Journal of Classification 25:125-136 (2008)
DOI: 10.1007/s00357-008-9006-8

On the Indeterminacy of Resemblance Measures for Binary (Presence/Absence) Data

Matthijs J. Warrens

Leiden University, The Netherlands

Abstract: Many similarity coefficients for binary data are defined as fractions. For certain resemblance measures the denominator may become zero. If the denominator is zero the value of the coefficient is indeterminate. It is shown that the seriousness of the indeterminacy problem differs with the resemblance measures. Following Batagelj and Bren (1995) we remove the indeterminacies by defining appropriate values in critical cases.

Keywords: Association coefficients; Indeterminate values; Critical cases.

1. Introduction

Association coefficients are measures that reflect in some way the similarity or agreement of two sequences. Resemblance measures for two binary sequences i and j can be found in Gower and Legendre (1986, Section 4.1), Baulieu (1989), and Batagelj and Bren (1995, Section 4) (see also the appendix). These so-called presence/absence coefficients are usually defined using the four dependent quantities

a = the number of attributes present in both i and j
b = the number of attributes present in i but absent j
c = the number of attributes absent in i but present in j
d = the number of attributes absent in both i and j.

The author would like to thank 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

Let $m = a + b + c + d$ denote the total number of attributes. A presence/absence coefficient $S(a, b, c, d)$ or S is defined to be a map $S : (\mathbb{Z}^+)^4 \to \mathbb{R}$ from the set, U, of all ordered quadruples of nonnegative integers into the reals (Baulieu 1989). Following Sokal and Sneath (1963), the convention is adopted of calling a coefficient S_{Name} by its originator or the first we know to propose it. An example is

$$S_{\text{Jac}} = \frac{a}{a+b+c} \qquad \text{(Jaccard 1912).}$$

Presence/absence can be either a nominal or an ordinal variable. In the latter case presence is 'more' in a sense than absence. Sokal and Sneath (1963) (among others) make a distinction between coefficients that do or do not include the quantity d. If a binary sequence is a coding of the presence or absence of a list of attributes or features, then d (usually) reflects the number of negative matches. In the field of numerical taxonomy quantity d is generally felt not to contribute to similarity. Measures that do not include the quantity d are coefficient S_{Jac} and

$$S_{\text{Kul1}} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \qquad \text{(Kulczyński 1927).}$$

If the data are nominal, coefficients for which the quantities a and d are equally weighted are appropriate. Examples are

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d} \qquad \text{(Sokal and Michener 1958)}$$

and $$S_{\text{Yule2}} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \qquad \text{(Yule 1912).}$$

Some coefficients do include both a and d but do not equally weight the two quantities. Examples are

$$S_{\text{RR}} = \frac{a}{a+b+c+d} \qquad \text{(Russel and Rao 1940)}$$

and $$S_{\text{BUB}} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}} \qquad \text{(Baroni-Urbani and Buser 1976).}$$

Measure S_{RR} is called a hybrid coefficient in Sokal and Sneath (1963). Measure S_{RR} and S_{BUB} may be used with ordinal data. For thirty coefficients, that are considered in this paper, the formulas are presented in the appendix.

Since many coefficients are defined as fractions, the denominator may become zero for certain quadruples. For example, it is well-known that if $d = m$ then the value of the Jaccard coefficient S_{Jac} is indeterminate. As

noted by Batagelj and Bren (1995, Section 4.2), this case of indeterminacy for some values of presence/absence coefficients has been given surprisingly little attention. The critical case of coefficient S_{Jac} implies a situation in which objects i and j possess none of the attributes. One may argue that it is highly unlikely that this occurs in practice. For example, in ecology it is unlikely to have an ordinal data table that has sites without species. Furthermore, the problem can be resolved by excluding zero vectors from the data. Although these may be valid arguments for S_{Jac}, it turns out that the number of cases in which the value of a coefficient is indeterminate, differs with the coefficients.

2. Critical Cases

Consider the set U of all ordered quadruples (a, b, c, d) of nonnegative integers. Since $a + b + c + d = m$, the number of different quadruples for given $m(m \geq 1)$ is given by the binomial coefficient

$$\binom{m+3}{3} = \frac{(m+3)!}{m!\,3!} = \frac{(m+3)(m+2)(m+1)}{6}.$$

Thus, for $m = 1, 2, 3, 4, 5, ..., U$ consists of 4, 10, 20, 35, 56, ... different quadruples. For each coefficient we may study for how many quadruples, given fixed m, the value of the coefficient is indeterminate.

