OLSZTYN ECONOMIC JOURNAL
Abbrev.: Olszt. Econ. J., 2011, 6(1)

APPLICATION OF KULLBACK-LEIBLER RELATIVE ENTROPY FOR STUDIES ON THE DIVERGENCE OF HOUSEHOLD EXPENDITURES STRUCTURES

Ewa Wędrowska

Chair of Quantitative Methods
University of Warmia and Mazury in Olsztyn

K e y w o r d s: Kullback-Leibler relative entropy, Shannon's entropy, similarity of structures, divergence of structures.

A b s t r a c t

The paper proposes the possibility of employing the methods defined on the grounds of the information theory to research on socioeconomic phenomena. The presented measures are Shannon's entropy and Kullback-Leibler relative entropy (divergence) applied for quantification of the degree of concentration of structures and the degree of divergence between structures analyzed in the dynamic approach respectively. The paper presents studies on the degree of divergence between structures of average monthly per capita expenditures in households in Poland during the years 2000–2008.

WYKORZYSTANIE ENTROPII WZGLĘDNEJ KULLBACKA-LEIBLERA DO BADANIA ROZBIEŻNOŚCI STRUKTUR WYDATKÓW GOSPODARSTW DOMOWYCH

Ewa Wędrowska

Katedra Metod Ilościowych
Uniwersytet Warmińsko-Mazurski w Olsztynie

S ł o w a k l u c z o w e: entropia względna Kullbacka-Leiblera, entropia Shannona, podobieństwo struktur, rozbieżność struktur.

A b s t r a k t

W artykule zaproponowano możliwość wykorzystania metod zdefiniowanych na gruncie teorii informacji do badania zjawisk społeczno-ekonomicznych. Prezentowane miary to entropia Shannona oraz entropia względna (dywergencja) Kullbacka-Leiblera wykorzystane odpowiednio do kwantyfikacji stopnia koncentracji struktur oraz stopnia rozbieżności między strukturami analizowanymi w ujęciu dynamicznym. W artykule zbadano stopień rozbieżności między strukturami przeciętnych miesięcznych wydatków na osobę w gospodarstwach domowych w latach 2000–2008 w Polsce.

Introduction

Analysis of socioeconomic phenomena is also frequently accompanied by comparison of the level of those phenomena during a certain period with the level of those phenomena during another period. Studies on similarity or dissimilarity of structures characterizing economic phenomena changing over time represent a special case of such analyses. Any comparative analyses concerning the dynamics of socioeconomic processes should be carried out by applying appropriate statistical methods allowing quantification that is methodologically correct and univocal for interpretation. At the same time the increase in the level of complexity of the phenomena investigated is continually contributing to the development of statistical methods applied to research on such phenomena. The wide spectrum of methods allowing comparison of structures is offered by the taxonomy of structures, although, in their majority they are measures of similarity (or dissimilarity) that are the functions of the metrics of the distance between the components of such structures. In this paper application of Kullback-Leibler entropy (divergence) for quantification of the level of divergence between structures according to the dynamic approach is proposed. Shannon's entropy was also used for investigating the level of concentration of the structures.

The article aims at presenting the potential for applying the methods defined on the grounds of the information theory in the studies on socioeconomic phenomena. The application goal of the paper is to investigate the level of divergence between structures of the average monthly per capita expenditures in households during the years 2000–2008 using the Kullback-Leibler measure.

Shannon's entropy

In this paper, according to the definition by Strahl (*Taksonomia struktur*... 1998), the structure will be understood as the object described by the structure (or share) indicators' vector. Determination of the S^n vector is justified in case when the characteristic X that is subject to the investigation satisfies the attribute of additivity that is when the sum of the values of the individual variants of the characteristic makes economic sense.

Indicators of structure (or indicators of share) α_i for $i = 1, 2,..., n$ that are respective components of the structure S^n, satisfy the following conditions:

(1) Normality: $0 \leq \alpha_i \leq 1$ $(i = 1, 2,..., n)$,

(2) Condition of unit sum: $\sum_{i=1}^{n} \alpha_i$ (i = 1, 2,..., n).

The indicators of structure α_i for $i = 1, 2,..., n$ represent the relative numbers of occurrences of specified variants of the characteristic X in the investigated population. Knowledge of the indicators of structure will be used in this paper for quantification of the level of concentration of the X characteristic value and quantification of the divergence and dissimilarity with other standardized structures in both spatial and dynamic format.

