PRZEGLĄD STATYSTYCZNY
R. LVIII – ZESZYT 1-2 – 2011

EMILIA TOMCZYK

APPLICATION OF MEASURES OF ENTROPY, INFORMATION CONTENT AND DISSIMILARITY OF STRUCTURES TO BUSINESS TENDENCY SURVEY DATA*

1. INTRODUCTION

Introduction of The Second Law of thermodynamics (the Law of Entropy) to mainstream economics is often attributed to H. Theil and N. Georgescu-Roegen (see [3], [10]). Since then, entropy and other measures of information content have been interpreted in a variety of applications. Recently, linked with the notion of sustainability, the concept of entropy enjoys a renaissance in economic literature. Sustainability, defined by the World Commission on Environment and Development [12] as a call for continued economic expansion without environmental degradation, focuses on issues of how large the economy should be relative to the environment and how to achieve an optimal inter-temporal allocation of resources. Current economic entropy literature includes elements of information theory, complex systems analysis, and environmental economics. However, few economic applications of entropy measures have been published in Polish literature. They include a study of propensity to smoke [2], correlation analysis [4], and application of relative entropy to evaluate expert forecasts [5].

Entropy of a probability distribution can be interpreted as a measure of uncertainty or, alternatively, as a measure of information content In this paper, I return to the original, information theory definition of entropy to evaluate similarities between *a priori* information supplied by the business tendency surveys (that is, expectations), and *a posteriori* information (that is, realizations). The idea is motivated by a paper by Wędrowska [13] who proposes to interpret the information content of a change of structure from its *a priori* to *a posteriori* form as a measure of degree of similarity (or dissimilarity) of structures. In this paper, *a priori* structure is defined by fractions of respondents expressing expectations, and *a posteriori* structure – by fractions of respondents declaring observed changes in economic variables (realizations).

There are two reasons for undertaking such analysis. First, measures of similarity may provide the means to evaluate accuracy of expectations (predictions) in the business tendency survey. The larger the similarity between expectations and realizations, the better predictive power of expectations. Second, the survey on which empirical

* Research underlying this publication was supported by the Warsaw School of Economics Research Grant No 03/E/0023/100. I would like to thank Dr Ewa Wędrowska for her valuable comments on the preliminary draft of this paper, and an anonymous Referee for helpful suggestions.

part of the paper is based defines expectations as changes "expected in the next 3-4 months". Information content may help to identify the actual forecast horizon used by respondents. This result may, in turn, be useful in other formal analyses of expectations, for example establishing appropriate number of lags in econometric models with expectations or defining dependent variables in quantification models.

In section 2, measures of entropy, information content and dissimilarity of structures are introduced. Business tendency survey data are described in section 3. Empirical results are reported in sections 4 (measures of entropy) and 5 (measures of dissimilarity of structures). Section 6 concludes.

2. MEASURES OF ENTROPY, INFORMATION CONTENT AND DISSIMILARITY OF STRUCTURES

Following Wędrowska [13], let us define structure S^n as a vector $S^n = [s_1, s_2, \ldots, s_n]^T \in \mathrm{R}^n$ which elements s_i ($i = 1, 2, \ldots, n$) fulfill two conditions:

$$0 \leqslant \quad s_i \quad \leqslant 1, \tag{1}$$

$$\sum_{i=1}^{n} s_i = 1. \tag{2}$$

Structure S^n is therefore fully described by a vector of fractions (structure elements) summing to a total of 1.

Amount of information provided by a message (that is, its information content) is defined in information theory in relation to the probability that a given message is received from the set of all possible messages: the less probable the message, the more information it carries. On the basis of the elements of S^n it is now possible to define the empirical measure of entropy introduced by C. E. Shannon in his classic paper *A mathematical theory of communication* as

$$H(S^n) = \sum_{i=1}^{n} s_i \log_2 \frac{1}{s_i}. \tag{3}$$

It is worth noting that value of $H(S^n)$ depends only on characteristics of the structure analyzed, that is, its elements s_i.

