Ewa Wędrowska

Classification of Objects on the Base of the Expected Information Value

Olsztyn Economic Journal 5/1, 78-89

2010

OLSZTYN ECONOMIC JOURNAL
Abbrev.: Olszt. Econ. J., 2010, 5(1)

DOI 10.2478/v10021-010-0007-1

CLASSIFICATION OF OBJECTS ON THE BASE OF THE EXPECTED INFORMATION VALUE

Ewa Wędrowska

Chair of Quantitative Methods
University of Warmia and Mazury in Olsztyn

K e y w o r d s: classification of structures, similarity of structures, expected value of information.

A b s t r a c t

In this paper applying the expected information value for determining the degree of dissimilarity of structures was proposed. The methodology proposed represents one of numerous attempts at employing the measures defined on the grounds of the information theory for investigating socio-economic phenomena. The expected value of information on transformation of the observed structure into another structure compared with it may be treated as the starting point in the central agglomeration procedure. The paper presents the hierarchic classification by means of full linking of counties in Warmińsko-Mazurskie voivodship according to the similarity of the structure of economic entities according to the PKD (Polish Classification of Business Activities) sections so that the possibility of employing the expected information value in the classification procedures was presented.

KLASYFIKACJA OBIEKTÓW NA PODSTAWIE WARTOŚCI OCZEKIWANEJ INFORMACJI

Ewa Wędrowska

Katedra Metod Ilościowych
Uniwersytet Warmińsko-Mazurski w Olsztynie

S ł o w a k l u c z o w e: klasyfikacja struktur, podobieństwo struktur, wartość oczekiwana informacji.

A b s t r a k t

W artykule zaproponowano wykorzystanie wartości oczekiwanej informacji do określania stopnia niepodobieństwa struktur. Zaproponowana metoda jest jedną z wielu prób wykorzystania miar zdefiniowanych na gruncie teorii informacji do badania zjawisk społeczno-ekonomicznych. Wartość oczekiwana informacji o transformacji obserwowanej struktury w inną, porównywaną z nią struk-

turą, można potraktować jako punkt wyjścia w centralnej procedurze aglomeracyjnej. Przeprowadzono klasyfikację hierarchiczną metodą pełnego wiązania powiatów województwa warmińsko-mazurskiego za względu na podobieństwo struktur podmiotów gospodarczych wg sekcji PKD, a zatem wskazano na możliwość wykorzystania wartości oczekiwanej informacji w procedurach klasyfikacyjnych.

Introduction

The increase in complexity of socio-economic phenomena contributes continually to the development of statistical methods used for investigating those phenomena. Classification procedures represent a rich group of tools used in socio-economic studies. In the methodology of sciences it is assumed that classification is the first objective among the fundamental objectives of a science being at the same time a tool and goal of cognition. The issues of classification have long been the subject of interest in numerous scientific disciplines and in modern science the development of classification methods was initially associated with biology. During the early 20th C., in natural sciences digital classification methods have been employed, which gave the beginning to the development of taxonomic methods. With the passage of years further classification procedures have been and still are developed that found applications in various disciplines of the knowledge. Polish scientists, starting with the outstanding anthropologist, demographer and statistician Jan Czekanowski, through the creators of the original taxonomic method (dendrite method) with Florek and Steinhaus as the leaders up to the works by Hellwig from the turn of the nineteen sixties and seventies, contributed significantly to the development of taxonomic methods. As concerns the new proposals, the taxonomy of structures proposed by Sokołowski (CHOMĄTOWSKI, SOKOŁOWSKI 1978), or the proposal for imparting of dynamism of the taxonomical methods presented by Grabiński (GRABIŃSKI 1984) should be mentioned. The consecutive years have brought a number of works concerning applications of taxonomical methods in socio-economic research. They included the proposals for applying the measures defined on the grounds of the information theory in the taxonomical procedures (WĘDROWSKA, ZAPOTOCZNA 2004, ROESKE-SŁOMKA 2008, CHEN, WANG 2008).

In this paper the methodology of objects classification on the base of the expected information value on structures characterizing the objects classified. The goal of the study was to show that the expected information value is the measure of dissimilarity of structures as a consequence of which it can be employed in taxonomical algorithms. The proposed methodology represents one of many attempts at implementation of methods and models defined on the grounds of the information theory in studies on socio-economic phenomena. The measures of the information theory found application, among others, in dynamic modeling or in the multidimensional data analysis.

