Journal of Classification 26:227-245 (2009)
DOI: 10.1007/s00357-009-9032-1

k-Adic Similarity Coefficients for Binary (Presence/Absence) Data

Matthijs J. Warrens

Leiden University, The Netherlands

Abstract: k-Adic formulations (for groups of objects of size k) of a variety of 2-adic similarity coefficients (for pairs of objects) for binary (presence/absence) data are presented. The formulations are not functions of 2-adic similarity coefficients. Instead, the main objective of the the paper is to present k-adic formulations that reflect certain basic characteristics of, and have a similar interpretation as, their 2-adic versions. Two major classes are distinguished. The first class is referred to as Bennani-Heiser similarity coefficients, which contains all coefficients that can be defined using just the matches, the number of attributes that are present and that are absent in k objects, and the total number of attributes. The coefficients in the second class can be formulated as functions of Dice's association indices.

Keywords: Indices of association; Resemblance measures; Simple matching coefficient; Jaccard coefficient; Dice/Sørenson coefficient; Rand index; Global order equivalence.

1. Introduction

A variety of data can be represented in strings of binary scores. In general, the binary scores reflect either the presence or absence of certain attributes of a specific object. For example, in psychology, the objects may be persons that may or may not posses certain traits; in ecology, the objects could be regions or districts in which certain species do or do not occur (or,

The author thanks Willem Heiser and three anonymous reviewers for their helpful comments and valuable suggestions on earlier versions of this article.

Author's Address: Psychometrics and Research Methodology Group, Leiden University Institute for Psychological Research, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands, e-mail: warrens@fsw.leidenuniv.nl.

Published online 6 June 2009

vice versa, the objects are species that coexist in a number of locations); in archaeology, the objects may be graves where specific artifact types can be found; finally, in chemical similarity searching, the objects may be target structures or queries and the attributes certain compounds in a database. A variety of *similarity coefficients* (SCs) have been introduced in the literature to measure the resemblance (association) between two objects. These SCs for presence/absence data can also be used to compare two clusterings (partitions) of a data set (Albatineh, Niewiadomska-Bugaj and Mihalko 2006; Hubert and Arabie 1985; Rand 1971). To obtain an overview of the SCs that have been proposed over the years, or which SCs are currently in use, the reader is referred to the following articles published in the Journal of Classification: Gower and Legendre (1986), Baulieu (1989), Batagelj and Bren (1995), Albatineh et al. (2006) and Warrens (2008a). Earlier reviews of SCs for binary data were presented in, among others, Sokal and Sneath (1963), Cheetham and Hazel (1969), Baroni-Urbani and Buser (1976), Janson and Vegelius (1981) and Hubálek (1982).

Many SCs for presence/absence data compare two objects or clusterings at a time. Let O be a finite set of objects (denoted by $j_1, j_2, j_3, ...$), and let the number of attributes be denoted by n ($n > 0$). A dyadic (2-adic) SC is defined as a mapping $S : O \times O \rightarrow \mathbb{R}$, into the reals, such that

$$S(j_1, j_1) \geq S(j_1, j_2), \quad \text{and} \quad S(j_1, j_2) = S(j_2, j_1), \quad \forall j_1, j_2 \in O.$$

Many SCs have the property $S(j_1, j_1) = 1$.

Instead of pairs of objects, SCs may also be defined on triples, quadruples or groups of k objects. A triadic (3-adic) SC is defined as a mapping $S : O \times O \times O \rightarrow \mathbb{R}$, such that $S(j_1, j_1, j_1) \geq S(j_1, j_1, j_2) \geq S(j_1, j_2, j_3)$ and there is 3-way symmetry,

$$\begin{aligned} S(j_1, j_2, j_3) = S(j_1, j_3, j_2) = S(j_2, j_1, j_3) \\ = S(j_2, j_3, j_1) = S(j_3, j_1, j_2) = S(j_3, j_2, j_1) \end{aligned}$$

$\forall j_1, j_2, j_3 \in O$. Furthermore, following Joly and Le Calvé (1995) and Heiser and Bennani (1997), a 3-adic SC must satisfy $S(j_1, j_1, j_2) = S(j_1, j_2, j_2)$ $\forall j_1, j_2 \in O$. With the latter condition we require that, if one of the objects is identical to one of the others, the similarity between the nonidentical objects should be the same, regardless of which two are the same.

The definition of a k-adic SC $S(j_1, j_2, ..., j_k)$ ($k \geq 2$), including k-way symmetry, is analogous to the definition of a 3-adic SC. Clearly, k-adic SCs for binary data can be used to compare k strings of binary scores at a time. Furthermore, if one wishes to compare k partitions in cluster analysis, instead of two partitions as in Albatineh et al. (2006) or Hubert and

Arabie (1985), it is required to know what quantities a k-adic SC consists of. Moreover, k-adic SCs can be used with the multi-way extensions of multidimensional scaling as considered in Cox, Cox and Branco (1991).

In this paper we consider k-adic formulations of various 2-adic presence/absence SCs. Many SCs were originally introduced as measures of similarity. It seems therefore natural to consider k-adic formulations of SCs, instead of their dissimilarity counterparts. The k-adic generalizations presented here, are not functions of 2-adic SCs, as is the case in, for example, Joly and Le Calvé (1995) or De Rooij and Gower (2003). Instead, the main objective of the paper is to present k-adic formulations that reflect certain basic characteristics of their 2-adic versions. For example, if it holds that $1 \geq S(j_1, j_2) \geq 0$, then we require that its k-adic version is at least on the same range. Another important characteristic is how the SC may be interpreted for pairs of objects, and how this may generalize to triples or groups of size k.

The paper is organized as follows. In the next section some well-known SCs are studied that are linear in both the numerator and the denominator. The SCs in Section 2 can be defined using just the matches, the number of attributes present in k objects as well as the number of attributes absent in k objects, and n, the total number of attributes. The SCs in Section 3 are functions of the association indices presented in Dice (1945). For the 2-adic case the two association indices are defined as the amount of similarity between any two species, relative to the occurrence of either. The functions considered in Section 3 include the Pythagorean means (harmonic, arithmetic and geometric means), the product and the minimum or maximum function. Section 4 contains a discussion.

