*Journal of Vegetation Science 12: 401-410, 2001*
*© IAVS; Opulus Press Uppsala. Printed in Sweden*

# An extension of presence/absence coefficients to abundance data: a new look at absence

**Tamás, Júlia[1,3]; Podani, János[1*] & Csontos, Péter[2]**

[1]*Department of Plant Taxonomy and Ecology, L. Eötvös University, Ludovika tér 2, H-1083 Budapest, Hungary;*
[2]*Ecological Modelling Research Group, Hungarian Academy of Sciences, Ludovika tér 2, H-1083 Budapest, Hungary;*
[3]*Research Institute for Botany and Ecology, Hungarian Academy of Sciences, H-2163 Vácrátót, Hungary:*
[*]*Corresponding author; Fax +361 3338764; E-mail podani@ludens.elte.hu*

**Abstract.** Alternative community analyses, based on quantitative and presence/absence data, are comparable logically if the data type is the only factor responsible for differences among results. For presence/absence indices that consider mutual absences, no quantitative alternatives are known. To facilitate such comparisons, a new family of similarity coefficients is proposed for abundance data. Formally, this extension is achieved by generalizing the four cells of the usual 2 × 2 contingency table to the quantitative case. This implies an expanded meaning of absence: for a given species at a given site it is understood as the difference between the actual value and the maximum detected in the entire study. The correspondence between 10 presence/absence coefficients and their quantitative counterparts is evaluated by graphical comparisons based on artificial data. The behaviour of the new functions is also examined using field data representing post-fire regeneration processes in grasslands and a chronosequence pertaining to forest regeneration after clear-cut. The examples suggest that the new coefficients are most informative for data sets with low beta-diversity and temporal background changes.

**Keywords:** Binary data; Mutual absence; Ordered comparison case series; Ordination; Resemblance; Site dissimilarity.

## Introduction

Ordinations and classifications of community data commonly start with symmetric matrices representing the resemblance structure of the study objects. *Resemblance*, as a general term, may refer to a wide range of measures, such as similarity, dissimilarity, proximity, distance, association or correlation (cf. Orlóci 1978; Ludwig & Reynolds 1988). The literature abounds with hundreds of propositions to measure ecological resemblance, and the choice among them is governed by several factors. The first property to consider is related to the measurement scale of the original variables describing the objects. Two scale types have received particular attention in ecological data analysis, namely the presence/absence and the ratio scale (for more details, see Anderberg 1973 and Orlóci 1978). The *presence/absence* scale, with its two possible states, is the simplest of all and is suitable for the floristic approach in which interest is focused on species lists for landscape units, study sites or sample plots (see, for example, Kuusinen & Siitonen 1998; Dalling & Denslow 1998; Roberts & Wuest 1999). The *ratio*-scale (also called 'relative-scale quantitative', Legendre & Legendre 1998) variables carry much more information. This measurement scale has (1) an infinite number of possible states, (2) a constant interval between any two adjacent units, and (3) a natural zero point. The second property implies that differences are meaningful, whereas the third property allows calculating the ratio of any two values. Species performance, if expressed as percentage cover, biomass or counts (henceforth collectively termed as *abundance*) is of this type; it provides a basis for potentially more sophisticated statistical analyses of ecological phenomena (e.g. Onipchenko et al. 1998). The choice between these two strikingly different levels of resolution is not always simple, and it is very often the case that one wishes to compare results, e.g. classifications or ordinations derived from both presence/absence and abundance versions of the same data set (for example, Stanek 1973; Neldner & Howitt 1991; Núñez-Olivera et al. 1995). Strong congruence of alternative results implies that data structure can be sufficiently represented on a presence/absence basis, and the use of finer scales of measurement is unnecessary. On the other hand, disagreement of results is an indication that the quantitative component in the data is significant. Conclusions derived from such comparisons may have far reaching consequences in pilot surveys before the main sampling is launched.

The question to be addressed first is under which circumstances are comparisons of results between the presence/absence and quantitative levels valid? In the methodological sequence from raw data to the final results, one is faced at each stage with a multitude of choices regarding the data type, data standardization, resemblance coefficient, and ordination and classification method, just to mention the most critical ones (see Podani 1989, 1992 for a review). The relative importance of decisions upon the final results and our conclusions can only be assessed by comparative analysis. If interest lies in detecting incongruence of results caused by differences in the input data type, then all other factors must be kept invariant. That is, if we wish to compare, say, ordinations based on the use of presence/absence and quantitative data, the resemblance function must be the same to eliminate the confounding effect that would arise from differences between the resemblance indices themselves (Kenkel & Orlóci 1986, 'elementary comparisons' of Podani 1989). For example, if the resemblance matrix is obtained by Faith's (1983) binary similarity index (Eq. 9a, below) for presence/absence data, then the ratio scale version of the same data set should be processed using a coefficient entirely consistent with our previous choice. A quantitative counterpart of this presence/absence coefficient is to be found, however, and we wish to emphasize that this problem is very far from being trivial. Very often, the coefficients are not selected carefully, because the presence/absence and quantitative coefficients used in the comparison are not logical counterparts of each other (e.g., Stamol 1991; Pinder & Rosso 1998), although there are lucky choices as well (e.g. Neldner & Howitt 1991; Jutila 1998). The difficulty is that a wide range of presence/absence coefficients used in ecological data analysis have no exact counterparts in the domain of quantitative coefficients, and this is particularly true of measures that consider negative matches ('mutual absences' or 'double zeros') in the calculation of similarity. Curran & Swithinbank (1981) were among the first to recognize this.

