ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVI 1 (1977)

M. KAROŃSKI (Gainesville, Florida)* and Z. PALKA (Poznań)

ON MARCZEWSKI-STEINHAUS TYPE DISTANCE BETWEEN HYPERGRAPHS

This paper is concerned with establishing a metric space of hypergraphs with the same set of vertices. In particular, the distance between hypergraphs generated by arborescences with the same set of terminal vertices is also given. The basis of the proposed distances is the Marczewski-Steinhaus distance between two sets (see [6] and [7]).

It is shown also that hypergraphs can represent a grouping of objects by different cluster analysis methods. An example of comparing different clusterings, using the above idea, is given.

1. INTRODUCTION

Marczewski and Steinhaus [6] generalized the Fréchet-Nikodym-Aronszajn distance between two sets and applied this distance in the analysis of numerical characterization of differences between biotopes (see [7]). In this paper we establish a metric space of hypergraphs with the same set of vertices. This problem is equivalent to the one of finding the distance between two families of subsets of the same set (see Ulam [9]). In our case we apply the Marczewski-Steinhaus distance as a basis for the distances between such families. We also consider a special case of hypergraph being generated by an arborescence and whose family of edges has special properties. The Marczewski-Steinhaus type metric which we propose is related to this property.

The idea, which leads us to present both distances together, arises from the fact that a hypergraph can be regarded as a representation of grouping of different objects. We can consider an object as a vertex in a representing hypergraph, and the established group (or cluster) as an edge of this hypergraph.

Suppose that different methods of clustering yield different results for the same set of objects. The problem which we attempt to solve, in the next stage of analysis, is the following: which methods give similar results.

* On leave from the Institute of Mathematics, Adam Mickiewicz University, Poznań, Poland.

The distance of hypergraphs representing these methods provides us with the basis to compare them. It can also be regarded as the first step to the cluster analysis of cluster analysis methods.

In addition, results of hierarchical clustering can be presented in the form of a binary tree with a root, which is a special case of an arborescence. Similarity of two such methods can be treated as a function of distance between trees or, more generally, as a distance between hypergraphs generated in a special way by these trees.

2. METRIC SPACES OF HYPERGRAPHS

2.1. Basic definitions. Let X be a finite set such that $|X| = n$, where $|\cdot|$ denotes cardinality of a set. Denote by E^* the class of all subsets of the set X, and by $\mu(E)$ the measure of a subset $E \in E^*$. Assume also that $\mu(E) < \infty$ for all elements of the class E^*. Then, according to the Marczewski-Steinhaus generalization of the so-called Fréchet-Nikodym-Aronszajn metric, the distance between two sets from E^* is

$$\sigma_\mu(E_1, E_2) = \begin{cases} \dfrac{\varrho(E_1, E_2)}{\mu(E_1 \cup E_2)} & \text{if } \mu(E_1 \cup E_2) > 0, \\ 0 & \text{if } \mu(E_1 \cup E_2) = 0, \end{cases} \tag{1}$$

where

$$\varrho(E_1, E_2) = \mu(E_1 \Delta E_2),$$

i.e., the measure of symmetric difference of sets E_1 and E_2. Note that $0 \leqslant \sigma_\mu(\cdot, \cdot) \leqslant 1$. In particular, if we assume that $\mu_c(E) = |E|$, then, setting

$$e_1 = |E_1|, \quad e_2 = |E_2| \quad \text{and} \quad d = |E_1 \cap E_2|,$$

the Marczewski-Steinhaus distance between E_1 and E_2 can be presented as

$$\sigma_{\mu_c}(E_1, E_2) = \frac{e_1 + e_2 - 2d}{e_1 + e_2 - d}. \tag{2}$$

The family $E = (E_i; i \in I)$ of subsets of X is said to be a *hypergraph* on X if

$$E_i \neq \emptyset \ (i \in I) \quad \text{and} \quad \bigcup_{i \in I} E_i = X.$$

The tuple $H = (X, E)$ is called a *hypergraph*, the elements $x_1, x_2, \ldots, x_n$ are its *vertices*, the sets $E_1, E_2, \ldots, E_m$ its *edges*, and n is called the *order* of the hypergraph.