An important property of $H(S^n)$ as a measure of entropy is that it reaches its maximum value of $H_{max} = \log_2 n$ if all structure elements s_i are equal (that is, $s_1 = s_2 = \ldots = s_n$; see [6], [8]). As $H(S^n)$ approaches its maximum value, differences between structure elements decrease, and for $H(S^n) = H_{max}$, distribution of structure elements becomes uniform. Also, $H(S^n) = H_{min} = 0$ if one of the elements s_i ($i = 1, 2, \ldots, n$) is equal to 1, and all the remaining structure elements are equal to 0 (that is, distribution is concentrated in one element of structure only). Value of $H(S^n)$ can be therefore interpreted as measure of concentration of elements s_i of structure S^n, and

can be used in empirical setting to evaluate information content of a structure. When several structures ordered in time are available, it is also possible to analyze their dynamics. Empirical values and dynamics of entropy measure $H(S^n)$ for expectations and realizations expressed in business tendency surveys are presented in section 4.

In practice, however, not only the degree of uncertainty associated with *a priori* and *a posteriori* structures may be economically interesting but also extent of changes detected between assumed (*a priori*) and observed (*a posteriori*) structures. In order to analyze the size of change between *a priori* structure S_p^n and *a posteriori* structure S_q^n, relative entropy (or Kullback-Leibler divergence; see [8]) is calculated:

$$I\left(S_q^n : S_p^n\right) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}. \tag{4}$$

Relative entropy is also known as information gain; it measures expected amount of "new" information provided by *a posteriori* structure. One of the properties of $I(S_q^n : S_p^n)$ states that it takes its minimum value of zero if both structures are identical (that is, $S_p^n = S_q^n$), and increases with the size of differences between the structures to infinity (see [8], [13]). $I(S_q^n : S_p^n)$ can be interpreted as degree of change between assumed (*a priori*) and observed (*a posteriori*) structures, and therefore serve as measure of dissimilarity of structures: the larger it is, the less similar the structures are.

In empirical setting, it is more convenient to apply a standardized coefficient defined on interval [0, 1] to facilitate interpretations and comparisons. Chomątowski and Sokołowski in [1] introduce a similarity measure to classify data into comparable phases, and employ it to define clusters of industrial production in Poland. They also provide a related dissimilarity measure that can be used to evaluate extent of change from *a priori* to *a posteriori* structure:

$$P\left(S_q^n : S_p^n\right) = 1 - \sum_{i=1}^{n} \min(q_i, p_i). \tag{5}$$

From the properties of structure defined by (1) and (2) it follows that $P(S_q^n : S_p^n) \in [0, 1]$. The lower limit is attained when analyzed structures are identical, that is, $S_p^n = S_q^n$. As dissimilarities between structures increase, value of $P(S_q^n : S_p^n)$ increases towards the upper limit of 1. Empirical values and dynamics of dissimilarity measure $P(S_q^n : S_p^n)$ employed to evaluate similarities between expectations and realizations expressed in business tendency surveys are described in section 5. They are introduced to supplement results obtained on the basis of entropy measure as both methods reflect structure change from its *a priori* to *a posteriori* state.

3. DATA

Empirical part of this paper focuses on evaluation of information content and dissimilarity between expectations and observed realizations, declared by Polish industrial enterprises in business tendency surveys. Since 1986, qualitative business tendency

surveys are conducted by the Research Institute for Economic Development (RIED) at the Warsaw School of Economics. Originally launched for manufacturing industry, currently they cover households, farming sector, exporters, construction industry, and banking sector as well. In the monthly survey addressed to industrial enterprises (see Table 1), respondents are asked to evaluate both current situation (as compared to last month) and expectations for the next 3 – 4 months by assigning them to one of three categories: increase / improvement, no change, or decrease / decline. Aggregated survey results are regularly published in RIED bulletins (see [7]).