The above question leads to the second problem to be discussed: how is mutual absence treated by different formulations of site dissimilarity? The presence/absence coefficients used in numerical ecology and related fields provide three different solutions. In one group of coefficients, attributes get equal weight for mutual presence and absence ('symmetric' coefficients, according to the terminology of Legendre & Legendre 1998), with simple matching coefficient as a typical example. In another group, mutual absences are completely ignored, such as the Jaccard index ('asymmetric' coefficients). There is a set of transitional forms providing opportunity to give mutual absences an intermediate weight (e.g. the Faith index, also 'asymmetric' in the above sense). Aside from these exceptions, most resemblance coefficients currently available for abundance data are insensitive to mutual absences. However, if we wish to make logical comparisons of results following our present argumentation, quantitative analogues of presence/absence coefficients must be found. Furthermore, considering mutual absence for abundance data could allow a refined analysis of vegetation structure in any situation where 'symmetric' presence/absence coefficients are also meaningful. The task is now to find the quantitative analogue of mutual absence.

In our approach, mutual absence is not related merely to species that are missing from both sites compared. For each species, '*potential abundance*' is defined as the maximum amount reached by that species in the study area, and is used as a reference basis to which all abundance values are compared. If a species has much lower quantities than this maximum in both sites in question, then the difference from the maximum is treated as *absence*. This allows extension of 'symmetric' presence/absence coefficients to the quantitative case. Formally, it is achieved by generalizing the well-known 2 × 2 contingency table. The properties of the new indices will be illustrated using a small artificial data set and actual vegetation data coming from rock grassland and oak forest communities.

## A new family of quantitative coefficients

### *Extension of the contingency table*

Presence/absence coefficients of similarity rely upon the well-known 2 × 2 contingency table. Its cells express the number of species present in both sites compared ($a$), the number of species present only in either of them ($b$ and $c$) and the number of species absent from both sites, but present in other sample plots ($d$). The sum $n = a + b + c + d$ is the total number of species present in the entire collection of sample plots. Similarity indices that use all these four frequencies include the simple matching coefficient, the Russell-Rao index, Baroni-Urbani - Buser's indices, etc. (for a more complete survey, see Lamont & Grant 1979; Huhta 1979; Janson & Vegelius 1981; Wolda 1981; Gower & Legendre 1986; Kenkel & Booth 1987; Sgardelis & Stamou 1990; Legendre & Legendre 1998). The quantitative indices operate completely differently: they are expressed in terms of the original raw data scores. The difference in formalism makes it difficult to find a one-to-one correspondence between members of these two families of coefficients so important for consistent comparisons. However, if the 2 × 2 contingency table is redefined for quantitative data such that each cell derives directly from the raw scores, then practically all presence/absence coefficients can be rewritten to conform with quantitative data - providing a new group of resemblance coefficients and facilitating logical comparison between results based on presence/absence and abundance data.

In the following extension of the contingency table approach, capital letters $A$, $B$, $C$ and $D$ are used to make reference to and distinction from the presence/absence variants *a, b, c* and *d*, respectively. Let $\mathbf{X} = \{x_{ij}\}$ denote the data matrix of abundances and its general element for row *i* and column *j*. Assume that the number of rows (variables, species) is *n*, and the number of columns (sites, objects, quadrats) is *m*. Then, for the pair of sites *j* and *k* we define the following cell 'frequencies' (without subscripts *j* and *k*, to follow the convention in the presence/absence case).

*A* is the amount of abundance in which the two sites agree, summed over all species, i.e.,

$$A = \sum_{i=1}^{n} \min\{x_{ij}, x_{ik}\} \tag{1}$$

*B* is the sum of abundances by which site *j* exceeds site *k*, that is

$$B = \sum_{i=1}^{n} \left(\max\{x_{ij}, x_{ik}\} - x_{ik}\right) \tag{2}$$

*C* is the sum of abundances by which site *k* exceeds site *j*, calculated as

$$C = \sum_{i=1}^{n} \left(\max\{x_{ij}, x_{ik}\} - x_{ij}\right) \tag{3}$$

(*B* + *C*) thus reflects the total amount of disagreement between the two sites being compared. Finally, *D* is obtained as the sum of differences from the highest values attained in the entire sample. For each species, the maximum value over all sites is determined; it is understood as 'potential' abundance that could be reached in the study area. The higher of the two abundances manifested in sites *j* and *k* is then subtracted from the potential value, giving the 'mutual absence' in quantitative terms. Then, summation over all species gives the desired quantity. Formally, it is obtained by the expression

$$D = \sum_{i=1}^{n} \left(\max_j\{x_{ij}\} - \max\{x_{ij}, x_{ik}\}\right) \tag{4}$$

One may verify easily that *A, B, C* and *D* reduce to the respective cell frequencies of the 2 × 2 contingency table if the data matrix contains only 1-s and 0-s.

Considering *d* in presence/absence coefficients has the obvious consequence that addition of a site in which new species occur modifies all previous resemblance values (Goodall 1973). This is also true of *D*, although no new species are needed to evoke this overall change in the quantitative case. The resemblance structure modifies if the newly added site supersedes the former maxima for at least one species. This phenomenon is not new in numerical community ecology: standardization by species maxima or species totals has similar effect. In some sense, therefore, incorporation of *D* in calculations of ecological resemblance is analogous to standardization. The rationale behind both the above definition of *D* and standardization is that the comparison of any two sites is 'embedded' in a reference basis provided by the entire set of sample sites.

In the presence/absence case, each species contributes to a single cell of the contingency table. This is not so with the above generalization, because a given species may contribute to two or even three cells simultaneously; appearance in one cell only is the exception rather than the rule. For example, if the abundance of a species is 8 in site *j* and 12 in site *k*, whilst its maximum in the entire sample is 32, then its contributions to *A*, *C* and *D* are 8, 4 and 20, respectively. Note also that the sum of the four cell values yields the total of 'potential abundances', that is

$$A{+}B{+}C{+}D = N = \sum_{i=1}^{n} \max_j\{x_{ij}\} \tag{5}$$

### *A categorization of resemblance coefficients*

The following short overview of presence/absence and quantitative coefficients will show that there are many new possibilities of treating abundance data based on the generalized contingency table notations. It becomes most obvious if we evaluate the correspondence between presence/absence and quantitative coefficients. When abundances are converted to presence/absence data, many operations involved in the resemblance functions become identical, therefore several quantitative indices simplify to the same presence/absence coefficient. The correspondence between indices in this direction is always unambiguous; i.e., there is a subjective mapping from the set of quantitative coefficients to the binary indices. The reverse is not true: expansion of presence/absence forms is often ambiguous or even meaningless. In the following categorization, the nomenclature of indices for which nor equations neither citations are presented here follows Orlóci (1978), Goodall (1973) and Kenkel & Booth (1987).