An *arborescence* is defined as a directed tree that has a root, i.e., a tree with the vertex v_0, such that all vertices of the tree can be reached by a path starting from v_0.

The *outer demi-degree* $d^+(v)$ of the vertex v is defined as the number of arcs that are incident out of v, and the *inner demi-degree* $d^-(v)$ of v is defined as the number of arcs that are incident into the vertex v. The vertices of an arborescence can be classified into three classes:

(i) root — a one-element class consisting of a vertex v_0 such that $d^-(v_0) = 0$ and $d^+(v_0) \geqslant 1$;

(ii) non-terminal vertices — $N = \{v\colon d^-(v) = 1, d^+(v) \geqslant 1\}$;

(iii) terminal vertices — $X = \{v\colon d^-(v) = 1,\ d^+(v) = 0\}$.

In defining a distance between arborescences with the same set X of terminal vertices, we use a hypergraph on X as its representation.

For other definitions we refer the reader to Berge [1] or to any other book on graph theory.

2.2. Hypergraphs. Let $\mathscr{H}$ be the class of all hypergraphs on the set $X = \{x_1, x_2, \ldots, x_n\}$. Suppose that $H_1 = (X, E_1)$ and $H_2 = (X, E_2)$ are two hypergraphs from this class, both having the same set of vertices and different sets of edges:

$$E_1 = (E_{1i};\ i \in I_1) \quad \text{and} \quad E_2 = (E_{2j};\ j \in I_2).$$

Then the distance between H_1 and H_2 can be defined as

$$\varrho_1(H_1, H_2) = \tfrac{1}{2}\Big[\max_{i\in I_1}\min_{j\in I_2}\sigma_\mu(E_{1i}, E_{2j}) + \max_{j\in I_2}\min_{i\in I_1}\sigma_\mu(E_{1i}, E_{2j})\Big], \tag{3}$$

where $\sigma_\mu(\cdot,\cdot)$ is the Marczewski-Steinhaus distance between two sets. It follows that:

(i) $(\mathscr{H}, \varrho_1)$ *is a metric space (when we identify any two edges of hypergraphs, the symmetric difference of which is of measure zero)*,

(ii) $\varrho_1(H_1, H_2) \leqslant 1$.

If we use in (3) the distance $\sigma_{\mu_c}(\cdot,\cdot)$ defined by (2), then

(iii) $\varrho_1(H_1, H_2) < 1$.

To prove (iii) note that, for $E_{1i} \in E_1$ and $E_{2j} \in E_2$, we have $\sigma_{\mu_c}(E_{1i}, E_{2j}) = 1$ iff $\mu_c(E_{1i} \cap E_{2j}) = 0$, i.e., $E_{1i} \cap E_{2j} = \emptyset$. In the hypergraph H_1, each edge has a non-empty intersection with at least one edge of H_2. This implies that

$$\max_{E_1}\min_{E_2}\sigma_{\mu_c}(\cdot,\cdot) < 1$$

as well as

$$\max_{E_2}\min_{E_1}\sigma_{\mu_c}(\cdot,\cdot) < 1,$$

which proves the assertion.

The distance $\varrho_1(\cdot,\cdot)$ given by formula (3) is a modification of Ulam's [9] generalization of the Hausdorff metric for sets.