Structure information expected value

In case of socio-economic issues the degree of similarity or dissimilarity of structures characterizing objects frequently is the subject of studies. In this paper, according to (*Taksonomia...* 1998). A structure will be interpreted as an object described by o vector of structure (or share) indicators. Determination of the S^n vector is justified only when the characteristic X that is the subject of study satisfies the property of additivity that is when the sum of values of the individual variants of that characteristic make economic sense.

The indicators of structure or indicators of share that are respectively the component of the structure S^n satisfy the following conditions:

(1) Normalizability: $0 \leq \alpha \leq 1$ (i= 1,...,n),

(2) Unit-sum condition: $\sum_{i=1}^{n} \alpha_i = 1$ (i = 1,...,n).

Investigation of changes and similarity between structures may be of statistical or dynamic character and as a consequence the analysis may cover the similarity of structures in an ndimensional space or testing the variability of structures over time.

We will consider the object of classification that is a countable, consisting of m-elements, set **O** of objects O_i, characterized by n-dimensional structures S_i^n. The criteria for division of the set of objects will be based on the function assigning to every pair of objects O_i, $O_j \in$ **O** the measure of mutual dissimilarity of structures S_i, S_j characterizing those objects. The criteria of classification are the functional defined on the set of all possible subsets G_1,..., G_p (where $p \leq m$) of the set **O** and defining the homogeneity of the individual subsets that are the effect of grouping and the degree of heterogeneity between the identified groups. The subsets G_1,..., G_p that result from the division should satisfy the conditions of:

(1) separability ($G_i \cap G_j = \emptyset$; $i \neq j$; i, j= 1,...,p),

(2) completeness $\bigcup_{i=1}^{p} G_i = \mathbf{O}$.

Choice of the appropriate measure of the distance (or similarity) between the classified structures is the starting point for the majority of taxonomical procedures. The choice of the measure with which the degree of similarity of structures in multidimensional space is determined is of major influence on the results of grouping or organizing (*Statystyczne metody...* 1999). Usually the measures of similarity of structures fulfill the function of the measures of the distance of their partial indicators. Because of the specificity of the problems of similarity of structures, that issue was a subject of numerous works. The

review of the major methods for measurement of similarity of structures was presented by MŁODAK (2006).

In this paper the measure defined on the grounds of the theory of information will be proposed that could be employed for determining the degree of dissimilarity of structures. In the paper by (THEIL 1979) the expected quantity of information contained in the message on transformation of the probabilities p_i ($i = 1,2,...,n$) into probabilities q_i for n mutually excluding events E_1, E_2,..., E_n is considered. Using that concept in the investigation of structures the quantity of information on the degree of dissimilarity of structures S_i^n. and S_j^n. can be determined. Let's the structure S_i^n. be expressed by the vector of structure (or share) indicators $[\alpha_i^1, \alpha_i^2,..., \alpha_i^n]$, and structure S_j^n. by the vector of indicators $[\alpha_j^1, \alpha_j^2,..., \alpha_j^n]$ satisfying the conditions of normalization and unit sum. The expected quantity of information about transformation of the structure S_i^n. treated as the base one into the structure S_j^n. is given by the formula:

$$I(S_j^n : S_i^n.) = \sum \alpha_i^k \log \frac{\alpha_j^k}{\alpha_i^k} \tag{1}$$

The expected value of information given by the formula (1) satisfies the following properties:

(a) $I(S_j^n : S_i^n) = 0$, if $\forall\ k = 1,2,...,n\ \alpha_i^k = \alpha_j^k$;

(b) $\alpha_i^k \log \frac{\alpha_j^k}{\alpha_i^k} > 0$ for $\alpha_i^k < \alpha_j^k$;

(c) $\alpha_i^k \log \frac{\alpha_j^k}{\alpha_i^k} < 0$ for $\alpha_i^k > \alpha_j^k$;

(d) $I(S_j^n : S_i^n) > 0$, if $S_i^n \neq S_j^n$

The value $I(S_j^n : S_i^n)$ informs about the degree of transformation between the base structure S_i^n and the structure S_j^n, that is the degree of similarity or dissimilarity of structures S_i^n and S_j^n. According to property (a) the expected information value $I(S_j^n : S_i^n)$ assumes the value equal to zero for two identical structures $S_i^n = S_j^n$, that is the structures for which each corresponding indicator $\alpha_i^k = \alpha_j^k$ for every $k = 1,2,...,n$. With appearance of increasing differences between the structures S_i^n and S_j^n the expected information value $I(S_j^n : S_i^n)$ is positive (property (d)) and increases to the infinity.