#### *Group 1. Indices with D or d disregarded*

Quantitative indices with summation over species as the primary operation can be readily expressed in terms of the new notation, and their correspondence to binary indices is straightforward. Consider first a well-known dissimilarity function, the Bray-Curtis dissimilarity measure given by

$$\mathrm{BC}_{jk} = \frac{\sum_{i=1}^{n} |x_{ij} - x_{ik}|}{\sum_{i=1}^{n} (x_{ij} + x_{ik})} \tag{6}$$

which can be rewritten in terms of (1) - (4) as

$$\mathrm{BC}_{jk} = \frac{\sum_{i=1}^{n} 2\max\{x_{ij}, x_{ik}\} - x_{ik} - x_{ij}}{\sum_{i=1}^{n}\left(x_{ij} + x_{ik}\right)} = (B+C)/(2A+B+C) \qquad (7)$$

Its complement is 2$A$/[2$A$ + $B$ + $C$], with the Sørensen index (2$a$/[2$a$ + $b$ + $c$]) as the binary counterpart. When the Sørensen index is expanded back to quantitative data, i.e., $a$, $b$ and $c$ are replaced by $A$, $B$ and $C$, the complement of the Bray-Curtis measure is obtained readily. Other indices with a similar behaviour include the Kulczyński's formulae; these are given by 1/2($A$/[$A$ + $B$] + $A$/[$A$ + $C$]) and 1/2($a$/[$a$ + $b$] + $a$/[$a$ + $c$]). The Manhattan metric ($B$ + $C$ and $b$ + $c$) is another example with clear relationship between the presence/absence and quantitative forms.

For elucidating an ambiguous case, consider the Ružička index (which is the complement of the Marczewski-Steinhaus dissimilarity coefficient) and Wishart's (1969) similarity ratio (see also van der Maarel 1979). Both of them reduce to the Jaccard index in the binary case ($a$/[$a$ + $b$ + $c$]). However, expansion of the Jaccard index to the quantitative case yields only one of these indices; according to formulae (1) - (4) the Ružička index is reproduced. The similarity ratio cannot be expressed using our min-max-based definitions of $A$, $B$, $C$ and $D$. For similar reasons, formally replacing $a$, $b$ and $c$ by $A$, $B$ and $C$ in the presence/absence versions of Euclidean, chord and geodesic distance cannot reproduce the original formula for the quantitative case.

*Group 2. New conversions for formulae considering d*

This group contains coefficient pairs for which only the presence/absence version was known (with one noted exception, see below) – the quantitative variants are derived according to the proposed new interpretation of the contingency table. The presence/absence versions are those that consider in some manner the number of mutual absences ($d$). The simplest of these is the Russell-Rao coefficient, $a$/[$a$ + $b$ + $c$ + $d$], whose counterpart is defined as

$$\mathrm{RRq}_{jk} = A/(A + B + C + D) =$$

$$\sum_{i=1}^{n}\min\left\{x_{ij}, x_{ik}\right\} \Big/ \sum_{i=1}^{n}\max{}_{j}\left\{x_{ij}\right\} \qquad (8)$$

The coefficient proposed by Faith (1983) considers both $a$ and $d$ of the contingency table, although asymmetrically. This function and its quantitative transcription are given by the following equations:

$$\mathrm{FA2}_{jk} = (a + 1/2d) / (a + b + c + d) \qquad (9a)$$

$$\mathrm{FA2q}_{jk} = (A + 1/2D) / (A + B + C + D) =$$

$$\left[\sum_{i=1}^{n}\min\left\{x_{ij}, x_{ik}\right\} + 0.5\sum_{i=1}^{n}\left(\max{}_{j}\left\{x_{ij}\right\} - \max\left\{x_{ij}, x_{ik}\right\}\right)\right] \Big/ \sum_{i=1}^{n}\max{}_{j}\left\{x_{ij}\right\} \qquad (9b)$$

Further coefficients, which consider double zeros and can be expanded to the quantitative case using 'frequencies' (1)-(4) include the simple matching coefficient, the Rogers - Tanimoto index, Anderberg's two formulae, the Sokal-Sneath index, Baroni-Urbani & Buser's and Yule's indices. It is left to the reader to derive the sometimes complicated formulae; we do not present them here to save space.

*Relationships to known formulae*

The Manhattan metric, mentioned already in Group 1, deserves some more attention here. Faith (1984) has expressed this formula by its similarity counterpart ($A$ + $D$) with the present denotations. That is, Faith implicitly considers mutual absences of abundances. Besides a small typographic error, corrected afterwards in Faith et al. (1987), Faith's expression is identical to the expansion of $A$ + $D$ based on formulae (1) and (4). The possibility that $D$ can be generalized to other coefficients escaped the attention of the authors, however, with one exception being their *intermediate coefficient*. In this, the Manhattan metric ($B$ + $C$) and the dissimilarity version of Kendall's (1970) minimum agreement measure ($B$ + $C$ + $D$) are dynamically averaged using a scale factor α:

$$INTC_{jk} = \alpha\,(B + C) + (1 - \alpha)\,(B + C + D) \qquad (10)$$

With α = 0.5, this coefficient can be written as $B$ + $C$ + 0.5$D$ which is apparently the numerator of the complement of the generalized version of Faith's binary index (Eq. 9a). This is the only quantitative formula published so far which can be mentioned in the context of our generalized contingency approach. Since the value of α can be modified from 0 to 1, in effect we obtain an infinite series of coefficients which implies a continuity from a formula in which $D$ does not contribute to the dissimilarity at all (α = 1) to another in which $D$ is equally weighted with $B$ and $C$ (α = 0). Thus, with fully balanced weighting we achieve a compromise, whose advantages in the presence/absence case have been discussed in detail by Faith (1983).

