ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XIII, 2 (1972)

D. CZERWIŃSKA (Wrocław)

ON THE SIMILARITY OF SETS

1. Introduction. In various practical questions we have to determine the similarity of sets. Botanists, for instance, regard as similar such plant complexes which differ insignificantly in their species. We may ask about the similarity of those plant complexes or, in general, about the similarity of two sets. Many authors tried to define similarity of sets. It is convenient, investigating properties of the similarity of sets, to introduce the distance of sets defined by the formula

$$\varrho(A, B) = 1-s, \tag{1}$$

where s denotes the similarity of the sets A and B.

In this paper we give a necessary and sufficient condition for the set B to lie between the sets A and C in the sense of triangle inequality and the analogous condition for the function g to lie between functions f and h. The question concerning these conditions has been raised in paper [1]. It is interesting from the view, point of applications, e.g. to biology, where the fact that the set B lies between the sets A and C can mean for instance the way of evolution of plant complexes A, B and C.

2. Properties of the similarity index of sets of Marczewski and Steinhaus. Marczewski and Steinhaus [1] investigated the similarity index

$$s = \frac{\omega}{a+b-\omega}, \tag{2}$$

of the sets A and B, where ω denotes the number of elements common to the sets A and B, and a and b are the numbers of elements of the sets A and B, respectively. Let $m(E)$ be the number of elements of the set E. In view of (1) and (2) the distance of the sets A and B is then equal to

$$\varrho(A, B) = \frac{m(A \dot{-} B)}{m(A \cup B)}, \tag{3}$$

where $\dot{-}$ denotes the symmetric difference. Since ϱ is a metric, the triangle inequality

$$\varrho(A, B)+\varrho(B, C) \geqslant \varrho(A, C) \tag{4}$$

is satisfied.

Let us assume that to every element A there corresponds on the plane a unit area. The sets A, B and C divide the plane into atoms which will be denoted by the Greek letters (cf. Fig. 1).

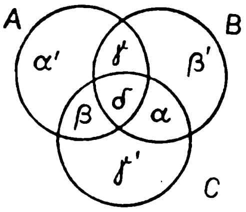

Fig. 1

Moreover, let us write: $\alpha+\alpha' = a$, $\beta+\beta' = b$, $\gamma+\gamma' = c$.

Lemma. *Condition* (4) *is equivalent to the inequality*

$$P+Q\delta+R\delta^2 \geqslant p+q\delta+r\delta^2, \tag{5}$$

where

$$\begin{aligned}
P &= (a+b)(b+c)(a+c)+2(a+b)(b+c)\beta+(a+b)\alpha\beta+(b+c)\beta\gamma,\\
Q &= 2(a+b)(b+c)+(a+b)(\alpha+\beta)+(b+c)(\beta+\gamma),\\
R &= a+2b+c,\\
p &= (a+c)\alpha\gamma,\\
q &= (a+c)(\alpha+\gamma),\\
r &= a+c.
\end{aligned}$$

Proof of this lemma may be found in paper [1].

As the criterion of the lying the set B between the sets A and C let us assume that the triangle inequality holds.

Theorem 1. *If* $\delta > 0$, *then*

$$\varrho(A, B)+\varrho(B, C) = \varrho(A, C)$$

if and only if $\beta = 0$, $\beta' = 0$ *and* 1° $\alpha' = 0$ *and* $\alpha = 0$ *either* 2° $\alpha' = 0$ *and* $\gamma' = 0$ *or* 3° $\gamma' = 0$ *and* $\gamma = 0$.

Proof. Sufficiency. Immediate by substituting in (4) the corresponding values.

Necessity. By the lemma, instead of (4) we can take inequality (5) in another form:

$$(6) \qquad (R-r)\delta^2+(Q-q)\delta+P-p \geqslant 0.$$

Note that $R \geqslant r$, $Q \geqslant q$ and $P \geqslant p$ (cf. [1]). If in (6) the equality holds, then the equalities

$$(7) \quad R = r, \qquad (8) \quad Q = q, \qquad (9) \quad P = p$$

are satisfied. Equality (7) implies

$$(10) \qquad \beta = 0 \quad \text{and} \quad \beta' = 0.$$

From (8) and (10) it follows that

$$(11) \qquad a\gamma' = 0 \text{ and } a'\gamma' = 0 \text{ and } a'\gamma = 0.$$

