Osservazioni comparative sul comportamento di tre indici di similarità per dati binari

MAURIZIO BIONDI
Dipartimento di Biologia Animale e dell'Uomo. Università di Roma

SUMMARY

Comparative observations about the behaviour of three similarity coefficients for binary data

In this work three different similarity coefficients for binary data are compared. The coefficients considered are those proposed by: Baroni Urbani & Buser (1976), Jaccard (1908) and Dice-Sørensen (cfr. Sørensen, 1948). This study has been carried out using a personal computer Olivetti M24. The results obtained are: a) the Jaccard's coefficient keeps always very inferior to the effective similarity percentage value; b) the Baroni Urbani & Buser's coefficient presents the better approssimation in the most part of the situations, c) the Dice-Sørensen's coefficient can be powerfully used in some particular situations, as: 1) comparisons between OTUs potentially very similar and sufficiently sampled (at least the 60% of the effective attributes present in every OTU); 2) comparison between OTUs potentially with low similarity, very carefully sampled (at least the 90% of the effective attributes present in every OTU), without consider the contribution due to the number of attributes absent in both OTUs compared.

Nella presente nota sono riportate alcune osservazioni comparative riguardanti il comportamento di tre indici utilizzati nel calcolo della similarità con dati binari, ed impiegati frequentemente nell'ambito di ricerche faunistiche ed ecologiche. Gli indici presi in esame sono quelli proposti da: Baroni Urbani e Buser (1976), Jaccard (1908) e Dice-Sørensen (cfr. Sørensen, 1948).

Le considerazioni di seguito esposte rappresentano il frutto di un approccio di carattere esclusivamente pratico e non intendono assolutamente entrare nel merito di questioni teorico-matematiche.

Questa nota riunisce quindi una serie di indicazioni, utili per un orientamento nella scelta appropriata dell'indice di similarità da utilizzare nelle diverse situazioni sperimentali, scelta molto spesso affidata al caso, all'abitudine o alla disponibilità del software.

MATERIALI E METODI

Un indice di similarità misura il grado di somiglianza esistente tra due unità statistiche (OTUs)([1]) descritte da un set comune di variabili (attributi).

([1]) Le unità statistiche saranno indicate nell'ambito di questa nota con l'acronimo inglese OTU (operational taxonomic unit), per conformità con gran parte della letteratura esistente sull'argomento.

Gli indici qui considerati assumono valori compresi tra 0 (massima diversità) e 1 (massima similarità), e sono applicabili a matrici booleane, nelle quali i dati qualitativi sono espressi da variabili di tipo «assenza-presenza», e codificati di norma con «0» per l'assenza (attributo negativo) e «1» per la presenza (attributo positivo). In genere, nelle ricerche a carattere ecologico e faunistico, le OTUs sono rappresentate da diversi tipi di ambienti o di aree geografiche e gli attributi da categorie tassonomiche, quali specie, generi, ecc.

La presente analisi è stata impostata creando una serie di matrici booleane randomizzate, ciascuna comprendente 5 OTUs con 100 attributi ciascuna, distribuiti secondo i seguenti 4 casi:

A) ciascuna OTU presenta il 100% dei suoi attributi in comune con le altre 4. Percentuale di similarità ipotizzata (espressa dal rapporto del numero di attributi positivi comuni a ciascuna coppia di OTUs, sul numero totale di attributi presenti in ciascuna OTU) = 100%; matrice 5x100;

B) ciascuna OTU presenta il 75% dei suoi attributi in comune con le altre 4. Percentuale di similarità ipotizzata per ciascuna delle coppie di OTUs confrontate = 75%; matrice 5×200;

TABELLA 1 - Valori medi e relative deviazioni standard degli indici di similarità calcolati per il caso A (per le spiegazioni vedere il testo).