For thirty similarity coefficients for both nominal and ordinal data, Table 1 presents the number of quadruples in U for which the denominator of the corresponding coefficient equals zero. For example, if $m = 5$, U has 56 elements and for 20 of these quadruples the value of the phi coefficient S_{Yule2} is indeterminate. The formulas of the coefficients in Table 1 can be found in the appendix. Note that in Table 1 the coefficients are placed in groups with the same number of critical cases. For coefficients with the most critical cases $(4m)$, the number of quadruples for which the value of the coefficient is indeterminate increases in a linear fashion as m becomes larger. Increases of the number of quadruples with the indeterminacy problem are not proportional to increases of m. Hence, the ratio

$$\frac{\text{number of critical cases in } U}{\text{total number of quadruples in } U} \qquad \text{decreases as } m \text{ becomes larger.}$$

Furthermore, for most coefficients indeterminacy only occurs in the case that at least two elements of quadruple (a, b, c, d) are zero.

3. Defining Appropriate Values

Instead of excluding vectors that result in zero denominators values, Batagelj and Bren (1995) proposed to eliminate the indeterminacies

Table 1. Table of thirty resemblance measures for binary data. The definitions of the coefficients are presented in the appendix. The quantity in the third column is the number of quadruples in U for given m ($m \geq 1$) for which the value of the coefficient is indeterminate.

Ordinal data	Nominal data	# of 4-tuples
S_{RR}	$S_{\text{SM}}, S_{\text{SS3}}, S_{\text{Mich}}, S_{\text{RT}}, S_{\text{Ham}}$	0
$S_{\text{Jac}}, S_{\text{Dice}}, S_{\text{BUB}}, S_{\text{BB}}, S_{\text{SS1}}$		1
	$S_{\text{GK}}, S_{\text{Scott}}, S_{\text{Cohen}}, S_{\text{HD}}$	2
	S_{MP}	4
S_{Kul2}	S_{SS5}	$m+1$
$S_{\text{Kul1}}, S_{\text{Och}}, S_{\text{Sim}}, S_{\text{Sorg}}, S_{\text{McC}}$		$2m+1$
	$S_{\text{Yule1}}, S_{\text{Yule2}}, S_{\text{Yule3}}, S_{\text{SS2}}, S_{\text{SS4}}, S_{\text{Fleiss}}, S_{\text{Loe}}$	$4m$

by appropriately defining values in critical cases. Some of the definitions presented in this section give the same results as definitions proposed in Batagelj and Bren (1995). The definitions presented here simplify the reading.

3.1 Coefficients for Ordinal Data

Let

$$K_x = \frac{a}{a+x} \quad \text{with} \quad x = b, c.$$

Coefficients S_{Sorg}, S_{BB}, S_{Dice}, S_{Och}, S_{Kul1}, and S_{Sim} are, respectively, the product, minimum function, harmonic mean, geometric mean, arithmetic mean, and maximum function of K_b and K_c. Consider the arithmetic mean of K_b and K_c

$$S_{\text{Kul1}} = \frac{K_b + K_c}{2} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right).$$

Suppose $a + c = 0$. Note that the value of S_{Kul1} is indeterminate. If we set $K_c = 0$, then S_{Kul1} becomes

$$S_{\text{Kul1}} = \frac{1}{2}\left(\frac{a}{a+b} + 0\right) = 0 \quad \text{since} \quad a = 0.$$

Alternatively, we may remove the part from the definition of S_{Kul1} that causes the indeterminacy. Coefficient S_{Kul1} becomes

$$S_{\text{Kul1}} = \frac{a}{a+b} = 0 \quad \text{since} \quad a = 0.$$

Thus, either setting $K_c = 0$ or removing the indeterminate part from the definition of the coefficient, leads to the same conclusion: $S_{\text{Kul1}} = 0$. We therefore define

$$S_{\text{Kul1}} = \begin{cases} 0 & \text{if} \quad a+b=0 \quad \text{or} \quad a+c=0 \\ \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Och}, S_{Sim}, and S_{Sorg}. Coefficient

$$S_{\text{McC}} = \frac{a^2-bc}{(a+b)(a+c)} = 2S_{\text{Kul1}} - 1.$$

Suppose $a+c=0$. The value of coefficient S_{McC} is indeterminate. Also the numerator $(a^2-bc)=0$. We define

$$S_{\text{McC}} = \begin{cases} 0 & \text{if} \quad a+b=0 \quad \text{or} \quad a+c=0 \\ \frac{a^2-bc}{(a+b)(a+c)} & \text{otherwise.} \end{cases}$$

