Article

Fl. Medit. 31 (Special Issue): 105-123
https://doi.org/10.7320/FlMedit31SI.105
Version of Record published online on 17 July 2021

János Podani

The wonder of the Jaccard coefficient: from alpine floras to bipartite networks

Abstract

Podani, J.: The wonder of the Jaccard coefficient: from alpine floras to bipartite networks. — Fl. Medit. 31 (Special Issue): 105-123. 2021. — ISSN: 1120-4052 printed, 2240-4538 online.

The similarity index suggested by Paul Jaccard 120 years ago has been one of the best-known coefficients in statistical ecology and other research fields in which the objects to be compared are described in terms of the presence or absence of many characters. Jaccard and his immediate followers used the coefficient for the comparison of floras of different localities and phytosociological relevés, based on the list of constituting species. A historical and mathematical overview reveals that, in addition to applications in ordinations and classifications, partitioning the coefficient into additive fractions opens unlimited opportunities for evaluating taxonomic, phylogenetic and functional diversity and related phenomena of ecological communities.

Key words: beta diversity, contingency table, data structure, similarity, simplex diagrams.

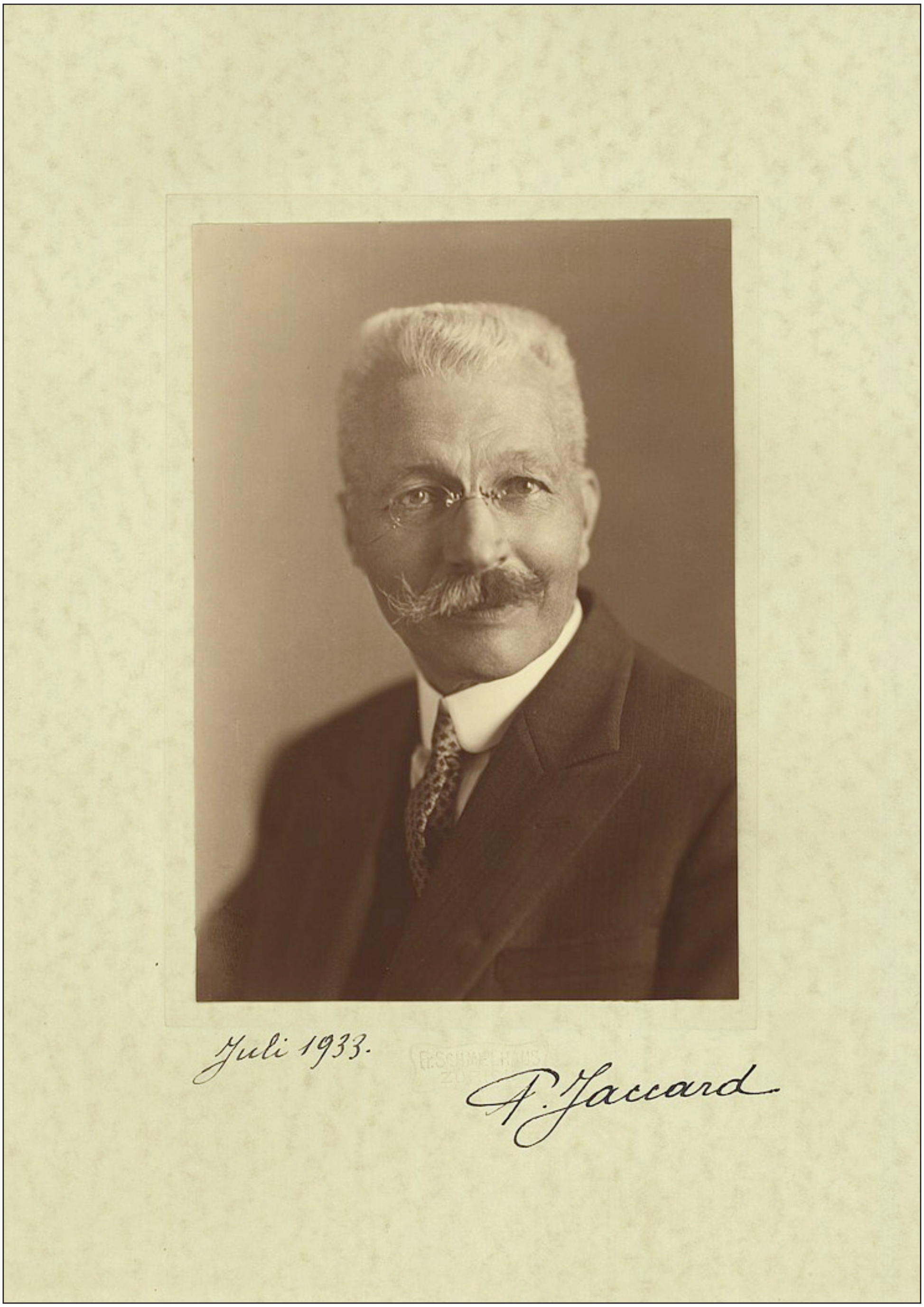
The inventor

Paul Jaccard, to whom we generally acknowledge the coefficient discussed in this paper, was born on 18 November 1868 in Sainte-Croix, canton Vaud, Switzerland (Frey-Wyssling 1944). During his school days, he started to collect fossils under the influence of his teacher of natural history, H. Golliez. At age 15 he visited the paleontological exhibition of the Swiss Federal Polytechnic (ETH) in Zurich; it was a trip which completely determined his future. Greatly impressed by what he saw, the young boy decided to study science, but his family could not provide sufficient financial support. Therefore, he first attended a teacher’s training college in Lausanne and then worked as a primary school teacher in the same town. His interest in natural sciences continued to increase in the meantime, and he developed contact with noted geologists and plant scientists of the time. As a second most influential event in his life, he made excursions to the Alps together with the famous botanist, L. Favrat. As Frey-Wyssling (1944) put it: “his enthusiasm knew no boundaries”. A position of plant preparator was offered to him in the Musée Botanique de Lausanne. While employed there, he worked hard and got his BSc degree. He then started to study natural sciences at the University of Lausanne in 1889, received the degree *licentia docendi* two years later, and finally obtained a doctoral degree at ETH, Zurich, in 1893.

His richly illustrated thesis on the embryology of *Ephedra helvetica* (Jaccard 1894) appeared the next year in the natural history journal of his beloved canton, *Bulletin de la Société Vaudoise des Sciences Naturelles*. Later, Jaccard published many of his papers in this local journal, which has become known worldwide much later, thanks primarily to his achievements.

These early papers reflect the expertise of a young, intelligent researcher whose interest extended from plant embryology, teratology, medicinal plants to floristics, and vegetation ecology. As a less widely known contribution to the history of evolutionary theory, he devoted his inaugural lecture held at the University of Lausanne to Darwinism (Jaccard 1895). In this, he takes the view that the theory cannot give satisfactory explanation to many aspects of plant evolution. The reason, according to him, lies primarily in the incompleteness of the highly fragmented paleontological material, and the lack of intermediate forms in the fossil record. However, he did not engage into further dispute over this subject and turned towards the study of alpine flora. Articles published around the turn of the 20th century clearly show this shift together with a new, and an even more important aspect of his research: the development of two coefficients by which he established statistical thinking in plant ecology. These are the *coefficient of floral community* ("coefficient de communauté florale" or "Gemeinschafts-coefficient") and the *generic coefficient* ("coefficient générique"). The first one is the subject matter of the present paper. The second formula is the ratio of the number of genera to the number of species, intended to reflect the ecological diversity of a given area. The generic coefficient evoked considerable interest and triggered long discussions over its applicability to vegetation ecology (Maillefer 1928, 1929; Williams 1949). Although it has not been used any longer, thanks partly to the obvious arbitrariness of genus level taxonomy, the idea that taxonomic (and then phylogenetic) relationships should be accounted for in diversity calculations persists in more complex conceptual frameworks.

In addition to the high mountains in his homeland, Jaccard travelled to other countries in Europe and participated in expeditions to exotic places, such as the Caucasus and Turkestan. In 1903, he was appointed as a professor of botany at ETH and remained there until his retirement in 1938. As a professor, he taught microscopy for 2500 students, and supervised 15 dissertations (Frey-Wyssling 1944). His portrait reproduced in Figure 1 was taken during these happy times. At ETH he developed interest in sylviculture and forest trees and the majority of his publications dealt with wood anatomy, mycorrhiza, plant morphology and physiology. Although many of these studies were reported in *Journal Forestier Suisse*, he continued to publish in the *Bulletin* as well as in other journals from his narrow homeland. One of these papers (Jaccard 1926) reports application of the generic coefficient to the Moroccan flora he investigated in a field trip organized by J. Braun-Blanquet. According to the obituary by Frey-Wyssling (1944), Jaccard has written 126 papers and book chapters, 89 in French and 37 in German. This list is obviously incomplete, however, because two papers that are extremely relevant to the present paper are missing, namely Jaccard (1907) and its translation, Jaccard (1912), supposedly his only publication in English. Jaccard continued to work hard after his retirement and took care of the wood collection of the university. His later years were made difficult not only by deteriorating health conditions but also by the escalating World War II. He passed away on 5 May 1944 in Zurich.



