Comparing Resemblance Measures*

Vladimir Batagelj
University of Ljubljana
Department of Mathematics
Jadranska 19, 61 111 Ljubljana
Slovenia

Matevž Bren
University of Maribor
FOV Kranj
Prešernova 11, 64 000 Kranj
Slovenia

August 23, 1993

Abstract

In the paper some types of equivalences over resemblance measures and some basic results about them are given. Based on induced partial orderings on the set of unordered pairs of units a dissimilarity between two resemblance measures over finite set of units can be defined. As an example, using this dissimilarity standard association coefficients between binary vectors are compared both theoretically and computationally.

Keywords: dissimilarity spaces, metric spaces, association coefficients, profile measures of resemblance.

AMS Subj. Class. (1991): 54 E, 62 H 30.

1 Introduction

In the first part of the paper we introduce some types of equivalences over resemblance measures and we present some general facts about them. A dissimilarity between two resemblance measures over finite set of units is defined. The rest of the paper is mainly devoted to applications of this dissimilarity for comparison of different association coefficients between binary vectors.

We believe that the notion of equivalence is a key to better understanding and organizing different resemblance measures encountered in applications. It also provides a framework to study the *invariance* and *stability* problems in data analysis: for which resemblance measures will a given algorithm produce the same or similar results?

*Extended version of the paper presented at DISTANCIA'92, June 22-26, 1992, Rennes, France.

2 Resemblance measures

Let $\mathcal{E}$ be a set of units (objects, OTUs, cases, individuals, ...). Quantitatively we describe the *resemblance* (association, similarity) between units by a function (*resemblance measure*)

$$r: (X, Y) \mapsto \mathbb{R}$$

which assigns to each pair of units $X, Y \in \mathcal{E}$ a real number. Several examples of resemblances for different types of units can be found in any book on data analysis and related topics (Sneath and Sokal 1973; Anderberg 1973; Lerman 1971; Späth 1977; Liebetrau 1983; Gower and Legendre 1986).

For r to be a resemblance, we require that it is *symmetric*:

P1. $\forall X, Y \in \mathcal{E} : \; r(X, Y) = r(Y, X)$

and that it has either the property:

P2.a $\forall X, Y \in \mathcal{E} : \; r(X, X) \leq r(X, Y)$,

or the property:

P2.b $\forall X, Y \in \mathcal{E} : \; r(X, X) \geq r(X, Y)$.

A resemblance which satisfies condition P2.a is called *forward* (straight) and denoted by d; it is called *backward* (reverse) and denoted by s if it satisfies condition P2.b.

In the set of unordered pairs of units

$$\mathcal{E}_2 = \{[X, Y] : X, Y \in \mathcal{E}\}, \quad [X, Y] = [Y, X],$$

a resemblance r induces the ordering $\ll_r$ in the following way:

$$[X, Y] \ll_r [U, V] \;\equiv\; r(X, Y) < r(U, V)$$

The unordered pair $[X, Y]$ is in relation $\ll_r$ with unordered pair $[U, V]$ whenever X and Y are closer (with respect to resemblance r) to each other than U and V.

The relation $\ll_r$ is a strict partial order. On the basis of this ordering we can define the notion of equivalent resemblances. Resemblances r and s are *(order) equivalent*, $r \cong s$, iff: $\ll_r = \ll_s$ or $\ll_r = \ll_s^{-1}$. It is easy to verify that $\cong$ is an equivalence relation. Also:

THEOREM 1 *Let $f: r(\mathcal{E} \times \mathcal{E}) \to \mathbb{R}$ be a strictly increasing/decreasing function and r a resemblance. Then*

$$s(X, Y) = f(r(X, Y)) \qquad \text{for all } X, Y \in \mathcal{E}$$

is also a resemblance and $s \cong r$.

And conversely: *Let r, s be resemblances and $r \cong s$. Then the function f: $r(\mathcal{E} \times \mathcal{E}) \to \mathbb{R}$, which is defined by*

$$f(t) = s(X, Y), \qquad \text{for } t = r(X, Y)$$

is well-defined, strictly increasing/decreasing and $s(X, Y) = f(r(X, Y))$ holds.

Proof: The first part of the theorem is trivial, so let us prove only the second part. Let r and s be resemblances and $r \cong s$. From the definition of order equivalence we get

$$\forall X, Y, U, V \in \mathcal{E} : (r(X,Y) = r(U,V) \ \Leftrightarrow\ s(X,Y) = s(U,V)).$$

Therefore, since

$$r(X,Y) = r(U,V) = t \ \Rightarrow\ s(X,Y) = f(r(X,Y)) = f(t) = f(r(U,V)) = s(U,V)$$

the function f: $r(\mathcal{E} \times \mathcal{E}) \to \mathbb{R}$ is well-defined by

$$f(t) = s(X,Y), \qquad \text{for} \quad t = r(X,Y)$$

To prove the strict monotonicity of f, let us choose any two real numbers $t, w \in r(\mathcal{E} \times \mathcal{E})$. Then there exist $X, Y, U, V \in \mathcal{E}$ such that $t = r(X,Y)$ and $w = r(U,V)$. Suppose that r and s are of the same type. Then we have

$$t < w \ \Leftrightarrow\ r(X,Y) < r(U,V) \ \Rightarrow\ s(X,Y) < s(U,V) \ \Leftrightarrow\ f(t) < f(w).$$

Function f is strictly increasing. In the same way we can see that in the case when r and s are of different type the function f is strictly decreasing. $\square$

An important consequence of this theorem is that every backward resemblance measure s can always be transformed by $d(X,Y) = -s(X,Y)$ into an order equivalent forward resemblance measure d. Therefore in the following we can limit our discussion to forward resemblances.

Other types of equivalences can also be defined on $\mathcal{E}_2$:

Resemblances r and s are *weakly equivalent* , $r \simeq s$, iff

$$\forall X, Y, U, V \in \mathcal{E} : (r(X,Y) = r(U,V) \ \Leftrightarrow\ s(X,Y) = s(U,V)).$$

It is easy to verify that $\simeq$ is also an equivalence relation and $\cong \subset \simeq$.