Condition (11) is equivalent to

$$(12) \quad (a' = 0 \text{ and } a = 0) \text{ or } (a' = 0 \text{ and } \gamma' = 0) \text{ or } (\gamma' = 0 \text{ and } \gamma = 0).$$

$$(a = 0 \vee \gamma' = 0) \wedge (a' = 0 \vee \gamma' = 0) \wedge (a' = 0 \vee \gamma = 0)$$
$$\equiv [(a' = 0 \wedge a = 0) \vee (a = 0 \wedge a' = 0 \wedge \gamma = 0)] \vee (\gamma' = 0 \wedge a' = 0) \vee$$
$$\vee (\gamma' = 0 \wedge \gamma = 0) \equiv (a' = 0 \wedge a = 0) \vee (a' = 0 \wedge \gamma' = 0) \vee$$
$$\vee (\gamma' = 0 \wedge \gamma = 0).$$

Thus, in view of (7) and (8), we obtained the condition contained in the thesis of the theorem. It is easy to see that condition (9) gives nothing new, which completes the proof of Theorem 1.

Fig. 2 shows cases 1°, 2° and 3° of Theorem 1.

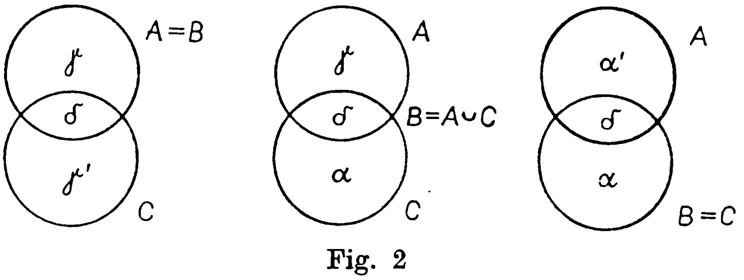

Fig. 2

If the fact that the set B is lying between A and C in the sense of the triangle inequality will be interpreted as the way of evolution of plant complexes A, B and C, then condition (10) is obvious (cf. Figs. 2 and 3). Unfortunately, this condition is not sufficient that the equality holds in (4). There are several cases where condition (10) is satisfied although the equality in (4) does not hold. In Fig. 3 some of these cases are shown (the stroken set B).

Theorem 2. *If $\delta = 0$, then the equality in* (4) *holds if and only if $\beta = 0$, $\beta' = 0$ and* 1° $\alpha > 0$ *and* $\big((\alpha' = 0$ *and* $\gamma' = 0)$ *or* $(\gamma = 0$ *and* $\gamma' = 0)\big)$ *or* 2° $\gamma > 0$ *and* $\big((\alpha' = 0$ *and* $\gamma' = 0)$ *or* $(\alpha = 0$ *and* $\alpha' = 0)\big)$.

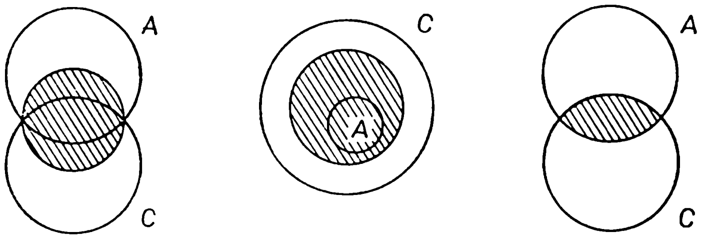

Fig. 3

Proof. Sufficiency follows the same lines as for Theorem 1.

Necessity. Condition (4) takes on the form

$$\frac{\alpha+\alpha'+\beta+\beta'}{\alpha+\alpha'+\beta+\beta'+\gamma+\delta}+\frac{\beta+\beta'+\gamma+\gamma'}{\beta+\beta'+\gamma+\gamma'+\alpha+\delta}\geqslant\frac{\alpha+\alpha'+\gamma+\gamma'}{\alpha+\alpha'+\gamma+\gamma'+\beta+\delta},$$

where the Greek letters denote the atoms from Fig. 1. Obviously, the denominators in this formula are greater than zero.

Since $\delta = 0$, we have

$$(13)\qquad a+b+\gamma > 0, \quad b+c+\alpha > 0, \quad a+c+\beta > 0.$$

Instead of condition (4) we can consider condition (5). Hence in (4) the equality holds if condition (9) and three inequalities (13) are simultaneously satisfied.

Equality (9) implies that $\beta = 0$. From (9) again and from the inequality $a+c > 0$ it follows that

$$\alpha(\beta'+\gamma')+\gamma(\alpha'+\beta')+(\beta'+\gamma')(\alpha'+\beta') = 0,$$

which implies that $\beta' = 0$. We have thus inferred that $\beta = 0$, $\beta' = 0$ and $\alpha\gamma'+\alpha'\gamma+\alpha'\gamma' = 0$ with $\alpha+\gamma > 0$ and $a+c > 0$, which completes the proof of Theorem 2.