% CONOSCENZA	DICE-SØRENSEN		JACCARD		BARONI-BUSER	
	$\bar{X}$	DS	$\bar{X}$	DS	$\bar{X}$	DS
10	.070	.067	.037	.037	.160	.142
20	.225	.072	.128	.046	.341	.070
30	.245	.070	.141	,045	.348	.069
40	.465	.055	.304	.046	.496	.053
50	.484	.037	.320	.042	.481	.038
60	.623	.047	.458	.049	.576	.054
70	.679	.034	.515	.039	.599	.046
80	.788	.015	.650	.021	.699	.023
90	.901	.010	.820	.016	.832	.024
100	1.000	.000	1.000	.000	1.000	.000

TABELLA 2 - Valori medi e relative deviazioni standard degli indici di similarità calcolati per il caso B (per le spiegazioni vedere il testo).

% CONOSCENZA	DICE-SØRENSEN		JACCARD		BARONI-BUSER	
	$\bar{X}$	DS	$\bar{X}$	DS	$\bar{X}$	DS
10	.110	.099	.061	.057	.217	.165
20	.182	.126	.105	.078	.293	.153
30	.197	.074	.111	.047	.317	.075
40	.317	.054	.190	.038	.407	.050
50	.420	.052	.267	.042	.479	.047
60	.468	.063	.308	.055	.508	.059
70	.504	.041	.338	.037	.540	.039
80	.589	.029	.418	.029	.609	.028
90	.694	.024	.532	.028	.696	.024
100	.750	.000	.600	.000	.750	.000

C) ciascuna OTU presenta il 50% dei suoi attributi in comune con le altre 4. Percentuale di similarità ipotizzata per ciascuna delle coppie di OTUs confrontate = 50%; matrice 5×300;

D) ciascuna delle OTU presenta il 25% dei suoi attributi in comune con le altre 4. Percentuale di similarità ipotizzata per ciascuna delle coppie di OTUs confrontate = 25%; matrice 5×400.

Per ognuno dei casi sopra esposti, sono state estratte da ogni OTU, con l'ausilio di uno specifico software, dieci serie di numeri random, pari rispettivamente ai livelli percentuali del 10, 20, 30, 40, 50, 60, 70, 80, 90, 100% del numero di attributi presenti in ciascuna OTU, come definito per ipotesi.

In altre parole si immagina una situazione sperimentale, nella quale si vuole valutare il grado di somiglianza esistente tra 5 differenti aree geografiche, comprendenti ciascuna 100 specie animali. Per ipotesi sappiamo che: nel caso A, ciascuna area ha esattamente le stesse specie presenti nelle altre 4, (100% di similarità); nel caso B, soltanto 75 specie di ogni area sono in comune con le altre 4, mentre le restanti 25 sono esclusive (75% di similari-

TABELLA 3 - Valori medi e relative deviazioni standard degli indici di similarità calcolati per il caso C (per le spiegazioni vedere il testo).

% CONOSCENZA	DICE-SØRENSEN		JACCARD		BARONI-BUSER	
	$\bar{X}$	DS	$\bar{X}$	DS	$\bar{X}$	DS
10	.040	.070	.022	.038	.088	.146
20	.090	.061	.048	.034	.214	.102
30	.150	.080	.083	.049	.286	.088
40	.220	.035	.124	.023	.352	.034
50	.236	.044	.134	.029	.366	.043
60	.297	.037	.175	.026	.415	.035
70	.372	.027	.229	.021	.475	.024
80	.400	.032	.250	.025	.493	.028
90	.447	.009	.287	.007	.534	.008
100	.500	.000	.333	.000	.577	.000

TABELLA 4 - Valori medi e relative deviazioni standard degli indici di similarità calcolati per il caso D (per le spiegazioni vedere il testo).