For a given resemblance r and $0 < \varepsilon \in \mathbb{R}$ we can define an open ball

$$K_r(X, \varepsilon) = \{Y \in \mathcal{E} : |r(X,Y) - r(X,X)| < \varepsilon\}$$

Using it, we can introduce some types of refinement relations $\preceq$:

topological	$r \preceq_t s \ \equiv\ \forall X \in \mathcal{E}\ \forall \varepsilon \in \mathbb{R}^+\ \exists \delta \in \mathbb{R}^+ : (K_r(X,\delta) \subseteq K_s(X,\varepsilon))$
uniform topological	$r \preceq_u s \ \equiv\ \forall \varepsilon \in \mathbb{R}^+\ \exists \delta \in \mathbb{R}^+\ \forall X \in \mathcal{E} : (K_r(X,\delta) \subseteq K_s(X,\varepsilon))$

For each type of refinement we can define a corresponding type of equivalence: Resemblances r and s are *(uniformly) topologically equivalent*, $r \sim s$, iff $(r \preceq s) \wedge (s \preceq r)$. It holds $\sim_u \subset \sim_t$.

3 Dissimilarities

Forward resemblances usually have the property:

P3.a $\quad \exists r^* \in \mathbb{R}\ \forall X \in \mathcal{E} :\ r(X,X) = r^*$.

In this case we can define a new resemblance d: $d(X,Y) = r(X,Y) - r^*$ which is order equivalent to r and has the properties:

R1. $\forall\, X, Y \in \mathcal{E} : \quad d(X, Y) \geq 0;$
R2. $\forall\, X \in \mathcal{E} : \quad d(X, X) = 0;$
R3. $\forall\, X, Y \in \mathcal{E} : \quad d(X, Y) = d(Y, X).$

A resemblance d satisfying properties R1, R2 and R3 is called a *dissimilarity*. Many data analysis algorithms deal with dissimilarities.

For some dissimilarities, additional properties hold:

R4. *evenness*: $d(X, Y) = 0 \;\Rightarrow \forall\, Z : \; d(X, Z) = d(Y, Z);$
R5. *definiteness*: $d(X, Y) = 0 \;\Rightarrow\; X = Y;$
R6. *triangle inequality*: $d(X, Y) \leq d(X, Z) + d(Z, Y);$
R7. *ultrametric inequality*: $d(X, Y) \leq \max(d(X, Z), d(Z, Y));$
R8. *Buneman's inequality* or *four-points condition*:
$d(X, Y) + d(U, V) \leq \max(d(X, U) + d(Y, V), d(X, V) + d(Y, U));$
R9. *translation invariance*: Let $(\mathcal{E}, +)$ be a group
$d(X, Y) = d(X + Z, Y + Z).$

These properties are related in the following way: R7 $\Rightarrow$ R6 $\Rightarrow$ R4 $\Leftarrow$ R5 and R8 $\Rightarrow$ R6. Dissimilarity d which has also the properties R5 and R6 is called a *distance*. Monotone hierarchical clustering algorithms transform dissimilarities into ultrametric dissimilarities. Dissimilarities satisfying Buneman's inequality are *tree distances* – distances between units are the shortest path lengths in some tree (Batagelj, Pisanski and Simões-Pereira 1990; Bandelt 1990).

When the space of units $\mathcal{E}$ is finite we can define a dissimilarity between resemblances r and s as follows (Lerman 1971):

$$D(r, s) = \begin{cases} \frac{1}{|\mathcal{E}_2|^2} \mid \ll_r \oplus \ll_s \mid & r \text{ and } s \text{ are both forward or both backward;} \\ \frac{1}{|\mathcal{E}_2|^2} \mid \ll_r \oplus \ll_s^{-1} \mid & \text{otherwise;} \end{cases}$$

where $\oplus$ denotes the symmetric difference of sets $A \oplus B \equiv (A \cup B) \setminus (A \cap B)$. Therefore the dissimilarity $D(r, s)$ equals to the number of pairs of pairs that are ordered differently by r and s, normalized by the total number of pairs of pairs.

Resemblance D thus defined has properties P2a, R1, R2 and R3; therefore D is a dissimilarity. D has also properties R4, R6 and:

$$D(r, s) = 0 \;\Leftrightarrow\; r \cong s.$$

Therefore D is a distance over order equivalence classes set of resemblances.

Dissimilarities usually take values in the interval $[0, 1]$ or in the interval $[0, \infty]$. They can be transformed one into the other by mappings:

$$\frac{d}{1 - d} : [0, 1] \to [0, \infty]$$

and

$$\frac{d}{1 + d} : [0, \infty] \to [0, 1],$$

or in the case $d_{max} < \infty$ by

$$\frac{d}{d_{max}} : [0, d_{max}] \to [0, 1].$$

To transform distance into distance we often use the mappings:

$$\log(1 + d), \quad \min(1, d) \quad \text{and} \quad d^r, \ 0 < r < 1.$$

Not all resemblances are dissimilarities. For example, the correlation coefficient has the interval $[1, -1]$ as its range. We can transform it to the interval $[0, 1]$ by mappings:

$$\tfrac{1}{2}(1 - d), \quad \sqrt{1 - d^2}, \quad 1 - |d|, \ \ldots$$

When applying these transformations to a measure d we wish that the nice properties of d were preserved. In this respect the following theorems should be mentioned:

PROPOSITION 2 *Let d be a dissimilarity on $\mathcal{E}$ and let a mapping $f\colon d(\mathcal{E} \times \mathcal{E}) \to \mathbb{R}_0^+$ has the property $f(0) = 0$, then $d'(X, Y) = f(d(X, Y))$ is also a dissimilarity. If f is also injective then $d' \simeq d$.*

PROPOSITION 3 *Let d be a distance on $\mathcal{E}$ and let the mapping $f\colon d(\mathcal{E} \times \mathcal{E}) \to \mathbb{R}$ has the properties:*

(a) $f(x) = 0 \Leftrightarrow x = 0$,
(b) $x < y \Rightarrow f(x) < f(y)$,
(c) $f(x + y) \leq f(x) + f(y)$,

then $d'(X, Y) = f(d(X, Y))$ is also a distance and $d' \cong d$.

It is easy to verify that all concave functions have also the sub-additivity property (c).