Measure $I(S_j^n : S_i^n)$ does not satisfy the condition of symmetry, which means that for different structures S_i^n and S_j^n the situation: $I(S_j^n : S_i^n) \neq I(S_i^n : S_j^n)$ occurs. The measures used for testing the similarity of structures are the most often

the measures representing the function of the distance between indicators of structures, as a consequence of which they satisfy the property of symmetry. There are however a few approaches that consider measures of similarity of structures showing lack of symmetry. In the literature critical opinions concerning the measures not satisfying the condition of symmetry can be found (MŁODAK 2006), but in the opinion of the author, in the studies on the similarity of structures the situations exist in which it is useful to apply measures of hat type. The conditions in which we accept that transformation of the structure S_i^n into the structure S_j^n is not equivalent to the transformation of the structure S_j^n into the structure S_i^n, that is in the situations when we treat one of the structures as the base structure or when the cost or weight of transformation is important can be the example. Moreover, use of the value of the expected quantity of information can be useful for investigating transformations of structures according to the dynamic approach.

Classification of the counties of Warmińsko-Mazurskie voivodship according to the structure of economic entities according to the PKD sections

Counties and cities possessing the rights of counties in Warmińsko-Mazurskie voivodship characterized by the number of entities of the national economy according to the selected sections of the PKD as at the 31st of December 2007 are elements of the set **O**. The X characteristic, which is the number of entities of the national economy, consists of nine variants A_k ($k = 1,...,9$), resulting from belonging of those entities to the specific sections of the Polish Classification of Business Activities (PKD).

The aim of the study is to determine the degree of dissimilarity between the structures covered. One of the methods of such studies is the graphic method of presenting the multidimensional data using the so-called symbolic graphs. The solution allows presenting each structure in the form of a symbolic drawing through which similarity of structures or individual properties of the structures can be identified. The star graphs are an example of symbolic graphs drawing of which for each structure starts in point P, in which all the rays have their beginning while maintaining identical angles between them (*Statystyczne metody...* 1999). Each of the rays symbolizes one of the components of the structure vector and the length of the ray is proportional to the value of the k coordinate of the vector S_i^n ($k = 1,2,...,9$) (Fig. 1). The symbolic graph offers the possibility of the initial assessment of similarity of the structures investigated.

Table 1

Structure of the entities of the national economy registered with the REGON register according to the selected sections of the PKD in 2007 (status as at the 31st of December)