% CONOSCENZA	DICE-SØRENSEN		JACCARD		BARONI-BUSER	
	$\bar{X}$	DS	$\bar{X}$	DS	$\bar{X}$	DS
10	.020	.042	.010	.022	.052	.110
20	.040	.032	.021	.016	.135	.097
30	.067	.035	.035	.019	.200	.055
40	.077	.032	.040	.017	.215	.053
50	.096	.031	.051	.017	.240	.041
60	.140	.038	.076	.022	.293	.042
70	.143	.021	.077	.012	.297	.025
80	.192	.021	.107	.013	.346	.021
90	.225	.015	.127	.009	.377	.015
100	.250	.000	.143	.000	.400	.000

tà); nel caso C, le specie in comune sono 50 e le esclusive 50 (50% di similarità); nel caso D, le specie in comune sono 25 e le esclusive 75 (25% di similarità). Viene quindi effettuata, in ogni area per ciascuno dei casi presi in esame, una serie di 10 campionamenti, durante i quali si cattura, per ipotesi, rispettivamente il 10, 20, 30, 40, 50, 60, 70, 80, 90 e 100% delle specie presenti.

In base ai dati così ottenuti, sono state quindi «costruite» un totale di 40 matrici binarie, 10 per ciascun caso, 1 per ciascun livello di conoscenza considerato. Su queste matrici sono stati quindi calcolati i valori degli indici di similarità di Baroni Urbani e Buser, Jaccard e Dice-Sørensen, dei quali più avanti sono riportate le rispettive formule. Per ciascuna matrice, formata da 5 OTUs sono stati effettuati 10 confronti, pari alla metà del numero delle possibili disposizioni semplici di classe 2, sottratto il numero dei confronti di ciascuna OTU con se stessa.

Per le operazioni sopra descritte è stato utilizzato un personal computer Olivetti M24, corredato da adeguato software.

Le formule degli indici di similarità utilizzati sono le seguenti:

Baroni Urbani e Buser: $(\sqrt{cd}+c)/(\sqrt{cd}+c+a+b)$

Jaccard: $c/(a+b+c)$

Dice-Sørensen: $2c/(a+b+2c)$

a = numero di attributi presenti nella prima OTU ed assenti nella seconda;
b = numero di attributi presenti nella seconda OTU ed assenti nella prima;
c = numero di attributi comuni alle due OTUs;
d = numero di attributi assenti sia nella prima che nella seconda OTU, ma presenti in almeno una delle altre OTUs considerate.

Rispetto alle situazioni di tipo reale, le principali limitazioni presentate dall'utilizzo di questo modello randomizzato, sono:
- ogni OTU ha per ipotesi lo stesso numero di attributi positivi rispetto alle altre;
- ogni attributo presenta la stessa possibilità di essere «raccolto».

RISULTATI

Per ciascuno dei 4 casi presi in esame, sono riportati nelle tabb. 1-4, i valori medi e le relative deviazioni standard dei tre indici di similarità calcolati, ordinati rispetto ai dieci livelli di conoscenza considerati. Nelle figg. 1-4 sono riportati gli andamenti grafici relativi.

Dall'esame dei grafici si può osservare:
- in tutti i casi considerati, l'indice di Jaccard sottostima notevolmente il valore percentuale di similarità ipotizzato;
- nel caso A (fig. 1) (percentuale di similarità ipotizzata = 100%), si ha per valori di conoscenza di ciascuna OTU inferiori al 50%, una miglio-