Table 1

Monthly RIED questionnaire in industry

		Observed within last month	Expected for next 3 – 4 months
01	Level of production (value or physical units)	up unchanged down	will increase will remain unchanged will decrease
02	Level of orders	up normal down	will increase will remain normal will decrease
03	Level of export orders	up normal down not applicable	will increase will remain normal will decrease not applicable
04	Stocks of finished goods	up unchanged down	will increase will remain unchanged will decrease
05	Prices of goods produced	up unchanged down	will increase will remain unchanged will decrease
06	Level of employment	up unchanged down	will increase will remain unchanged will decrease
07	Financial standing	improved unchanged deteriorated	will improve will remain unchanged will deteriorate
08	General situation of the economy regardless of situation in your sector and enterprise	improved unchanged deteriorated	will improve will remain unchanged will deteriorate

Source: the RIED database

For empirical analysis, four survey questions have been selected, namely, those pertaining to changes in production, prices, employment and general business conditions. These variables have been analyzed previously, and they have precisely defined counterparts in official statistics necessary for purposes of quantitative analysis (see [11]). *A priori* structure is defined as percentages of respondents who expect increase / no change / decline, and *a posteriori* structure as percentages of respondents who observe increase / no change / decline three and four months later. This definition fulfills conditions (1) and (2). All variables are analyzed for three ownership types: public,

private and total. Since forecast horizon is not precisely defined, both alternatives ($k = 3$ and $k = 4$) are analyzed when calculating measure of dissimilarity $P(S_q^n : S_p^n)$.

Available data cover 161 observations from March 1997 to July 2010. However, since expectations have to be matched with observed realizations to calculate the measure of dissimilarity $P(S_q^n : S_p^n)$, length of time series is reduced either by three (for 3-month forecast horizon) or by four observations (for 4-month forecast horizon). For the purposes of clarity of presentation, all results are reported for the core time period of June 1997 to April 2010 (155 observations).

4. EMPIRICAL RESULTS: ENTROPY

Table 2 presents summary statistics for entropy measure $H(S^n)$ given by formula (3), calculated for all four variables, separately for expectations and observed changes, across ownership sectors [1].

To facilitate comparisons, Figures 1-4 present values of entropy measure $H(S^n)$ for expectations and realizations (for all firms only, to make graphs more legible) against the same scale.

Table 2

Summary statistics for entropy measures

	production						
	expectations				realizations		
	all	public	private		all	public	private
min	1.2650	1.1956	1.2851		1.3289	1.2638	1.2790
max	1.5515	1.5773	1.5679		1.5702	1.5803	1.5751
avg	1.4680	1.4455	1.4741		1.5054	1.4913	1.5093
median	1.4736	1.4588	1.4832		1.5106	1.5079	1.5212
std dev	0.0530	0.0738	0.0583		0.0444	0.0604	0.0485
	prices						
	expectations				realizations		
	all	public	private		all	public	private
min	0.7713	0.6295	0.7566		0.7674	0.5938	0.8155
max	1.3140	1.3355	1.2962		1.2962	1.4447	1.2882
avg	0.9916	0.9755	0.9970		1.0301	1.0224	1.0306
median	0.9821	0.9555	0.9933		1.0272	0.9976	1.0316
std dev	0.1065	0.1331	0.1118		0.1042	0.1630	0.1035

[1] Detailed results are available from author upon request.

c.d. Table 2

	employment						
	expectations				realizations		
	all	public	private		all	public	private
min	0.9398	0.8669	0.8676		1.0486	0.8241	1.0404
max	1.3822	1.4246	1.4101		1.3940	1.4194	1.3901
avg	1.1884	1.1678	1.1925		1.2434	1.2065	1.2605
median	1.1939	1.1723	1.1981		1.2485	1.2228	1.2564
std dev	0.0751	0.0934	0.0895		0.0696	0.1063	0.0713
	general business conditions						
	expectations				realizations		
	all	public	private		all	public	private
min	0.8663	0.8703	0.8778		0.6238	0.7749	0.5354
max	1.4407	1.4902	1.4796		1.3857	1.4478	1.4308
avg	1.2921	1.2575	1.3151		1.1741	1.1334	1.2018
median	1.3020	1.2706	1.3303		1.1929	1.1566	1.2319
std dev	0.0827	0.1080	0.0885		0.1116	0.1138	0.1295