County	Structure	Agriculture, forestry, hunting	Industry	Construc-tion	Trade and repair services	Hotels and restaurants	Transport, wareho-using and communica-tion	Financial services	Services for real properties and companies	Others
Braniewski	1.	0.0704	0.0752	0.0691	0.2786	0.0305	0.0372	0.0335	0.2041	0.2014
Działdowski	2.	0.0431	0.0980	0.1538	0.3000	0.0192	0.0575	0.0311	0.1206	0.1767
Elbląski	3.	0.0741	0.1407	0.0859	0.2706	0.0307	0.0600	0.0226	0.2092	0.1061
Iławski	4.	0.0570	0.1139	0.1046	0.3065	0.0206	0.0718	0.0322	0.1216	0.1718
Nowomiejski	5.	0.0696	0.1284	0.1269	0.2739	0.0139	0.0487	0.0317	0.1025	0.2043
Ostródzki	6.	0.0523	0.0958	0.1140	0.2757	0.0265	0.0583	0.0390	0.1679	0.1704
m. Elbląg	7.	0.0091	0.0928	0.0800	0.2676	0.0280	0.0714	0.0425	0.2331	0.1754
Bartoszycki	8.	0.0520	0.0834	0.0958	0.3009	0.0261	0.0630	0.0551	0.1354	0.1883
Kętrzyński	9.	0.0354	0.0752	0.0863	0.2915	0.0243	0.0632	0.0512	0.1811	0.1917
Lidzbarski	10.	0.0402	0.0956	0.1038	0.2940	0.0268	0.0505	0.0414	0.1768	0.1710
Mrągowski	11.	0.0511	0.0863	0.1149	0.2838	0.0688	0.0714	0.0282	0.1453	0.1501
Nidzicki	12.	0.0674	0.0922	0.1347	0.2900	0.0277	0.0463	0.0311	0.1233	0.1873
Olsztyński	13.	0.0616	0.1000	0.1197	0.2658	0.0287	0.0652	0.0300	0.1590	0.1700
m. Olsztyn	14.	0.0077	0.0742	0.0944	0.2652	0.0196	0.0778	0.0459	0.2218	0.1934
Szczycieński	15.	0.0885	0.0946	0.1233	0.2903	0.0339	0.0669	0.0267	0.1297	0.1462
Ełcki	16.	0.0298	0.0680	0.1098	0.2999	0.0291	0.0872	0.0349	0.1530	0.1883
Giżycki	17.	0.0484	0.0815	0.0972	0.2816	0.0516	0.0553	0.0280	0.1744	0.1820
Gołdapski	18.	0.0984	0.1144	0.1238	0.2250	0.0207	0.0551	0.0268	0.1662	0.1695
Olecki	19.	0.0628	0.0859	0.1333	0.2840	0.0321	0.0635	0.0321	0.1542	0.1522
Piski	20.	0.0780	0.0794	0.1006	0.2845	0.0560	0.0610	0.0246	0.1454	0.1705
Węgorzewski	21.	0.0886	0.0744	0.0815	0.2626	0.0433	0.0427	0.0304	0.1611	0.2154

Source: Central Statistical Office, www.stat.gov.pl

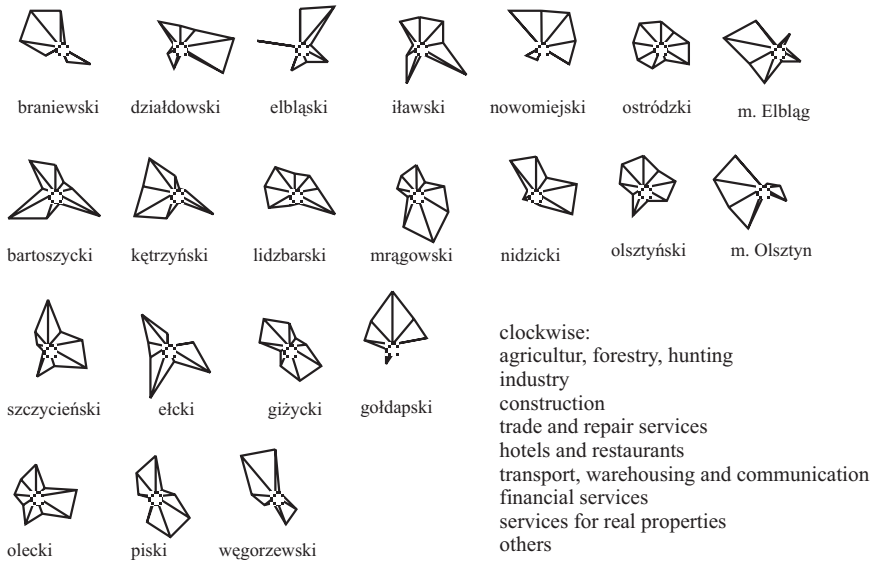

Fig. 1. Star graphs for the investigated structures
Source: own work using the STATISTICA software package

With the data on the structures S_i^9, S_j^9 (i, j = 1,2,...,21) available, the expected information values $I(S_j^n : S_i^n)$ were determined. In that way the asymmetric matrix of dissimilarities of values $I(S_j^n : S_i^n)$ assuming the structure with indicator i recorded in line i for the base structure (Tab. 2). As a consequence, the values recorded under the dissimilarities matrix diagonal are the values of information of the structure S_i^n into the structure S_j^n, where i is the indicator of the line corresponding to the structure S_i^n and j the indicator of the column corresponding to the structure S_j^n. With the increase in the degree of dissimilarity of structures the values of the expected quantity of information $I(S_j^n : S_i^n)$ increase. This confirms the hypothesis that the expected information value can be used for assessment of dissimilarity of structures. In case of the identical structures the measure $I(S_j^n : S_i^n)$ assumes the value of zero. Investigating the degree of dissimilarity of structures using the expected value of information on the transformation of structures may apply to transformations of dynamic character, i.e. where transformations of structures over time and dissimilarities of structures in statistical categories are observed, when the investigated structures are observed in an $n$multidimensional space.