The following concave functions satisfy the last theorem:

(a) $f(x) = \alpha x, \ \alpha > 0$, (b) $f(x) = \log(1 + x), \ x \geq 0$,
(c) $f(x) = \frac{x}{1+x}, \ x \geq 0$, (d) $f(x) = \min(1, x)$,
(e) $f(x) = x^\alpha, \ 0 < \alpha \leq 1$, (f) $f(x) = \arcsin x, \ 0 \leq x \leq 1$.

PROPOSITION 4 *Let $d\colon \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ has the property* Ri, *$i = 1, \ldots, 7$, then $f(d)$, $f \in$* (a)-(f) *also has this property.*

From the theory of metric spaces we know for example:

PROPOSITION 5 *Let $\mathcal{E}$ be a finite dimensional vector space over $\mathbb{R}$ or $\mathbb{C}$. Then any two translation invariant distances over $\mathcal{E}$ are topologically equivalent.*

Some operations preserve properties Ri, $i = 1, \ldots, 7$:

PROPOSITION 6 *Let* $d_1: \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ *and* $d_2: \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ *have property* Ri*, then* $d_1 +_p d_2 = \sqrt[p]{d_1^p + d_2^p}$ *also has property* Ri*,* $i = 1, \ldots, 5, 7$ *for* $p > 0$ *and also has property* R6 *for* $p \geq 1$.

PROPOSITION 7 *Let* $d_1: \mathcal{E}_1 \times \mathcal{E}_1 \to \mathbb{R}$ *and* $d_2: \mathcal{E}_2 \times \mathcal{E}_2 \to \mathbb{R}$ *have property* Ri*, then* $(d_1 +_p d_2)((X_1, X_2), (Y_1, Y_2)) = \sqrt[p]{d_1(X_1, Y_1)^p + d_2(X_2, Y_2)^p}$ *also has property* Ri*,* $i = 1, \ldots, 5, 7$ *for* $p > 0$ *and also has property* R6 *for* $p \geq 1$ *over* $\mathcal{E}_1 \times \mathcal{E}_2$.

$d_1 +_1 d_2$ *is a distance iff* $d_1 +_p d_2$ *is a distance for some* $p \geq 1$.

4 Resemblances on binary vectors

In the case, when all the m properties measured on each unit are of presence/absence type, a description of an unit X has the form $X = [x_1, x_2, \ldots, x_m]$, $x_i \in \mathbb{B} = \{0, 1\}$, where $x_i = 1$, if unit X has the i-th property, and $x_i = 0$, if X lacks the i-th property, $1 \leq i \leq m$.

With XY we denote the scalar product $XY = \sum_{i=1}^{m} x_i y_i$ of units $X, Y \in \mathcal{E}$, and with $\overline{X}$ the complementary vector of X: $\overline{X} = \mathbf{1} - X = [1 - x_i]$. It holds $\overline{\overline{X}} = X$. Now, for any two units $X, Y \in \mathcal{E}$, we define counters:

$a = XY$	– numbers of properties which X and Y share,
$b = X\overline{Y}$	– numbers of properties which X has and Y lacks,
$c = \overline{X}Y$	– numbers of properties which Y has and X lacks,
$d = \overline{X}\overline{Y}$	– numbers of properties which both X and Y lack,

where $a + b + c + d = m$, and with them several resemblances on binary vectors (see Table 1 Lerman 1971; Hubálek 1982; Liebetreau 1983; Gower and Legendre 1986; Baulieu 1989). We assume here that all properties are of the same importance.

4.1 Order equivalent association coefficients

Gower and Legendre (1986) introduced two families of similarities

$$S_\theta = \frac{a + d}{a + d + \theta(b + c)} \quad \text{and} \quad T_\theta = \frac{a}{a + \theta(b + c)},$$

where $\theta > 0$ to avoid negative values. So $s_2 = S_1$, $s_3 = S_2$, $s_4 = 2S_1 - 1$, and $s_6 = T_1$, $s_8 = T_{\frac{1}{2}}$, $s_9 = T_2$. See Table 1 for the meaning of s_i.

Functions $f(x) = \frac{1}{1+x}$ and $\theta(x) = \theta x$ are strictly de/increasing and since $S_\theta = f \circ \theta(\frac{b+c}{a+d})$ by Theorem 1, we have for every θ: $S_\theta \cong \frac{b+c}{a+d} = S$. Also $T_\theta \cong \frac{b+c}{a} = T$. Therefore (Gower and Legendre 1986) for every $\theta, \varrho > 0$: $S_\theta \cong S_\varrho$ and $T_\theta \cong T_\varrho$.

We have: $s_2 \cong s_3 \cong s_4 \cong S$, $s_6 \cong s_7 \cong s_8 \cong s_9 \cong T$, and $s_{13} \cong s_{14} \cong Q_0$. These results were obtained independently also by Beninel (1987).

Table 1: Association coefficients

measure		definition	range	class
Russel and Rao (1940)	s_1	$\frac{a}{m}$	$[1,0]$	
Kendall, Sokal-Michener (1958)	s_2	$\frac{a+d}{m}$	$[1,0]$	S
Rogers and Tanimoto (1960)	s_3	$\frac{a+d}{m+b+c}$	$[1,0]$	S
Hamann (1961)	s_4	$\frac{a+d-b-c}{m}$	$[1,-1]$	S
Sokal & Sneath (1963), un_3^{-1}, S	s_5	$\frac{b+c}{a+d}$	$[0,\infty]$	S
Jaccard (1900)	s_6	$\frac{a}{a+b+c}$	$[1,0]$	T
Kulczynski (1927), T^{-1}	s_7	$\frac{a}{b+c}$	$[\infty,0]$	T
Dice (1945), Czekanowski (1913)	s_8	$\frac{a}{a+\frac{1}{2}(b+c)}$	$[1,0]$	T
Sokal and Sneath	s_9	$\frac{a}{a+2(b+c)}$	$[1,0]$	T
Kulczynski	s_{10}	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$	$[1,0]$	
Sokal & Sneath (1963), un_4	s_{11}	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$	$[1,0]$	
Q_0	s_{12}	$\frac{bc}{ad}$	$[0,\infty]$	Q
Yule (1912), ω	s_{13}	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$[1,-1]$	Q
Yule (1927), Q	s_{14}	$\frac{ad-bc}{ad+bc}$	$[1,-1]$	Q
– bc –	s_{15}	$\frac{4bc}{m^2}$	$[0,1]$	
Driver & Kroeber (1932), Ochiai (1957)	s_{16}	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$[1,0]$	
Sokal & Sneath (1963), un_5	s_{17}	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1,0]$	
Pearson, ϕ	s_{18}	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$[1,-1]$	
Baroni-Urbani, Buser (1976), S^{**}	s_{19}	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$[1,0]$	
Braun-Blanquet (1932)	s_{20}	$\frac{a}{\max(a+b,a+c)}$	$[1,0]$	
Simpson (1943)	s_{21}	$\frac{a}{\min(a+b,a+c)}$	$[1,0]$	
Michael (1920)	s_{22}	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$[1,-1]$	

4.2 Indeterminacy problem

Surprisingly little attention is given in the literature to the problem of the values of association coefficients in the case of indeterminacy (expressions of the form $\frac{0}{0}$). Also in computer programs it is ignored (Anderberg 1973) or reported as an error (Jambu and Lebeaux 1983).