Example 1. Let us consider three hypergraphs given in Fig. 1. The set of vertices has $|X| = 6$ and

$$H_1 = (X, E_1), \quad H_2 = (X, E_2), \quad H_3 = (X, E_3),$$

where

$$E_1 = \{\{1, 2, 3, 6\}, \{3, 5, 6\}, \{4\}\}, \quad E_2 = \{\{1, 2, 4\}, \{3, 5, 6\}, \{2, 3\}, \{4, 5\}\}$$

and

$$E_3 = \{\{1, 2, 3\}, \{3, 5, 6\}, \{4, 5\}\}.$$

The distances between these hypergraphs are the following:

$$\varrho_1(H_1, H_2) = .55, \quad \varrho_1(H_1, H_3) = .50 \quad \text{and} \quad \varrho_1(H_2, H_3) = .42.$$

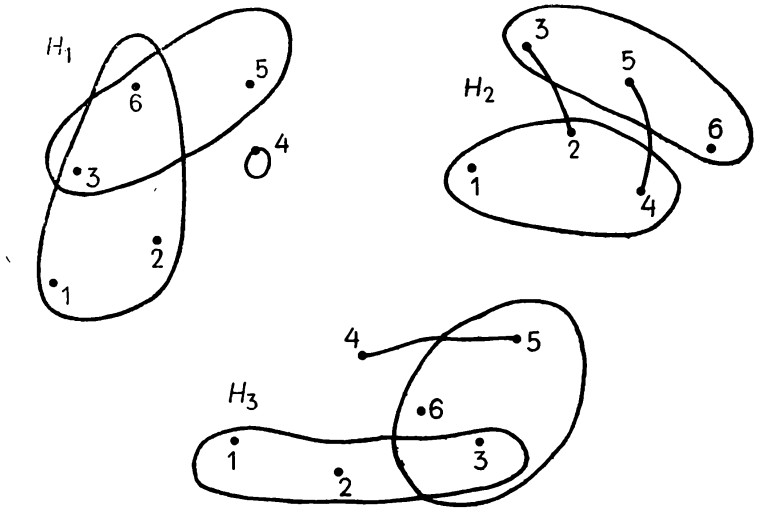

Fig. 1. Hypergraphs with $|X| = 6$

2.3. Hypergraphs generated by arborescences. Let $X = \{x_1, x_2, \ldots, x_n\}$ be the set of terminal vertices of an arborescence, i.e., $d^-(x_i) = 1$, $d^+(x_i) = 0$ for $i = 1, 2, \ldots, n$, and let $\mathscr{A}$ denote the class of all arborescences with X as the set of terminal vertices. Let $A \in \mathscr{A}$ be represented by the hypergraph $H_A = (X, E_A)$, where the class of edges E_A is defined in the following way. Each $v \notin X$ (i.e., each non-terminal vertex in the arborescence A) generates $d^+(v) - 1$ the same edges in E_A. Such an edge consists of these elements of the set X which are the terminal vertices of the subarborescence generated by a vertex v. The subarborescence generated by a vertex v is obtained by treating this vertex as a root, i.e., by assuming that $d^-(v) = 0$.

The method of construction of the edges of H_A leads us to the following obvious assertions:

(i) *If* $H_A = (X, E_A)$ *is the hypergraph generated by an arborescence* $A \in \mathscr{A}$ *in the described way, then* $|E_A| = n - 1$.

Proof. Let N denote the number of all vertices of an arborescence $A \in \mathscr{A}$. Let us denote the set of all vertices of the arborescence A by

$$V_A = \{v_0, v_1, \ldots, v_{N-n-1}, x_1, \ldots, x_n\},$$

where v_0 is the root, v_j $(j = 1, 2, \ldots, N-n-1)$ a non-terminal vertex, and x_i $(i = 1, 2, \ldots, n)$ a terminal vertex.

It is easy to see that

$$\sum_{j=0}^{N-n-1} d^+(v_j) = N-1,$$

i.e., the sum of the outer demi-degrees of the root and of all non-terminal vertices is equal to the number of all arcs in the arborescence A. Each edge of H_A occurs in E_A $d^+(v)-1$ times, so the total number of edges in E_A is

$$\sum_{j=0}^{N-n-1} [d^+(v_j)-1] = n-1,$$

which proves the assertion.