The matrix of dissimilarities offers the starting point in the conducted clustering procedure. The structures describing the counties of Warmińsko-Mazurskie voivodship considering the entities of the national economy according to the PKD sections were classified according to the hierarchic method assuming the distance of the most distant neighborhood for the base. In the

method of the most distant neighborhood (complete connection, *complete-link*), the clusters were identified on the base of the highest expected information value $I(S_j^n : S_i^n)$ from among all the values for the structures belonging to the clusters connected. Two classifications were made where in the first one the expected information values recorded under the diagonal of the matrix of dissimilarities were considered so that the structure S_i^n. was treated as the base one. The values of $I(S_j^n : S_i^n)$ provide information on the transformation of the structure S_i^n into the structure S_j^n. The central agglomeration procedure was represented graphically in Figure 2 in the form of the dendrogram (tree of connections) indicating the order of connections between clusters. The hierarchy obtained allows presenting individual classes and structures contained in them.

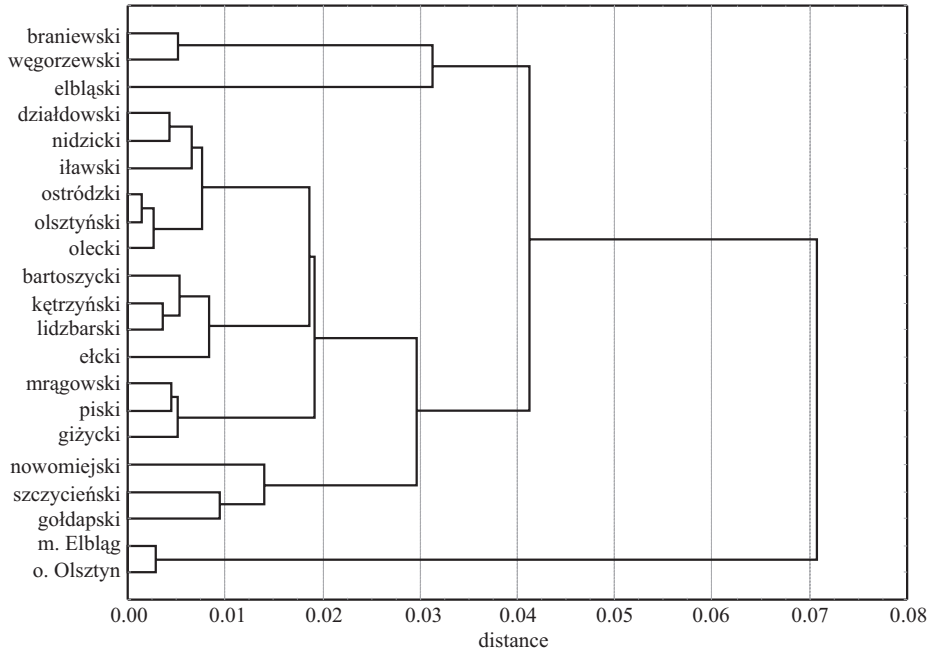

Fig. 2. Tree of connections in the hierarchic method with complete link on the base of the matrix of dissimilarities (expected information values from under the matrix diagonal)
Source: *own work using the STATISTICA software package.*

The selected hierarchic method belongs to the basic methods of classification. The following are listed as the major advantages of that method (*Statystyczna analiza...* 2009):

1. it functions according to one procedure,
2. the classification results are presented in the form of a sequence of classification, which allows controlling the classification process,
3. the classification results can be presented in the graphic form.

