confornta il valore osservato con una distribuzione empirica dei valori della stessa statistica ottenuti permutando aleatoriamente righe e colonne della matrice analizzata un numero molto elevato di volte (ciò equivale a riassegnare ciascuna

osservazione ad un gruppo a caso). Se il valore osservato di R è maggiore del 95% o del 99% dei valori ottenuti con le permutazioni casuali della matrice analizzata, si può concludere che esso sia significativo e che di conseguenza lo siano le differenze fra i grupi di osservazioni.

Il test viene effettuato abitualmente ad una coda perché è discutibile il significato di un valore negativo di R (che pure a volte si osserva). Infatti, ciò implicherebbe una eterogeneità intra-gruppo maggiore di quella inter-gruppo, ma è evidente che questa condizione non può che originare da condizioni fortuite. Infatti, nel caso di osservazioni assegnate a caso ai diversi gruppi, il valore di R dovrebbe tendere a zero, perché non ci sarebbero differenze fra le distanze o dissimilarità intra-gruppo e quelle inter-gruppo.

Clarke, K.R. 1993. Non-parametric multivariate analysis of changes in community structure. Australian Journal of Ecology 18:117-143.

Il coeffciente di Spearman

Nel caso in cui due variabili siano legate fra loro da una relazione non lineare (o nel caso in cui si ritenga verosimile questa ipotesi su base teorica), l'uso di un coefficiente di correlazione di rango è preferibile rispetto a quello del consueto coefficiente di correlazione lineare di Bravais-Pearson. Ovviamente, la relazione soggiacente i dati da analizzare deve essere, per quanto non lineare, almeno monotona, ovvero al variare di una delle due variabili deve corrispondere una variazione dell'altra che abbia sempre lo stesso segno indipendentemente dal valore considerato.

Un coefficiente di correlazione di rango frequentemente utilizzato è quello di Spearman, che è strettamente imparentato con quello di Bravais-Pearson.

Il coefficiente di Spearman, se non ci sono casi di ranghi uguali nelle serie da analizzare, si ottiene come:

$$r = \frac{\sum_{i=1}^{n}\left[R(x_i) - \frac{n+1}{2}\right]\left[R(y_i) - \frac{n+1}{2}\right]}{n(n^2-1)/12}$$

dove $R(x_i)$ ed $R(y_i)$ sono i ranghi delle osservazioni *i*-me per le due variabili *x* e *y*.

Nel caso in cui la frequenza dei dati con uguale rango sia bassa (<20%), si può utilizzare invece la seguente formulazione:

$$r = 1 - \frac{6\sum_{i=1}^{n}\left[R(x_i) - R(y_i)\right]^2}{n(n^2-1)}$$

Infine, nel caso di una frequenza più elevata dei dati con uguale rango (>20%), una stima più corretta si otterrà facendo riferimento ad una formulazione che equivale al calcolo del coefficiente di Bravais-Pearson sui ranghi centrati rispetto ai ranghi medi [cioè a $(n+1)/2$]:

$$r = \frac{\sum_{i=1}^{n}\left[R(x_i) - \frac{n+1}{2}\right]\left[R(y_i) - \frac{n+1}{2}\right]}{\sqrt{\sum_{i=1}^{n}\left[R(x_i) - \frac{n+1}{2}\right]^2 \cdot \sum_{i=1}^{n}\left[R(y_i) - \frac{n+1}{2}\right]^2}}$$

In tutti i casi si potrà verificare la significatività della correlazione (con $H_0: r=0$) mediante apposite tavole o, per $n \geq 10$, sapendo che

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

è distribuito come un *t* di Student con n-2 gradi di libertà.

SIMPER

SIMPER (Similarity Percentage) è un semplice metodo per stimare quali taxa siano responsabili più di altri delle differenze fra due o più gruppi di osservazioni (Clarke, 1993). Nella implementazione più

spesso utilizzata SIMPER sottende una distanza di Bray-Curtis (1957) ed il contributo δ_i alla distanza complessiva fra due campioni della specie *i*-ma è dunque:

$$\delta_i = \frac{\left| x_{ij} - x_{ik} \right|}{\sum_{i=1}^{s} \left(x_{ij} + x_{ik} \right)}$$

I contributi δ_i possono poi essere mediati su tutti i campioni che costituiscono due o più gruppi di campioni per identificare le specie che globalmente pesano di più nel edeterminare la distanza complessiva fra i gruppi.