It is also straightforward that

(ii) *The hypergraph H_A generated by an $A \in \mathscr{A}$ is not simple if, for at least one vertex $v \in A$, $d^+(v) > 2$.*

By definition, a hypergraph is simple if its edges are distinct. A binary tree, which is a special case of an arborescence, has the representation by a simple hypergraph.

Suppose that A_1 and A_2 are elements of $\mathscr{A}$ and are represented by $H_{A_1} = (X, E_{A_1})$ and $H_{A_2} = (X, E_{A_2})$, respectively.

The distance between these hypergraphs, we propose, takes into consideration the specific way of construction of edges. We can also say that the following equation describes the distance between arborescences:

$$\varrho_2(H_{A_1}, H_{A_2}) = d(A_1, A_2) = \frac{1}{n-1} \min_{p \in P} \sum_{i=1}^{n-1} \sigma_\mu(E^i_{A_1}, E^{p_i}_{A_2}), \tag{4}$$

where p_i is the i-th element of the permutation p of the first $n-1$ integers, P is the set of all such permutations, the distance $\sigma_\mu(\cdot, \cdot)$ is given by (1), and $E^i_{A_1} \in E_{A_1}$, $E^{p_i}_{A_2} \in E_{A_2}$ $(i = 1, 2, \ldots, n-1)$.

The following facts are implied by the above definition.

(i) $(\mathscr{A}, d)$ is a metric space (when we identify two edges of the representing hypergraphs the symmetric difference of which is of measure zero).

(ii) $d(A_1, A_2) \leqslant 1$, $A_1, A_2 \in \mathscr{A}$. The distance $d(\cdot, \cdot)$ is strictly less than one if instead of $\sigma_\mu(\cdot, \cdot)$ we use in (4) the distance $\sigma_{\mu_c}(\cdot, \cdot)$ given by formula (2).

The distance $\varrho_2(\cdot, \cdot)$ defined by (4) is a generalization of the Boorman-Olivier [2] distance for rooted binary trees.

Example 2. Let us consider four arborescences with $|X| = 4$ given in Fig. 2.

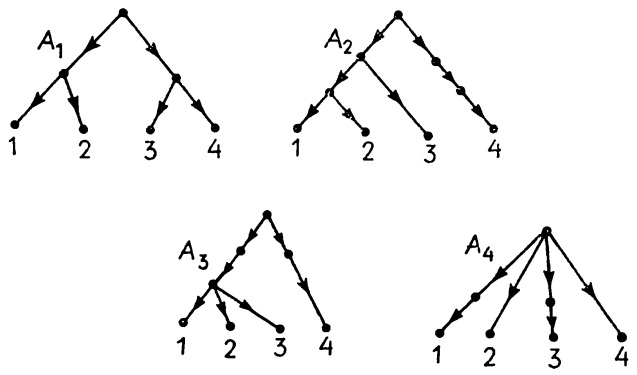

Fig. 2

The families of edges of the representing hypergraphs are

$$E_{A_1} = \{\{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}, \quad E_{A_2} = \{\{1, 2\}, \{1, 2, 3\}, \{1, 2, 3, 4\}\},$$

$$E_{A_3} = \{\{1, 2, 3\}, \{1, 2, 3\}, \{1, 2, 3, 4\}\},$$

$$E_{A_4} = \{\{1, 2, 3, 4\}, \{1, 2, 3, 4\}, \{1, 2, 3, 4\}\},$$

and the distances are given in Table 1.

TABLE 1

	A_1	A_2	A_3
A_2	.25		
A_3	.36	.11	
A_4	.33	.25	.16

3. APPLICATIONS

One of the interesting problems in cluster analysis is the evaluation of similarity of two different groupings of the same data obtained by different cluster analysis methods. In this approach we deal with some kind of second stage cluster analysis in the sense that we want to cluster cluster analysis methods. It can also be used, in a more general sense, to evaluate stability of cluster analysis. First, we have to establish a measure of similarity between two different results of grouping. Rand [8] proposed such a measure based on averaging the number of pairs of grouped elements which occur or not occur together in the same cluster. In other words,

using graph theory concepts, Rand assumes that each group forms a complete subgraph, which is not always an adequate model in clustering.