2. Bennani-Heiser SCs

Many 2-adic SCs are written as functions of the four dependent variables

$$a = \text{the number of attributes present in both } j_1 \text{ and } j_2$$
$$b = \text{the number of attributes present in } j_1 \text{ but absent in } j_2$$
$$c = \text{the number of attributes present in } j_2 \text{ but absent in } j_1$$
$$d = \text{the number of attributes absent in both } j_1 \text{ and } j_2$$

where $a + b + c + d = n$ (cf. Baulieu, 1989, p. 234). It is interesting to note that, although b and c are two separate variables, many (well-known) SCs are defined to be symmetric in b and c. As noted by Heiser and Bennani (1997, p. 195), a large number of 2-adic SCs are characterized by the number of positive matches (a), negative matches (d), and mismatches (b, c). This is especially the case for SCs that are linear in both numerator and denominator.

Instead of variables a, b, c, and d, we define for k binary n-vectors the three variables

$$\begin{aligned} x^{(k)} &= \text{the number of attributes present in } j_1, j_2, ..., j_k \\ z^{(k)} &= \text{the number of attributes absent in } j_1, j_2, ..., j_k \\ y^{(k)} &= n - x^{(k)} - z^{(k)}, \text{ the number of mismatches.} \end{aligned}$$

We have $x^{(2)} = a$, $z^{(2)} = d$ and $y^{(2)} = b + c$. It should be noted that there is no special reason for the use of symbols x, y and z.

SCs that can be defined using only the variables $x^{(k)}$, $y^{(k)}$ and $z^{(k)}$ will be named after Bennani-Dosse (1993) and Heiser and Bennani (1997), who first presented these SCs for triples of objects. Note that, in the following definition, S is a k-adic function of the three quantities $x^{(k)}$, $y^{(k)}$ and $z^{(k)}$, not a function of three objects.

Definition. A *Bennani-Heiser* SC (BHSC) is a mapping

$$S\left(x^{(k)}, y^{(k)}, z^{(k)}\right) : (\mathbb{Z}^+)^3 - \{(0,0,0)\} \rightarrow \mathbb{R}$$

from the set of all ordered 3-tuples of non-negative integers other than the origin into the reals.

Although many well-known BHSCs are linear in both numerator and denominator, it is not a necessary property (see Section 2.3).

With respect to BHSCs, we may reformulate the concept of order equivalence, originally coined by Sibson (1972). Note that, in the following definition, S and T are k-adic functions of the three quantities $x^{(k)}$, $y^{(k)}$ and $z^{(k)}$, instead of functions of three objects.

Definition. Two BHSCs, S and T, are said to be *globally order equivalent* (GOE) provided

$$\forall \left(x_1^{(k)}, y_1^{(k)}, z_1^{(k)}\right), \left(x_2^{(k)}, y_2^{(k)}, z_2^{(k)}\right) \in (\mathbb{Z}^+)^3 - \{(0,0,0)\}$$

we have

$$\begin{aligned} &S\left(x_1^{(k)}, y_1^{(k)}, z_1^{(k)}\right) > S\left(x_2^{(k)}, y_2^{(k)}, z_2^{(k)}\right) \quad \text{iff} \\ &T\left(x_1^{(k)}, y_1^{(k)}, z_1^{(k)}\right) > T\left(x_2^{(k)}, y_2^{(k)}, z_2^{(k)}\right). \end{aligned}$$

If two coefficients are GOE, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations.

2.1 The Jaccard Coefficient

Paul Jaccard (1912) studied the distribution of certain flora in the Alpine zone. In his particular field of interest, the objects were three different Alpine districts and the attributes were species of plants. To measure the resemblance or similarity of two districts in terms of species, Jaccard used the ratio

$$S_{\text{Jac}}^{(2)} = \frac{\text{Number of species common to the two districts}}{\text{Total number of species in the two districts}} = \frac{a}{a+b+c} = \frac{x^{(2)}}{x^{(2)}+y^{(2)}} = \frac{x^{(2)}}{n-z^{(2)}}. \tag{1}$$

A seemingly proper and straightforward 3-adic formulation of the Jaccard coefficient would be

$$S_{\text{Jac}}^{(3)} = \frac{\text{Number of species common to the three districts}}{\text{Total number of species in the three districts}} = \frac{x^{(3)}}{x^{(3)}+y^{(3)}}.$$

The dissimilarity formulation $1-S_{\text{Jac}}^{(3)}$ was presented in Cox, Cox and Branco (1991, p. 200). Furthermore, the k-adic formulation of (1) is then given by

$$S_{\text{Jac}}^{(k)} = \frac{x^{(k)}}{x^{(k)}+y^{(k)}} = \frac{x^{(k)}}{n-z^{(k)}}.$$

The coefficient in (1) is a member of a family of SCs with a positive parameter θ, which was, according to Heiser and Bennani (1997, p. 197), first studied by both Fichet (1986) and Gower (1986). This family is given by

$$S_{\text{F-G}}^{(2)}(\theta) = \frac{a}{a+\theta(b+c)} = \frac{x^{(2)}}{x^{(2)}+\theta\, y^{(2)}}, \tag{2}$$

where θ is a positive parameter that modifies the number of mismatches in (2). The Fichet-Gower family can be generalized to the k-adic family

$$S_{\text{F-G}}^{(k)}(\theta) = \frac{x^{(k)}}{x^{(k)}+\theta\, y^{(k)}}.$$

A SC with $0 < \theta < 1$ gives more weight to $x^{(k)}$. For $x^{(2)}$ this is regularly done in the case that there are only a few positive matches relative to the number of mismatches: $x^{(2)}$ is much smaller than $y^{(2)}$. Similar arguments can be used for the opposite case and $\theta > 1$. Note that $1 \geq S_{\text{F-G}}^{(k)}(\theta) \geq 0$, $\forall\, \theta\, \forall\, k$, where 1 is obtained iff $y^{(k)} = 0$, and 0 is obtained iff $x^{(k)} = 0$.

For an arbitrary ordinal comparison with respect to $S^{(k)}_{\text{F-G}}(\theta)$, we have

$$\frac{x_1^{(k)}}{x_1^{(k)} + \theta\, y_1^{(k)}} > \frac{x_2^{(k)}}{x_2^{(k)} + \theta\, y_2^{(k)}} \quad \text{iff} \quad \frac{x_1^{(k)}}{y_1^{(k)}} > \frac{x_2^{(k)}}{y_2^{(k)}}.$$

Since an arbitrary ordinal comparison with respect to $S^{(k)}_{\text{F-G}}(\theta)$ does not depend on the value of θ, any two members of (2) are GOE. Table 1 presents several members of (2), their corresponding k-adic formulations, and a GOE SC that is not a member of $S^{(k)}_{\text{F-G}}(\theta)$.