In some cases this problem can be resolved by excluding disturbing units from the set of units $\mathcal{E}$. For example: zero vector in the case of Jaccard coefficient.

In this paper we propose an alternative solution – to eliminate the indeterminacies by appropriately defining values in critical cases. This solution substantially simplifies our study and also permits writing robust computer programs (which can still produce a warning message in the indeterminate cases) for calculation of association coefficients.

We define the Jaccard's coefficient by the expression

$$s_6 = \begin{cases} 1 & d = m \\ \frac{a}{a+b+c} & \text{otherwise} \end{cases}$$

thus ensuring $s_6(X, X) = 1$. In the same way we resolve also the indeterminate cases for s_8 and s_9.

To preserve the monotonic connection between Kulczynski's and Jaccard's coefficients $T = \frac{1}{s_6} - 1$ we set

$$s_7^{-1} = T = \begin{cases} 0 & a = 0,\ d = m \\ \infty & a = 0,\ d < m \\ \frac{b+c}{a} & \text{otherwise} \end{cases}$$

Let us denote

$$K_x = \frac{a}{a+x} \qquad K'_x = \frac{d}{d+x}$$

We cover the indeterminate cases by setting for $x = b, c$

$$x = 0 \Rightarrow K_x = K'_x = 1$$

Using these quantities we can express

$$s_{10} = \frac{1}{2}(K_b + K_c)$$

$$s_{11} = \frac{1}{4}(K_b + K_c + K'_b + K'_c)$$

$$s_{16} = \sqrt{K_b K_c}$$

$$s_{17} = \sqrt{K_b K_c K'_b K'_c}$$

$$s_{18} = \phi = \begin{cases} s_{17} & bc = 0 \\ \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} & \text{otherwise} \end{cases}$$

For the coefficients of type Q we set

$$s_{12} = Q_0 = \begin{cases} 1 & ad = bc \\ \frac{bc}{ad} & \text{otherwise} \end{cases}$$

that implies by order equivalence $s_{13} = s_{14} = 0$ for $ad = bc$.

For Baroni-Urbani's and Braun-Blanquet's coefficients we set $s_{19} = s_{20} = 1$ whenever $b + c = 0$; for Simpson's coefficient $s_{21} = 1$ whenever $bc = 0$.

4.3 Complementary measures

Let $\overline{s}$ denotes the resemblance *complementary* to s defined as

$$\overline{s}(X, Y) = s(\overline{X}, \overline{Y}) \qquad \text{for each pair } X, Y \in \mathcal{E}.$$

Since $a(\overline{X}, \overline{Y}) = d(X, Y)$, $b(\overline{X}, \overline{Y}) = c(X, Y)$, $\ldots$, we have $\overline{s}_i = s_i$ for $i = 2, 3, 4, 5, 11, 12, 13, 14, 15, 17, 18, 22$. We shall call such measures *selfcomplementary.*

Note, that for any property L (P1, P2a,b, P3a, R1–R8) defined in previous section, it holds: *Resemblance measure $\overline{s}$ has the property L iff s has the property L.*

In the space of units $\mathcal{E} = \mathbb{B}^m$ we shall prove the following statements about dissimilarity D introduced in section 3:

STATEMENT 8 *For any pair of resemblances p and r it holds* $\quad D(p, r) = D(\overline{p}, \overline{r})$.

Proof: Let p and r denote resemblance of the same kind (otherwise we can take $-r$ instead of r, because $D(p, r) = D(p, -r)$) and $t = [X, Y]$, $w = [U, V]$, $\overline{t} = [\overline{X}, \overline{Y}]$, $\overline{w} = [\overline{U}, \overline{V}]$. Immediate consequences of the definition of resemblance $\overline{p}$ are $\overline{p}(t) = p(\overline{t})$, $\overline{p}(\overline{t}) = p(t)$, $\ldots$

Let us show, that $(t, w) \in \ll_p \oplus \ll_r \Leftrightarrow (\overline{t}, \overline{w}) \in \ll_{\overline{p}} \oplus \ll_{\overline{r}}$ holds:

$$(\overline{t}, \overline{w}) \in \ll_{\overline{p}} \oplus \ll_{\overline{r}} \Leftrightarrow (\overline{p}(\overline{t}) < \overline{p}(\overline{w})) \veebar (\overline{r}(\overline{t}) < \overline{r}(\overline{w})) \Leftrightarrow$$
$$\Leftrightarrow (p(t) < p(w)) \veebar (r(t) < r(w)) \Leftrightarrow (t, w) \in \ll_p \oplus \ll_r .$$

Since the mapping $X \rightarrow \overline{X}$ is a bijection on $\mathcal{E} = \mathbb{B}^m$ we have:

$$| \ll_p \oplus \ll_r | = | \ll_{\overline{p}} \oplus \ll_{\overline{r}} |.$$

□

STATEMENT 9 *Let p be any resemblance on $\mathcal{E}$ and r a resemblance defined by*

$$r(X, Y) = \varphi(p(X, Y), \overline{p}(X, Y)),$$

where function φ: $\mathbb{R}^2 \rightarrow \mathbb{R}$ satisfies conditions:

$$a < b \wedge c < d \quad \Rightarrow \quad \varphi(a, c) < \varphi(b, d)$$

$$a \leq b \wedge c \leq d \quad \Rightarrow \quad \varphi(a, c) \leq \varphi(b, d);$$

then

$$D(p, r) + D(r, \overline{p}) = D(p, \overline{p}).$$

Proof: Evidently resemblance r is of the same type as p.