There are further connections to existing formulations. The complement of the Russell - Rao index, [$B$ + $C$ + $D$]/$N$, is closely related to Kendall's dissimilarity measure: only the range of the latter is normalized to [0,1]. Similar relationship holds true for the complement of the simple matching coefficient, $[B + C]/N$, and the Manhattan metric. This illustrates an important fact that normalization of a coefficient introduces scale conversion only. An advantage of normalization is that clustering results become more directly comparable based on hierarchical levels if all coefficients are normalized.

At first glance, one may suggest that incorporation of the $D$ cell is similar in effect to data standardization with the maximum (or the standard deviation) for each species. This is not the case, however, because standardization equalizes the range (or the variance) of all species so that the variables become entirely commensurable. Calculating $D$ based on raw data, as proposed here, retains the original differences between the variables.

## Some selected features of the new coefficients

Two topics associated with the present approach have particular importance for the practitioner. (1) The similarity of the new forms to the presence/absence counterparts and the question whether the properties of the former can be deducted from those of the latter. (2) The relationship between the newly described formulae and existing quantitative coefficients, and the important requirement that the former ones should yield information from the data which cannot be obtained otherwise. These are investigated in some detail using artificial and actual vegetation data.

### *Artificial data*

The properties of similarity coefficients have been revealed efficiently and extensively by graphical evaluation (Lamont & Grant 1979; *ordered comparison case series* of Hajdu 1981; Gower & Legendre 1986; Shi 1993; Podani 2000). The procedure involves the use of a standard object to which all other objects, changed in a regular way to represent some meaningful trend in the data, are compared. The sample data set given in Table 1 was selected here as a basis for the comparative evaluation of indices. In this matrix, nine sample sites are described in terms of the abundances of 16 variables (species). The species are assumed to have simple unimodal response to an hypothetical underlying gradient, the objects gradually change from site 1 to site 9 such that the optima of the species are shifted at each step. Table 2 summarizes the values of the contingency table for both presence/absence and abundance-based comparisons. Table 3 presents the presence/absence coefficients used for the artificial example.

**Table 1.** Artificial data for the graphical evaluation of indices. Note the apparent gradient from site 1 to site 9.

| | | OBJECTS | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| R | 5 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 |
| I | 6 | 3 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 |
| A | 7 | 2 | 3 | 4 | 4 | 3 | 2 | 1 | 0 | 0 |
| B | 8 | 1 | 2 | 3 | 4 | 4 | 3 | 2 | 1 | 0 |
| L | 9 | 0 | 1 | 2 | 3 | 4 | 4 | 3 | 2 | 1 |
| E | 10 | 0 | 0 | 1 | 2 | 3 | 4 | 4 | 3 | 2 |
| S | 11 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 4 | 3 |
| | 12 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 | 4 |
| | 13 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 | 4 |
| | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Overall, the presence/absence and quantitative forms of the same coefficient exhibit essentially the same response to systematic changes in the data (Fig. 1). The curves fitted to the points take similar shapes, and the rank order of values for a given coefficient is identical for its both versions. The discrepancy between the alternatives becomes more increased in the direction of 1/1→1/9, due to the increased importance of cell $D$ in the calculations (cf. Table 2). Since this cell cannot be zero in this sample data set, many similarities do not fall to zero, even though there are no species in common for the pair 1/9 (Fig. 1b, d). When the coefficient's value entirely depends on the multiple of $A$ and $D$ (A1q and BB2q) or $D$ is not used in the numerator (RRq), however, the minimum of zero is reached.

**Table 2.** Cell frequencies of the 2 × 2 contingency table according to presence/absence (a) and quantitative (b) approaches for the comparison of site 1 with itself and all the others in Table 1.

| **a**. | 1/1 | 1/2 | 1/3 | 1/4 | 1/5 | 1/6 | 1/7 | 1/8 | 1/9 |
|---|---|---|---|---|---|---|---|---|---|
| *a* | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| *b* | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| *c* | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| *d* | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |

| **b.** | 1/1 | 1/2 | 1/3 | 1/4 | 1/5 | 1/6 | 1/7 | 1/8 | 1/9 |
|---|---|---|---|---|---|---|---|---|---|
| *A* | 20 | 16 | 12 | 9 | 6 | 4 | 3 | 1 | 0 |
| *B* | 0 | 4 | 8 | 11 | 14 | 16 | 18 | 19 | 20 |
| *C* | 0 | 4 | 8 | 11 | 14 | 16 | 18 | 19 | 20 |
| *D* | 32 | 28 | 24 | 21 | 18 | 16 | 14 | 13 | 12 |

**Table 3.** Presence/absence coefficients extended to abundance data in this paper.

| Abbreviation | Name | Formula |
|---|---|---|
| SM | Simple matching coefficient | $\frac{a+d}{a+b+c+d}$ |
| RT | Rogers & Tanimoto | $\frac{a+d}{a+2b+2c+d}$ |
| SS1 | Sokal & Sneath | $\frac{2a+2d}{2a+b+c+2d}$ |
| A1 | Anderberg 1 | $\left(\frac{a}{a+b}\cdot\frac{a}{a+c}\cdot\frac{d}{b+d}\cdot\frac{d}{c+d}\right)^{1/2}$ |
| A2 | Anderberg 2 | $\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$ |
| FA2 | Faith 2 | $\frac{a+0.5d}{a+b+c+d}$ |
| RR | Russell & Rao | $\frac{a}{a+b+c+d}$ |
| BB2 | Baroni-Urbani & Buser | $\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$ |
| Y1 | Yule 1 | $\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$ |
| Y2 | Yule 2 | $\frac{ad-bc}{ad+bc}$ |

### *Field data*

Post-fire successional data collected in an Austrian pine (*Pinus nigra*) plantation on the south-facing dolomite slopes of the Buda Hills, north-central Hungary (Tamás & Csontos 1998; Podani et al. 2000) serve as the first actual example. Percentage species cover scores were recorded in five repetitions of 2 m × 4 m permanent plots examined annually for four years after the fire of 1993. A control plot examined in a neighbouring grassland stand untouched by the fire was also included as a reference basis. The sampling survey yielded a 112 × 25 matrix, with species as rows and sample plot/year combinations as columns.