Table 2

Matrix of dissimilarity of structures

Counties	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.
1.	0.0000	0.0336	0.0261	0.0241	0.0318	0.0126	0.0274	0.0155	0.0116	0.0112	0.0276	0.0204	0.0159	0.0313	0.0245	0.0252	0.0099	0.0215	0.0198	0.0152	0.0052
2.	0.0317	0.0000	0.0366	0.0061	0.0086	0.0079	0.0390	0.0108	0.0180	0.0105	0.0225	0.0046	0.0078	0.0355	0.0133	0.0121	0.0190	0.0212	0.0072	0.0220	0.0294
3.	0.0267	0.0365	0.0000	0.0228	0.0378	0.0182	0.0364	0.0335	0.0328	0.0204	0.0256	0.0311	0.0164	0.0485	0.0186	0.0380	0.0229	0.0177	0.0189	0.0232	0.0330
4.	0.0239	0.0065	0.0237	0.0000	0.0070	0.0063	0.0357	0.0064	0.0146	0.0092	0.0191	0.0063	0.0052	0.0358	0.0077	0.0116	0.0158	0.0156	0.0074	0.0152	0.0212
5.	0.0333	0.0081	0.0389	0.0072	0.0000	0.0148	0.0575	0.0154	0.0287	0.0199	0.0378	0.0063	0.0126	0.0549	0.0147	0.0279	0.0304	0.0142	0.0164	0.0276	0.0261
6.	0.0117	0.0077	0.0175	0.0060	0.0137	0.0000	0.0200	0.0047	0.0058	0.0015	0.0130	0.0060	0.0013	0.0210	0.0093	0.0085	0.0062	0.0113	0.0027	0.0109	0.0139
7.	0.0437	0.0433	0.0546	0.0435	0.0685	0.0290	0.0000	0.0342	0.0146	0.0192	0.0428	0.0572	0.0391	0.0028	0.0727	0.0190	0.0284	0.0776	0.0444	0.0600	0.0667
8.	0.0153	0.0111	0.0321	0.0061	0.0152	0.0047	0.0271	0.0000	0.0049	0.0054	0.0171	0.0083	0.0074	0.0255	0.0132	0.0077	0.0105	0.0217	0.0084	0.0135	0.0147
9.	0.0120	0.0186	0.0327	0.0145	0.0289	0.0062	0.0114	0.0048	0.0000	0.0036	0.0221	0.0185	0.0116	0.0105	0.0262	0.0060	0.0095	0.0311	0.0140	0.0213	0.0204
10.	0.0114	0.0104	0.0203	0.0090	0.0189	0.0016	0.0147	0.0054	0.0035	0.0000	0.0157	0.0101	0.0053	0.0167	0.0167	0.0088	0.0062	0.0202	0.0068	0.0155	0.0179
11.	0.0248	0.0170	0.0250	0.0143	0.0284	0.0103	0.0331	0.0152	0.0192	0.0130	0.0000	0.0136	0.0087	0.0369	0.0104	0.0129	0.0052	0.0244	0.0069	0.0047	0.0191
12.	0.0197	0.0042	0.0308	0.0065	0.0060	0.0062	0.0438	0.0085	0.0179	0.0100	0.0162	0.0000	0.0050	0.0425	0.0059	0.0162	0.0122	0.0116	0.0047	0.0101	0.0129
13.	0.0143	0.0075	0.0158	0.0050	0.0114	0.0013	0.0263	0.0080	0.0111	0.0051	0.0108	0.0047	0.0000	0.0275	0.0051	0.0101	0.0067	0.0070	0.0016	0.0078	0.0126
14.	0.0502	0.0409	0.0702	0.0471	0.0707	0.0319	0.0029	0.0351	0.0148	0.0225	0.0516	0.0587	0.0429	0.0000	0.0789	0.0165	0.0350	0.0844	0.0472	0.0687	0.0739
15.	0.0238	0.0114	0.0198	0.0071	0.0140	0.0089	0.0474	0.0136	0.0239	0.0149	0.0109	0.0058	0.0050	0.0491	0.0000	0.0188	0.0135	0.0091	0.0035	0.0058	0.0156
16.	0.0245	0.0123	0.0413	0.0131	0.0296	0.0089	0.0178	0.0084	0.0059	0.0083	0.0153	0.0169	0.0115	0.0142	0.0236	0.0000	0.0102	0.0365	0.0115	0.0198	0.0278
17.	0.0093	0.0172	0.0226	0.0138	0.0254	0.0055	0.0207	0.0104	0.0089	0.0056	0.0052	0.0115	0.0063	0.0241	0.0141	0.0098	0.0000	0.0208	0.0074	0.0049	0.0093
18.	0.0197	0.0187	0.0167	0.0150	0.0131	0.0103	0.0442	0.0215	0.0267	0.0173	0.0279	0.0113	0.0065	0.0461	0.0094	0.0297	0.0206	0.0000	0.0104	0.0174	0.0159
19.	0.0178	0.0067	0.0191	0.0071	0.0157	0.0027	0.0308	0.0086	0.0131	0.0064	0.0084	0.0045	0.0016	0.0312	0.0036	0.0103	0.0075	0.0106	0.0000	0.0069	0.0154
20.	0.0144	0.0189	0.0235	0.0128	0.0225	0.0097	0.0393	0.0134	0.0190	0.0136	0.0045	0.0092	0.0070	0.0424	0.0055	0.0161	0.0045	0.0152	0.0066	0.0000	0.0066
21.	0.0052	0.0286	0.0312	0.0207	0.0243	0.0132	0.0401	0.0145	0.0174	0.0156	0.0189	0.0134	0.0128	0.0424	0.0157	0.0245	0.0083	0.0159	0.0157	0.0067	0.0000