We shall use the fact, that for any finite sets A, B, C

$$|A \oplus B| + |B \oplus C| = |A \oplus C|$$

iff $A \cap C \subseteq B \subseteq A \cup C$.

In our case we must prove that $\ll_p \cap \ll_{\overline{p}} \subseteq \ll_r \subseteq \ll_p \cup \ll_{\overline{p}}$.

The first inclusion follows by the first condition on φ:

$$\begin{aligned} &(t,w) \in \ll_p \cap \ll_{\overline{p}} \Leftrightarrow (p(t) < p(w)) \wedge (\overline{p}(t) < \overline{p}(w)) \quad \Rightarrow \\ \Rightarrow \quad &\varphi(p(t), \overline{p}(t)) < \varphi(p(w), \overline{p}(w)) \;\Leftrightarrow\; r(t) < r(w) \;\Leftrightarrow\; (t,w) \in \ll_r \;. \end{aligned}$$

For the second inclusion we must show the implication:

$$\begin{aligned} &(t,w) \in \ll_r \Leftrightarrow r(t) < r(w) \quad \Rightarrow \\ \Rightarrow \quad &(p(t) < p(w)) \vee (\overline{p}(t) < \overline{p}(w)) \;\Leftrightarrow\; (t,w) \in \ll_p \cup \ll_{\overline{p}}, \end{aligned}$$

Or instead, if we consider that $P \Rightarrow Q \equiv \neg Q \Rightarrow \neg P$, the equivalent implication:

$$\begin{aligned} &(p(t) \geq p(w)) \wedge (\overline{p}(t) \geq \overline{p}(w)) \quad \Rightarrow \\ \Rightarrow \quad &\varphi(p(w), \overline{p}(w)) \leq \varphi(p(t), \overline{p}(t)) \;\Leftrightarrow\; r(w) \leq r(t). \end{aligned}$$

which follows by the second condition on φ. □

An immediate consequence of Statement 8 is:

STATEMENT 10 *For any resemblance p on $\mathcal{E}$ and for a selfcomplementary resemblance r, it holds:*

$$D(p, r) = D(r, \overline{p}).$$

Two examples of the function $\varphi(u, v)$, that satisfy the conditions of the Statement 9 are $c(u + v)$ and $(uv)^c$, for $u, v \geq 0$, where $c > 0$ is a constant. Therefore, for $r = c(p + \overline{p})$ and $r = (p\,\overline{p})^c$, we have:

$$D(p, r) = D(r, \overline{p}) = \frac{1}{2} D(p, \overline{p}).$$

Table 2: Values of association coefficients s_1 and s_{15} for $m = 2$

		s_1				s_{15}			
		00	10	01	11	00	10	01	11
1	00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	10	0.0	0.5	0.0	0.5	0.0	0.0	1.0	0.0
3	01	0.0	0.0	0.5	0.5	0.0	1.0	0.0	0.0
4	11	0.0	0.5	0.5	1.0	0.0	0.0	0.0	0.0

Table 3: Trace of computation of $D(s_1, s_{15})$ for $m = 2$

k	$[X,Y]$	$[U,V]$	$s_1(X,Y)$	$s_1(U,V)$	$\ll_{s_1}$	$-s_{15}(X,Y)$	$-s_{15}(U,V)$	$\ll_{s_{15}}^{-1}$	$\cap$	$\oplus$
1	$[1,1]$	$[1,1]$	0.0	0.0		0.0	0.0			
3	$[1,1]$	$[1,2]$	0.0	0.0		0.0	0.0			
5	$[1,1]$	$[1,3]$	0.0	0.0		0.0	0.0			
7	$[1,1]$	$[1,4]$	0.0	0.0		0.0	0.0			
9	$[1,1]$	$[2,2]$	0.0	0.5	1	0.0	0.0			1
11	$[1,1]$	$[2,3]$	0.0	0.0		0.0	-1.0	1		1
13	$[1,1]$	$[2,4]$	0.0	0.5	1	0.0	0.0			1
15	$[1,1]$	$[3,3]$	0.0	0.5	1	0.0	0.0			1
17	$[1,1]$	$[3,4]$	0.0	0.5	1	0.0	0.0			1
19	$[1,1]$	$[4,4]$	0.0	1.0	1	0.0	0.0			1
20	$[1,2]$	$[1,2]$	0.0	0.0		0.0	0.0			
22	$[1,2]$	$[1,3]$	0.0	0.0		0.0	0.0			
24	$[1,2]$	$[1,4]$	0.0	0.0		0.0	0.0			
26	$[1,2]$	$[2,2]$	0.0	0.5	1	0.0	0.0			1
28	$[1,2]$	$[2,3]$	0.0	0.0		0.0	-1.0	1		1
30	$[1,2]$	$[2,4]$	0.0	0.5	1	0.0	0.0			1
32	$[1,2]$	$[3,3]$	0.0	0.5	1	0.0	0.0			1
34	$[1,2]$	$[3,4]$	0.0	0.5	1	0.0	0.0			1
36	$[1,2]$	$[4,4]$	0.0	1.0	1	0.0	0.0			1
37	$[1,3]$	$[1,3]$	0.0	0.0		0.0	0.0			
39	$[1,3]$	$[1,4]$	0.0	0.0		0.0	0.0			
41	$[1,3]$	$[2,2]$	0.0	0.5	1	0.0	0.0			1
43	$[1,3]$	$[2,3]$	0.0	0.0		0.0	-1.0	1		1
45	$[1,3]$	$[2,4]$	0.0	0.5	1	0.0	0.0			1
47	$[1,3]$	$[3,3]$	0.0	0.5	1	0.0	0.0			1
49	$[1,3]$	$[3,4]$	0.0	0.5	1	0.0	0.0			1
51	$[1,3]$	$[4,4]$	0.0	1.0	1	0.0	0.0			1
52	$[1,4]$	$[1,4]$	0.0	0.0		0.0	0.0			
54	$[1,4]$	$[2,2]$	0.0	0.5	1	0.0	0.0			1
56	$[1,4]$	$[2,3]$	0.0	0.0		0.0	-1.0	1		1
58	$[1,4]$	$[2,4]$	0.0	0.5	1	0.0	0.0			1
60	$[1,4]$	$[3,3]$	0.0	0.5	1	0.0	0.0			1
62	$[1,4]$	$[3,4]$	0.0	0.5	1	0.0	0.0			1
64	$[1,4]$	$[4,4]$	0.0	1.0	1	0.0	0.0			1
65	$[2,2]$	$[2,2]$	0.5	0.5		0.0	0.0			
67	$[2,2]$	$[2,3]$	0.5	0.0	1	0.0	-1.0	1	1	
69	$[2,2]$	$[2,4]$	0.5	0.5		0.0	0.0			
71	$[2,2]$	$[3,3]$	0.5	0.5		0.0	0.0			
73	$[2,2]$	$[3,4]$	0.5	0.5		0.0	0.0			
75	$[2,2]$	$[4,4]$	0.5	1.0	1	0.0	0.0			1
76	$[2,3]$	$[2,3]$	0.0	0.0		-1.0	-1.0			
78	$[2,3]$	$[2,4]$	0.0	0.5	1	-1.0	0.0	1	1	
80	$[2,3]$	$[3,3]$	0.0	0.5	1	-1.0	0.0	1	1	
82	$[2,3]$	$[3,4]$	0.0	0.5	1	-1.0	0.0	1	1	
84	$[2,3]$	$[4,4]$	0.0	1.0	1	-1.0	0.0	1	1	
85	$[2,4]$	$[2,4]$	0.5	0.5		0.0	0.0			
87	$[2,4]$	$[3,3]$	0.5	0.5		0.0	0.0			
89	$[2,4]$	$[3,4]$	0.5	0.5		0.0	0.0			
91	$[2,4]$	$[4,4]$	0.5	1.0	1	0.0	0.0			1
92	$[3,3]$	$[3,3]$	0.5	0.5		0.0	0.0			
94	$[3,3]$	$[3,4]$	0.5	0.5		0.0	0.0			
96	$[3,3]$	$[4,4]$	0.5	1.0	1	0.0	0.0			1
97	$[3,4]$	$[3,4]$	0.5	0.5		0.0	0.0			
99	$[3,4]$	$[4,4]$	0.5	1.0	1	0.0	0.0			1
100	$[4,4]$	$[4,4]$	1.0	1.0		0.0	0.0			
					29			9	5	28