La deviazione standard dei contributi δ_i indica, quando i suoi valori sono piccoli, che una specie ha un ruolo coerente in tutti i campioni di ciascun gruppo. Inoltre, i contributi standardizzati (divisi quindi per la loro deviazione standard) possono essere utilizzati per meglio valutare quali siano le specie diagnostiche ai fini della discriminazione di due gruppi. La differenza complessiva fra i gruppi può essere stimata a sua volta mediante ANOSIM (Clarke 1993).

Runs test

La successione di due stati di un sistema, di due livelli di risposta, o di qualsiasi cosa possa essere ricondotta ad una sequenza di dati binari può essere sottoposta ad un *runs test* per verificare se la sequenza osservata è compatibile con l'ipotesi nulla di una generazione casuale o meno. Una applicazione molto frequente è quella per cui si assegna 1 ad ogni osservazione in una seri in cui il valore sia maggiore della mediana e 0 ad ognuna di quelle al di sotto della mediana. Poiché, in un caso del genere, bisognerà stabilire cosa fare se un valore è esattamente uguale alla mediana, è evidente che questo approccio non è ideale se il numero dgli stati possibili è piccolo.

In pratica, si deve innanzitutto determinare il numero U di sequenze di valori identici nella serie considerata. Se, ad esempio, la serie fosse 100010001111001, allora si avrebbero sei sequenze di valori identici, cioè U=6 (1 000 1 000 1111 00 1). Se i due stati ricorrono rispettivamente in n_1 ed n_2 casi (7 volte 1 e 8 volte 0, nell'esempio), allora il numero medio $\overline{U}$ di sequenze dello stesso stato che è atteso per una serie generata casualmente sarà pari a:

$$\overline{U} = \frac{2n_1n_2}{n_1 + n_2} + 1$$

mentre l'errore standard σ_U si stima come:

$$\sigma_U = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

Se sia n_1 che n_2 sono maggiori di 10, allora la distribuzione di U può essere approssimata da una distribuzione normale, per cui è possibile utilizzare il seguente test *Z*:

$$Z = \frac{U - \overline{U}}{\sigma_U}$$

Si rigetta dunque l'ipotesi nulla H_0: $U = \overline{U}$ se il valore di *Z* è maggiore in valore assoluto di 1.96 (per un livello di probabilità P=0.05). Nel caso, peraltro non infrequente nel caso di dati lepidocronologici, in cui n_1 o n_2 sono minori o uguali a 10 si fa ricorso ad apposite tavole di riferimento.

Cross-association

In alcuni casi è necessario confrontare fra loro due sequenze di stati di un sistema che non possono essere tradotti in termini quantitativi o ordinali, ma che possono invece essere codificati nominalmente (es. ABBACAABBBBCABCC). In questo caso, si può verificare l'associazione fra le due serie, eventualmente decalate fra loro, medianta il calcolo della *cross-association*.

La probabilità di osservare una concordanza fra due serie sarà:

$$P = \frac{\sum_{k=1}^{m} X_{1k} X_{2k}}{n^2}$$

dove X_{ik} è il numero di casi in cui il *k*-mo valore ricorre nella *i*-ma serie ed *m* è il numero dei possibili stati delle serie (pari a 3 nella serie mostrata come esempio). La probabilità *P'* di osservare una discordanza, ovviamente è il complemento a 1 di *P*.

La statistica da testare è un X^2 con un solo grado di libertà (ν=1), a cui si deve applicare la correzione di Yates (cioè sottrarre 1/2) nel caso di serie brevi. Ovvero:

$$\mathrm{X}^2 = \frac{\left(O - E - \frac{1}{2}\right)^2}{E} + \frac{\left(O' - E' - \frac{1}{2}\right)^2}{E'}$$

dove *O* ed *O'* sono rispettivamente le concordanze e le discordanze osservate ed *E* ed *E'* sono i corrispondenti valori attesi, che si ottengono moltiplicando rispettivamente *P* e *P*' per *n*.

Il valore tabulare di X^2 per ν=1 è pari a 3.84 per p=0.05 ed a 6.63 per p=0.01. Se il valore della statistica è superiore a queste soglie si può concludere che il numero di concordanze è superiore a quello atteso nel caso di una generazione casuale delle serie: in questo caso si può rigettare l'ipotesi nulla di indipendenza delle serie a confronto.