The characteristics of the distribution of structure indicators $S^n = [\alpha_1, \alpha_2,..., \alpha_n]$ concerning the degree of diversification and concentration may be investigated by means of Shannon's entropy. The Shannon's entropy of S^n structure is defined as follows:

$$H_S(S^n) = H_S(\alpha_1,\alpha_2,...,\alpha_n) = \sum_{i=1}^{n} \alpha_i \log_2 \frac{1}{\alpha_i} \tag{1}$$

The Shannon's entropy H_S given by the formula (1) satisfies the characteristics specified, e.g. in the works by (PRZYBYSZEWSKI, WĘDROWSKA 2005, LAVENDA 2005):

1. it is a non-negative value, $\forall \alpha_i \in [0,1]\ H_S(S^n) \geq 0$,
2. it assumes the value of 0, when one of the structure coefficients $\alpha_i = 1$ for a certain i ($i = 1, 2,..., n$), the remaining coefficients are equal to 0,
3. satisfy the characteristic of symmetry: $H(\alpha_1,\alpha_2,...,\alpha_n) = H(\alpha_{(1)},\alpha_{(2)},...,\alpha_{(n)}$,
4. assumes the highest value equal to $H_S(S^n) = \log_2 n$, when all the structure coefficients α_i are equal to each other for i = 1, 2,..., n:

$$\alpha_1 = \alpha_2 = ... = \alpha_n$$

5. it is concave: $\forall \alpha_i \in [0, 1]\ \frac{\delta^2}{\delta {\alpha_i}^2} H_S(x) \leq 0$.

Shannon's entropy H_S of the structure $S^n = [\alpha_1,\alpha_2,...,\alpha_n]$ is treated as the measure of uncertainty related to the distribution of the coefficients of structure α_i for $i = 1, 2,..., n$. The value of entropy H_S depends exclusively on the frequency of appearance of the i-variant of X characteristic, i.e. the indicators of structure (or share). If structure S^n has the form of [0, 0,..., 1] this means that the fund of the investigated X characteristic is concentrated in a single variant. Entropy $H_S(0, 0,..., 1) = 0$, which means that there is no uncertainty related to achievement of characteristic X, and the distribution of the characteristic is determined. The attribute of symmetry that Shannon's entropy possesses causes that component $\alpha_i = 1$ ($i = 1, 2,..., n$) may be any i coordinate of structure S^n. On the other hand the maximum uncertainty as concerns obtaining one of the variants of the X characteristic is linked to the

presence of the structure $\left[\frac{1}{n}, \frac{1}{n},...,\frac{1}{n}\right]$. The total deconcentration taking place then accompanies the situation when the entropy $H_S\left(\frac{1}{n}, \frac{1}{n},...,\frac{1}{n}\right)$ is not maximal for the structure vector with n components. This means that the distribution of the investigated characteristic transforms into an even one. The values of entropy $H(S^n)$ are standardized within the range of $H(S^n) \in [0, \log_2 n]$, and as a consequence, knowledge of $H(S^n)$ may be useful for identification of the level of concentration of the characteristic. Understanding of concentration using entropy applies to concentration of units around certain values.

Kullback-Leibler relative entropy

Knowledge of measures of similarity of structures characterizing the investigated objects or phenomena is the starting point for the majority of taxonomic procedures. The measure of similarity of structures usually is a function of the measures of the distance of their partial indicators. As a consequence of involvement in numerous studies on similarity of structures undertaken in relation to socioeconomic phenomena that issue has been presented in many publications. The review of the most important methods for measurement of the similarity of structures has been presented in the works by, e.g. NOWAK (1990) and MŁODAK (2006).

In this paper the measure defined on the grounds of the information theory will be used for quantification of the level of dissimilarity of structures. The relative entropy also referred to as the Kullback-Leibler (KL) divergence was proposed by Kullback and Leibler in 1951 and found numerous applications, in particular for investigating the "distance" between two distributions of probability $\{p(x_i)\}$ and $\{q(x_i)\}$ (DHILLON et al. 2003, ZHANG, JIANG 2008) or two models: actual $f(x)$ and theoretical $g(x, \theta)$ (ASADI et al. 2005, PIŁATOWSKA 2009).