Juli 1933.

P. Jaccard

Fig. 1. Photograph and signature of Paul Jaccard from July 1933. Source: https://de.wikipedia.org/wiki/Paul_Jaccard.

Early history of a remarkable idea

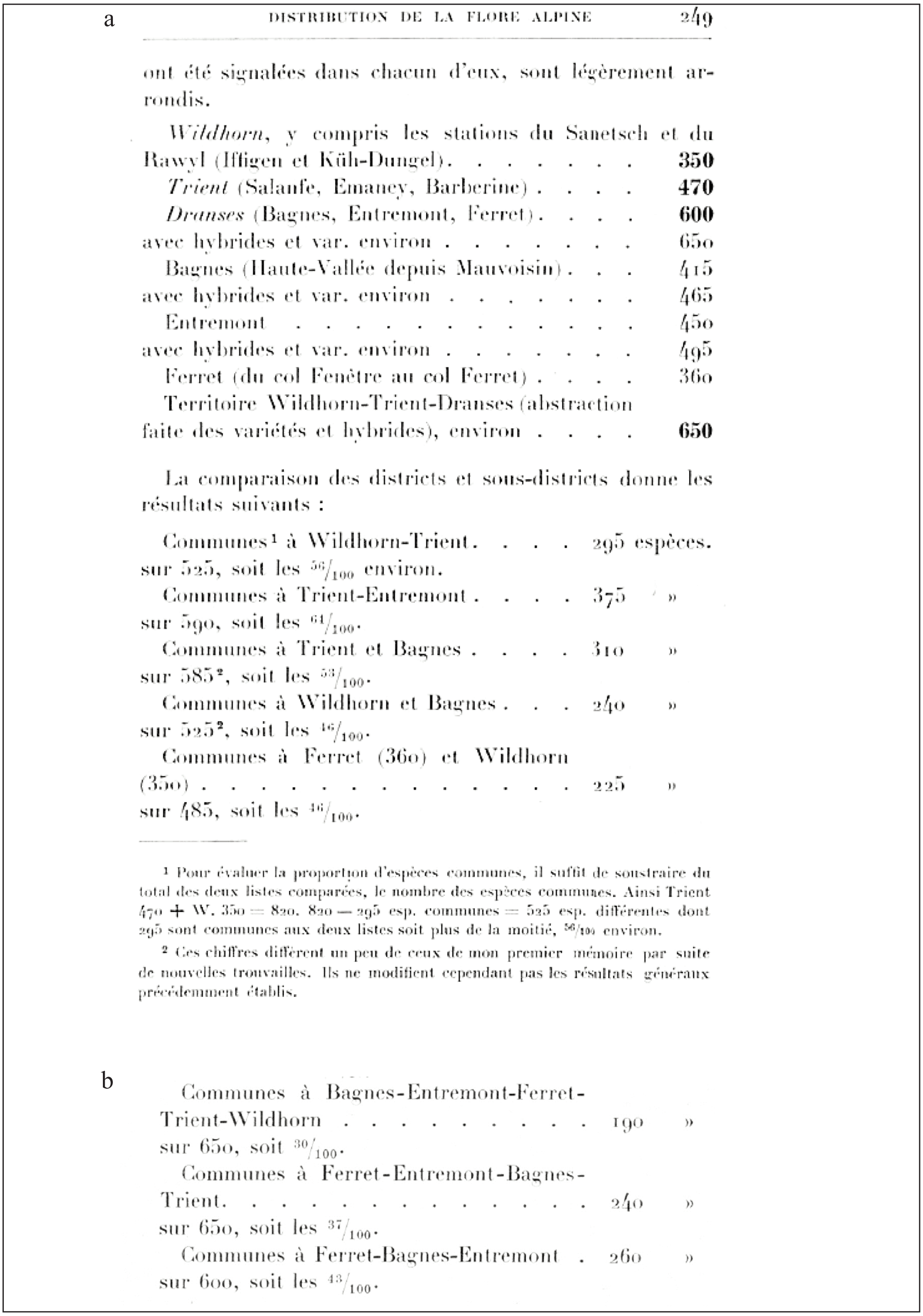
References to the origin of the Jaccard coefficient are inconsistent in the ecological and statistical literature, by mentioning his papers from 1900 to 1912 almost haphazardly. This is because the *Bulletin* and other journals in which his early papers appeared were hardly available to most authors, especially in the 20th century[1]. Given that more recently Jaccard's all publications have been available electronically for the wide public, thanks especially to ETH Bibliotek in Zurich (http://retro.seals.ch), we can trace back precisely how and where his coefficient was developed.

The idea that the flora of two localities may be compared by counting the number of species occurring in both places arose first in Jaccard's mind in a study of the alpine flora of three regions in the Swiss Alps: the Wildhorn Massif (W), the Trient Basin (T) and the Bagnes Valley (B) (Jaccard 1900). In addition to three tables that list and count the shared species for the three possible pairs of these localities, no other numerical manipulations were made with the data. Yet. In a subsequent publication, Jaccard (1901a) goes one step further and describes shortly – in a footnote (!) – the method of how to compare two floras in a standardized way. On p. 249-250, he compares several districts and subdistricts in various combinations, and calculates the ratio of the number of shared species by the number of all species for T and W (Fig. 2a), with explanation given in his Footnote 1: T has 470 species, W has 350, whereas the number of common species is 295. Distracting the latter number from 470+350 = 820 yields the total number of species present in T and W, namely 525. Then, 295 divided by 525 provides the relative proportion of shared species, which is around 56/100. Actually, a more precise value of this ratio, the first ever published, is 0.5619 for four decimal digits. He gives the name of this ratio, "*coefficient de communauté florale*" only on p. 251, without presenting an explicit formula, and uses it as a percentage similarity throughout the paper.

Initially, Jaccard did not restrict the use of his new method to pairwise comparisons. He goes further on and calculates the ratio for several localities simultaneously, implicitly giving the first instance of a multiple site similarity coefficient. On p. 250, five localities are first evaluated together, producing a ratio of 0.30, which increases to 0.37 and 0.43 after successive removal of two localities (Fig. 2b)[2]. The occurrence of the first semimatrix of similarities is another remarkable achievement of this paper. On pages 253-257, the author gives complete details for the comparison of ten localities in every possible pair (partly reproduced in Fig. 3a). Based on this information it is easy to compile the semimatrix, which is presented here in dissimilarity form (Fig. 3b) for further analysis which Jaccard could not even dream of. The results of principal coordinates analysis and group average clustering (Fig. 3c-d) demonstrate the relationships between the ten regions.

[1] Even Francey (1941) in the same *Bulletin* does not refer properly to the first occurrence of the coefficient, by mentioning Jaccard (1902a).

[2] Unfortunately, this latter idea was never expanded any further. Instead, Jaccard adapted the practice of averaging pairwise indices calculated for relevé data.



a

DISTRIBUTION DE LA FLORE ALPINE 249

ont été signalées dans chacun d'eux, sont légèrement arrondis.

Wildhorn, y compris les stations du Sanetsch et du Rawyl (Iffigen et Küh-Dungel)	**350**
Trient (Salanfe, Emaney, Barberine)	**470**
Dranses (Bagnes, Entremont, Ferret)	**600**
avec hybrides et var. environ	650
Bagnes (Haute-Vallée depuis Mauvoisin)	415
avec hybrides et var. environ	465
Entremont	450
avec hybrides et var. environ	495
Ferret (du col Fenêtre au col Ferret)	360
Territoire Wildhorn-Trient-Dranses (abstraction faite des variétés et hybrides), environ	**650**

La comparaison des districts et sous-districts donne les résultats suivants :

Communes[1] à Wildhorn-Trient . . . . 295 espèces.
sur 525, soit les $^{56}/_{100}$ environ.

Communes à Trient-Entremont . . . . 375 »
sur 590, soit les $^{64}/_{100}$.

Communes à Trient et Bagnes . . . . 310 »
sur 585[2], soit les $^{53}/_{100}$.

Communes à Wildhorn et Bagnes . . . 240 »
sur 525[2], soit les $^{46}/_{100}$.

Communes à Ferret (360) et Wildhorn (350) . . . . 225 »
sur 485, soit les $^{46}/_{100}$.

[1] Pour évaluer la proportion d'espèces communes, il suffit de soustraire du total des deux listes comparées, le nombre des espèces communes. Ainsi Trient 470 + W. 350 = 820. 820 — 295 esp. communes = 525 esp. différentes dont 295 sont communes aux deux listes soit plus de la moitié, $^{56}/_{100}$ environ.