Source: own calculations on the basis of RIED data

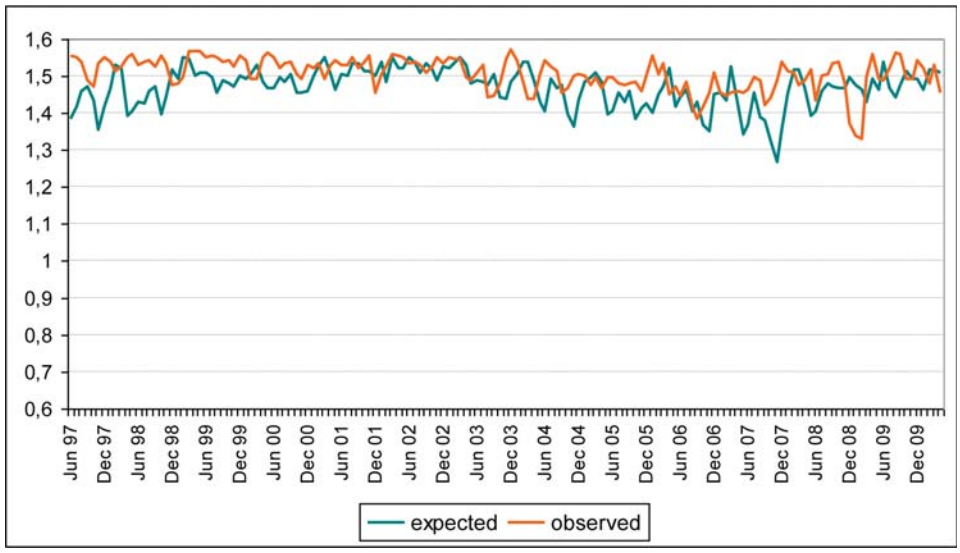

Figure 1. Entropy of production (all firms)

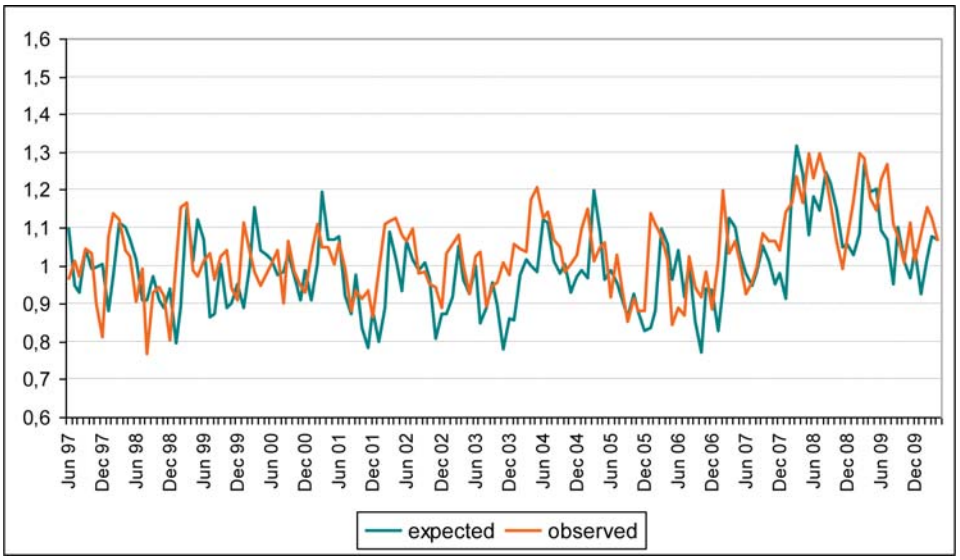

Figure 2. Entropy of prices (all firms)