4.4 Computational results

For small values of m we can compute the dissimilarity $D(p, q)$ between given resemblances p and q exactly by complete enumeration. In Table 2 values of association coefficients s_1 (Russel and Rao) and s_{15} ($-bc-$) over binary vectors of length $m = 2$ are presented. In Table 3 a trace of computation of $D(s_1, s_{15})$ is given. Since s_1 and s_{15} are of different types we compare s_1 and $-s_{15}$. Note that whenever $[X, Y] \neq [U, V]$ at most one of pairs $([X, Y], [U, V])$ and $([U, V], [X, Y])$ contributes to dissimilarity D.

From Table 3 we can see:

$$|\mathcal{E}| = 2^m = 4, \qquad |\mathcal{E}_2| = \binom{|\mathcal{E}|+1}{2} = 10$$

$$|\ll_{s_1}| = 29, \qquad |\ll_{s_{15}}^{-1}| = 9, \qquad |\ll_{s_1} \cap \ll_{s_{15}}^{-1}| = 5, \qquad |\ll_{s_1} \oplus \ll_{s_{15}}^{-1}| = 28$$

Therefore for $m = 2$

$$D(s_1, s_{15}) = \frac{|\ll_{s_1} \oplus \ll_{s_{15}}^{-1}|}{|\mathcal{E}_2|^2} = \frac{28}{100} = 0.28$$

In Table 4 dissimilarities between (complementary) association coefficients are given for $m = 6$. Values are multiplied with 10000. From the table we can see many confirmations of the above statements:

$$D(s_6, s_2) = D(\overline{s}_6, s_2) = \tfrac{1}{2}D(s_6, \overline{s}_6),\ D(s_{10}, s_{11}) = D(\overline{s}_{10}, s_{11}) = \tfrac{1}{2}D(s_{10}, \overline{s}_{10}),$$
$$D(s_{16}, s_{17}) = D(\overline{s}_{16}, s_{17}) = \tfrac{1}{2}D(s_{16}, \overline{s}_{16});$$

$$D(\overline{s}_1, s_i) = D(s_1, s_i),\ D(s_6, s_i) = D(\overline{s}_6, s_i),\ D(s_{10}, s_i) = D(\overline{s}_{10}, s_i),$$
$$D(s_{16}, s_i) = D(\overline{s}_{16}, s_i),\ i = 2, 11, 14, 17, 18;$$

$$D(s_i, s_j) = D(\overline{s}_i, \overline{s}_j),\ D(s_i, \overline{s}_j) = D(\overline{s}_i, s_j),$$
$$(i, j) = (1, 6), (1, 10), (1, 16), (6, 10), (6, 16), (10, 16).$$

In the upper triangle of Table 5 dissimilarities between 14 selected association coefficients are given for $m = 6$. Since for order equivalent p and q, we have $D(p, q) = 0$ and $D(p, s) = D(q, s)$, we considered in our study only one coefficient from each equivalence class (S, T, Q).

For larger m we can obtain good approximations of $D(p, q)$ by Monte Carlo method ($m = 15$, lower triangle of Table 5). We were repeating the Monte Carlo method until the results stabilized at the fourth decimal. We used $5 \cdot 10^6$ runs.

All three dissimilarity matrices are summarized by dendrograms presented in Figures 1, 2, and 3. Note that the three main clusters in Figure 1 are: selfcomplementary coefficients, nonselfcomplementary coefficients and complementary coefficients to nonselfcomplementary coefficients. The top division in Figures 3 and 2 (with exception Simpson's coefficient) is again selfcomplementary/nonselfcomplementary coefficients.