2.2 The Simple Matching Coefficient

Instead of positive matches only, one may also be interested in a SC that involves the negative matches. The simple matching coefficient

$$\begin{aligned} S^{(2)}_{\text{SM}} &= \frac{\text{Number of attributes present and absent in two objects}}{\text{Total number of attributes}} \\ &= \frac{a+d}{a+b+c+d} = \frac{x^{(2)} + z^{(2)}}{x^{(2)} + y^{(2)} + z^{(2)}} = \frac{x^{(2)} + z^{(2)}}{n} \end{aligned} \tag{3}$$

(or Rand index in cluster analysis) has a slightly different formulation compared to the Jaccard coefficient. Possible 3-adic and k-adic formulation of (3) are

$$S^{(3)}_{\text{SM}} = \frac{x^{(3)} + z^{(3)}}{x^{(3)} + y^{(3)} + z^{(3)}} \quad \text{and} \quad S^{(k)}_{\text{SM}} = \frac{x^{(k)} + z^{(k)}}{x^{(k)} + y^{(k)} + z^{(k)}}.$$

For a different but interesting extension of (3), see Gower and Hand (1996, p. 66).

The simple matching coefficient is a member of a second parameter family, that can be found in Gower and Legendre (1986, p. 13). This family is given by

$$S^{(2)}_{\text{G-L}}(\theta) = \frac{a+d}{a+\theta(b+c)+d} = \frac{x^{(2)} + z^{(2)}}{x^{(2)} + \theta\, y^{(2)} + z^{(2)}}. \tag{4}$$

The k-adic formulation of the parameter family in (4) is

$$S^{(k)}_{\text{G-L}}(\theta) = \frac{x^{(k)} + z^{(k)}}{x^{(k)} + \theta\, y^{(k)} + z^{(k)}}.$$

For $0 < \theta < 1$, the SC gives more weight to both $x^{(k)}$ and $z^{(k)}$; for $\theta > 1$ more weight is assigned to $y^{(k)}$. Note that $1 \geq S^{(k)}_{\text{G-L}}(\theta) \geq 0$, $\forall\ \theta\ \forall\ k$, where 1 is obtained iff $y^{(k)} = 0$, and 0 is obtained iff $x^{(k)} = z^{(k)} = 0$.

Table 1. Fichet-Gower SCs and a GOE SC by Kulczyński.

Source	2-adic	θ	k-adic
Kulczyński (1927)	$\frac{a}{b+c}$	–	$\frac{x^{(k)}}{y^{(k)}}$
Jaccard (1912)	$\frac{a}{a+b+c}$	1	$\frac{x^{(k)}}{x^{(k)}+y^{(k)}}$
Sokal and Sneath (1963)	$\frac{a}{a+2(b+c)}$	2	$\frac{x^{(k)}}{x^{(k)}+2y^{(k)}}$
Gleason (1920), Dice (1945), Sørenson (1948), Czekanowski (1932), Nei and Li (1979)	$\frac{2a}{2a+b+c}$	1/2	$\frac{2x^{(k)}}{2x^{(k)}+y^{(k)}}$

For an arbitrary ordinal comparison with respect to $S^{(k)}_{\text{G-L}}(\theta)$, we have

$$\frac{x_1^{(k)}+z_1^{(k)}}{x_1^{(k)}+\theta\, y_1^{(k)}+z_1^{(k)}} > \frac{x_2^{(k)}+z_2^{(k)}}{x_2^{(k)}+\theta\, y_2^{(k)}+z_2^{(k)}} \quad \text{iff}$$
$$\frac{x_1^{(k)}+z_1^{(k)}}{y_1^{(k)}} > \frac{x_2^{(k)}+z_2^{(k)}}{y_2^{(k)}}.$$

Since an arbitrary ordinal comparison with respect to $S^{(k)}_{\text{G-L}}(\theta)$ does not depend on the value of θ, any two members of (4) are GOE. Table 2 presents several members of (4), the corresponding k-adic formulations and two other GOE SCs.

2.3 Miscellaneous SCs

In its 2-adic form, a SC by Russel and Rao (1940) is given by

$$S^{(2)}_{\text{R-R}} = \frac{a}{a+b+c+d} = \frac{x^{(2)}}{x^{(2)}+y^{(2)}+z^{(2)}} = \frac{x^{(2)}}{n}.$$

A straightforward k-adic generalization of this SC is

$$S^{(k)}_{\text{R-R}} = \frac{x^{(k)}}{x^{(k)}+y^{(k)}+z^{(k)}} = \frac{x^{(k)}}{n}.$$

Baroni-Urbani and Buser (1976, p. 258) introduced the two SCs

$$S^{(2)}_{\text{B-B}} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}} = \frac{x^{(2)}+\sqrt{x^{(2)}z^{(2)}}}{x^{(2)}+y^{(2)}+\sqrt{x^{(2)}z^{(2)}}}$$

and

Table 2. Gower-Legendre SCs and two other GOE SCs.

Source	2-adic	θ	k-adic
Sokal and Sneath (1963)	$\frac{a+d}{b+c}$	–	$\frac{x^{(k)}+z^{(k)}}{y^{(k)}}$
Sokal and Michener (1958), Rand(1971)	$\frac{a+d}{a+b+c+d}$	1	$\frac{x^{(k)}+z^{(k)}}{x^{(k)}+y^{(k)}+z^{(k)}}$
Rogers and Tanimoto (1960)	$\frac{a+d}{a+2(b+c)+d}$	2	$\frac{x^{(k)}+z^{(k)}}{x^{(k)}+2y^{(k)}+z^{(k)}}$
Gower and Legendre (1986), Sokal and Sneath (1963)	$\frac{2(a+d)}{2a+b+c+2d}$	$1/2$	$\frac{2(x^{(k)}+z^{(k)})}{2(x^{(k)}+z^{(k)})+y^{(k)}}$
Hamann (1961), Hubert (1977), Holley and Guilford (1964)	$\frac{a-b-c+d}{a+b+c+d}$	–	$\frac{x^{(k)}-y^{(k)}+z^{(k)}}{x^{(k)}+y^{(k)}+z^{(k)}}$

$$S^{(2)}_{\text{B-B2}} = \frac{a-b-c+\sqrt{ad}}{a+b+c+\sqrt{ad}} = \frac{x^{(2)}-y^{(2)}+\sqrt{x^{(2)}z^{(2)}}}{x^{(2)}+y^{(2)}+\sqrt{x^{(2)}z^{(2)}}}.$$

Possible k-adic formulations of these SCs are

$$S^{(k)}_{\text{B-B}} = \frac{x^{(k)}+\sqrt{x^{(k)}z^{(k)}}}{x^{(k)}+y^{(k)}+\sqrt{x^{(k)}z^{(k)}}}$$

and

$$S^{(k)}_{\text{B-B2}} = \frac{x^{(k)}-y^{(k)}+\sqrt{x^{(k)}z^{(k)}}}{x^{(k)}+y^{(k)}+\sqrt{x^{(k)}z^{(k)}}}.$$

3. Dice's Association Indices

Denote by

$$n_{j_i} = \text{the total number of attributes in object } j_i.$$

Then, for the 2-adic case we have

$$a+b=n_{j_1}, \quad b+d=n-n_{j_2}, \quad a+c=n_{j_2} \quad \text{and} \quad c+d=n-n_{j_1}.$$

For the 2-adic SCs in this section, these quantities are essential. Since the k-adic formulations presented in the following are no longer only based on the matches, $x^{(k)}$ and $z^{(k)}$, and the total number of attributes n, we need to

reformulate Sibson's (1972) concept of GOE for k-adic SCs that are not BH-SCs. In the following definition, $h_1, h_2, ..., h_k$ are, similar to $j_1, j_2, ..., j_k$, elements of the set O.