The study of similarity of structures is of static or dynamic nature and as a consequence the analysis of similarity of structures is considered in the n-dimensional space or the variability of structures over time is investigated. In this paper the similarity of structures according to the dynamic approach will be studied. The structures with n components will be considered: structure S_t^n characterizing the investigated phenomenon at the time t expressed by the vector of structure (or share) indicators $S_t^n = [\alpha_{1t}, \alpha_{2t},..., \alpha_{nt}]$ structure S_τ^n characterized the investigated phenomenon at the time τ expressed by the vector of components $S_\tau^n = [\alpha_{1\tau}, \alpha_{2\tau},..., \alpha_{n\tau}]$. The components of vectors $[\alpha_{1t}, \alpha_{2t},..., \alpha_{nt}]$ and $[\alpha_{1\tau}, \alpha_{2\tau},..., \alpha_{n\tau}]$ satisfy the conditions of standardization and unit sum.

Kullback-Leibler relative entropy for the pair of structures S_t^n and S_τ^n is defined by the formula (DHILLON et al. 2003):

$$\mathrm{KL}(S_t^n,S_\tau^n) = \sum_{i=1}^{n}\alpha_{it} \log \frac{\alpha_{it}}{\alpha_{i\tau}} \tag{2}$$

The KL divergence is the measure of divergence, dissimilarity between two structures. In the formula format defined in that way structure S_τ^n defined in the time τ is treated as the base structure. In the literature the term of Kullback-Leibler "distance" appears frequently but that is a misleading term as KL relative entropy does not satisfy the characteristics of distance metrics, i.e. the conditions of symmetry and inequality of triangle (DHILLON et al. 2003).

Kullback-Leibler relative entropy may be expressed as the difference between the so-called cross entropy of structures S_t^n and S_τ^n and Shannon's entropy of structure S_t^n (HUN, YANG 2007):

$$\mathrm{KL}(S_t^n,S_\tau^n) = \sum_{i=1}^{n}\alpha_{it} \log \frac{\alpha_{it}}{\alpha_{i\tau}} =$$

$$= \alpha_{1t} \log \frac{\alpha_{1t}}{\alpha_{1\tau}} + \alpha_{2t} \log \frac{\alpha_{2t}}{\alpha_{2\tau}} +\ldots+ \alpha_{nt} \log \frac{\alpha_{nt}}{\alpha_{n\tau}} =$$

$$=\alpha_{1t} (\log \alpha_1 - \log \alpha_{1\tau}) + \alpha_{2t} (\log \alpha_{2t} - \log \alpha_{2\tau} +\ldots+ \alpha_{nt} (\log \alpha_{nt} - \log \alpha_{n\tau}) =$$

$$= \sum_{i=1}^{n} \alpha_{it} \log \alpha_{it} - \sum_{i=1}^{n} \alpha_{it} \log \alpha_{i\tau} =$$

$$= \sum_{i=1}^{n} \alpha_{it} \log \frac{1}{\alpha_{1\tau}} - \sum_{i=1}^{n} \alpha_{it} \log \frac{1}{\alpha_{1t}} =$$

$$= H_S(S_t^n, S_\tau^n) - H_S(S_t^n) \tag{3}$$

H_S (S_t^n, S_τ^n) entropy is called the cross entropy (ZHANG, JIANG 2008). The more similar the structures S_t^n and S_τ^n are the more the cross entropy aims at Shannon's entropy $H_S(S_t^n)$, hence the difference in formula (3) aims at zero. For identical structures $S_t^n = S_\tau^n$ the equality of cross entropy and Shannon's entropy takes place $H_S(S_t^n, S_\tau^n)$, which means that for identical structures the Kullback-Leibler relative entropy is zero. Formula (3) allows intuitive cognition of the KL divergence as the "cost" of identifying the indefiniteness of the distribution of structures S_t^n when the indefiniteness of the distribution of structure S_τ^n is known.

The values of the KL measure are always non-negative and unlimited, which means that with appearance of increasing differences between structures S_t^n and S_τ^n they increase to infinity (DHILLON et al. 2003). The KL relative entropy is asymmetric, which means that $KL(S_t^n, S_\tau^n) \neq KL(S_\tau^n, S_t^n)$ for $S_t^n \neq S_\tau^n$, and that is why that measure should not be treated as the distance between the structures but as the divergence while considering one of the structures to be the base structure (WĘDROWSKA 2010). In the literature proposals of a symmetric measure being a function of Kullback-Leibler divergence exist (CAVANAUGH 1999, HUNG, YANG 2007).

Study of the divergence of household expenditures structures

Studies on the budgets of households play an important role in the analyses concerning the living standards of the people. Next to the information on incomes and expenditures of specific population groups it also provides the information on the level and structure of expenditures. In the study of the households; expenditures structure it is important to investigate whether divergences in the observed structure over a certain period of time exist.