Consider the harmonic mean of K_b and K_c

$$S_{\text{Dice}} = \frac{2}{K_b^{-1}+K_c^{-1}} = \frac{2a}{2a+b+c}.$$

Suppose $a+c=0$. The value of K_c and K_c^{-1} is indeterminate. However, $2a/(2a+b+c)=0$. Similar to S_{Kul1} we define

$$S_{\text{Dice}} = \begin{cases} 0 & \text{if} \quad d=m \\ 2a/(2a+b+c) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Jac}, S_{SS1}, S_{BB}, and S_{BUB}.

Note that the definitions of S_{Kul1} and S_{Dice} presented here do not ensure that $S_{\text{Kul1}} = 1$ or $S_{\text{Dice}} = 1$ if i is compared with itself. If $i = j = \overbrace{(0,0,...,0)}^{m}$, that is, the two sequences have nothing in common, $S_{\text{Kul1}} = S_{\text{Dice}} = 0$. Furthermore, if $i = \overbrace{(0,0,...,0)}^{m}$ is compared with itself, $S_{\text{Kul1}} = S_{\text{Dice}} = 0$. Since these coefficients are appropriate for ordinal data, it is a moot point what the value of the coefficient should be if sequences i and j, or just sequence i if i is compared with itself, are zero vectors. From a philosophical point of view it might be better to leave the coefficients for ordinal data undefined for the critical case $d = m$.

To eliminate indeterminacies, coefficient S_{Kul2} may be defined as

$$S_{\text{Kul2}} = \begin{cases} \infty & \text{if} \quad b + c = 0,\ d < m \\ 0 & \text{if} \quad d = m \\ a/(b+c) & \text{otherwise.} \end{cases}$$

An analogous definition may be formulated for coefficient S_{SS5}.

3.2 Coefficients for Nominal Data

Consider coefficient

$$S_{\text{HD}} = \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right).$$

The value of S_{HD} is indeterminate if either $a = m$ or $d = m$. If $a = m$ then variables i and j are unit vectors; if $d = m$ then variables i and j are zero vectors. If both variables are zero vectors or unit vectors, we may speak of perfect agreement if i and j are nominal variables. We therefore define

$$S_{\text{HD}} = \begin{cases} 1 & \text{if} \quad a = m \quad \text{or} \quad d = m \\ \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right) & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{Cohen}, S_{GK} and S_{Scott}. We also define

$$S_{\text{MP}} = \begin{cases} 1 & \text{if} \quad a = m \quad \text{or} \quad d = m \\ 0 & \text{if} \quad b = m \quad \text{or} \quad c = m \\ \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)} & \text{otherwise.} \end{cases}$$

Consider the phi coefficient

$$S_{\text{Yule2}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

The value of S_{Yule2} is indeterminate if $a + b = 0$, $a + c = 0$, $b + d = 0$, or $c + d = 0$. For these critical cases the covariance $(ad - bc) = 0$. We define

$$S_{\text{Yule2}} = \begin{cases} 1 & \text{if} \quad a = m \quad \text{or} \quad d = m \\ 0 & \text{if} \quad a + b = 0, \quad a + c = 0, \quad b + d = 0 \quad \text{or} \quad c + d = 0 \\ \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} & \text{otherwise.} \end{cases}$$

Analogous definitions may be formulated for coefficients S_{SS4}, S_{Yule1}, S_{Yule3}, S_{Fleiss}, and S_{Loe}.

Let

$$K_x = \frac{a}{a+x} \quad \text{and} \quad K_x^* = \frac{d}{x+d} \quad \text{with} \quad x = b, c.$$

Consider the arithmetic mean of K_b, K_c, K_b^* and K_c^*

$$S_{\text{SS2}} = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right).$$

Suppose $c+d=0$. Note that the value of K_c^* is indeterminate. To eliminate the critical case, we may set $K_c^* = 0$, and S_{SS2} becomes

$$S_{\text{SS2}} = \frac{1}{4}\left(\frac{a}{a+b} + 1 + 0 + 0\right) = \frac{2a+b}{4(a+b)}. \tag{1}$$