Definition. Two k-adic SCs, S and T, are said to be GOE if

$$S(j_1, j_2, ..., j_k) > S(h_1, h_2, ..., h_k) \quad \text{iff}$$
$$T(j_1, j_2, ..., j_k) > T(h_1, h_2, ..., h_k).$$

Dice (1945, p. 298) proposed 2-adic association indices that consist of the amount of co-occurrence between any two species j_1 and j_2, relative to the occurrence of either j_1 or j_2. Hence, for every pair of objects there are two indices, namely

$$\text{index } j_2/j_1 = \frac{a}{a+b} = \frac{x^{(2)}}{n_{j_1}} \quad \text{and} \quad \text{index } j_1/j_2 = \frac{a}{a+c} = \frac{x^{(2)}}{n_{j_2}}.$$

Albatineh et al. (2006) report Wallace (1983) as a source for these indices.

3.1 The Harmonic Mean

Dice (1945) recognized that in some ecologic studies it would be desirable to have an index that does not change depending on which species is used as a base in the denominator. What became know as the Dice coefficient is Dice's coincidence index, which has a value intermediate between the reciprocal association indices. The coincidence index is given by

$$S^{(2)}_{\text{Dice}} = \frac{2a}{2a+b+c} = \frac{2\,x^{(2)}}{n_{j_1} + n_{j_2}}. \tag{5}$$

This SC has been proposed independently by multiple authors (see Table 1). Bray (1956) reports Gleason (1920) as one of the first to introduce $S^{(2)}_{\text{Dice}}$. The SC can be interpreted as the number of joint occurrences of two species $x^{(2)}$, divided by the average frequency of occurrence of the two species $(n_{j_1} + n_{j_2})/2$, that is,

$$S^{(2)}_{\text{Dice}} = \frac{a}{\frac{a+b}{2} + \frac{a+c}{2}} = \frac{x^{(2)}}{\frac{n_{j_1}+n_{j_2}}{2}}.$$

In addition, $S^{(2)}_{\text{Dice}}$ may be interpreted as the harmonic mean of the two association indices, which is given by

$$S^{(2)}_{\text{Dice}} = \frac{2}{\frac{a+b}{a} + \frac{a+c}{a}} = \frac{2}{\frac{n_{j_1}}{x^{(2)}} + \frac{n_{j_2}}{x^{(2)}}}.$$

We have $1 \geq S^{(2)}_{\text{Dice}} \geq 0$, which we already knew, because the coefficient in (5) is a member of $S^{(2)}_{\text{F-G}}(\theta)$, for $\theta = 1/2$.

Dice (1945, p. 300) already noted that the indices he proposed could be easily expanded to measure the amount of association between three or more species. Thus, for every triple of objects there are three indices, namely

$$\text{index} j_2 j_3 / j_1 = \frac{x^{(3)}}{n_{j_1}}, \quad \text{index} j_1 j_3 / j_2 = \frac{x^{(3)}}{n_{j_2}} \quad \text{and} \quad \text{index} j_1 j_2 / j_3 = \frac{x^{(3)}}{n_{j_3}}.$$

The 3-adic and k-adic formulations of (5) are

$$S^{(3)*}_{\text{Dice}} = \frac{3}{\frac{n_{j1}}{x^{(3)}} + \frac{n_{j2}}{x^{(3)}} + \frac{n_{j3}}{x^{(3)}}} = \frac{3\, x^{(3)}}{n_{j_1} + n_{j_2} + n_{j_3}}$$

and

$$S^{(k)*}_{\text{Dice}} = \frac{k\, x^{(k)}}{\sum_{i=1}^{k} n_{j_i}}$$

where $*$ is used to denote that $S^{(k)*}_{\text{Dice}}$ is a different k-adic SC compared to the k-adic generalization in Table 1. Both $S^{(3)*}_{\text{Dice}}$ and $S^{(k)*}_{\text{Dice}}$ are harmonic means of 3, respectively k, association indices. Furthermore, $S^{(3)*}_{\text{Dice}}$ (and $S^{(k)*}_{\text{Dice}}$) can be interpreted as the number of joint occurrences of three (k) species $x^{(3)}$, divided by the average frequency of occurrence of the three (k) species $(n_{j_1} + n_{j_2} + n_{j_3})/3$. Similar to $S^{(2)}_{\text{Dice}}$, we have $1 \geq S^{(k)*}_{\text{Dice}} \geq 0$.

3.2 The Arithmetic Mean

For pairs of objects, Kulczyński (1927) proposed the SC

$$S^{(2)}_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) = \frac{1}{2}\left(\frac{x^{(2)}}{n_{j_1}} + \frac{x^{(2)}}{n_{j_2}}\right) \tag{6}$$

which is the arithmetic mean (or average) of Dice's indices. Hence, straightforward 3-adic and k-adic formulations of (6) are

$$S^{(3)}_{\text{Kul}} = \frac{1}{3}\left(\frac{x^{(3)}}{n_{j_1}} + \frac{x^{(3)}}{n_{j_2}} + \frac{x^{(3)}}{n_{j_3}}\right) \quad \text{and} \quad S^{(k)}_{\text{Kul}} = \frac{1}{k}\sum_{i=1}^{k} \frac{x^{(k)}}{n_{j_i}}.$$

Both formulations are arithmetic means of 3, respectively k, association indices. However, $S^{(2)}_{\text{Kul}}$ can also be written as

$$S^{(2)}_{\text{Kul}} = \frac{a\,(2a+b+c)}{2(a+b)(a+c)} = \frac{x^{(2)}(n_{j_1} + n_{j_2})}{2 n_{j_1} n_{j_2}}. \tag{7}$$

Possible 3-adic and k-adic formulations of (7) are respectively

$$S_{\text{Kul}}^{(3)*} = \frac{\left[x^{(3)}\right]^2 (n_{j_1} + n_{j_2} + n_{j_3})}{3n_{j_1}n_{j_2}n_{j_3}}$$

and

$$S_{\text{Kul}}^{(k)*} = \frac{\left[x^{(k)}\right]^{k-1} \left(\sum_{i=1}^{k} n_{j_i}\right)}{k \prod_{i=1}^{k} n_{j_i}},$$

where $*$ in $S_{\text{Kul}}^{(k)*}$ is used to denote that this is an alternative k-adic formulation compared to $S_{\text{Kul}}^{(k)}$. Although $S_{\text{Kul}}^{(k)*}$ is not the arithmetic mean of k association indices, this SC (and not $S_{\text{Kul}}^{(k)}$) is GOE to a coefficient in Section 3.4.