[2] Ces chiffres diffèrent un peu de ceux de mon premier mémoire par suite de nouvelles trouvailles. Ils ne modifient cependant pas les résultats généraux précédemment établis.

b

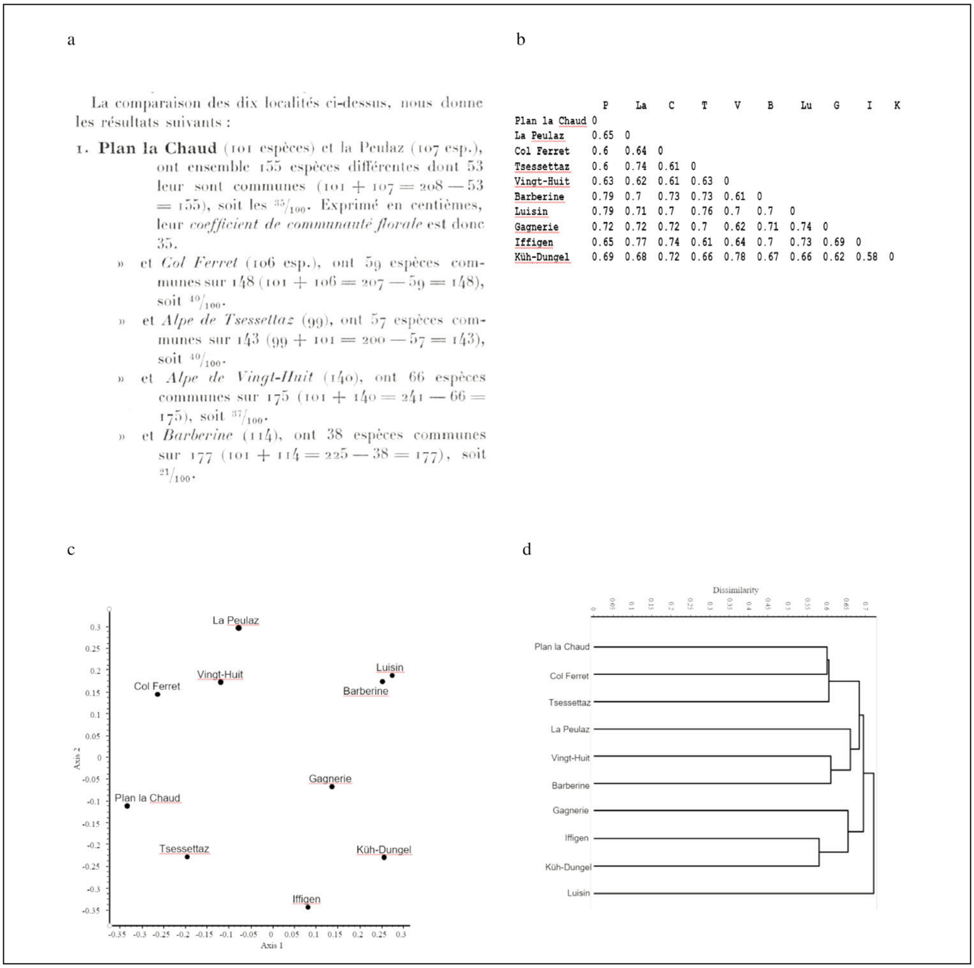
Communes à Bagnes-Entremont-Ferret-Trient-Wildhorn . . . . 190 »
sur 650, soit $^{30}/_{100}$.

Communes à Ferret-Entremont-Bagnes-Trient. . . . . 240 »
sur 650, soit les $^{37}/_{100}$.

Communes à Ferret-Bagnes-Entremont . 260 »
sur 600, soit les $^{43}/_{100}$.

Fig. 2. Extracts from Jaccard (1901a) illustrating the first example of his "coefficient de communauté florale" (**a**) and its application to multiple sites (**b**).

The ordination axes have relatively low explanatory power, with percentage eigenvalues decreasing gradually (the first three being 19, 18 and 14%). The points are dispersed relatively evenly in the ordination space, and the closeness of Barberine and Luisin is misleading, as the dendrogram demonstrates. The dissimilarities fall into a narrow range, the smallest being 0.58, which suggest high floristic diversity of these mountainous areas. In subsequent publications (Jaccard 1901b, 1902a, b, d), calculations were demonstrated partly on the same dataset, that is, still at the biogeographical level. In Jaccard (1902c, p. 362) the same upper semimatrix is listed by rows. These papers also use artificial examples to demonstrate the calculations.



a

La comparaison des dix localités ci-dessus, nous donne les résultats suivants :

1. **Plan la Chaud** (101 espèces) et la Peulaz (107 esp.), ont ensemble 155 espèces différentes dont 53 leur sont communes (101 + 107 = 208 — 53 = 155), soit les $^{35}/_{100}$. Exprimé en centièmes, leur *coefficient de communauté florale* est donc 35.

» et *Col Ferret* (106 esp.), ont 59 espèces communes sur 148 (101 + 106 = 207 — 59 = 148), soit $^{40}/_{100}$.

» et *Alpe de Tsessettaz* (99), ont 57 espèces communes sur 143 (99 + 101 = 200 — 57 = 143), soit $^{40}/_{100}$.

» et *Alpe de Vingt-Huit* (140), ont 66 espèces communes sur 175 (101 + 140 = 241 — 66 = 175), soit $^{37}/_{100}$.

» et *Barberine* (114), ont 38 espèces communes sur 177 (101 + 114 = 225 — 38 = 177), soit $^{21}/_{100}$.

b

	P	La	C	T	V	B	Lu	G	I	K
Plan la Chaud	0									
La Peulaz	0.65	0								
Col Ferret	0.6	0.64	0							
Tsessettaz	0.6	0.74	0.61	0						
Vingt-Huit	0.63	0.62	0.61	0.63	0					
Barberine	0.79	0.7	0.73	0.73	0.61	0				
Luisin	0.79	0.71	0.7	0.76	0.7	0.7	0			
Gagnerie	0.72	0.72	0.72	0.7	0.62	0.71	0.74	0		
Iffigen	0.65	0.77	0.74	0.61	0.64	0.7	0.73	0.69	0	
Küh-Dungel	0.69	0.68	0.72	0.66	0.78	0.67	0.66	0.62	0.58	0

Fig. 3. Extract from Jaccard (1901a) showing the first part of his calculations of the semimatrix of similarities for ten alpine localities in Switzerland (**a**), the entire lower semimatrix converted to dissimilarities (**b**), results of PCoA from this matrix (**c**) and the dendrogram obtained by group average clustering (**d**).

Metamorphosis of an index

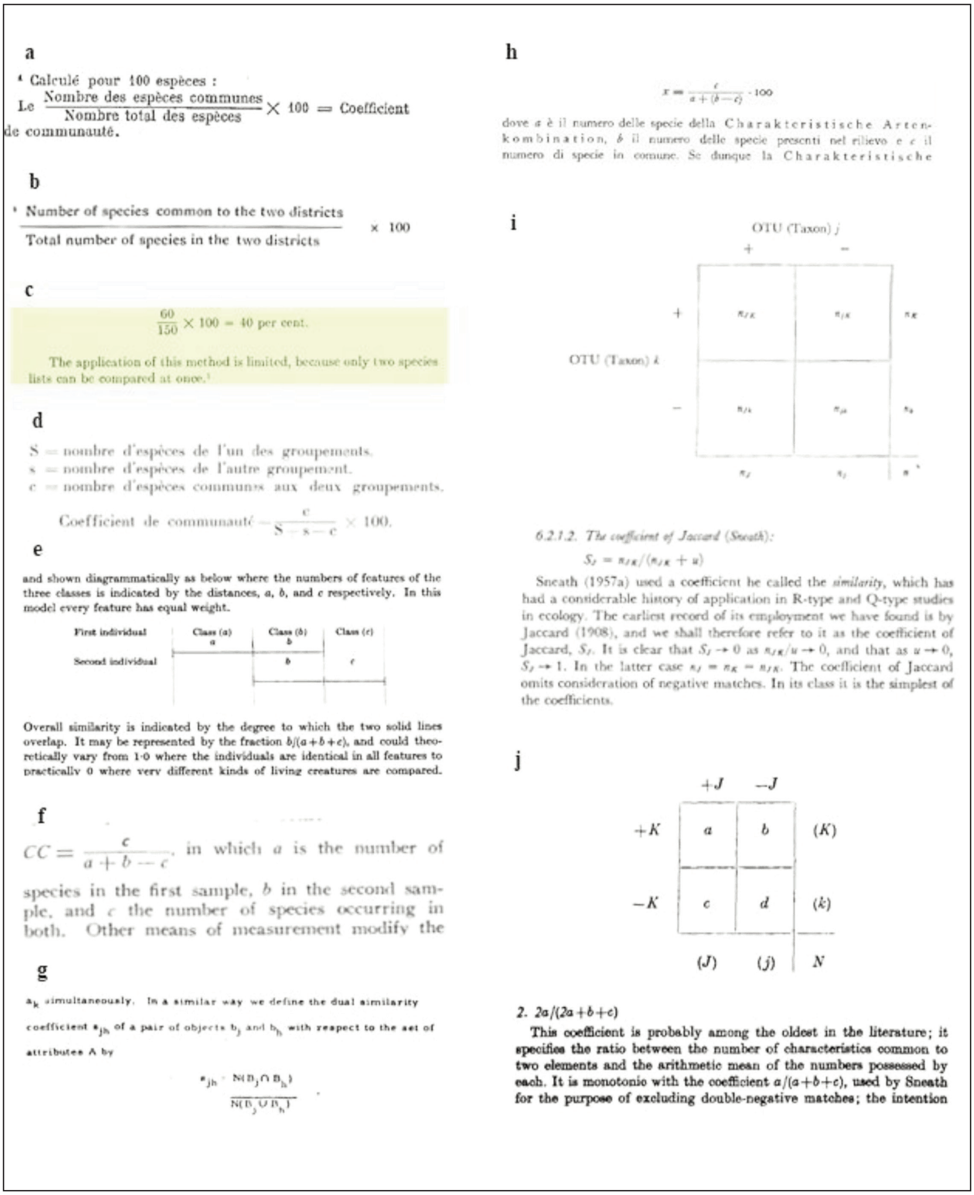
Having investigated Jaccard's publications thoroughly, I can safely say that he never gave an explicit mathematical formula (i.e., one with symbols) for the *coefficient de communauté florale*. At best, Jaccard (1907, p. 962) presents a verbal version of his index in a footnote (Fig. 4a), and the text explains that this index is understood for a pair of localities. In the English translation (Jaccard 1912, p. 39), this information is given in the ratio itself (Fig. 4b).