A variety of different metrics on a space of finite trees is also given by Boorman and Olivier [2]. They use these metrics to evaluate differences between hierarchical cluster analysis methods.

As was mentioned earlier, the results of both non-hierarchical and hierarchical groupings can be represented by a hypergraph. In the case of non-hierarchical grouping, with a fixed number of groups, it is a direct representation. In the case of step-by-step hierarchical grouping (divisive or agglomerative type) each method can be represented as a binary tree which is a special case of an arborescence and, moreover, can be represented by a hypergraph of this binary tree (as described in Section 2.3).

We want to present two different applications of Marczewski-Steinhaus type distances between hypergraphs to compare results of non-hierarchical and hierarchical clusterings.

3.1. Non-hierarchical grouping. Jardine and Sibson in their book (see [3], p. 250) analyzed the well-known Mahalanobis, Majumdar and Rao data concerning anthropometric measurements on individuals from 23 local caste and tribal populations in India. To this goal they used two methods, so-called 1- and 2-clustering. In graph theory interpretation, having a valued complete graph with 23 vertices, they removed all edges with "length" greater than some critical value. The first method (1-clustering), denoted by A, determines clusters as maximal connected subgraphs of this graph. The second method (2-clustering), denoted by B, determines a cluster as a maximal connected subgraph of all vertices incident to at least two other vertices of this subgraph. A vertex not included in one of the above subgraphs forms either an isolated cluster or a two-element cluster with a vertex incident to it.

For three different critical levels, Jardine and Sibson applied both methods A and B to analyze the anthropometric data. We denote the results of method A (B) at the first level by A_1 (B_1) and at other levels by A_2, A_3 (B_2, B_3). In our interpretation we deal with 6 hypergraphs on the same set of vertices X, where $|X| = 23$ (number of castes). The resulting families of edges are given below, as well as the table of distances $\varrho_1(\cdot,\cdot)$ (see Table 2). The distance $\varrho_1(\cdot,\cdot)$ is defined by formula (3), and the Marczewski-Steinhaus distance between two sets is based on the measure $\mu_c(\cdot)$ given by formula (2). Families of edges are

$$E_{A_1} = \{\{1,2\},\{3\},\{4,5,10,11,12,13,15\},\{6\},\{7\},\{8\},\{9\},\{14\},\{16,17,18,19\},\{20\},\{21\},\{22\},\{23\}\},$$

$$E_{A_2} = \{\{1,2,3,4,5,6,7,10,11,12,13,15\},\{8\},\{9\},\{14,16,17,18,19,22\},\{20\},\{21\},\{23\}\},$$

$$E_{A_3} = \{\{1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22\}, \{9\}, \{23\}\},$$

$$E_{B_1} = \{\{1, 2\}, \{3\}, \{4, 5\}, \{5, 12\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10, 11, 12, 13, 15\}, \{14\}, \{16, 17\}, \{16, 18, 19\}, \{20\}, \{21\}, \{22\}, \{23\}\},$$

$$E_{B_2} = \{\{1, 2, 4, 5, 7, 10, 11, 12, 13, 15\}, \{2, 3\}, \{5, 6\}, \{8\}, \{9\}, \{14, 16, 17, 18, 19\}, \{18, 22\}, \{20\}, \{21\}, \{23\}\},$$

$$E_{B_3} = \{\{1, 2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15\}, \{2, 3\}, \{9\}, \{14, 15, 16, 17, 18, 19\}, \{18, 20\}, \{18, 22\}, \{20, 21\}, \{23\}\}.$$

TABLE 2. Distances $\varrho_1(\cdot, \cdot)$ between cluster analysis methods

	A_1	A_2	A_3	B_1	B_2
A_2	.666				
A_3	.800	.690			
B_1	.500	.750	.857		
B_2	.700	.500	.738	.700	
B_3	.708	.875	.666	.750	.708

Analysis of the table of these distances leads to the conclusion that critical levels, which play important roles in forming structures of the representing graph, are more decisive in establishing the results than methods A and B themselves.