Note that coefficient S_{SS2} in (1) has a range $\left[\frac{1}{4}, \frac{1}{2}\right]$. We may define

$$S_{\text{SS2}} = \begin{cases} \frac{2a+b}{4(a+b)} & \text{if} \quad c+d=0 \\ \frac{2a+c}{4(a+c)} & \text{if} \quad b+d=0 \\ \frac{b+2d}{4(b+d)} & \text{if} \quad a+c=0 \\ \frac{c+2d}{4(c+d)} & \text{if} \quad a+b=0 \\ \frac{1}{2} & \text{if} \quad a=m \quad \text{or} \quad d=m \\ 0 & \text{if} \quad b=m \quad \text{or} \quad c=m \\ \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) & \text{otherwise.} \end{cases}$$

As an alternative to the above robust definition of S_{SS2}, we propose to eliminate the critical case by removing the part from the definition of S_{SS2} that causes the indeterminacy. Suppose $c+d=0$. The arithmetic mean of K_b, K_c and K_b^* is given by

$$S_{\text{SS2}}^* = \frac{1}{3}\left(\frac{a}{a+b} + 0 + 1\right) = \frac{2a+b}{3(a+b)}. \tag{2}$$

Note that coefficient S_{SS2}^* in (2) has a range $\left[\frac{1}{3}, \frac{2}{3}\right]$. We define

$$S_{\text{SS2}}^* = \begin{cases} \frac{2a+b}{3(a+b)} & \text{if} \quad c+d=0 \\ \frac{2a+c}{3(a+c)} & \text{if} \quad b+d=0 \\ \frac{b+2d}{3(b+d)} & \text{if} \quad a+c=0 \\ \frac{c+2d}{3(c+d)} & \text{if} \quad a+b=0 \\ 1 & \text{if} \quad a=m \quad \text{or} \quad d=m \\ 0 & \text{if} \quad b=m \quad \text{or} \quad c=m \\ \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) & \text{otherwise.} \end{cases}$$

4. Discussion

Because many association coefficients for binary data are defined as fractions, the denominator may become zero for certain resemblance measures. Some similarity coefficients have more indeterminate cases than others. Following Batagelj and Bren (1995), the indeterminacies may be eliminated by appropriately defining values in critical cases. For instance, Batagelj and Bren (1995, p. 81) defined the Jaccard coefficient as

$$S_{\text{Jac}} = \begin{cases} 1 \quad \text{if} \quad d = m \\ a/(a+b+c) \qquad \text{otherwise.} \end{cases} \tag{3}$$

Definition (3) ensures that $S_{\text{Jac}} = 1$ if sequence i is compared with itself, that is, the coefficient matrix of all pairwise S_{Jac} has unit elements on the diagonal. Note that if sequences i and j have nothing in common, that is, $i = j = \overbrace{(0, 0, ..., 0)}^{m}$, then $S_{\text{Jac}} = 1$ using (3). Because S_{Jac} may be used for ordinal data, it is a moot point what the value of S_{Jac} should be when i and j are zero vectors.

The alternative definition of the Jaccard coefficient proposed here is given by

$$S^*_{\text{Jac}} = \begin{cases} 0 \quad \text{if} \quad d = m \\ a/(a+b+c) \qquad \text{otherwise.} \end{cases} \tag{4}$$

Similar definitions are proposed for other coefficients for ordinal data. To ensure that the coefficient matrix has unit elements on the main diagonal we may include in (4) the statement, $S^*_{\text{Jac}} = 1$ if sequence i is compared with itself.

Appendix

Measures for ordinal data:

Jaccard (1912):

$$S_{\text{Jac}} = \frac{a}{a+b+c}$$

Kulczyński (1927):

$$S_{\text{Kul1}} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right) \quad \text{and} \quad S_{\text{Kul2}} = \frac{a}{b+c}$$

Braun-Blanquet (1932):

$$S_{\text{BB}} = \frac{a}{a + \max(b, c)}$$

Russel and Rao (1940):

$$S_{\text{RR}} = \frac{a}{a+b+c+d}$$

Simpson (1943):

$$S_{\text{Sim}} = \frac{a}{a+\min(b,c)}$$

Dice (1945), Sørenson (1948):