Sokal and Sneath (1963) presented a SC which is the arithmetic mean (or average) of Dice's indices and the quantities $d/(b+d)$ and $d/(c+d)$. The SC is given by

$$\begin{aligned} S_{\text{S-S}}^{(2)} &= \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) \\ &= \frac{1}{4}\left(\frac{x^{(2)}}{n_{j_1}} + \frac{x^{(2)}}{n_{j_2}}\right) + \frac{1}{4}\left(\frac{z^{(2)}}{n-n_{j_1}} + \frac{z^{(2)}}{n-n_{j_2}}\right) \end{aligned} \tag{8}$$

and extends (6) in that it includes negative matches. Possible 3-adic and k-adic formulations of (8) are

$$S_{\text{S-S}}^{(3)} = \frac{1}{6}\left(\frac{x^{(3)}}{n_{j_1}} + \frac{x^{(3)}}{n_{j_2}} + \frac{x^{(3)}}{n_{j_3}}\right) + \frac{1}{6}\left(\frac{z^{(3)}}{n-n_{j_1}} + \frac{z^{(3)}}{n-n_{j_2}} + \frac{z^{(3)}}{n-n_{j_3}}\right)$$

and

$$S_{\text{S-S}}^{(k)} = \frac{1}{2k}\sum_{i=1}^{k}\frac{x^{(k)}}{n_{j_i}} + \frac{1}{2k}\sum_{i=1}^{k}\frac{z^{(k)}}{n-n_{j_i}}.$$

Similar to $S_{\text{Kul}}^{(2)}$ and $S_{\text{S-S}}^{(2)}$, we have $1 \geq S_{\text{Kul}}^{(k)}, S_{\text{Kul}}^{(k)*}, S_{\text{S-S}}^{(k)} \geq 0$.

3.3 The Geometric Mean

The geometric mean of Dice's indices

$$\begin{aligned} S_{\text{Och}}^{(2)} &= \sqrt{\frac{a}{a+b} \times \frac{a}{a+c}} = \frac{a}{\sqrt{(a+b)(a+c)}} \\ &= \frac{x^{(2)}}{n_{j_1}^{1/2}\, n_{j_2}^{1/2}} \end{aligned} \tag{9}$$

is considered in Ochiai (1957) and Fowlkes and Mallows (1983). The 3-adic and k-adic formulations of (9) are given by

$$S_{\text{Och}}^{(3)} = \frac{x^{(3)}}{n_{j_1}^{1/3}\, n_{j_2}^{1/3}\, n_{j_3}^{1/3}} \quad \text{and} \quad S_{\text{Och}}^{(k)} = \frac{x^{(k)}}{\prod_{i=1}^{k} n_{j_i}^{1/k}}.$$

Both formulations are geometric means of 3, respectively k, association indices.

An extension of (9) presented in Sokal and Sneath (1963) that includes the negative matches between objects j_1 and j_2, can be written as

$$\begin{aligned} S_{\text{S-S2}}^{(2)} &= \sqrt{\frac{a}{a+b} \times \frac{a}{a+c} \times \frac{d}{b+d} \times \frac{d}{c+d}} \\ &= \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\ &= \frac{x^{(2)} z^{(2)}}{\left[n_{j_1}(n-n_{j_1})\right]^{1/2} \left[n_{j_2}(n-n_{j_2})\right]^{1/2}}. \end{aligned} \tag{10}$$

The SC in (10) is not a geometric mean. The SC may be interpreted as a product of two geometric means, or as the square of the geometric mean of Dice's indices and the quantities $d/(b+d)$ and $d/(c+d)$. Possible 3-adic and k-adic formulations of (10) are

$$S_{\text{S-S2}}^{(3)} = \frac{x^{(3)} z^{(3)}}{\left[n_{j_1}(n-n_{j_1})\right]^{1/3} \left[n_{j_2}(n-n_{j_2})\right]^{1/3} \left[n_{j_3}(n-n_{j_3})\right]^{1/3}}$$

and

$$\begin{aligned} S_{\text{S-S2}}^{(k)} &= \frac{x^{(k)}}{\prod_{i=1}^{k} n_{j_i}^{1/k}} \times \frac{z^{(k)}}{\prod_{i=1}^{k} (n-n_{j_i})^{1/k}} \\ &= \frac{x^{(k)} z^{(k)}}{\prod_{i=1}^{k} \left[n_{j_i}(n-n_{j_i})\right]^{1/k}}. \end{aligned}$$

Similar to (10), these formulations are products of two geometric means. Similar to $S_{\text{Och}}^{(2)}$ and $S_{\text{S-S2}}^{(2)}$, we have $1 \geq S_{\text{Och}}^{(k)} \geq S_{\text{S-S2}}^{(k)} \geq 0$.

3.4 The Product

The product of Dice's association

$$S_{\text{Sorg}}^{(2)} = \frac{a^2}{(a+b)(a+c)} = \frac{\left[x^{(2)}\right]^2}{n_{j_1} n_{j_2}} \tag{11}$$

is also called the correlation ratio. Cheetham and Hazel (1969, p. 1131) report Sorgenfrei (1959) as one of the first to use this SC. Straightforward 3-adic and k-adic formulations of the SC in (11) are

$$S^{(3)}_{\text{Sorg}} = \frac{\left[x^{(3)}\right]^3}{n_{j_1}\, n_{j_2}\, n_{j_3}} \quad \text{and} \quad S^{(k)}_{\text{Sorg}} = \frac{\left[x^{(k)}\right]^k}{\prod_{i=1}^k n_{j_i}}.$$

In words, $S^{(3)}_{\text{Sorg}}$ ($S^{(k)}_{\text{Sorg}}$) is the product of 3 (k) association indices. Since we have $S^{(k)}_{\text{Och}} = \sqrt[k]{S^{(k)}_{\text{Sorg}}}$, the two coefficients are GOE.