3.2. Hierarchical clustering. Results of hierarchical clustering methods, both agglomerative and divisive, can be represented by a rooted tree. For example, in an agglomerative method, n objects are treated first as single one-element clusters (terminal vertices in the tree), then the two most similar are merged together forming a new cluster (non-terminal vertex in the tree), and so on. The process is completed, in $n-1$ consecutive steps, when all objects form one cluster (root of the tree). A similar process is involved in finding groups in divisive clustering methods.

As an illustration of a possible application of the proposed distance $\varrho_2(\cdot, \cdot)$ between directed rooted trees, we have analyzed results of grouping of 12 soil samples from different parts of Poland, described by 9 characteristics mainly of chemical type, by the use of 6 agglomerative hierarchical clustering methods. The data and the computer algorithm of these methods are presented with more details in the papers by Karoński and Caliński [4], [5], and Wishart [10]. The methods applied to group the data are those of

I. nearest neighbour,

II. furthest neighbour,

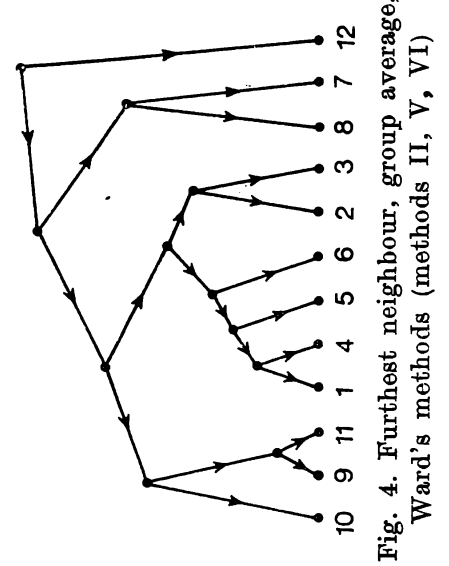

Fig. 4. Furthest neighbour, group average, Ward's methods (methods II, V, VI)

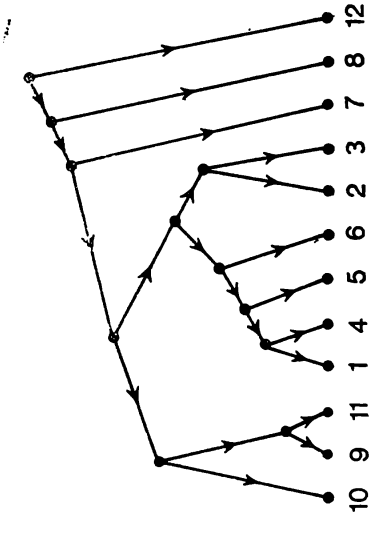

Fig. 6. Centroid method (method IV)

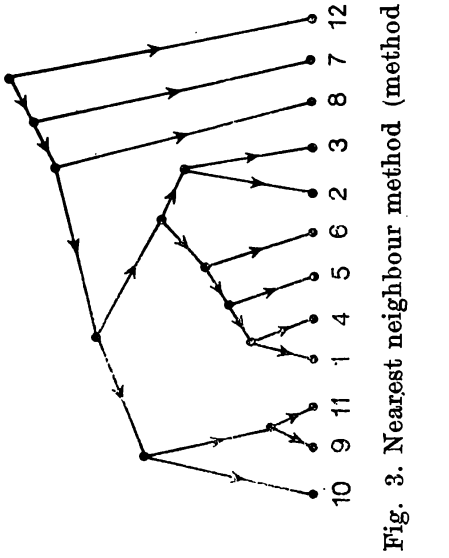

Fig. 3. Nearest neighbour method (method I)