Table 4: Dissimilarities between (complementary) association coefficients

	s_2	s_6	$\overline{s}_6$	s_1	$\overline{s}_1$	s_{10}	$\overline{s}_{10}$	s_{11}	s_{16}	$\overline{s}_{16}$	s_{17}	s_{18}	s_{14}
s_2	0	1679	1679	2784	2784	1776	1776	1175	1826	1826	1538	1197	1859
s_6		0	3357	1105	4463	749	3335	2001	289	3505	1803	1973	2311
$\overline{s}_6$			0	4463	1105	3335	749	2001	3505	289	1803	1973	2311
s_1				0	5568	1676	4392	3036	1275	4610	2909	3013	3091
$\overline{s}_1$					0	4392	1676	3036	4610	1275	2909	3013	3091
s_{10}						0	2828	1414	460	3143	1799	1472	1765
$\overline{s}_{10}$							0	1414	3143	460	1799	1472	1765
s_{11}								0	1766	1766	1094	197	1026
s_{16}									0	3403	1701	1738	2097
$\overline{s}_{16}$										0	1701	1738	2097
s_{17}											0	897	945
s_{18}												0	829
s_{14}													0

CLUSE – minimum $[0.00, 0.15]$

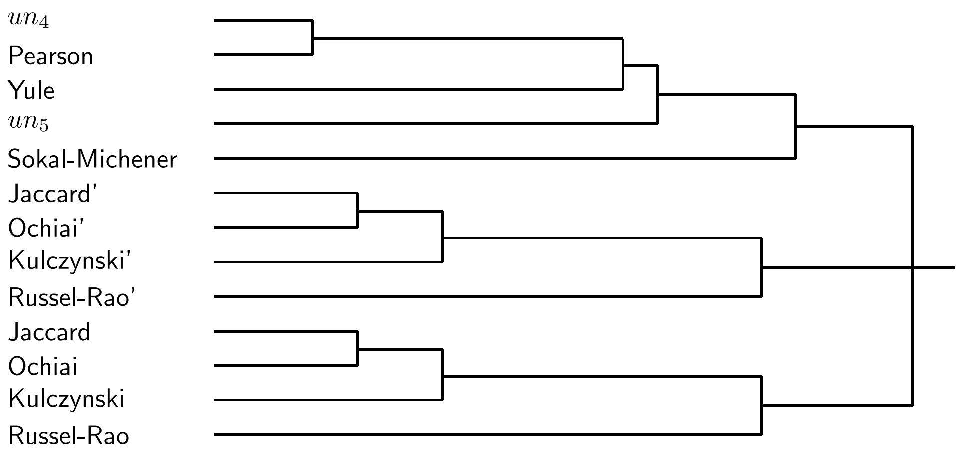

Figure 1: Selfcomplementary association coefficients

Table 5: Association coefficients, upper $m = 6$ / lower $m = 15$

	s_1	s_2	s_6	s_{10}	s_{11}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}	s_{21}	s_{22}
s_1	0	2784	1105	1676	3036	3091	3200	1275	2909	3013	1713	1082	2139	3039
s_2	2948	0	1679	1776	1175	1859	1432	1826	1538	1197	1179	2068	2388	1275
s_6	1076	1872	0	749	2001	2311	2508	289	1803	1973	607	628	1913	2060
s_{10}	1306	1971	413	0	1414	1765	1813	460	1799	1472	1059	1377	1164	1559
s_{11}	3069	889	2021	1819	0	1026	1083	1766	1094	197	1513	2513	1773	375
s_{14}	3082	976	2051	1830	150	0	1154	2097	945	829	1818	2786	1476	880
s_{15}	3150	912	2197	1886	726	724	0	2260	2087	1218	2205	3044	1073	1243
s_{16}	1219	1941	224	189	1856	1888	2012	0	1701	1738	755	917	1624	1825
s_{17}	3042	944	1969	1837	339	338	1062	1839	0	897	1196	2234	2409	1031
s_{18}	3068	885	2020	1819	5	154	726	1855	339	0	1442	2447	1908	178
s_{19}	1865	1103	780	940	1270	1311	1546	857	1193	1268	0	1042	2047	1542
s_{20}	1204	2193	789	1202	2447	2492	2726	1013	2336	2445	1217	0	2541	2509
s_{21}	1717	2290	1366	954	1908	1858	1716	1143	2060	1911	1558	2156	0	1957
s_{22}	3070	921	2025	1825	132	265	729	1854	363	127	1281	2452	1921	0