A SC by McConnaughey (1964) extends the SC in (11) by subtracting the 2-adic mismatches in the numerator. The SC is given by

$$S^{(2)}_{\text{McC}} = \frac{a^2 - b\,c}{(a+b)(a+c)} = \frac{x^{(2)}(n_{j_1} + n_{j_2}) - n_{j_1} n_{j_2}}{n_{j_1} n_{j_2}}. \tag{12}$$

Possible 3-adic and k-adic formulations of the SC in (12) are respectively

$$S^{(3)}_{\text{McC}} = \frac{\frac{2}{3}\left[x^{(3)}\right]^2 (n_{j_1} + n_{j_2} + n_{j_3}) - n_{j_1} n_{j_2} n_{j_3}}{n_{j_1} n_{j_2} n_{j_3}}$$

and

$$S^{(k)}_{\text{McC}} = \frac{\frac{2}{k}\left[x^{(k)}\right]^{k-1} \left(\sum_{i=1}^k n_{j_i}\right) - \prod_{i=1}^k n_{j_i}}{\prod_{i=1}^k n_{j_i}}.$$

Similar to $S^{(2)}_{\text{McC}}$, it holds that $1 \geq S^{(k)}_{\text{McC}} \geq -1$.

Denote by

$$q^{(k)} = \text{the number of attributes present in objects } h_1, h_2, ..., h_k.$$

For an arbitrary ordinal comparison with respect to $S^{(k)}_{\text{McC}}$, we have

$$\frac{\frac{2}{k}\left[x^{(k)}\right]^{k-1} \sum_{i=1}^k n_{j_i} - \prod_{i=1}^k n_{j_i}}{\prod_{i=1}^k n_{j_i}} > \frac{\frac{2}{k}\left[q^{(k)}\right]^{k-1} \sum_{i=1}^k n_{h_i} - \prod_{i=1}^k n_{h_i}}{\prod_{i=1}^k n_{h_i}}$$

$$\text{iff} \quad \frac{\left[x^{(k)}\right]^{k-1} \sum_{i=1}^k n_{j_i}}{\prod_{i=1}^k n_{j_i}} > \frac{\left[q^{(k)}\right]^{k-1} \sum_{i=1}^k n_{h_i}}{\prod_{i=1}^k n_{h_i}}. \tag{13}$$

For an arbitrary ordinal comparison with respect to $S^{(k)*}_{\text{Kul}}$ from Section 3.2, we also obtain (13), which implies that $S^{(k)}_{\text{McC}}$ and $S^{(k)*}_{\text{Kul}}$ are GOE.

3.5 The Minimum/Maximum

Suppose that one is not interested in both of Dice's (1945) 2-adic association indices. Instead, one may only be interested in the SC that reflects the amount of similarity between species j_1 and j_2, relative to the most abundant species. On the other hand, one may also be interested in the SC that reflects the amount of similarity between object j_1 and j_2, relative to the object that occurs the least. In the former case, one obtains

$$\begin{aligned} S_{\text{BB}}^{(2)} &= \min\left(\frac{a}{a+b}, \frac{a}{a+c}\right) = \frac{a}{\max(a+b, a+c)} \\ &= \frac{x^{(2)}}{\max(n_{j_1}, n_{j_2})}, \end{aligned} \tag{14}$$

which is a SC considered in Braun-Blanquet (1932). Straightforward 3-adic and k-adic formulations of (14) are respectively

$$S_{\text{BB}}^{(3)} = \min\left(\frac{x^{(3)}}{n_{j_1}}, \frac{x^{(3)}}{n_{j_2}}, \frac{x^{(3)}}{n_{j_3}}\right) = \frac{x^{(3)}}{\max(n_{j_1}, n_{j_2}, n_{j_3})}$$

and

$$S_{\text{BB}}^{(k)} = \frac{x^{(k)}}{\max(n_{j_1}, n_{j_2}, ..., n_{j_k})}.$$

In the latter case, we may use

$$\begin{aligned} S_{\text{Sim}}^{(2)} &= \max\left(\frac{a}{a+b}, \frac{a}{a+c}\right) = \frac{a}{\min(a+b, a+c)} \\ &= \frac{x^{(2)}}{\min(n_{j_1}, n_{j_2})}, \end{aligned} \tag{15}$$

which is a SC described in Simpson (1943). Straightforward 3-adic and k-adic formulations of (15) are respectively

$$S_{\text{Sim}}^{(3)} = \frac{x^{(3)}}{\min(n_{j_1}, n_{j_2}, n_{j_3})} \quad \text{and} \quad S_{\text{Sim}}^{(k)} = \frac{x^{(k)}}{\min(n_{j_1}, n_{j_2}, ..., n_{j_k})}.$$

Clearly, similar to $S_{\text{BB}}^{(2)}$ and $S_{\text{Sim}}^{(2)}$, it holds that $1 \geq S_{\text{Sim}}^{(k)} \geq S_{\text{BB}}^{(k)} \geq 0$.

4. Discussion

As pointed out by Gower and Legendre (1986, p. 31) for 2-adic SCs, a SC has to be considered in the context of the descriptive statistical analysis

of which it is a part. Furthermore, the choice of a SC is strongly influenced by the nature of the data and the intended type of analysis. Clearly, the same arguments apply for the k-adic generalizations of various 2-adic SCs for binary (presence/absence) data that were presented in this paper. Cox, Cox and Branco (1991) pointed out that k-adic SCs, for example, 3-adic or 4-adic SCs instead of 2-adic SCs, can be used to detect possible higher-order relations between the objects. A similar argument was made by Daws (1996) in the context of free-sorting data. Daws showed convincingly that an analysis that uses 3-adic information may be more informative than an analysis based on 2-adic information only.

Consider the data matrix for five binary strings on fourteen attributes in Table 3. For these data it can be verified that the ten 2-adic Jaccard SCs between the five objects are all equal ($S_{\text{Jac}}^{(2)} = 3/11$), and that the ten 3-adic Jaccard SCs are all equal ($S_{\text{Jac}}^{(3)} = 1/13$), giving no discriminative information about the five objects. However, the 4-adic Jaccard SC between objects 2, 3, 4 and 5 ($S_{\text{Jac}}^{(4)} = 1/13$) differs from the other four 4-adic Jaccard SCs ($S_{\text{Jac}}^{(4)} = 0$). This artificial example shows that higher-order information can put objects 2, 3, 4 and 5 in a group separated from object 1. Of course, one can also argue that the wrong 2-adic SC is specified.