The coefficient gained popularity relatively rapidly, thanks to Braun-Blanquet's seminal book on plant sociology (in German: 1927, in English: 1932), in which an example, rather than a general formula (Fig. 4c) was used to introduce the method. As the author adds, the index is applicable to pairwise comparisons. It was perhaps Francey (1941) who first suggested a truly mathematical formula for the coefficient, based on the same logic as the original: the number of common species (c) is divided by the sum of the species richness values of the two localities ($S + s$) from which the number of shared species is subtracted (Fig. 4d). Later, symbols *a, b* and *c* started to appear with various meanings and combinations. For Sneath (1957) *b* was the number of shared attributes in numerical taxonomic context, with an illustration which is almost set-theoretical (Fig. 4e). Whittaker and Fairbanks (1958) used *c* again to refer to the number of common species, whereas *a* is the number of species in the first sample unit, *b* in the second (Fig. 4f). Tanimoto (1958), apparently independently from Jaccard, suggested the same coefficient with set theoretical definition. If sets B_j and B_h contain the attributes possessed by objects b_j and b_h, respectively, then the index is the number of shared attributes (intersection of the two sets) divided by the cardinality of the union of the two sets (Fig. 4g).

Pignatti & Mengarda (1962) were the first to break with the traditional usage of the coefficient. Rather than contrasting the flora of two localities or sample units, they proposed to use the index for the comparison of each sample unit (relevé) to a set of characteristic species typical of the community ("Charakteristiche Artenkombination"). That is, one object was a real observation while the other was an abstraction. The formula itself did not change, however, the authors used the same symbols as Whittaker & Fairbanks (1958) with *c* being the number of characteristic species occurring in the relevé (Fig. 4h).

The notations of a 2×2 contingency table were first adapted to this index by Sokal & Sneath (1963), with n_{JK} referring to the number of attributes shared by objects (OTU's, in their terminology) *j* and *k*, and *u* denoting the number of attributes possessed only by either object *j* or *k*, that is, $u = n_{jK} + n_{Jk}$ (Fig. 4i).

In a milestone communication, Williams & Dale (1965) proposed the use of parameters *a, b, c* and *d* of the 2×2 contingency table, a simple scheme previously used in statistical texts, for writing presence-absence resemblance coefficients. In this, *a* is the number of characteristics common to both objects, *J* and *K*, being compared, *b* is the number of attributes pertaining only to *K*, *c* is the number of attributes present only in *J*, and *d* is the number of characters missing from both (double zeros). This is the first paper that published the Jaccard index written in terms of these symbols. However, the formula appears only in a commentary to another index (Sörensen's) and the authors attribute the formula to Sneath (Fig. 4j) even though Sokal & Sneath (1963, cited also by Williams and Dale) already realized that Jaccard was the originator.



a

[1] Calculé pour 100 espèces :

Le $\frac{\text{Nombre des espèces communes}}{\text{Nombre total des espèces}} \times 100 =$ Coefficient de communauté.

b

[1] $\frac{\text{Number of species common to the two districts}}{\text{Total number of species in the two districts}} \times 100$

c

$\frac{60}{150} \times 100 = 40$ per cent.

The application of this method is limited, because only two species lists can be compared at once.[1]

d

S = nombre d'espèces de l'un des groupements.
s = nombre d'espèces de l'autre groupement.
c = nombre d'espèces communes aux deux groupements.

Coefficient de communauté $= \frac{c}{S + s - c} \times 100.$

e

and shown diagrammatically as below where the numbers of features of the three classes is indicated by the distances, *a*, *b*, and *c* respectively. In this model every feature has equal weight.

	Class (*a*)	Class (*b*)	Class (*c*)
First individual	*a*	*b*	
Second individual		*b*	*c*

Overall similarity is indicated by the degree to which the two solid lines overlap. It may be represented by the fraction $b/(a+b+c)$, and could theoretically vary from 1·0 where the individuals are identical in all features to practically 0 where very different kinds of living creatures are compared.

f

$CC = \frac{c}{a + b - c}$, in which *a* is the number of species in the first sample, *b* in the second sample, and *c* the number of species occurring in both. Other means of measurement modify the

g

a_k simultaneously. In a similar way we define the dual similarity coefficient s_{jh} of a pair of objects b_j and b_h with respect to the set of attributes A by

$$s_{jh} = \frac{N(B_j \cap B_h)}{N(B_j \cup B_h)} .$$

h

$$x = \frac{c}{a + (b - c)} \cdot 100$$

dove *a* è il numero delle specie della Charakteristische Artenkombination, *b* il numero delle specie presenti nel rilievo e *c* il numero di specie in comune. Se dunque la Charakteristische

i

		OTU (Taxon) *j*		
		+	−	
OTU (Taxon) *k*	+	n_{JK}	n_{jK}	n_K
	−	n_{Jk}	n_{jk}	n_k
		n_J	n_j	n

6.2.1.2. The coefficient of Jaccard (Sneath):

$$S_J = n_{JK}/(n_{JK} + u)$$

Sneath (1957a) used a coefficient he called the *similarity*, which has had a considerable history of application in R-type and Q-type studies in ecology. The earliest record of its employment we have found is by Jaccard (1908), and we shall therefore refer to it as the coefficient of Jaccard, S_J. It is clear that $S_J \to 0$ as $n_{JK}/u \to 0$, and that as $u \to 0$, $S_J \to 1$. In the latter case $n_J = n_K = n_{JK}$. The coefficient of Jaccard omits consideration of negative matches. In its class it is the simplest of the coefficients.

j

	$+J$	$-J$	
$+K$	a	b	(K)
$-K$	c	d	(k)
	(J)	(j)	N

2. $2a/(2a+b+c)$

This coefficient is probably among the oldest in the literature; it specifies the ratio between the number of characteristics common to two elements and the arithmetic mean of the numbers possessed by each. It is monotonic with the coefficient $a/(a+b+c)$, used by Sneath for the purpose of excluding double-negative matches; the intention

Fig. 4. A brief history of different textual and algebraic expressions of the Jaccard index. **a.** The first occurrence in Jaccard (1907, p. 962), **b.** the first version in English (Jaccard 1912, p. 39), **c.** Jaccard index was still "defined" by an example by Braun-Blanquet (1932, p. 363), **d.** the first mathematical formula (Francey 1941, p. 298), **e.** graphical illustration by Sneath (1957, p. 202), **f.** the formula of Whittaker & Fairbanks (1958, p. 54), **g.** set-theoretical definition by Tanimoto (1958, p. 5), **h.** description of the formula by Pignatti & Mengarda (1962, p. 216), **i.** denotation based on the 2×2 contingency table (Sokal & Sneath 1963, p. 133), **j.** the use of symbols *a, b, c,* and *d* as suggested by Williams & Dale (1965, p. 36 and p. 49).

Later, thanks to Goodall (1973) and Sneath & Sokal (1973), this notation and the term *Jaccard index* have become widely known in statistical ecology and numerical taxonomy, respectively. A rapid check of most influential textbooks and review papers of vegetation ecology demonstrates that Jaccard index has been treated as a similarity coefficient in the form $S = a / (a + b + c)$ and, less commonly, in dissimilarity form $DIS = (b + c) / (a + b + c)$ as well. Since the coefficient has been used in a very wide variety of disciplines, it is not surprising that several alternatives continue to exist and even new versions arise. For example, Verma & Aggarwal (2020) list five different, set theoretical alternatives, one of which being identical to Tanimoto's ratio.