It easily seen that Theorems 1 and 2 can be written in the form of the following theorem (cf. Fig. 2):

Theorem 3. *The equality in condition* (4) *holds if and only if $A = B$ either $B = C$ or $B = A \cup C$.*

3. Another properties of the similarity index of sets. Measure m used in Section 2 may be generalized to any abstract measure defined on subsets of a space X. Let two functions f and g be defined on X. The distance of functions f and g is definde by the formula

$$(14)\qquad \varrho_m(f, g) = \frac{\int |f-g|\,dm}{\int \max(|f|, |g|, |f-g|)\,dm}.$$

In particular, if f and g are non-negative, formula (14) can be reduced to

$$(15)\qquad \varrho_m(f, g) = \frac{\int |f-g|\,dm}{\int \max(|f|, |g|)\,dm}.$$

The distance of sets, defined by formula (3), may be regarded as the distance (15) of characteristic functions of those sets,

$$\varrho_m(A, B) = \varrho_m(\chi_A, \chi_B),$$

where χ_A and χ_B are characteristic functions of the sets A and B, respectively. But the distance (14) of functions may be regarded as the distance (3) of some sets, i.e.

$$\varrho_m(f, g) = \varrho_\nu(C_f, C_g),$$

where C_f and C_g are the sets of points lying between the diagrams of functions f and g, respectively, and the X-axis, whereas ν is a measure in the sense of simple product of Lebesgue measure and measure m. The presentation of full analogy between the metric spaces (M_0, ϱ) and $(\mathscr{L}_m, \varrho_m)$ where M_0 is the family of subsets A of a space X such that $m(A) < \infty$ and $\mathscr{L}_m$ is the class of all m-integrable real functions defined on X, may be found in [2].

In Section 2 we have considered the cases where the equality in triangle condition holds for sets. For functions, as the analogue to Theorem 3, we have the following

THEOREM 4. *In the triangle inequality*

$$(4')\qquad \varrho_m(f, g) + \varrho_m(g, h) \geqslant \varrho_m(f, h)$$

the equality holds if and only if 1° $f = g - m$ *almost everywhere either* 2 $g = h - m$ *almost everywhere or* 3° $g(x) = \operatorname{sgn} f(x) \max(|f(x)|, |h(x)|)$ *for almost every* x *and the functions* f *and* h *are of the same sign.*

We assume the criterion that the function g lies between f and h if there holds the equality in condition (4') with definition (14) of the distance. If $\beta = 0$ and $\beta' = 0$, then the function g lies between f and h in the usual meaning, i.e.

$$f(x) \leqslant g(x) \leqslant h(x) \quad \text{or} \quad h(x) \leqslant g(x) \leqslant f(x).$$

We see that the fact that function g lies between f and h in the usual meaning is not a sufficient condition for the equality in (4') with the distance defined by (14), but it is only a necessary one.

The following example shows an interpretation of the distance of sets for plant complexes. Let m, as previously, denote the number of elements of a set, i.e. the number of species in the given plant complex. Let f and g characterize numerically two forests with respect to 5 species of trees. In the first forest every 10 trees contain: 4 pines, 2 oaks, 3 birches and 1 alder; in the second forest: 2 oaks, 1 birch, 2 alders and 5 spruces (the data are taken from paper [1]). The following table contains these values and gives the intermediate results:

	pine	oak	birch	alder	spruce	δ
f	4	2	3	1	0	
g	0	2	1	2	5	
$\lvert f-g\rvert$	4	0	2	1	5	12
$\max(\lvert f\rvert, \lvert g\rvert)$	4	2	3	2	5	16

Thus we have $\varrho(f, g) = 12/16 = 0,75$.

References

[1] E. Marczewski and H. Steinhaus, *O odległości systematycznej biotopów*, Zastos. Matem. 4 (1959), p. 195-203.

[2] — *On a certain distance of sets and the corresponding distance of functions*, Colloquium Mathematicum 6 (1958), p. 319-327.

MATHEMATICAL INSTITUTE
UNIVERSITY OF WROCŁAW

Received on 26. 1. 1972

CZERWIŃSKA (Wrocław)

O PODOBIEŃSTWIE ZBIORÓW

STRESZCZENIE

W pracy podaje się warunek konieczny i dostateczny na to, aby w nierówności trójkąta, dla odległości zbiorów zdefiniowanej wzorem (3), była spełniona równość. Warunek ten może służyć do definiowania własności leżenia zbioru między dwoma innymi zbiorami [1]. Ponieważ jest analogia między przestrzenią (M_0, ϱ) zbiorów miary m-skończonej z odległością zbiorów (3) i przestrzenią $(\mathscr{L}_m, \varrho_m)$ funkcji m-całkowalnych z odległością (14), więc podano również warunek konieczny i dostateczny na to, aby nierówność trójkąta dla odległości funkcji była równością.