The study covered the structure of the total average monthly per capita expenditures in the household during the years 2000–2008. The data considered originate from the publication by the Central Statistical Office concerning the budgets of households containing results of studies for 2008 (*Budżety*... 2009).

Food and non-alcoholic beverages have the highest share in the structure of expenditures in each individual year although it can be noticed that the share decreases systematically and 2008 was the lowest (at 25.56% of total expenditures). Expenditures related to the use of the apartment that range from 17.88% to 21.01% of total expenditures represent another important item in the structure of expenditures.

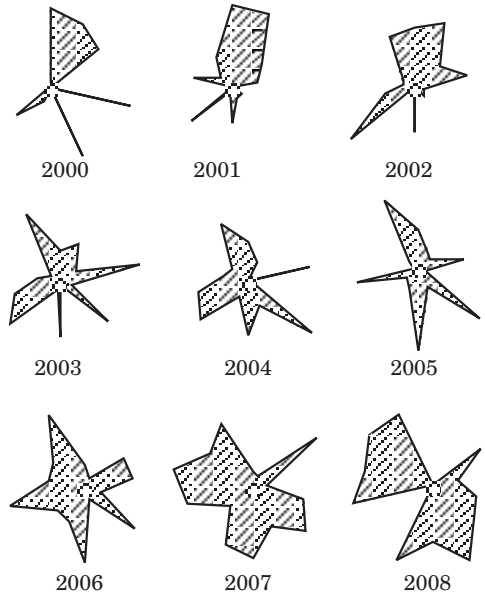
Quantification of the divergences between the structures of expenditures during the years 2000–2008 will be done applying the KL relative entropy. The symbolic graph that is the graphic illustration of the multidimensional data offers the possibility of the initial assessment of similarity of the investigated structures (Fig. 1). Elements of the star graph corresponding to structures of expenditures during consecutive years differ from each other, which is indicated by the radii representing the identified elements of the structures.

Table 1

Structure of average total monthly per capita expenditures if households during the years 2000–2008

Item	Year								
	2000	2001	2002	2003	2004	2005	2006	2007	2008
Expenditures of households	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Food and non-alcoholic beverages	0.3082	0.3096	0.2954	0.2777	0.2808	0.2812	0.2714	0.2664	0.2556
Alcoholic beverages, tobacco products and narcotic drugs	0.0300	0.0305	0.0299	0.0285	0.0274	0.0273	0.0268	0.0270	0.0263
Clothing and shoeing	0.0552	0.0528	0.0525	0.0512	0.0493	0.0507	0.0539	0.0571	0.0550
Use of apartment	0.1788	0.1885	0.1992	0.2101	0.2026	0.1965	0.1973	0.1841	0.1889
Apartment equipment and running the household	0.0594	0.0488	0.0500	0.0502	0.0490	0.0497	0.0510	0.0553	0.0546
Health	0.0444	0.0452	0.0453	0.0490	0.0505	0.0503	0.0491	0.0494	0.0483
Transport	0.0994	0.0878	0.0855	0.0857	0.0907	0.0891	0.0877	0.0932	0.1007
Communication	0.0351	0.0430	0.0450	0.0468	0.0468	0.0531	0.0515	0.0502	0.0475
Recreation and culture	0.0669	0.0652	0.0644	0.0655	0.0677	0.0684	0.0714	0.0760	0.0795
Education	0.0144	0.0148	0.0161	0.0153	0.0151	0.0131	0.0140	0.0137	0.0125
Restaurants and hotels	0.0140	0.0139	0.0162	0.0173	0.0176	0.0185	0.0196	0.0190	0.0200
Other goods and services	0.0494	0.0511	0.0498	0.0504	0.0505	0.0496	0.0510	0.0529	0.0523
Pocket money	0.0084	0.0096	0.0096	0.0095	0.0101	0.0095	0.0119	0.0132	0.0156
Other expenditures	0.0365	0.0394	0.0413	0.0429	0.0418	0.0429	0.0433	0.0424	0.0431

Source: Central Statistical Office.



Left to right
Food and non-alcoholic beverages
Alcoholic beverages, tobacco products, narcotic drugs
Clothing and shoeing
Use of apartment
Apartment equipment and running the household
Health
Transport
Communic
Recreation and culture
Education
Restaurants and hotels
Other goods and services
Pocket money
Other expenditures

Fig. 1. Star graph of the structures investigated

Source: own work with assistance of the STATISTICA software package

On the base of formula (1) the value of Shannon's entropy was determined for each of the structures of the total average per capita expenditures in households during the years 2000–2008, and next the values of the Kullback-Leibler relative entropy were computed assuming the structure from the period immediately preceding the analyzed period and the structure of 2000 as the base structure. The results of computations are presented in Table 2.