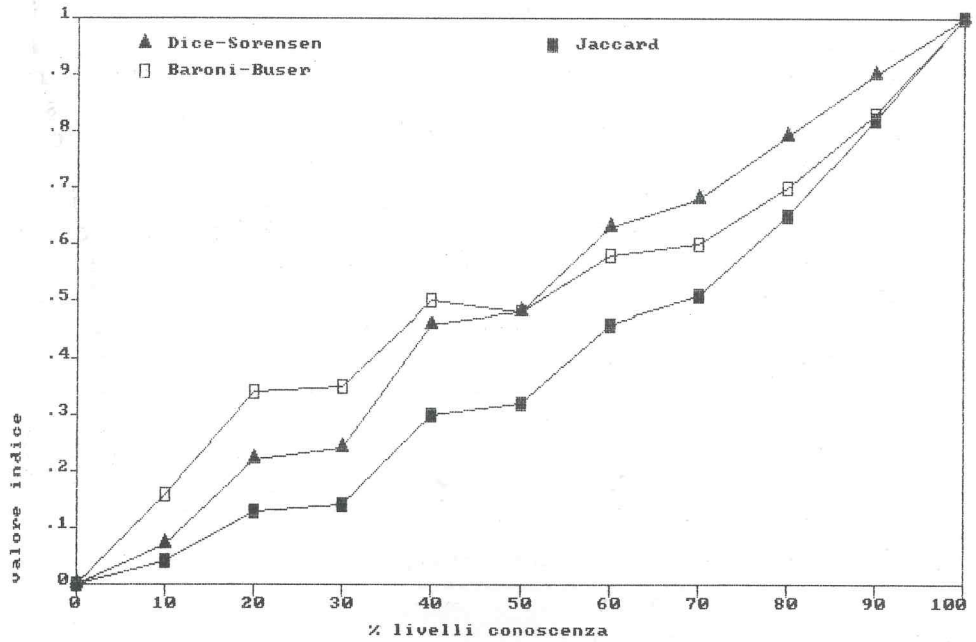

FIG. 1 - Valori medi degli indici di similarità ordinati rispetto ai livelli di conoscenza considerati. Caso A (vedere spiegazione nel testo).

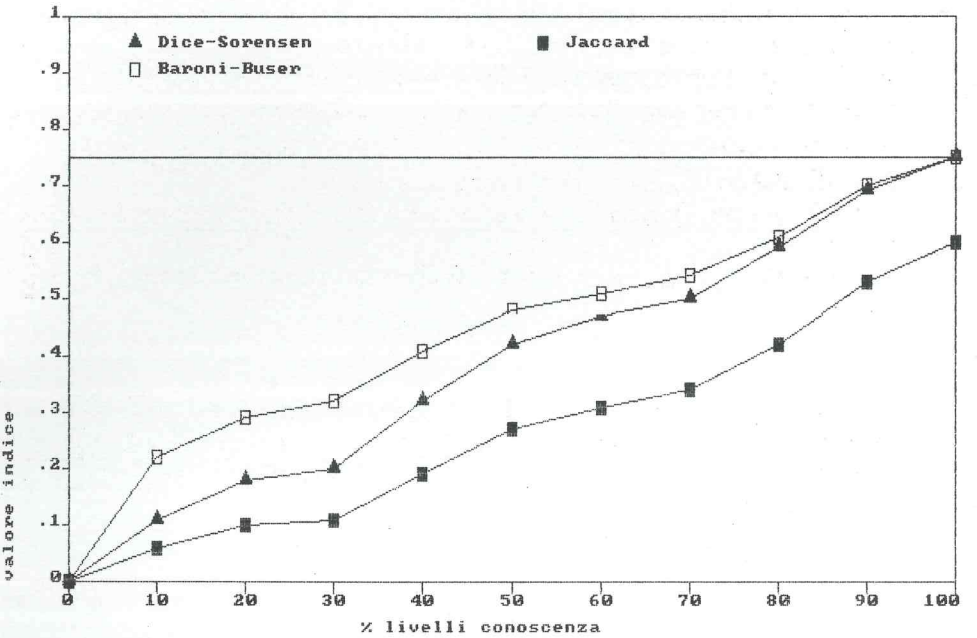

FIG. 2 - Valori medi degli indici di similarità ordinati rispetto ai livelli di conoscenza considerati. Caso B (vedere spiegazione nel testo).