Irrespective of mathematical formalism, as said, the index has been widely known not only in vegetation ecology but also in computer science, genomics, humanities, and other fields in which data may take binary (presence/absence) form. Widespread usage was greatly facilitated by commercial ordination and cluster analysis packages: practically all of them offer the option of the index for calculating (dis)similarity. In the Web of Science database, which goes back to 1975, the search terms "Jaccard index" and "Jaccard coefficient" appear in the title of 18 articles, and in the abstract or keywords of 1155 papers. Jaccard's early works from 1901-1910 have been cited by at least 7000 times, so they have become true citation classics. Remarkably, the citing papers are not always mere applications. The index was incorporated by Gower (1971) into his general coefficient of similarity. Being the ratio of the intersection and the union of two sets, the coefficient was chosen by Feoli & Lagonegro (1979) to maximize the monothetic criterion in clustering via intersection analysis. The mathematical and statistical properties of Jaccard's formula were evaluated by Gower & Legendre (1986), Li (2015), Chung & al. (2019), Kosub (2019) and Verma & Aggarwal (2020), among others. Most importantly, it has been shown by several authors that several coefficients developed to abundance data reduce to Jaccard similarity (Ruzicka index and the similarity ratio) or Jaccard dissimilarity (Marczewski-Steinhaus coefficient) in the binary case.

Dismantling the coefficient

Jaccard index did not remain intact, its components were modified or partitioned in several ways to exhaust more information and to increase interpretability of results. More precisely, only parameters b and c have been subject to change in two conceptual schemes in ecology since there is not much to do with parameter a reflecting absolute agreement or overlap.

Incorporating taxonomic, phylogenetic or functional distinctness

Jaccard index, just like all other (dis)similarity coefficients developed for presence-absence data consider every species (attribute, in general) equally important. Izsák & Price (2001) raised first the idea that the taxonomic relationships among species should also be considered in calculating similarity. They suggested to redefine the Sörensen index ($2a/(2a + b + c)$) with parameters b and c diminished according to the position of the nearest differential species in the Linnaean hierarchy. Species i present in site X and absent from site

Y contributes to the value of *b* by 1 only if it is separated from all species of site *Y* at the maximum rank in the entire sample (e.g., if it is a fern in site *X*, whereas site *Y* has only angiosperms which is a phylum-level separation in the Linnaean system). In all other cases, these contributions are proportionally smaller depending on the taxonomic distinctness (for example, 0.2 at genus, 0.4 at family, 0.6 at order, and 0.8 at class level) of species *i* from the closest relative species present in site *Y*. That is, *b* as well as *c* are replaced by the sum of taxonomic distinctness values, leading to an increased similarity (decreased dissimilarity) between the sites. Ricotta & al. (2016b) proposed that any other meaningful form of distinctness between species may also be considered in modulating *b* and *c*, so that functional and phylogenetic relationships may also be incorporated in the calculations. More formally, the Jaccard dissimilarity index is rewritten as

$$DIS' = \frac{B+C}{a+b+c} \qquad \text{where} \qquad B = \sum_{i \in X, i \notin Y} \min_{j \in Y}\{d_{ij}\} \text{ and } C = \sum_{j \in Y, j \notin X} \min_{i \in X}\{d_{ij}\}$$

In the above equations, $0 \leq d_{ij} \leq 1$ represents the taxonomic, phylogenetic or functional distinctness between species *i* and *j*, whereas *a, b* and *c* in the denominator retain their original meaning. It follows that $A = a + b + c - B - C$ is the absolute taxonomic, phylogenetic or functional overlap of the two sites. This general scheme applies to all other presence-absence resemblance coefficients and the modified forms can be used in ordinations and classifications that are not taxon- (mostly species-) based.

Beta diversity and its partitioning

The dissimilarity form of the index ($DIS = (b + c) / (a + b + c)$) was considered first by Colwell & Coddington (1994) as the β-diversity of a pair of sample units, starting a new field of its application and new possibilities of interpretation[3]. Two years later, Williams (1996) proposed another measure with the same denominator, $\min\{b, c\}/(a + b + c)$ for the same purpose. Since the latter coefficient is not bounded between 0 and 1, which is otherwise the case for dissimilarity coefficients, Cardoso & al. (2009) suggested to multiply the numerator by 2. The index thus obtained reflects the relative proportion of the number of species that are replaced by each other in a comparison of two sites. Podani & Schmera (2011) have shown that this is only part of beta diversity, a measure of relative species turnover, whereas the other component is $|b - c|/(a + b + c)$ which was called the relative richness difference. These two terms together comprise Jaccard dissimilarity. Podani and Schmera called attention to the obvious relationship:

$$1 = a/(a+b+c) + |b-c|/(a+b+c) + 2\min\{b,c\})/(a+b+c) = S + D + R.$$

In words, similarity, richness difference and replacement, if relativized by the total number of species, always add to 1, and as such, these three components can be illustrated using a 2D simplex diagram, an equilateral triangle. In this, a pair of sites is represented by a point, its position within the triangle depending on the three additive components. For

[3] In fact, Whittaker (1960) was the first to suggest the use of a presence-absence dissimilarity coefficient (again, the Sörensen index) as a measure of beta diversity.

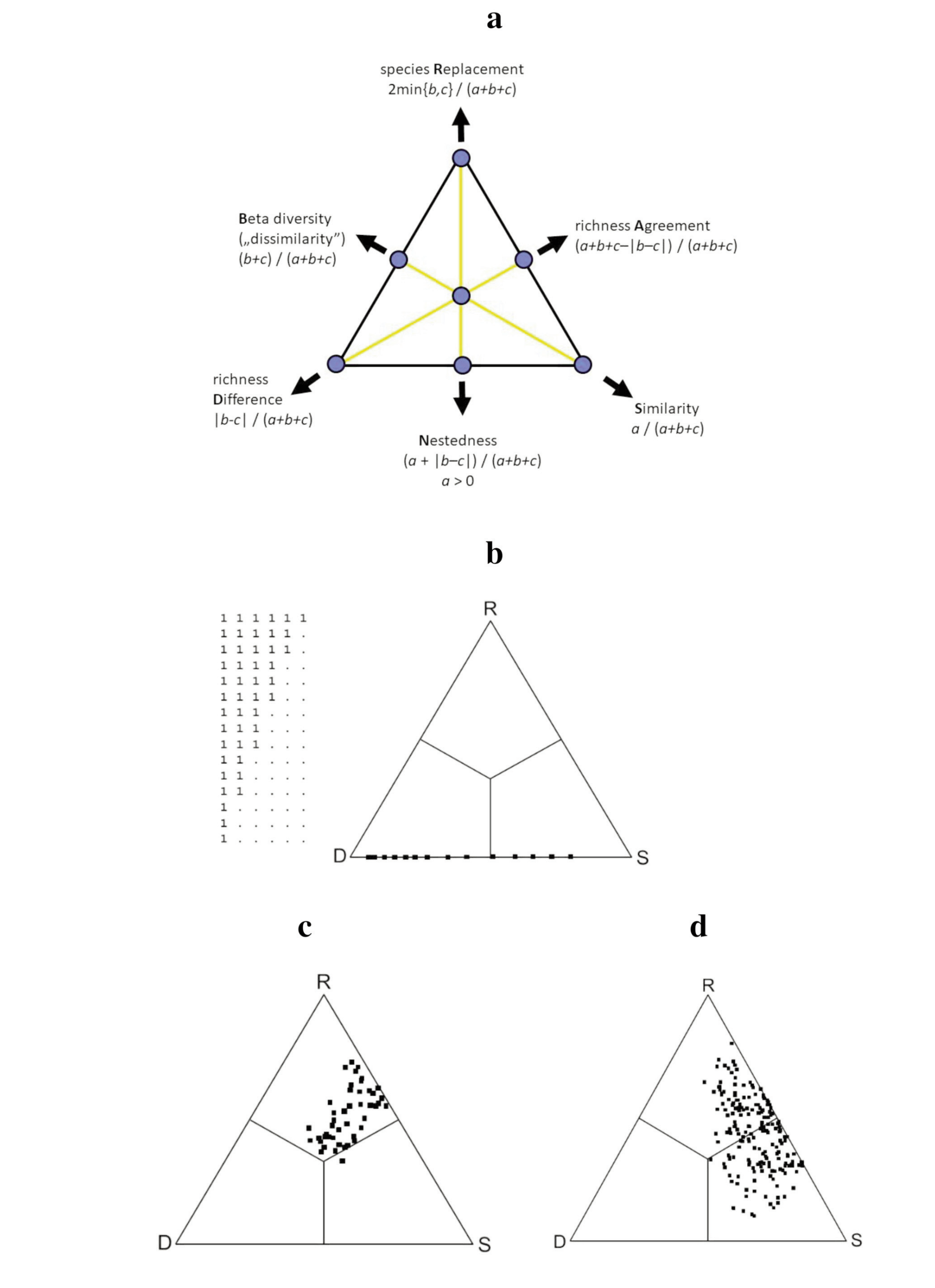
example, if $S = D = R = 0.33$, then the corresponding point will be in the centroid of the plot. If the two sites are identical in species composition ($b + c = 0$) then $S = 1$ and the point will be on the *S*, bottom right corner of the triangle. If $a = 0$ and $b = c$, there is only turnover and the point will fall onto *R*, the top corner of the triangle. The third corner, *D*, is taken when richness difference is the maximum, i.e. $a = b = 0$, which means that one site is empty, a situation usually avoided deliberately during vegetation surveys (e.g., pure sand or rock surfaces). In addition to the corners, the three edges are also meaningful. A point falls onto the left edge if $S = 0$, so $D + R = 1$, indicating maximum beta diversity. If there is no replacement, i.e. $\min\{b,c\} = 0$, then $S + D = 1$ and one site is nested within the other (with the condition that $a > 0$) and the point will be on the bottom edge. If the third combination of the two components, $S + R$ is 1, then there is complete richness agreement and the point falls onto the right edge (see Fig. 5a for illustration). In general, the distance of a point from a corner is inversely proportional to the value of the respective index.