Two major classes of k-adic SCs were distinguished in this paper. The first class is referred to as Bennani-Heiser SCs, which contains all SCs that can be defined using only the positive matches $x^{(k)}$, the negative matches $z^{(k)}$ and the total number of attributes n. Many BHSCs are fractions that are linear in both numerator and denominator. As it turned out, a second class was formed by SCs that could be formulated as functions of association indices first presented in Dice (1945). These functions included the Pythagorean means (harmonic, arithmetic and geometric means). New coefficients in the second class can be created by considering other type of means, like the Heronian mean and the root mean square (see, for example, Mays, 1983). The Heronian mean of

$$\frac{a}{a+b} \text{ and } \frac{a}{a+c} \text{ is given by } \frac{1}{3}\left(\frac{a}{a+b} + \frac{a}{\sqrt{(a+b)(a+c)}} + \frac{a}{a+c}\right),$$

whereas the root mean square equals

$$\sqrt{\frac{1}{2}\left(\frac{a}{a+b}\right)^2 + \frac{1}{2}\left(\frac{a}{a+c}\right)^2}.$$

New coefficients can also be created by including the quantities $d/(b+d)$ and $d/(c+d)$. For example, the function

Table 3. Hypothetical binary scores of five objects on fourteen attributes.

objects	attributes													
1	1	1	1	1	1	1	0	0	0	0	0	0	0	1
2	1	1	1	0	0	0	1	1	1	1	0	0	0	0
3	1	0	0	1	1	0	1	1	0	0	1	1	0	0
4	0	1	0	0	1	1	1	0	1	0	1	0	1	0
5	0	0	1	1	0	1	1	0	0	1	0	1	1	0

$$\frac{4ad}{4ad + (a+d)(b+c)} \quad \text{is the harmonic mean of}$$

$$\frac{a}{a+b}, \frac{a}{a+c}, \frac{d}{b+d} \quad \text{and} \quad \frac{d}{c+d}.$$

The reader may have noted that we have failed to present k-adic versions of SCs that involve the covariance $(ad - bc)$ between a pair of objects, for example, the phi coefficient or Cohen's kappa, given by respectively

$$S^{(2)}_{\text{Phi}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

and

$$S^{(2)}_{\text{Cohen}} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}.$$

The definition of covariance between triples of objects is already quite complex and the topic is outside the scope of the present study. We also have not considered k-adic versions of the odds ratio ad/bc or coefficients that are transformations of ad/bc to a $[-1, 1]$ scale, for example,

$$S^{(2)}_{\text{Yule}} = \frac{\frac{ad}{bc} - 1}{\frac{ad}{bc} + 1} = \frac{ad - bc}{ad + bc}.$$

A completely different way of formulating k-adic SCs for binary data, including a k-adic generalization of $S^{(2)}_{\text{Cohen}}$, can be found in Warrens (2008b). The SCs in that paper are studied in the context of correction for chance.

We end this paper with the following problem. Two k-adic formulations of the Dice's coincidence index S_{Dice} were considered in this paper, namely

$$S^{(k)}_{\text{Dice}} = \frac{2x^{(k)}}{2x^{(k)} + y^{(k)}} \quad \text{and} \quad S^{(k)*}_{\text{Dice}} = \frac{kx^{(k)}}{\sum_{i=1}^{k} n_{j_i}}.$$

The first, $S^{(k)}_{\text{Dice}}$, is a BHSC (Section 2) and belongs to the parameter family $S^{(k)}_{\text{F-G}}(\theta) = x^{(k)}/(x^{(k)}+\theta y^{(k)})$. All the members of this parameter family are GOE and the following question arises: are there any SCs that are GOE with respect to $S^{(k)*}_{\text{Dice}}$? Instead of the BHSC-formulation, let the 2-adic version of the Jaccard coefficient be written in the notation of Section 3, that is,

$$S^{(2)}_{\text{Jac}} = \frac{x^{(2)}}{n_{j_1} + n_{j_2} - x^{(2)}}. \tag{16}$$

Ignoring the interpretation of the SC in (1), up to three possible k-adic versions of (16), that use similar generalizations compared to $S^{(k)*}_{\text{Dice}}$, can be found:

$$S^{(k)*}_{\text{Jac}} = \frac{(k-1)\, x^{(k)}}{\sum_{i=1}^{k} n_{j_i} - x^{(k)}}, \quad S^{(k)**}_{\text{Jac}} = \frac{x^{(k)}}{\frac{2}{k}\sum_{i=1}^{k} n_{j_i} - x^{(k)}}$$

$$\text{and} \quad S^{(k)***}_{\text{Jac}} = \frac{x^{(k)}}{\sum_{i=1}^{k} n_{j_i} - (k-1)\, x^{(k)}}.$$

Similar to $S^{(2)}_{\text{Jac}}$, we have $1 \geq S^{(k)*}_{\text{Jac}},\ S^{(k)**}_{\text{Jac}},\ S^{(k)***}_{\text{Jac}} \geq 0$, but neither of them can be interpreted as a k-adic formulation in terms of the SC in (1). However, for an arbitrary ordinal comparison with respect to $S^{(k)*}_{\text{Dice}}$, we have

$$\frac{k\, x^{(k)}}{\sum_{i=1}^{k} n_{j_i}} > \frac{k\, q^{(k)}}{\sum_{i=1}^{k} n_{h_i}} \quad \text{iff} \quad \frac{x^{(k)}}{\sum_{i=1}^{k} n_{j_i}} > \frac{q^{(k)}}{\sum_{i=1}^{k} n_{h_i}}. \tag{17}$$

For an arbitrary ordinal comparison with respect to either $S^{(k)*}_{\text{Jac}}$, $S^{(k)**}_{\text{Jac}}$ or $S^{(k)***}_{\text{Jac}}$, we also obtain (17), which implies that all four SCs are GOE. Thus, multiple k-adic SCs can be presented that are GOE to $S^{(k)*}_{\text{Dice}}$, but no SC has the clear interpretation that holds for the class of BHSCs.

References

ALBATINEH, A.N., NIEWIADOMSKA-BUGAJ, M., and MIHALKO, D. (2006), "On Similarity Indices and Correction for Chance Agreement," *Journal of Classification, 23,* 301–313.

BARONI-URBANI, C. and BUSER, M.W. (1976), "Similarity of Binary Data," *Systematic Zoology, 25,* 251–259.

BATAGELJ, V. and BREN, M. (1995), "Comparing Resemblance Measures," *Journal of Classification, 12,* 73–90.