Data from a chronosequence representing the regeneration succession of sessile oak-turkey oak forest following clear-cut offer a possibility for further illustrations. Percentage cover scores were recorded from the herb layer using 20 m × 20 m quadrats. Each quadrat was taken from a different watershed of the same geographic region (Visegrádi Mts., Hungary) with practically uniform climate and very similar soil properties. Stand age varied from 4 to 30 years since clear-cut, and within this time span three consecutive stages of regeneration succession can be identified (Csontos 1996):

A. *Early stage* of forest regeneration with a mosaic pattern of grassy patches and groups of young shrub-like trees. Relative irradiation in the herb layer varies between 22 % and 59 % (12 quadrats).

B. *Intermediate stage* with a very dense thicket of young trees. Relative irradiation in the herb layer is as low as 2 % (8 quadrats).

C. *Final stage.* The tree canopy opens up again thus allowing 12-16 % relative irradiation penetrating to the herb layer (five quadrats).

This study yielded a 199 species × 25 quadrats data matrix with species as rows and quadrats as columns.

Both data sets were analysed by the same methodology. Resemblance matrices for sites were calculated using the complements (dissimilarity forms) of four similarity indices: the Jaccard index (a, in Figs. 2 and 3) and its quantitative counterpart, the Ružička index (c) represent formulae utilizing only three cells of the contingency table, whereas the simple matching coefficient (b) and its newly established quantitative version (d) were chosen for inspecting the effect of four cells. The dissimilarity matrices were subjected to metric multidimensional scaling (alias principal coordinates analysis, Gower 1966) to derive an ordination of plots and years. The ordinations were performed by the SYN-TAX 5.1 program package (Podani 1993).

## Results

*Post-fire succession.* The ordinations along the first two axes (Fig. 2a-d) account for 31.4, 46.8, 24.3 and 42% of variance, respectively, showing that variance extraction is more efficient when the *d(D)* cell is also considered in calculating dissimilarity. This higher variance is expected for coefficients considering *d* or *D* (P. Legendre, pers. comm.). The four scatter diagrams agree in that the plots representing the unburnt area separate along the first, and most significant axis from the other sample plots taken in the burnt stands. There are, however, differences in the arrangement of burnt sites, allowing alternative interpretations of post-fire successional changes.

On the basis of presence/absence data (Fig. 2a, b), the Jaccard index and the simple matching coefficient produced remarkably similar ordinations. The two scattergrams apparently suggest similar trends and groups in the data indicating that there is low beta diversity (i.e., short gradient) in the data. This is not so with the cover data! In the ordination diagram obtained by the extended version of the simple matching coefficient (SMq, Fig. 2d), the plots from the first study year are positioned very close to one another near the origin, as a manifestation of their small vegetation cover relative to the subsequent years (0.9% versus the average of 40-65%). This arrangement reflects well the high overall similarity of sites from early regeneration stages. The

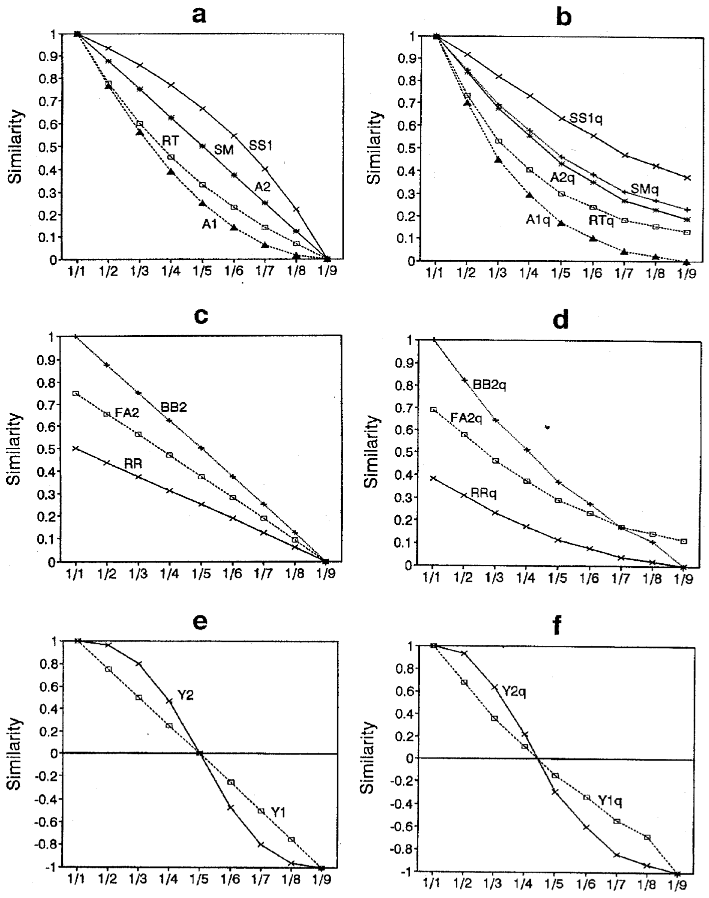

**Fig. 1.** Graphical comparison of presence/absence coefficients (**a**, **c** and **e**) with their quantitative counterparts (**b**, **d** and **f**, respectively) based on the artificial data of Table 1. Object 1 is compared with itself and with all other objects along an imaginary gradient (ordered comparison case series). Abbreviations: A1 and A2 = Anderberg indices, BB2 = Baroni-Urbani & Buser's second index, FA2 = Faith's index, RR = Russell & Rao index, RT = Rogers & Tanimoto's index, SM = simple matching coefficient, SS1 = Sokal & Sneath coefficient, Y1 and Y2 = Yule coefficients; q refers to the 'quantitative' variant of the index. See Table 3, for formulae of presence/absence forms.