An advantage of the SDR simplex approach is its ability to demonstrate pattern in presence/absence data matrices. The coefficients are calculated for all pairs of objects, and the shape of the point cloud in the SDR plot will be informative about internal data structure. For example, if the matrix is entirely nested, then all pairs of objects (columns) will exhibit zero replacement, and all points will fall onto the nestedness side (Fig. 5b). In case of high beta diversity, the point cloud will be near the left edge. Data for the pairwise comparisons presented by Jaccard (1901a) for ten localities in the Swiss Alps allowed to calculate the *S, D* and *R* scores, and the results – after a latency of 120 years! – are shown in Fig. 5c. Almost all points (each of them representing a pair of localities) fall into the upper third of the triangle demonstrating that replacement (species turnover) is the dominant process in affecting beta diversity the alpine flora at the geographic scale used by Jaccard: the mean of 45 *R* values is 0.52. The second major feature is overall similarity, with $\bar{S} = 0.32$ while richness difference is relatively small, $\bar{D} = 0.16$.

This agrees well with the ordination and classification results (Fig. 3c-d): the ten points are evenly arranged in the ordination space and the fusion levels in the dendrogram fall into a narrow range of high dissimilarity values. Similar result is obtained at an even higher geographical level. If we compare the flora of the 20 regions if Italy based on a total of more than 6000 species (Pignatti 1982; Conti & al. 2005), we get a point cloud more closely attached to the richness agreement side of the plot, with a balanced contribution by replacement and similarity (Fig. 5d). This pattern is typical when the flora of the localities follows a gradient, in this case a biogeographical one largely from north to the south, with the alpine regions at one end and the two big islands (Sicily and Sardinia) at the other.

Combined approach

A most natural suggestion is to combine the above two approaches by constructing a simplex diagram based on the Jaccard coefficient as modified to incorporate taxonomic, phylogenetic or functional distinctness of species (Podani & al. 2018a). In this way, the conceptual scheme developed for taxon-based data (i.e., species by localities matrices) is extended to three other areas of application. In addition to raw data, we also need three species by species matrices of distinctness values standardized to the range [0,1]. This condition must satisfy for the modified Jaccard index to allow comparison of data structures



```
1 1 1 1 1 1
1 1 1 1 1 .
1 1 1 1 1 .
1 1 1 1 . .
1 1 1 1 . .
1 1 1 1 . .
1 1 1 . . .
1 1 1 . . .
1 1 1 . . .
1 1 . . . .
1 1 . . . .
1 1 . . . .
1 . . . . .
1 . . . . .
1 . . . . .
```

Fig. 5. **a.** Components of the SDR simplex diagram (with explanations in text); **b.** the SDR simplex plot for an artificial data matrix with complete nestedness; **c.** the SDR simplex for the alpine flora data of Jaccard (1901a); **d.** the SDR simplex for the flora of 20 Italian regions.

based on four different schemes. The approach is illustrated using plot data from pioneer successional stages of alpine meadows above the timberline in the Aosta Valley, Italian Alps (Ricotta & al. 2016b). The simplex diagrams (Fig. 6a-d) demonstrate relatively high beta diversity at the species level, with replacement as the dominating background factor. Its average $\bar{S} = 0.52$ is almost identical to the value presented above for Jaccard's data from the Swiss Alps even though the scale of the two surveys is radically different! In this case, however, the point cloud has different shape, many pairs of plots show remarkable similarity and richness difference as well – reflecting extreme heterogeneity of sample plots. After switching to taxonomy, similarity becomes dominant because during succession many species are replaced by close relatives in the Linnaean system. When phylogeny is accounted for, similarity is further increased because phylogenetic relationships are in fact closer than what the use of Linnaean ranks forces upon the system. The points become highly concentrated near the S corner when functional distinctness is considered in calculating the Jaccard indices. This means that species in the alpine meadow differ very little functionally along the entire successional sere. We found similar trends in SDR simplices in further two cases (rock grasslands, coastal marshes) and therefore put forward the *beta-redundancy hypothesis*: beta diversity of ecological communities decreases in the following order: taxon (species)-level – taxonomic – phylogenetic – functional. Further studies in a wide range of community types would be welcome to confirm this proposition.

Multiple comparison of several data matrices

The SDR diagrams reflect internal structure of presence-absence data matrices, offering the possibility for comparing data matrices indirectly, through the comparison of point patterns within the plots (Podani & al. 2018b). Such comparisons are necessary to extract information on background factors that influence community pattern or beta diversity at different ecological or biogeographical scales. The summary of the method is as follows. Each diagram is dissected into 100 small equilateral triangles (inset in Fig. 7), within which the number of points is calculated. Then, these numbers are converted into relative frequencies, and two simplex diagrams are compared by the Manhattan distance function based on the corresponding relative frequency values. The advantage of the approach is that two matrices can be compared even though they differ in the number of rows and columns.

It is illustrated by presence-absence data of various animal groups of animals (butterflies, centipedes, isopods, reptiles, land snails, tenebrionids) from two groups of islands in the Mediterranean Sea (Anatolian Islands and the Cyclades). Sources of information are described in the Electronic Supplement to Podani & al. (2018a). Data matrices pertaining to the 12 combinations of animal and island groups were compared in every possible pair, providing a distance matrix of data matrices, which was in turn subjected to group average clustering (Fig. 7). The dendrogram reveals succinctly the similarities and dissimilarities in the distributional pattern of these animal groups, as influenced by their dispersal ability, local extinction, and past colonization through land bridges between islands. Note, for example, the closeness of the two butterfly faunas, as well as that of the land snail faunas, and the disparity of the reptile composition of the two island groups.

Bipartite networks

All applications of the Jaccard index discussed above are based on standard species by localities presence-absence matrices. The coefficient also applies to computing interaction similarity in pollination networks as suggested by Olesen & al. (2007). Consequently, the SDR scheme can also be extended and generalized to explore and quantify structure in any type of bipartite ecological networks (Podani & al. 2014). In these systems, the mutualistic relationships between two groups of organisms are summarized by data matrices in which 1 indicates the presence of mutualism (link in the graph), and 0 refers to absence (no link). In addition to plant–pollinator schemes, further examples are host–parasite, plant–disperser or plant–ant networks (Bascompte 2009).

Dissimilarities between all species pairs in either group are calculated based on their interactions with the other group using the Jaccard index, and then decomposed into additive fractions. These components are derived mathematically in the same way as for regular presence-absence data, while the meaning of the components is different, for example, species replacement changes to link replacement (Fig. 8). An advantage of the approach is that interaction pattern is visualized better by the SDR plots than by bipartite graphs or the data matrices themselves, especially if the data set is large.