CLUSE – maximum $[0.00, 0.34]$

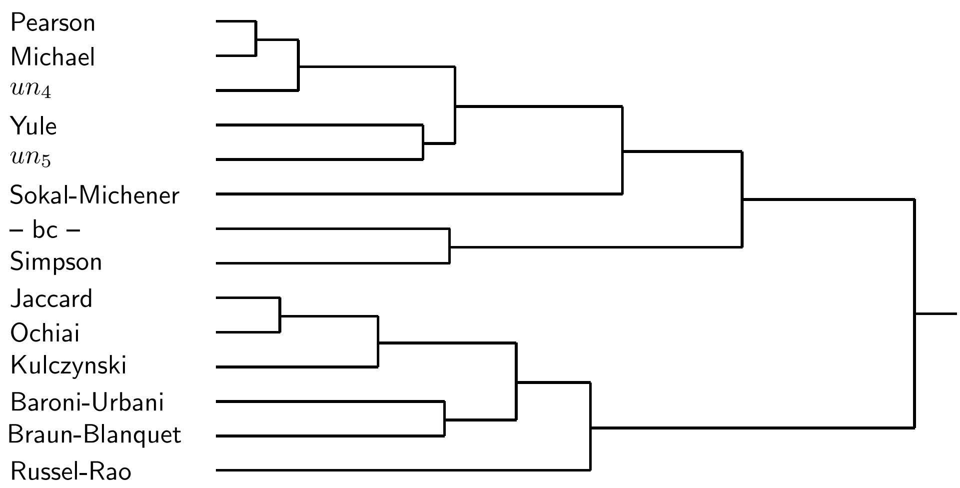

Figure 2: Association coefficients, enumeration, $m = 6$

CLUSE – maximum $[0.00, 0.33]$

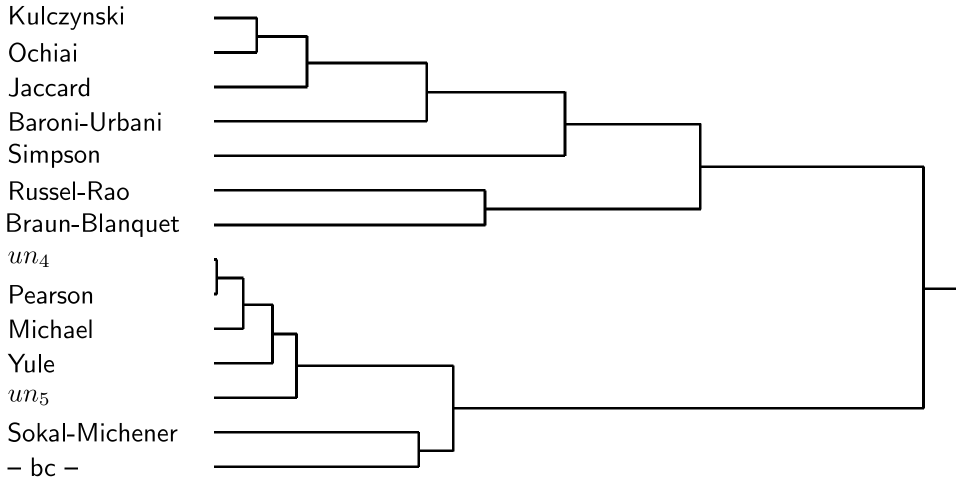

Figure 3: Association coefficients, Monte Carlo, $m = 15$

5 Conclusion

We believe that further study of different types of equivalences of resemblances can give a better understanding of data analysis methods based on them and some guidelines for their (correct) applications. In this paper we presented only some special results in this direction. We expect that a more comprehensive and elaborate picture can be produced.

Also some problems about dissimilarities between resemblances remain open. The most important is how to extend the dissimilarty D to other types of resemblances; for example, to the case $\mathbb{R}^m$.

For association coefficients we can pose the following questions:

- What is a behavior of $D(p,q)$ over $\mathcal{E} = \mathbb{B}^m$ when $m \to \infty$? We expect that for some coefficients p and q an explicit formula for $D(p,q)$ can be derived.
- Other types of normalization of $\mid \ll_p \oplus \ll_q \mid$ can be given. A dissimilarity, with the property that the upper bound in $0 \le D(p,q) \le 1$ is attained, can be based on the solution of the (unsolved) problem

$$\max\{\mid \ll_p \oplus \ll_q \mid : p, q \in \text{forward coefficients over } \mathcal{E} = \mathbb{B}^m\}.$$

Another interesting measure is given, for p and q of the same type, by the semidistance (Kaufmann 1975):

$$D_2(p,q) = \frac{\mid \ll_p \oplus \ll_q \mid}{\mid \ll_p \cup \ll_q \mid}$$

and yet another by

$$D_3(p,q) = \frac{\max(\mid \ll_p \setminus \ll_q \mid, \mid \ll_q \setminus \ll_p \mid)}{\max(\mid \ll_p \mid, \mid \ll_q \mid)}.$$

What can be said about these dissimilarities?

Acknowledgments

We would like to thank the editor and two anonymous referees for numerous remarks and suggestions that significantly improved the presentation of the material.

This work was supported in part by the Ministry for Science and Technology of Slovenia.

References

[1] ANDERBERG, M.R. (1973), *Cluster analysis for applications.* New York: Academic Press.

[2] BANDELT, H-J. (1990), "Recognition of tree metrics", *SIAM Journal on Discrete Mathematics, 3/1*, 1-6.

[3] BATAGELJ, V. (1989), *Similarity measures between structured objects*, in A. Graovac (Ed.), Proceedings MATH/CHEM/COMP 1988, Dubrovnik, Yugoslavia 20-25 June 1988, Studies in Physical and Theoretical Chemistry. Vol 63, pp. 25-40, Amsterdam: Elsevier.

[4] BATAGELJ, V. (1992), *CLUSE/TV – clustering programs*, Manual, Ljubljana.

[5] BATAGELJ, V., PISANSKI, T., and SIMÕES-PEREIRA, J.M.S. (1990), "An algorithm for tree-realizability of distance matrices", *International Journal of Computer Mathematics, 34*, 171-176.

[6] BAULIEU, F.B. (1989), "A classification of presence/absence based dissimilarity coefficients", *Journal of Classification, 6*, 233-246.

[7] BENINEL, F. (1987), *Problemes de representations spheriques des tableaux de dissimilarite*, Thesis, Université de Rennes I, (in French).

[8] DIEUDONNÉ, J. (1960), *Foundations of modern analysis*, New York: Academic Press.

[9] GORDON, A.D. (1981), *Classification*, London: Chapman and Hall.

[10] GOWER, J.C., and LEGENDRE, P. (1986), "Metric and Euclidean properties of dissimilarity coefficients", *Journal of Classification, 3*, 5-48.

[11] GOWER, J.C. (1971), "A general coefficient of similarity and some of its properties", *Biometrics 27*, 857-871.

[12] HUBÁLEK, Z. (1982), "Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation", *Biological Review 57*, 669-689.

[13] JAMBU, M., and LEBEAUX, M-O. (1983), *Cluster Analysis and Data Analysis*, Amsterdam: North-Holland.

[14] JOLY, S., and LE CALVE, G. (1986), "Etude des puissances d'une distance", *Statistique et Analyse des Données, 11/3*, 30-50.

[15] KAUFMANN, A. (1975), *Introduction a la théorie des sous-ensembles flous*, Vol. III, 153-155, Paris: Masson.

[16] KRANTZ, D.H., LUCE, R.D., SUPPES, P., and TVERSKY, A. (1971), *Foundations of Measurement*, Vol. I., New York: Academic Press.

[17] KRUSKAL, J.B. (1983), "An overview of sequence comparison: time warps, string edits and macromolecules", *SIAM Review 25/2*, 201-237.

[18] LERMAN, I.C. (1971), *Indice de similarité et préordonnance associée*, Ordres. Travaux du séminaire sur les ordres totaux finis, Aix-en-Provence, 1967; Paris: Mouton.

[19] LIEBETRAU, A.M. (1983), *Measures of association*, Newbury Park, CA: Sage Publications.

[20] SNEATH, P.H.A., and SOKAL, R.R. (1973), *Numerical taxonomy*, San Francisco: W.H. Freeman.

[21] SPÄTH, H. (1977), *Cluster Analyse Algorithmen zur Objekt-Klassifizierung und Datenreduction*, München: R. Oldenbourg.