BAULIEU, F.B. (1989), "A Classification of Presence/Absence Based Dissimilarity Coefficients," *Journal of Classification, 6,* 233–246.

BENNANI-DOSSE, M. (1993), *Analyses Métriques á Trois Voies*, Ph.D. Dissertation, Université de Haute Bretagne Rennes II, France.

BRAUN-BLANQUET, J. (1932), *Plant Sociology: The Study of Plant Communities*, Authorized English translation of Pflanzensoziologie, New York: McGraw-Hill.

BRAY, J.R. (1956), "A Study of Mutual Occurrence of Plant Species," *Ecology, 37,* 21–28.

CHEETHAM, A.H. and HAZEL, J.E. (1969), "Binary (Presence-Absence) Similarity Coefficients," *Journal of Paleontology, 43,* 1130–1136.

COX, T.F., COX, M.A.A., and BRANCO, J.A. (1991), "Multidimensional Scaling of n-Tuples," *British Journal of Mathematical and Statistical Psychology, 44,* 195–206.

CZEKANOWSKI, J. (1932), "Coefficient of Racial Likeliness und Durchschnittliche Differenz," *Anthropologischer Anzeiger, 9,* 227–249.

DAWS, J.T. (1996), "The Analysis of Free-sorting Data: Beyond Pairwise Comparison," *Journal of Classification, 13,* 57–80.

DE ROOIJ, M. and GOWER, J.C. (2003), "The Geometry of Triadic Distances," *Journal of Classification, 20,* 181–220.

DICE, L.R. (1945), "Measures of the Amount of Ecologic Association Between Species", *Ecology, 26,* 297–302.

FICHET, B. (1986), "Distances and Euclidean Distances for Presence-Absence Characters and Their Application to Factor Analysis," in *Multidimensional Data Analysis*, Eds. J. de Leeuw, W.J. Heiser, J.J. Meulman and F. Critchley, Leiden: DSWO Press, 23–46.

FOWLKES, E.B. and MALLOWS, C.L. (1983), "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association, 78,* 553–569.

GLEASON, H.A. (1920), "Some Applications of the Quadrat Method," *Bulletin of the Torrey Botanical Club, 47,* 21–33.

GOWER, J.C. (1986), "Euclidean Distance Matrices," in *Multidimensional Data Analysis*, Eds. J. de Leeuw, W.J. Heiser, J.J. Meulman and F. Critchley, Leiden: DSWO Press, 11–22.

GOWER, J.C. and LEGENDRE, P. (1986), "Metric and Euclidean Properties of Dissimilarity Coefficients," *Journal of Classification, 3,* 5–48.

GOWER, J.C. and HAND, D.J. (1996), *Biplots*, London: Chapman and Hall.

HAMANN, U. (1961), "Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen," *Willdenowia, 2,* 639–768.

HEISER, W.J. and BENNANI, M. (1997), "Triadic Distance Models: Axiomatization and Least Squares Representation," *Journal of Mathematical Psychology, 41,* 189–206.

HOLLEY, J.W. and GUILFORD, J.P. (1964), "A Note on the G Index of Agreement," *Educational and Psychological Measurement, 24,* 749–753.

HUBÁLEK, Z. (1982), "Coefficients of Association and Similarity Based on Binary (Presence-Absence) Data: An Evaluation," *Biological Reviews, 57,* 669–689.

HUBERT, L.J. (1977), "Nominal Scale Response Agreement as a Generalized Correlation," *British Journal of Mathematical and Statistical Psychology, 30,* 98–103.

HUBERT, L.J. and ARABIE, P. (1985), "Comparing Partitions," *Journal of Classification, 2,* 193–218.

JACCARD, P. (1912), "The Distribution of the Flora in the Alpine Zone," *The New Phytologist, 11,* 37-5-0.

JANSON, S. and VEGELIUS, J. (1981), "Measures of Ecological Association," *Oecologia, 49,* 371–376.

JOLY, S. and LE CALVÉ, G. (1995), "Three-way Distances," *Journal of Classification, 12,* 191–205.

KULCZYŃSKI, S. (1927), "Die Pflanzenassociationen der Pienenen," *Bulletin International de L'Académie Polonaise des Sciences et des Letters, classe des sciences mathematiques et naturelles, Serie B, Supplément II, 2,* 57–203.

MAYS, M.E. (1983), "Functions Which Parametrize Means," *The American Mathematical Monthly, 90,* 677–683.

MCCONNAUGHEY, B.H. (1964), "The Determination and Analysis of Plankton Communities," *Marine Research, Special No, Indonesia,* 1–40.

NEI, M. and LI, W.-H. (1979), "Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases," *Proceedings of the National Academy of Sciences of the United States of America, 76,* 5269–5273.

OCHIAI, A. (1957), "Zoogeographic Studies on the Soleoid Fishes Found in Japan and Its Neighboring Regions," *Bulletin of the Japanese Society for Fish Science, 22,* 526–530.

RAND, W. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association, 66,* 846–850.

ROGERS, D.J. and TANIMOTO, T.T. (1960), "A Computer Program for Classifying Plants," *Science, 132,* 1115–1118.

RUSSEL, P.F. and RAO, T.R. (1940), "On Habitat and Association of Species of Anopheline Larvae in South-Eastern Madras," *Journal of Malaria Institute India, 3,* 153–178.

SIBSON, R. (1972), "Order Invariant Methods for Data Analysis," *Journal of the Royal Statistical Society, Series B, 34,* 311–349.

SIMPSON, G.G. (1943), "Mammals and the Nature of Continents," *American Journal of Science, 241,* 1–31.

SOKAL, R.R. and MICHENER, C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships", *University of Kansas Science Bulletin, 38,* 1409–1438.

SOKAL, R.R. and SNEATH, R.H. (1963), *Principles of Numerical Taxonomy*, San Francisco: W. H. Freeman and Company.

SØRENSON, T. (1948), "A Method of Stabilizing Groups of Equivalent Amplitude in Plant Sociology Based on the Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons," *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter, 5,* 1–34.

SORGENFREI, T. (1958), *Molluscan Assemblages from the Marine Middle Miocene of South Jutland and Their Environments*, Copenhagen: Reitzel.

WALLACE, D.L. (1983), "A Method for Comparing Two Hierarchical Clusterings: Comment," *Journal of the American Statistical Association, 78,* 569–576.

WARRENS, M.J. (2008a), "On the Indeterminacy of Resemblance Measures for Binary (Presence/Absence) Data," *Journal of Classification, 25,* 125–136.

WARRENS, M.J. (2008b), "On Similarity Coefficients for 2×2 Tables and Correction for Chance," *Psychometrika, 73*, 487-502.