explanation is fairly obvious: the denominator of SMq is constant over all comparisons and the small individual differences between first year sample plots are overwhelmed by the absence of large quantities that appear in later years in the sample. However, when mutual absence is disregarded (Ružička index, Fig. 2c), the small differences among first year quadrats are more emphasized because the denominator is also small. In this case, the first year plots clearly separate from the other burnt sites on axis 2. These are the most essential differences between the two groups of indices. As mentioned above, considering mutual absences implies the same reference basis for all comparisons of quadrats. If mutual absences are not included, we disregard quantities that are absent from the two quadrats being compared, but are potentially manifested in the study region. Therefore, the two possibilities of expressing ecological similarity are complementary to each other.

*Regeneration succession after clear-cut.* As far as ordination efficiency is concerned (18%, 24%, 18% and 27% for dimensions 1-2 in Fig. 3a-d), the situation is the same as in the previous example: when *d(D)* is disregarded, the efficiency is lower. The ordinations based on presence/absence data (Fig. 3a, b) are muc more similar to each other, as in the previous data set. Again, it shows the presence of a short background gradient. The early phase of regeneration (stage A) is separated from the two older stages (B and C) along the first axis, while stages B-C overlap to some extent in the ordination plane. An explanation is that in stage A the high relative irradiation supports a well-developed species rich herb layer with several light demanding species. In the next stage (B), under the shade of the thicket the herb layer impoverishes, only shade tolerant species and a few highly tolerant oak wood species can survive. In stage C, as a result of increased relative irradiation the herb layer becomes more vigorous. The species that survived in stage B now attain higher cover and only a few species appear, hence the pronounced floristic relationship between stages B and C.

Ordinations based on quantitative data produced a markedly different arrangement of groups (Fig. 3c, d): stage B is separated from A and C along the first axis. This separation is caused by the low total cover of species in this intermediate successional stage. For the Ružička index (Fig. 3c), quadrats from stages A and C show a slight tendency to segregate, whereas the new quantitative form of the simple matching coefficient places them into practically the same cluster. The closeness of the starting and closing phases is the manifestation of higher

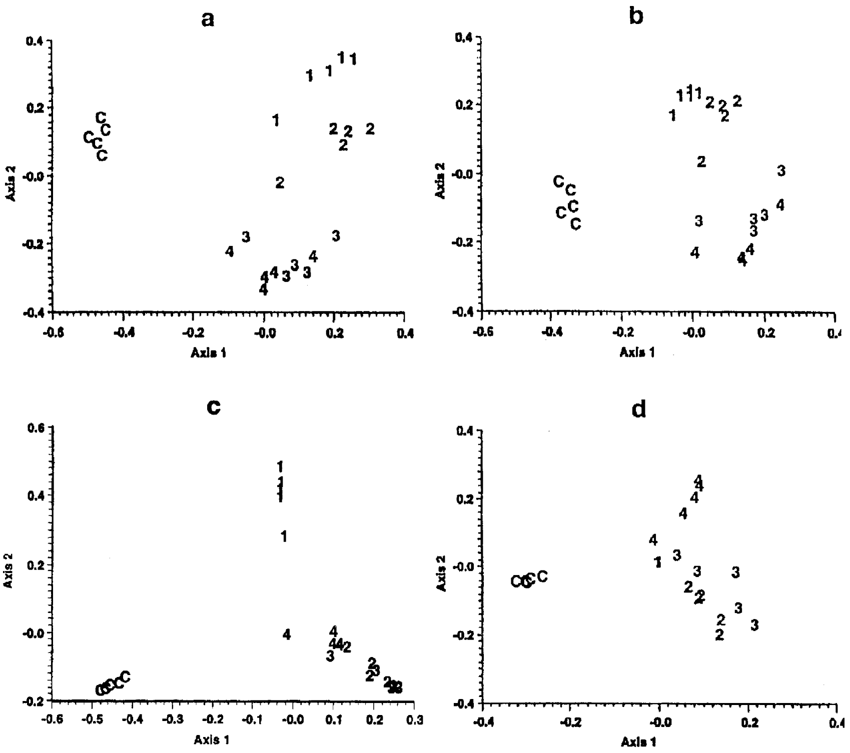

**Fig. 2.** Principal coordinates analysis of post-fire succession data over four consecutive years (numbered 1 to 4) using five replicate plots from each year. C refers to the control (unburnt) area. Presence/absence coefficients are Jaccard (**a**) and simple matching coefficient (**b**), quantitative indices are the Ružička (**c**) and the new, quantitative form of simple matching coefficient (**d**).

cover values (due to elevated relative irradiation) notwithstanding that stage C is much poorer in species than stage A. The high similarity between A and C is more pronounced by the newly established quantitative form of the simple matching coefficient (Fig. 3d). Note also that stage B quadrats form a more compact group with the new index. These sites are all characterized by the absence of potential abundance of their species, as a consequence of the very shady thicket stage of the forest regeneration succession. The new index seems to be a powerful tool to detect such lack of abundances.