Source: own computations.

In the hierarchic methods, the rule of clustering is not clearly determined. To solve that problem it is proposed to investigate the dendrogram as concerns the differences in the distance between the consecutive nodes (*Statystyczne metody*... 1999). Computing the threshold value determining the optimum concentration in hierarchic methods is also proposed. The proposal presented in the work (JABŁOŃSKI, ROBASZEK 2000), where the threshold value is the sum of the arithmetic average and two standard deviations from the minimum values obtained from all the columns is also a popular approach. In this paper that approach was assumed and the computed threshold value for all the expected information was 0.0109. The results of the first classification are presented in Table 3.

Table 3
Results of clustering using the hierarchic clustering procedure by means of complete linking method

Cluster	Counties belonging to the cluster
I	działdowski, nidzicki, iławski, olsztyński, ostródzki, olecki
II	bartoszycki, kętrzyński, lidzbarski, ełcki
III	mrągowski, piski, giżycki
IV	m. Elbląg, m. Olsztyn
V	braniewski, węgorzewski
VI	szczycieński, gołdapski
VII	nowomiejski
VIII	elbląski

Source: own work.

Each of the clusters contains the most uniform objects, i.e. counties of Warmińsko-Mazurskie voivodship similar in the structure of the entities of national economy according to the PKD sections. Such division coupled with the relatively low threshold value resulted in seven clusters with small populations. There are also single element clusters because there are structures that are dissimilar to the others. Such a division results from specific economic conditions in the counties covered.

On the base of the expected information values recorded above the diagonal of the matrix of dissimilarities the next classification was prepared. In that second classification the structure S_j^n. was treated as the base structure and the value $I(S_i^n : S_j^n)$ represented the quantity of information on the transformation of the S_j^n structure into the S_i^n structure. The central agglomeration procedure for that classification is presented graphically in figure 3 in the form of the tree of connections indicating the order of connections between clusters. The obtained hierarchy allows presenting individual clusters and structures con-

tained in them at the assumed, as previously, threshold value indicating the dissimilarity of structures at the level of 0.0109.