Table 2

Shannon's entropy and KL relative entropy for the structures of average monthly expenditures during the years 2000–2008

Year	Shannon's entropy $H_S(S^n)$	Kullback-Leibler relative entropy $H_S(S_t^n, S_\tau^n)$	
		base structure for the period $\tau = t - 1$	base structure for the period $\tau = 2000$
2000	3.1736	–	–
2001	3.1727	0.00452	0.00452
2002	3.1973	0.00146	0.00755
2003	3.2192	0.00170	0.01310
2004	3.2221	0.00067	0.01115
2005	3.2311	0.00103	0.01364
2006	3.2615	0.00112	0.01551
2007	3.2919	0.00179	0.01427
2008	3.3024	0.00158	0.01897

Source: own computations.

Knowledge of Shannon's entropy may serve quantification of the level of concentration of the distribution of the average monthly expenditures funds. The values of entropy $H_S(S_t^n)$ Sfor structures representing consecutive years are similar and show minor differences from the maximum value of the entropy for the structure with 14 components at $\log_2 14 = 3{,}8074$. As a consequence we can talk about poor concentration of the average monthly expenditures in one of the variants that is the expenditures on food and non-alcoholic beverages.

Investigation of the level of divergence between the structures applying the chain approach assuming the structure from the period immediately preceding the analyzed period as the base structure indicates low level of divergences. The highest value of the KL measure occurred in the divergence between the structure of the average monthly expenditures in 2001 and the structure of those expenditures in 2000. Consecutive values of Kullback-Leibler relative entropy indicate minor divergences between the analyzed structures.

Quantification of the level of divergence between the structure of the average per capita expenditures in households during the consecutive years and the structure of 2000 indicate slightly larger differences. The structure of expenditures in 2001 was the closest to that of 2000. The consecutive years indicate increasing divergences of structures from that of the year 2000. The largest divergence occurred in case of the structure for 2008, which was a consequence of the decrease in share of the expenditures on food and non-alcoholic beverages by ca. 5 percent points as compared to the year 2000.

Conclusion

The paper presents the possibility of applying Shannon's entropy and Kullback-Leibler relative entropy in studies on properties of the structures of average monthly per capita expenditures in households during the years 2000–2008. The results obtained indicate that the studied structures are characterized by poor level of concentration and small changes during the period considered. The divergences showing the largest variability in the structure of expenditures were observed for the structures of 2008 as compared to the structure of 2000 although the level of those changes is minor as indicated by the value of the Kullback-Leibler measure.

Translated by JERZY GOZDEK

Accepted for print 14.12.2010

References

ASADI M., EBRAHIMI N., SOOFI E.S. 2005. *Dynamic generalized information measures*. Statistics & Probability Letters, 71: 85–98.

Budżety gospodarstw domowych w 2008 roku. 2009. Informacje i opracowania statystyczne, GUS, Warszawa.

CAVANAUGH J. 1999. *A large-sample selection criterion based on Kullback's symmetric divergence*. Statistics & Probability Letters, 42: 333–343.

DHILLON I.S., MALLELE S., KUMAR R. 2003. *A divisive information – theoretic feature clustering algorithm for text classification*. Journal of Machine Learning Research, 3: 1265–1287.

HUNG W.L., YANG M.S. 2007. *On the J-divergence of intuitionistic fuzzy seta with its application to pattern recognition*, Information Sciences 178: 1641–1650.

LAVENDA B.H. 2005. *Mean Entropies*. Open System Infor. Dyn., 12: 289–302.

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

NOWAK E. 1990. *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych*. PWE, Warszawa.

PIŁATOWSKA M. 2009. *Prognozy kombinowane z wykorzystaniem wag Akaike'a*. Acta Universitatis Nicolai Copernici. Ekonomia, XXXIX: 51–62.

PRZYBYSZEWSKI R., WĘDROWSKA E. 2005. *Algorytmiczna teoria entropii*. Przegląd Statystyczny, 2(52): 85–102.

Taksonomia struktur w badaniach regionalnych. 1998. Red. D. Strahl. Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław.

WĘDROWSKA E. 2010. *Classification of objects on the base of the expected information value*. Olsztyn Economic Journal, 5(1): 78–89.

ZHANG Q-S, JIANG Y-J. 2008. *A note on information entropy measures for vague sets and its applications*. Information Sciences, 178: 4184–4191.