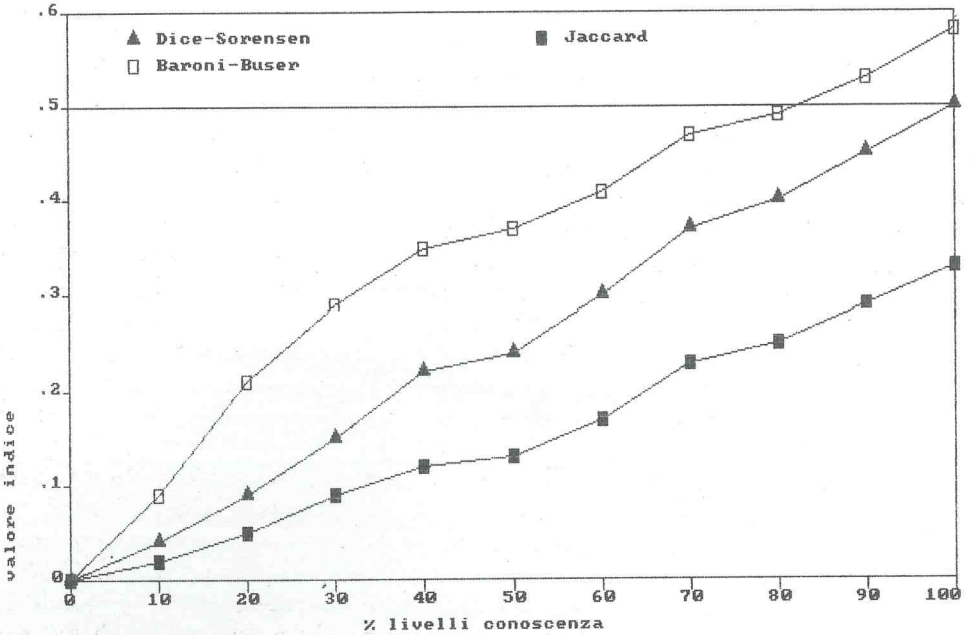

FIG. 3 - Valori medi degli indici di similarità ordinati rispetto ai livelli di conoscenza considerati. Caso C (vedere spiegazione nel testo).

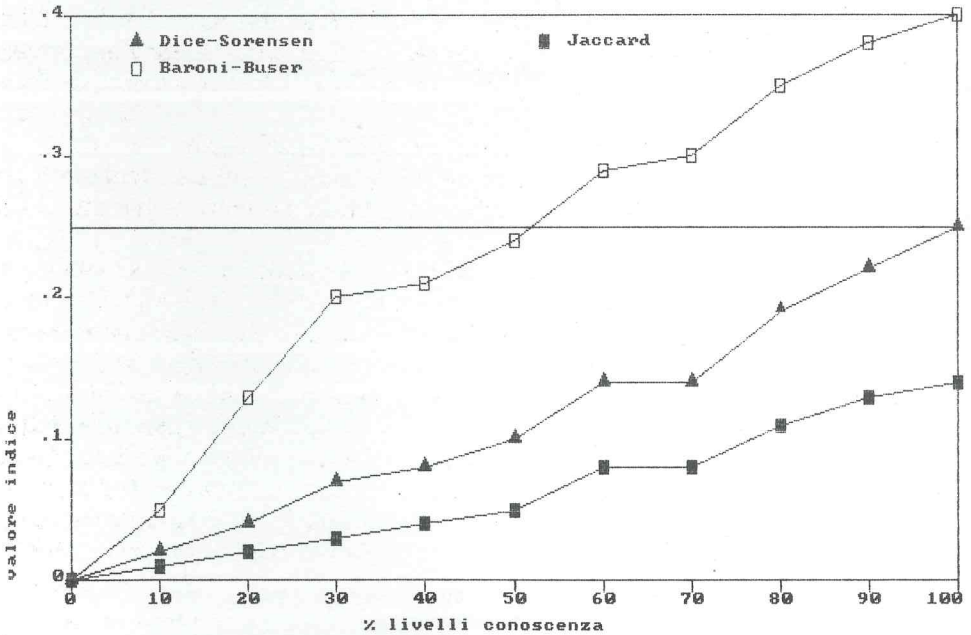

FIG. 4 - Valori medi degli indici di similarità ordinati rispetto ai livelli di conoscenza considerati. Caso D (vedere spiegazioni nel testo).