The method is illustrated using plant–pollinator network data taken by Bartomeus & al. (2008) in coastal plant communities in Catalonia to compare undisturbed sites with those invaded by either *Carpobrotus affine* or *Opuntia stricta,* both species with large, attractive flowers. The data sets were obtained from the Interaction Web DataBase (https://iwdb.nceas.ucsb.edu). The numbers of plant species versus pollinator taxa were 27

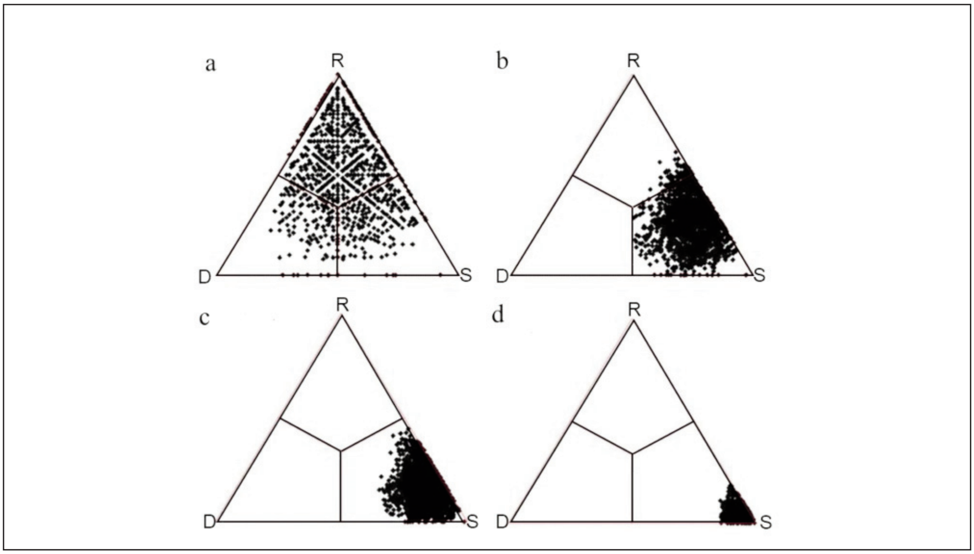


Fig. 6. SDR simplex diagrams for the alpine meadow vegetation from the Italian Alps. **a.** species, **b.** taxonomic, **c.** phylogenetic, **d.** functional. Redrawn after Podani & al. (2018b).

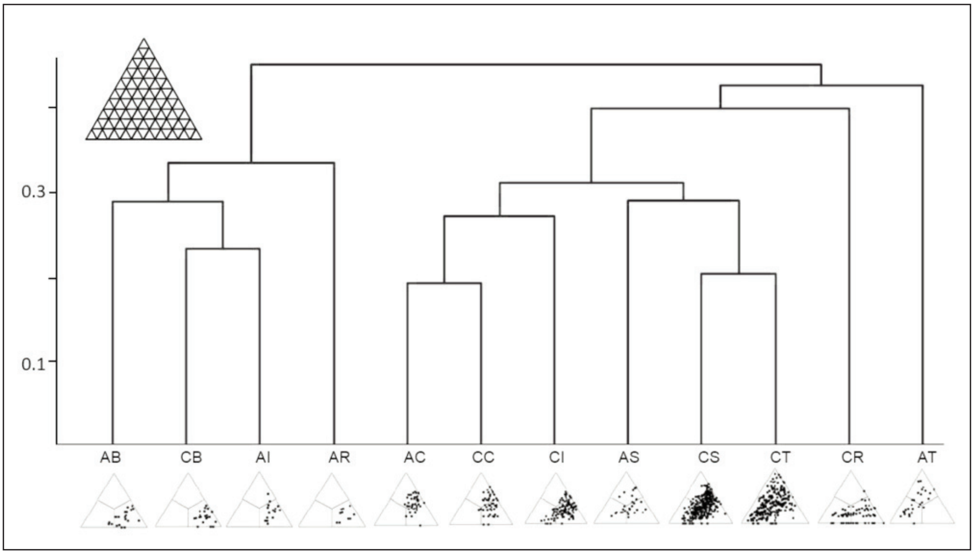


Fig. 7. Classification of biogeographical presence-absence data matrices from two groups of Mediterranean islands (first letter: A-Anatolian islands, C-Cyclades) for six taxonomic groups (second letter: B-butterflies, C-centipedes, I-isopods, R-reptiles, S-land snails, T-tenebrionids) (modified from Podani & al. 2018a). SDR simplex diagrams pertaining to each data matrix are shown on bottom. Inset: ternary plot subdivided into 100 equilateral triangles.

× 85, 18 × 70 and 13 × 47, respectively. Although the number of points in the plots differ and many of them overlap one another, the decrease of pollinator interaction diversity in that direction is clear from the SDR plots (Fig. 9), confirming the current hypothesis on the influence of invasive plants. Numerically, the fractions are 91%, 88% and 86%, respectively. Correspondingly, isolatedness of pollinators (i.e., when two species have no pollinated plant in common) also decreases in this direction (20%, 18% and 14%) while nestedness increases (22%, 30% and 29%) when spectacular invader plants occur in the community, which certainly have homogenizing effect on the plant – pollinator networks.

Conclusions

Jaccard never published a bona fide mathematical formula, an equation with symbols, for his coefficient; he introduced his index by examples and later presented a ratio with verbal terms. Nevertheless, the correct reference to the first use of the method is Jaccard (1901a). In the past century, the coefficient appeared in various forms with different systems of algebraic symbols. It seems now that the use of the parameters of the 2 × 2 contingency table, namely *a, b*, and *c* has been generally accepted and used.

Although originally suggested to compare two items described by the number of shared and differentiating attributes, the index offers many other opportunities. Quantities reflecting

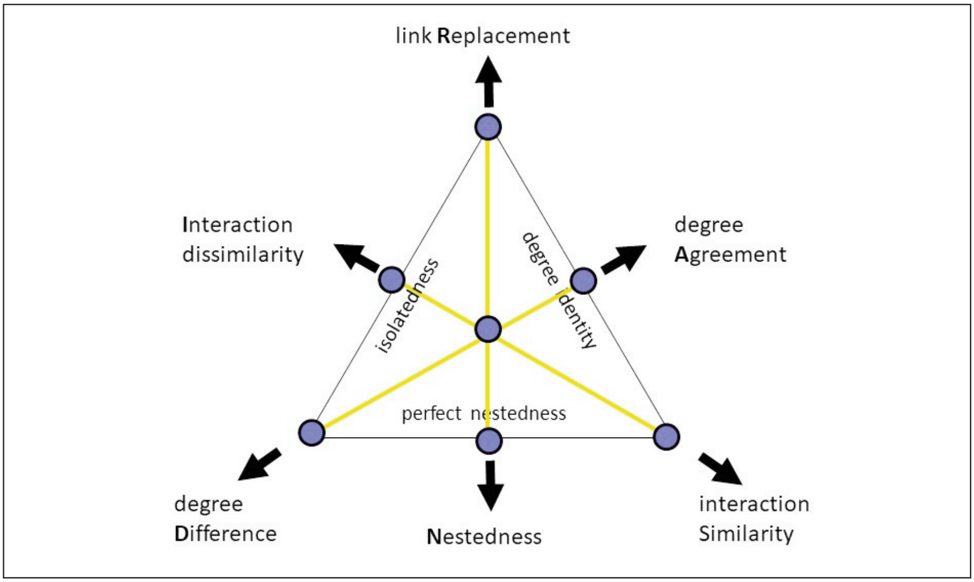


Fig. 8. Adapting the SDR simplex to bipartite interaction networks.

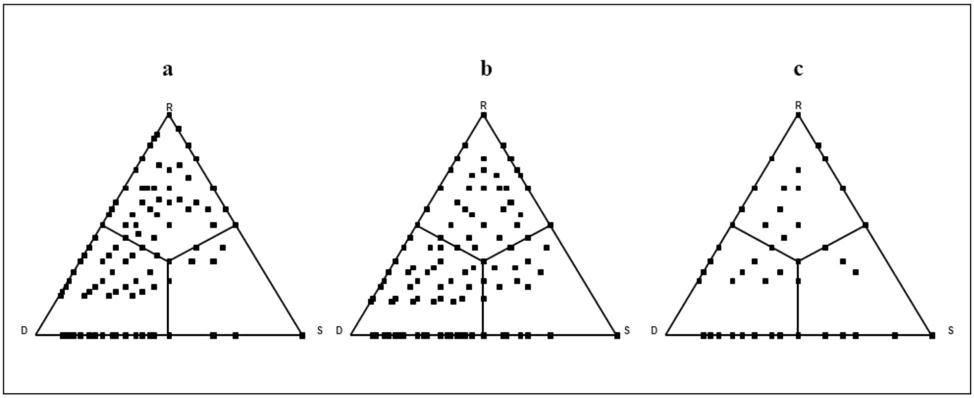


Fig. 9. SDR simplex plots for insects in plant-pollinator networks of uninvaded (**a**), *Carpobrotus*-invaded (**b**) and *Opuntia*-invaded (**c**) Mediterranean coastal plant communities.

disagreement between the two items (i.e., b and c) can be modified to consider taxonomic, phylogenetic or functional distinctness between species in calculating dissimilarity between sites. Furthermore, $b + c$ can be decomposed into additive components, $|b–c| + 2\min\{b,c\}$ which may be used in relativized forms to explore pattern in presence-absence data via twodimensional simplex diagrams. In this way, well-known ecological phenomena, such as beta diversity, nestedness, similarity, richness difference and species turnover are integrated into the same conceptual scheme. These two techniques may be combined into a single, multifaceted approach to the evaluate data structure in a more complex way. The most indirect use of the Jaccard coefficient is in the comparison of data matrices of different size, which reduces this problem to the comparison of point patterns within the ternary plots. Another promising field of application is the study of bipartite ecological networks, in which links between two sets of organisms (usually species) correspond to presence – allowing generalization of the SDR simplexes to the analysis of these complex ecological systems.