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

Ochiai (1957):

$$S_{\text{Och}} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Sorgenfrei (1958):

$$S_{\text{Sorg}} = \frac{a^2}{(a+b)(a+c)}$$

Sokal and Sneath (1963):

$$S_{\text{SS1}} = \frac{a}{a+2(b+c)}$$

McConnaughey (1964):

$$S_{\text{McC}} = \frac{a^2-bc}{(a+b)(a+c)}$$

Baroni-Urbani and Buser (1976):

$$S_{\text{BUB}} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}.$$

Measures for nominal data:

Yule (1900):

$$S_{\text{Yule1}} = \frac{ad-bc}{ad+bc}$$

Yule (1912):

$$S_{\text{Yule2}} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad \text{and} \quad S_{\text{Yule3}} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$$

Michael (1920):

$$S_{\text{Mich}} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$$

Loevinger (1948):

$$S_{\text{Loe}} = \frac{ad - bc}{\min[(a+b)(b+d), (a+c)(c+d)]}$$

Goodman and Kruskal (1954):

$$S_{\text{GK}} = \frac{2\min(a,d) - b - c}{2\min(a,d) + b + c}$$

Scott (1955):

$$S_{\text{Scott}} = \frac{4(ad - bc) - (b-c)^2}{(2a+b+c)(b+c+2d)}$$

Sokal and Michener (1958):

$$S_{\text{SM}} = \frac{a+d}{a+b+c+d}$$

Rogers and Tanimoto (1960):

$$S_{\text{RT}} = \frac{a+d}{a+2(b+c)+d}$$

Cohen (1960):

$$S_{\text{Cohen}} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}$$

Hamann (1961):

$$S_{\text{Ham}} = \frac{a-b-c+d}{a+b+c+d}$$

Sokal and Sneath (1963):

$$\begin{aligned}
S_{\text{SS2}} &= \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right) \\
S_{\text{SS3}} &= \frac{2(a+d)}{2a+b+c+2d} \\
S_{\text{SS4}} &= \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\
S_{\text{SS5}} &= \frac{a+d}{b+c}.
\end{aligned}$$

Maxwell and Pilliner (1968):

$$S_{\text{MP}} = \frac{2(ad - bc)}{(a+b)(c+d) + (a+c)(b+d)}$$

Fleiss (1975):

$$S_{\text{Fleiss}} = \frac{(ad - bc)[(a+b)(b+d) + (a+c)(c+d)]}{2(a+b)(a+c)(b+d)(c+d)}$$

Hawkins and Dotson (1975):

$$S_{\text{HD}} = \frac{1}{2}\left(\frac{a}{a+b+c} + \frac{d}{b+c+d}\right).$$

References

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.

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

COHEN, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement, 20,* 37–46.

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

FLEISS, J.L. (1975), "Measuring Agreement between Two Judges on the Presence or Absence of a Trait," *Biometrics, 31,* 651–659.

GOODMAN, L.A. and KRUSKAL, W.H. (1954), "Measures of Association for Cross Classifications," *Journal of the American Statistical Association, 49,* 732–764.

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

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

HAWKINS, R.P. and DOTSON, V.A. (1968), "Reliability Scores That Delude: An Alice in Wonderland Trip Through Misleading Characteristics of Interobserver Agreement Scores in Interval Recording", in *Behavior Analysis: Areas of Research and Application*, eds. E. Ramp and G. Semb, Englewood Cliffs, N. J.: Prentice-Hall.

JACCARD, P. (1912), "The Distribution of the Flora in the Alpine Zone," *The New Phytologist, 11,* 37–50.

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.

LOEVINGER, J.A. (1948), "The Technique of Homogeneous Tests Compared with Some Aspects of Scale Analysis and Factor Analysis," *Psychological Bulletin, 45,* 507–530.

MAXWELL, A.E. and PILLINER, A. E. G. (1968), "Deriving Coefficients of Reliability and Agreement for Ratings," *British Journal of Mathematical and Statistical Psychology, 21,* 105-116.

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

MICHAEL, E.L. (1920), "Marine Ecology and the Coefficient of Association: A Plea in Behalf of Quantitative Biology," *The Journal of Ecology, 8,* 54-59.

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.

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.

SCOTT, W.A. (1955), "Reliability of Content Analysis: The Case of Nominal Scale Coding," *Public Opinion Quarterly, 19,* 321-325.

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.

YULE, G.U. (1900), "On the Association of Attributes in Statistics," *Philosophical Transactions of the Royal Society of London, 194,* 257–319.

YULE, G.U. (1912), "On the Methods of Measuring the Association between Two Attributes," *Journal of the Royal Statistical Society, 75,* 579–652.