re approssimazione da parte dell'indice di Baroni Urbani e Buser, dovuta al contributo fornito dalle assenze comuni di attributi alla misura della similarità([2]). Al di sopra di questa percentuale la migliore approssimazione è fornita dall'indice di Dice-Sørensen; ciò è dovuto al fatto che in questo indice si dà un peso relativamente maggiore al numero di attributi positivi comuni alle due OTUs confrontate;

- nel caso B (fig. 2) (percentuale di similarità ipotizzata é 75%), si ha per ciascun livello di conoscenza considerato, una migliore approssimazione dell'indice di Baroni Urbani e Buser, mentre l'indice di Dice-Sørensen presenta valori paragonabili al precedente soltanto ai livelli compresi tra l'80% ed il 100%;
- anche per il caso C (fig. 3) (percentuale di similarità ipotizzata = 50%), si possono fare osservazioni simili a quelle esposte per il caso precedente, precisando che ai livelli compresi tra l'85% ed il 100%, l'indice di Baroni Urbani e Buser eccede il valore percentuale di similarità ipotizzato per l'apporto dovuto alle assenze comuni, mentre l'indice Dice-Sørensen risulta migliore a livelli di conoscenza superiori al 90%, nei casi in cui non si vogliono prendere in considerazione le assenze comuni;
- nel caso D (fig. 4) (percentuale di similarità ipotizzata = 25%), l'indice di Baroni Urbani e Buser è sicuramente da preferire sino a livelli di conoscenza del 70%. Oltre questa percentuale si presentano due possibilità di scelta differenti: la prima in cui si è interessati a calcolare la percentuale di similarità tra OTUs basandosi esclusivamente sugli attributi positivi comuni, nella quale è preferibile utilizzare l'indice Dice-Sørensen; la seconda in cui si desidera sottolineare anche l'importanza delle assenze comuni, nella quale è preferibile utilizzare l'indice di Baroni Urbani e Buser.

In sintesi si può osservare come in casi particolari quali:

- confronti tra OTUs potenzialmente molto simili e sufficientemente campionate (almeno il 60% degli effettivi attributi presenti in ciascun OTU) (fig. 1),
- confronti tra OTUs con bassa similarità potenziale, dove si sia effettuato un campionamento molto accurato (almeno il 90% degli attributi positivi effettivi presenti in ciascuna OTU) e dove non si voglia considerare il contributo dato alla similarità dalle assenze comuni (figg. 3-4),

è da preferire il calcolo della similarità mediante l'indice di Dice-Sørensen, mentre nella maggior parte delle situazioni sperimentali, sia di carattere faunistico che ecologico, la migliore stima della effettiva similarità tra OTUs e di matrici binarie, viene fornita dall'indice di Baroni Urbani e Buser, perlomeno tra quelli presi in esame. Da rilevare infine la notevole sottostima rispetto alla percentuale di similarità effettiva, fornita dall'indice di Jaccard nei vari casi presi in considerazione.

([2]) Per una discussione riguardante il significato delle coassenze di attributi nel calcolo della similarità rimando a Badaloni e Vinci (1983).

BIBLIOGRAFIA

BADALONI M., VINCI E., (1983) - *Osservazioni sugli indici di similarità.* - Metron, **61** (1-2): 113-133.
BARONI URBANI C., BUSER M.W., (1976) - *Similarity of binary data.* - Syst. Zool., **25**: 251-259.
JACCARD P., (1908) - *Nouvelles recherches sur la distribution florale.* - Bull. Soc. Vaud. Sci. Nat., **44**: 223-270.
SØRENSEN T., (1948) - *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.* - Biol. Skr., **5** (4): 1-34.