References

Bartomeus, I., Vilà, M. & Santamaria, L. 2008: Contrasting effects of invasive plants in plant-pollinator networks. – Oecologia **155:** 761-770. https://doi.org/10.1007/s00442-007-0946-1

Bascompte, J. 2009: Mutualistic networks. – Frontiers Ecol. Environ. **2009:** 7. https://doi.org/10.1890/080026

Braun-Blanquet, J. 1927: Pflanzensoziologie. – Wien.

— 1932: Plant Sociology. – Translated, revised, and edited by Fuller, G. D. & Conard, H. S. – New York and London.

Cardoso, P., Borges, P. A. V. & Veech, J. A. 2009: Testing the performance of beta diversity measures based on incidence data: the robustness to undersampling. – Div. Distrib. **6:** 1081-1090. https://doi.org/10.1111/j.1472-4642.2009.00607.x

Chung, N. C., Miasojedow, B., Startek, M. & Gambin, A. 2019: Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. – BMC Bioinformatics **20:** 644. https://doi.org/10.1186/s12859-019-3118-5

Colwell, R. K. & Coddington, J. A. 1994: Estimating terrestrial biodiversity through extrapolation. – Philos. Trans. Roy. Soc., London B **345:** 101-118.

Conti, F., Abbate, G., Alessandrini, A. & Blasi, C. (eds) 2005: An annotated checklist of the Italian vascular flora. – Roma.

Feoli, E. & Lagonegro, M. 1979: Intersection analysis in phytosociology: computer program and application. – Vegetatio **40:** 55-59. https://doi.org/10.1007/bf00052016

Francey, P. 1941: Le coefficient de communauté de P. Jaccard. – Bull. Soc. Vaudoise Sci. Nat. **61:** 297-316.

Frey-Wyssling, A. 1944: Paul Jaccard. 1868-1944. – Verh. Schweiz. Naturf. Ges. **124:** 339-346.

Goodall, D. W. 1973. Sample similarity and species correlation. – Pp. 107-156 in: Whittaker, R. H (ed.), Ordination and Classification of Communities. – The Hague.

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

— & Legendre, P. 1986: Metric and Euclidean properties of dissimilarity coefficients. – J. Classification **3:** 5-48.

Izsak, C. & Price, A. R. G. 2001: Measuring β-diversity using a taxonomic similarity index, and its relation to spatial scale. — Marine Ecol. Progr. Ser. **215:** 69-77. https://doi.org/10.3354/meps215069

Jaccard, P. 1894: Recherches embryologiques sur l'*Ephédra helvetica* C. A. Meyer. – Bull. Soc. Vaudoise Sci. Nat. **30:** 46-84.

— 1895: Considérations critiques sur les bases du darwinisme appliquées au monde vegetal. – Bull. Soc. Vaudoise Sci. Nat. **31:** 295-311.

— 1900: Contribution au problème de l'immigration post-glaciaire de la flore alpine: étude comparative de la flore alpine du massif de Wildhorn, du haut bassin du Trient et de la haute vallée de Bagnes. – Bull. Soc. Vaudoise Sci. Nat. **36:** 87-130.

— 1901a: Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. – Bull. Soc. Vaudoise Sci. Nat. **37:** 241-272.

— 1901b: Étude comparative de la distribution florale dans une portion des Alpes et du Jura. – Bull. Soc. Vaudoise Sci. Nat. **37:** 547-579.

— 1902a: Lois de distribution florale dans la zone alpine. – Bull. Soc. Vaudoise Sci. Nat. **38:** 69-155.

— 1902b: Distribution comparée de la flore alpine dans quelques régions des Aples occidentales et orientales. – Bull. Murithienne Soc. Valaisanne Sci. Nat. **31:** 81-92.

— 1902c: Gesetze der Pflanzenvertheilung in der alpinen region: auf Grund statistich-floristischer Untersuchungen. – Flora **90:** 349-377.

— 1902d: Vergleichende Untersuchungen über die Verbreitung der alpinen Flora in einigen Regionen der westlichen und östlichen Alpen. – Jahresb. Naturf. Ges. Graubünden **45:** 121-132.

— 1907: La distribution de la flore dans la zone Alpine. – Rev. Gén. Sci. Pures Appl. **18(23):** 961-967.

— 1908: Nouvelles recherches sur la distribution florale. – Bull. Murithienne Soc. Valaisanne Sci. Nat. **44:** 223-270.

— 1912: The distribution of the flora in the Alpine zone. – New Phytologist **11(2):** 37-50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x [English translation of Jaccard (1907)].

— 1926: Le coefficient générique et le coefficient de communauté dans la flore marocaine. – Mém. Soc. Vaudoise Sci. Nat. **2:** 385-403.

Kosub, S. 2019: A note on the triangle inequality for the Jaccard distance. – Pattern Recognition Lett. **120:** 36-38. https://doi.org/10.1016/j.patrec.2018.12.007

Li, W. 2015: Estimating Jaccard index with missing observations: a matrix calibration approach. – In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R. (eds), Advances in Neural Information Processing Systems **28:** 1-7.

Maillefer, A. 1928: Les courbes de Willis: répartition des espèces dans les genres de différente etendue. — Bulletin de la Société Vaudoise des Sciences Naturelles **56:** 617-632.

— 1929: Le Coefficient générique de P. Jaccard et sa signification. — Mémoires de la Société Vaudoise de Sciences Naturelles **3:** 113-183.

Olesen, J. M., Bascompte, J., Dupont, Y. L. & Jordano, P. 2007: The modularity of pollination networks. – Proc. Nat. Acad. Sci. USA **104:** 19891-19896. https://doi.org/10.1073/pnas.0706375104

Pignatti, S. 1982: Flora d'Italia 1-3. — Bologna.

— & Mengarda, F. 1962: Un nuovo procedimento per l'elaborazione delle tabelle fitosociologiche. – Accad Naz. Lincei. Rend. Classe Sci. Fis. Mat. Nat., ser. 8, **32(2):** 215-222.

Podani, J. & Schmera, D. 2011: A new conceptual and methodological framework for exploring and explaining pattern in presence-absence data. – Oikos **120:** 1625-1638. https://doi.org/10.1111/j.1600-0706.2011.19451.x

—, Jordán, F. & Schmera, D. 2014: A new approach to exploring architecture of bipartite (interaction) ecological networks. – J. Complex Networks **2:** 168-186. https://doi.org/10.1093/comnet/cnu002

—, Pavoine, S. & Ricotta, C. 2018a: A generalized framework for analyzing taxonomic, phylogenetic, and functional community structure based on presence–absence data. – Mathematics **6(11):** 250. https://doi.org/10.3390/math6110250

—, Ódor, P., Fattorini, S., Strona, G., Heino, J. & Schmera, D. 2018b: Exploring multiple presence-absence data structures in ecology. – Ecol. Modelling **383:** 41-51. Electronic supplement available at: https://ars.els-cdn.com/content/image/1-s2.0-S0304380018301674-mmc1.doc

Ricotta, C., Podani, J. & Pavoine, S. 2016a: A family of functional dissimilarity measures for presence and absence data. – Ecol. Evol. **6(15):** 5383-5389. https://doi.org/10.1002/ece3.2214

—, Luzzaro, A., Pierce, S., Ceriani, R. M. & Cerabolini, B. 2016b: Measuring the functional redundancy of biological communities: A quantitative guide. – Meth. Ecol. Evol. **7:** 1386-1395.

Sneath, P. H. A. 1957: The application of computers to taxonomy. – J. Gen. Microbiol. **17:** 201-226.

— & Sokal, R. R. 1973: Numerical Taxonomy. – San Francisco.

Sokal, R. R. & Sneath, P. H. A. 1963: Numerical Taxonomy. – San Francisco.

Tanimoto, T. T. 1958: An elementary mathematical theory of classification and prediction. Internal IBM Technical Report. – New York.

Verma, V. & Aggarwal, R. K. 2020: A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective. – Social Network Analysis Mining 10, Article number: 43. https://doi.org/10.1007/s13278-020-00660-9

Whittaker, R. H. 1960: Vegetation of the Siskiyou Mountains, Oregon and California. – Ecol. Monogr. **30:** 279-338.

— & Fairbanks, C. W. 1958: A study of plankton copepod communities in the Columbia Basin, Southeastern Washington. – Ecology **39:** 46-65.

Williams, C. B. 1949: Jaccard's generic coefficient and coefficient of floral community, in relation to the logarithmic series and the index of diversity. – Ann. Bot. **13(49):** 53-58.

Williams, P. H. 1996: Mapping variations in the strength and breadth of biogeographic transition zones using species turnover. – Proc. Roy. Soc., London B, **263:** 579-588. https://doi.org/10.1098/rspb.1996.0087

Williams, W. T. & Dale, M. B. 1965: Fundamental problems in numerical taxonomy. – Advances Bot. Res. **2:** 35-68.

Address of the author:
János Podani,
Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös University, Pázmány P. s. 1.C, H-1117 Budapest, Hungary. E-mail: podani@ludens.elte.hu