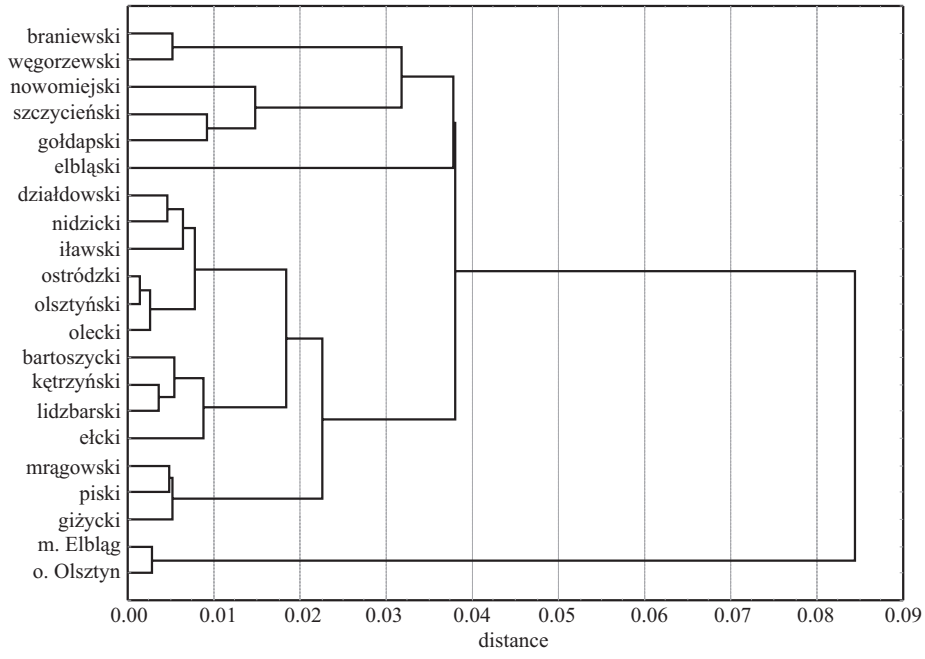

Fig. 3. Tree of connections in the hierarchic method with complete link on the base of the matrix of dissimilarities (expected information values from above the matrix diagonal)
Source: own work using the STATISTICA software package.

Both classifications were made according to the same methodology with the same central procedure and the same threshold value but based on the different expected information value. Despite the differences in the expected information values ($I(S_i^n : S_j^n) \neq I(S_j^n : S_i^n)$ for $S_i^n \neq S_j^n$) the same classification result, that is cluster of the same composition were obtained (Fig. 3). This means that the expected information value $I(S_j^n : S_i^n)$ can be treated as the measure of dissimilarity of structures and used as the starting point for the classification procedure despite not maintaining the property of symmetry by the measure proposed.

Conclusion

In the case of socio-economic issues there is a frequent need for identifying transformations of structures over time and assessment of similarity or

dissimilarity of structures of static nature. The paper presents the possibility of using the expected quantity of information on the transformation of the base structure S_j^n into the structure S_i^n for identification of the degree of dissimilarity of those structures. The methodology for assessment of the degree of dissimilarity of structures proposed in this paper may expand the set of taxonomical methods of multidimensional data analysis. Comparing structures from both the static and the dynamic perspective is not a new issue. The methodology for assessment of dissimilarity of structures proposed in this paper represents a modification of the measure defined on the grounds of the theory of information. The expected information value $I(S_j^n : S_i^n)$ may also be the starting point in the classification procedures and form the base for determination of the clusters of the most similar structures. The example of application of the methodology proposed presented in the paper indicates the possibility of employing the methodology presented in studies on socio-economic phenomena.

Translated by JERZY GOZDEK

Accepted for print 6.01.2010

References

CHEN C., WANG L. 2008. *Product platform design through clustering analysis and information theoretical approach*. International Journal of Production Research. 46(15): 4259–4284

CHOMĄTOWSKI S., SOKOŁOWSKI A. 1984. *Taksonomia struktur*. Przegląd Statystyczny, 2: 217–222.

GRABIŃSKI T. 1984. *Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych*. Seria specjalna: Monografie, 61, AE w Krakowie, Kraków.

JABŁOŃSKI R., ROBASZEK A. 2000. *Przestrzenne zróżnicowanie warunków środowiskowych w powiatach województwa łódzkiego*. Seria: Analizy statystyczne, Urząd Statystyczny w Łodzi, Łódź.

MŁODAK A. 2006. *Analiza taksonomiczna w statystyce regionalnej*. Wyd. Dyfin, Warszawa.

ROESKE-SŁOMKA I. 2008. *Klasyfikacja obiektów na podstawie entropii*. Wiadomości Statystyczne, 10: 22–29.

Statystyczna analiza danych z wykorzystaniem programu R. 2009. Ed. M. WALESIAK, E. GATNAR. PWN, Warszawa.

Statystyczne metody analizy danych. 1999. Ed. W. OSTASIEWICZ. AE we Wrocławiu, Wrocław.

Taksonomia struktur w badaniach regionalnych. 1998. Ed. D. STRAHL. AE we Wrocławiu, Wrocław.

THEIL H. 1979. *Zasady ekonometrii*. PWN, Warszawa.

WĘDROWSKA E., ZAPOTOCZNA M. 2004. *Zastosowanie metod taksonomicznych do oceny efektywności zagospodarowania substancji mieszkaniowej*. Prace Naukowe Akademii Ekonomicznej we Wrocławiu, 1023: 668–678.