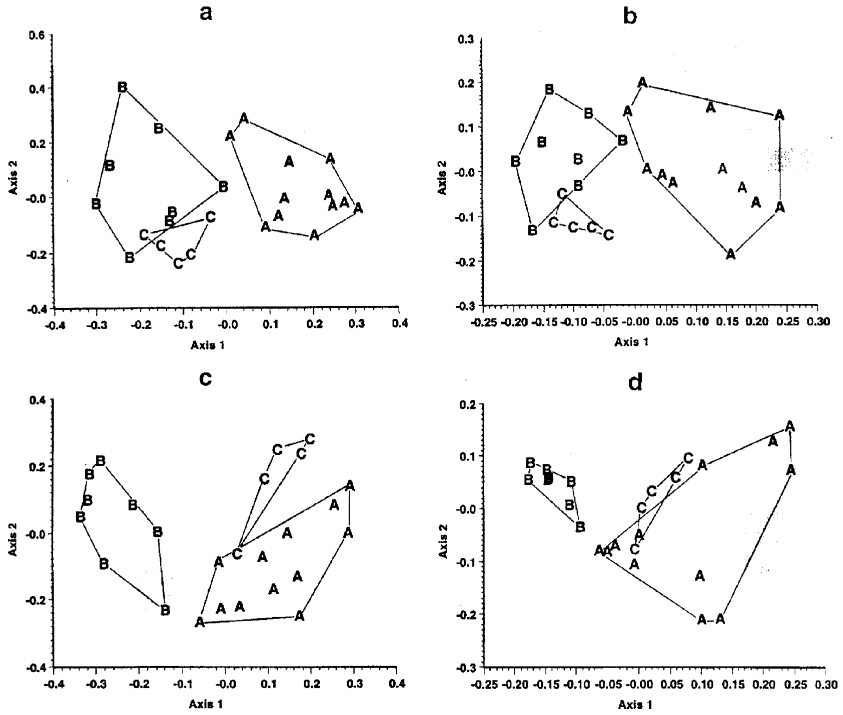

**Fig. 3.** Principal coordinates analysis of chronosequence data with three distinct successional stages (A: initial species rich stage, B: intermediate thicket stage, C: open wood). Presence/absence coefficients are Jaccard (**a**) and simple matching coefficient (**b**), quantitative indices are the Ružička (**c**) and the new, quantitative form of simple matching coefficient (**d**). Convex polygons showing the three stages are superimposed for clarity.

## Discussion

This paper presents an extension of presence/absence coefficients to abundance data. The inclusion of a 'quantitative' version of the *d* cell in the coefficients allows that maximum quantities of species attained by the sample are considered in the comparisons. As a result, all pairwise comparisons are referred to the same universal set, which is otherwise a natural requirement for presence/absence indices that consider mutual absences. Earlier, data standardization was practically the only possibility to place all pairwise comparisons to the same reference basis. Standardization and the incorporation of the *D* cell are, however, logically very different operations and therefore serve different purposes in data analysis.

There is considerable agreement among ecologists that the absence of a species from both sites being compared may be an indication of different background phenomena, contrary to mutual presence (Orlóci 1978; Green 1979; Legendre & Legendre 1998, to name only a few). This is especially true in communities with high beta-diversity. In case of long ecological gradients, for example, the measurement of ecological resemblance may be excessively biased if double absences are considered so that the asymmetric indices are favoured. Mutual absences do have ecological meaning, however, in sites where species absence is explained by factors other than environmental heterogeneity. Post-fire regeneration of vegetation is such a case, as we have demonstrated.

The new extensions imply more than a simple enlargement of the available arsenal of resemblance coefficients. A methodological importance of introducing the extended forms is that comparisons of analyses based on different data types can be brought to the same logical basis. If one wishes to evaluate a change from presence/absence to cover or abundance data, all aspects of the analysis (resemblance coefficient, method of ordination or classification) can be kept constant, so that differences among results are due *only to* data type changes. This allows exclusion of confounding effects, which may mask the trends that we are examining.

The examples demonstrated that the importance of the new extensions is not only theoretical. The quantitative forms reveal information that would remain undetectable by standard resemblance coefficients. When the underlying ecological gradient is short, as in the examples, the presence/absence alternatives yield very similar results, which is not the case with the quantitative forms. Such situations are expected to appear in any study concerned with temporal vegetation change over a study area with relatively narrow ecological variation. Under such circumstances, potential abundance, i.e., the maximum amount reached in the study area, does have ecological meaning to be considered in quantitative studies. It is noted that potential abundance is in effect even though no zeros appear in the data, that is, the expanded forms imply more than a simple extension of double absence for the quantitative case.

We acknowledge that the development or improvement of resemblance functions has been out of focus of contemporary vegetation science. Some could say that the number of coefficients is large enough, making the investigator's choice unnecessarily difficult. It happens very often in science that, whenever a large amount of knowledge has accumulated after an extensive research period and the overview of all results requires increased efforts, the attention of many researchers is diverted towards new, freshly emerged and more fashionable research topics. We agree with the late P. Juhász-Nagy (pers. comm.) that this is an undesirable situation, and feel that there are no research fields exhaustively exploited. As confirmed by the present study, this is the case with resemblance coefficients: the actual examples demonstrated that the use of logical counterparts of indices highlights new aspects inherent in vegetation data.

**Acknowledgements.** This research was supported by the Hungarian National Research Grant (OTKA) No. T029784 and European Commission Grant No. ENV4-CT97-0680 to the authors. We are grateful to P. Legendre, J. Oksanen and an anonymous referee for their constructive criticism of the manuscript.

## References

Anderberg, M.R. 1973. *Cluster analysis for applications*. Academic Press, New York, NY.

Csontos, P. 1996. *Regeneration succession of sessile oak-turkey oak forests: processes in the herb-layer.* Synbiologia Hungarica 2(2). Scientia, Budapest. (In Hungarian with English summary.)

Curran, P. & Swithinbank, P. 1981. The application of Gower's maximal predictive classification to vegetation data. *J. Biogeogr.* 8: 1-5.

Dalling, J.W. & Denslow, J.S. 1998. Soil seed bank composition along a forest chronosequence in seasonally moist tropical forest, Panama. *J. Veg. Sci.* 9: 669-678.