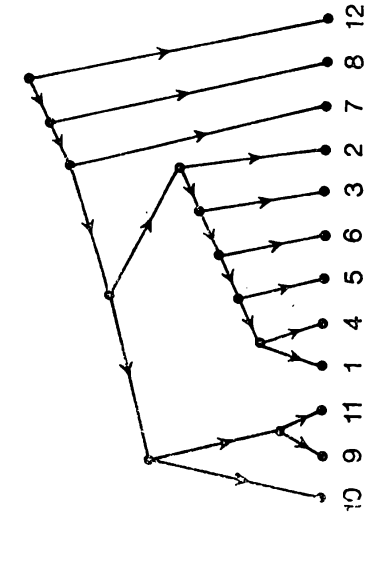

Fig. 5. Median method (method III)

III. median,
IV. centroid,
V. group average,
VI. Ward's.

The obtained trees are presented in Figs. 3-6. The results, if we consider only the tree representation, for furthest neighbour, group average and Ward's methods are identical.

We have applied the distance $\varrho_2(\cdot,\cdot)$ given by formula (4) using the Marczewski-Steinhaus distance between two edges of the appropriate hypergraph based on the measure $\mu_c(\cdot)$ (see formula (2)). The obtained distances are presented in Table 3 and indicate no major difference between the results of grouping with these 6 methods.

TABLE 3. Distances $\varrho_2(\cdot,\cdot)$ between 4 methods of hierarchical clustering

	I	II	III
II	.0826		
III	.0923	.1584	
IV	.0165	.0826	.0758

References

[1] C. Berge, *Graphs and hypergraphs*, North Holland, Amsterdam 1976.

[2] S. A. Boorman and D. C. Olivier, *Metrics on spaces of finite trees*, J. Math. Psychology 10 (1973), p. 26-59.

[3] N. Jardine and R. Sibson, *Mathematical taxonomy*, J. Wiley, New York 1971.

[4] M. Karoński and T. Caliński, *Grupowanie cech na podstawie współczynnika korelacji*, Algorytmy Biometryczne i Statystyczne 2 (1973), p. 95-103.

[5] — *Grupowanie obiektów wielocechowych na podstawie odległości euklidesowych*, ibidem 2 (1973), p. 117-129.

[6] E. Marczewski and H. Steinhaus, *On a certain distance of sets and the corresponding distance of functions*, Coll. Math. 6 (1958), p. 319-327.

[7] — *O odległości systematycznej biotopów*, Zastosow. Matem. 4 (1959), p. 195-203.

[8] W. M. Rand, *Objective criteria for the evaluation of clustering methods*, J. Amer. Statist. Assoc. 66 (1971), p. 846-850.

[9] S. Ulam, *Some ideas and prospects in biomathematics*, Annual Review of Biophysics and Bioengineering 1 (1972), p. 277-292.

[10] D. Wishart, *An algorithm for hierarchical classification*, Biometrics 25 (1969), p. 165-170.

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611, U.S.A.

INSTITUTE OF MATHEMATICS
ADAM MICKIEWICZ UNIVERSITY
60-769 POZNAŃ, POLAND

Received on 15. 9. 1976

M. KAROŃSKI (Gainesville, Florida) i Z. PALKA (Poznań)

O ODLEGŁOŚCI TYPU MARCZEWSKIEGO-STEINHAUSA MIĘDZY HIPERGRAFAMI

STRESZCZENIE

W pracy wprowadza się przestrzeń metryczną dla hipergrafów opartych na tym samym zbiorze wierzchołków. W szczególności rozpatrzono hipergrafy generowane przez drzewa zorientowane, mające identyczny zbiór wierzchołków wiszących. Jako podstawę rozważań przyjęto odległość między dwoma zbiorami, zaproponowaną przez Marczewskiego i Steinhausa. Pokazano także, że hipergrafy mogą odzwierciedlać grupowanie obiektów za pomocą różnych metod analizy skupień. Podano przykład porównania różnych metod grupowania.