Faith, D.P. 1983. Asymmetric binary similarity measures. *Oecologia* (*Berl.*) 57: 287-290.

Faith, D.P. 1984. Patterns of sensitivity of association measures in numerical taxonomy. *Math. Biosci.* 69: 199-207.

Faith, D.P., Minchin, P.R. & Belbin, L. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* 69: 57-68.

Goodall, D.W. 1973. Sample similarity and species correlation. In: Whittaker, R.H. (ed.) *Ordination and classifica-*

*tion of vegetation*, pp. 107-156. Junk, The Hague.

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

Gower, J.C. & Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 3: 5-48.

Green, R.H. 1979. *Sampling design and statistical methods for environmental biologists.* Wiley, New York, NY.

Hajdu, L. 1981. Graphical comparison of resemblance measures in phytosociology. *Vegetatio* 48: 47-59.

Huhta, V. 1979. Evaluation of different similarity indices as measures of succession in arthropod communities of the forest floor after clear-cutting. *Oecologia* (*Berl.*) 41: 11-23.

Janson, S. & Vegelius, J. 1981. Measures of ecological association. *Oecologia* (*Berl.*) 49: 371-376.

Jutila, H.M. 1998. Seed banks of grazed and ungrazed Baltic seashore meadows. *J. Veg. Sci.* 9: 395-408.

Kendall, D.G. 1970. A mathematical approach to seriation. *Philos. Trans. R. Soc. Lond. Ser. A.* 269: 125-135.

Kenkel, N.C. & Booth, T. 1987. A comparison of presence/absence coefficients for use in biogeographical studies. *Coenoses* 2: 25-30.

Kenkel, N.C. & Orlóci, L. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology* 67: 919-928.

Kuusinen, M. & Siitonen, J. 1998. Epiphytic lichen diversity in old-growth and managed *Picea abies* stands in southern Finland. *J. Veg. Sci.* 9: 283-292.

Lamont, B.B. & Grant, K.J. 1979. A comparison of twenty-one measures of site dissimilarity. In: Orlóci, L., Rao, C.R. & Stiteler, W.M. (eds.) *Multivariate methods in ecological work*, pp. 101-126. International Cooperative Publishing House, Burtonsville, MD.

Legendre, P. & Legendre, L. 1998. *Numerical ecology*. 2nd ed. Elsevier, Amsterdam.

Ludwig, J.A. & Reynolds, J.F. 1988. *Statistical ecology*. Wiley, New York, NY.

Neldner, V.J. & Howitt, C.J. 1991. Comparison of an intuitive mapping classification and numerical classifications of vegetation in south-east Queensland, Australia. *Vegetatio* 94: 141-152.

Núñez-Olivera, E., Martínez-Abaigar, J., Escudero, J.C. & García-Novo, F. 1995. A comparative study of *Cistus ladanider* shrublands in extremadura (CW Spain) on the basis of woody species composition and cover. *Vegetatio* 117: 123-132.

Onipchenko, V.G., Semenova, G.V. & van der Maarel, E. 1998. Population strategies in severe environments: alpine plants in the northwestern Caucasus. *J. Veg. Sci.* 9: 27-40.

Orlóci, L. 1978. *Multivariate analysis in vegetation research*. 2nd ed. Junk, The Hague.

Pinder, L. & Rosso, S. 1998. Classification and ordination of plant formations in the Pantanal of Brazil. *Plant Ecol.* 136: 151-165.

Podani, J. 1989. Comparison of ordinations and classifications of vegetation data. *Vegetatio* 83: 111-128.

Podani, J. 1992. Space series analysis of vegetation: processes reconsidered. *Abstr. Bot.* 16: 25-29.

Podani, J. 1993. SYN-TAX-pc. *Computer programs for multivariate data analysis in ecology and systematics*. Version 5.0. User's guide. Scientia, Budapest.

Podani, J. 2000. *Introduction to the exploration of multivariate biological data*. Backhuys, Leiden.

Podani, J., Csontos, P. & Tamás, J. 2000. Additive trees in the analysis of community data. *Community Ecol.* 1: 33-41.

Roberts, M.R. & Wuest, L.J. 1999. Plant communities of New Brunswick in relation to environmental variation. *J. Veg. Sci.* 10: 321-334.

Sgardelis, S.P. & Stamou, G.P. 1990. The effects of dominance, species ranking and species matching on some similarity indices. *J. Veg. Sci.* 1: 125-128.

Shi, G.R. 1993. Multivariate data analysis in palaeoecology and palaeobiogeography – a review. *Palaeogeogr. Palaeoclimat. Palaeoecol.* 105: 199-234.

Stamol, V. 1991. Coenological study of snails (Mollusca: Gastropoda) in forest phytocoenoses of Medvednica mountain (NW Croatia, Yugoslavia). *Vegetatio* 95: 33-54.

Stanek, W. 1973. A comparison of Braun-Blanquet's method with sum of squares agglomeration for vegetation classification. *Vegetatio* 27: 323-345.

Tamás, J. & Csontos, P. 1998. Early regeneration of dolomite vegetation after burning of *Pinus nigra* plantations. In: Csontos P. (ed.) *Sziklagyepek szünbotanikai kutatása*, pp. 231-264. Scientia, Budapest. (In Hungarian with English summary.)

van der Maarel, E. 1979. Multivariate methods in phytosociology, with reference to the Netherlands. In: Werger, M. J. A. (ed.) *The study of vegetation*, pp. 163-225. Junk, The Hague.

Wishart, D. 1969. An algorithm for hierarchical classifications. *Biometrics* 25: 165-170.

Wolda, H. 1981. Similarity indices, sample size and diversity. *Oecologia* (*Berl.*) 50: 296-302.

Received 18 October 1999;
Revision received 27 November 2000;
Accepted 9 February 2001.
Coordinating Editor: J. Oksanen.