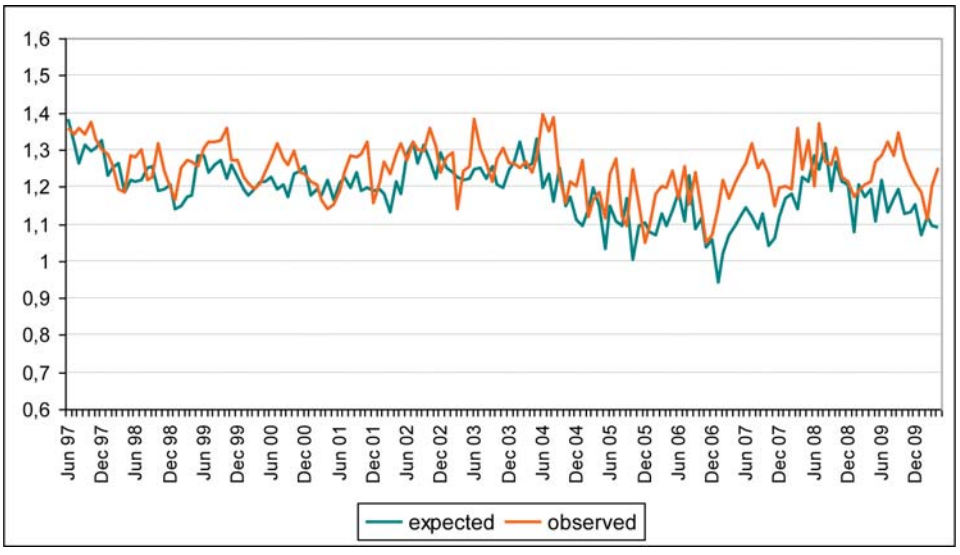

Figure 3. Entropy of employment (all firms)

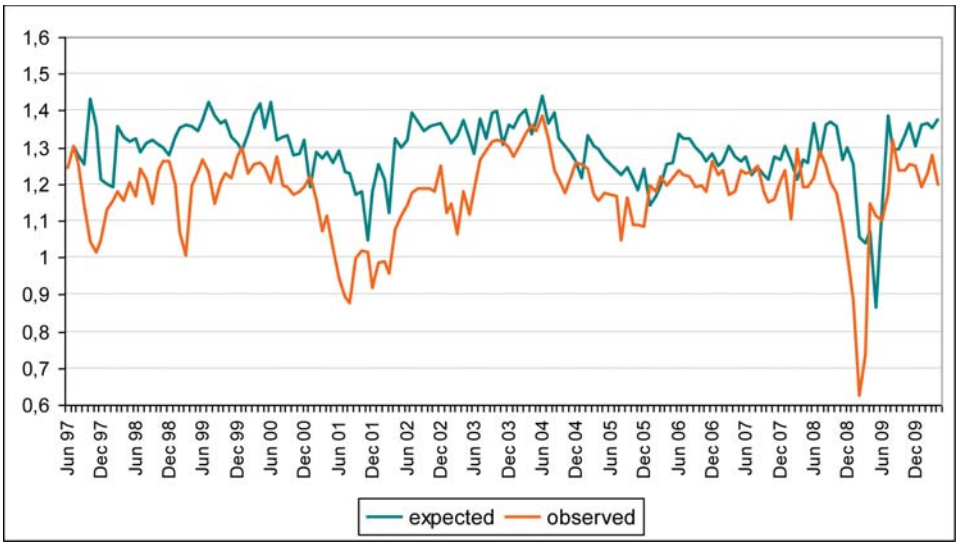

Figure 4. Entropy of general business conditions (all firms)

Source: own calculations on the basis of RIED data

High values obtained in case of production, both in comparison to other variables and in absolute terms, seem to be the most striking result. The maximum value of measure of entropy is $H_{max} = \log_2 n = \log_2 3 = 1.5850$; the closer empirical entropy of a structure to its maximum value, the more uniform the structure is, and therefore the less informative *a priori* structure becomes in relation to *a posteriori* structure. Maximum values obtained for production (1.5515 for expectations and 1.5702 for realizations) are close to the upper limit, and average values (1.4680 for expectations and 1.5054 for realizations) are also very high in comparison to entropy of prices, employment and general business conditions. On the other hand, entropy is equal to zero if one of the elements of a structure is equal to 1, that is, there is no uncertainty associated with distribution of outcomes. Value of zero is not attained for any of the variables analyzed, but the lowest values are observed for prices and realizations of general business conditions.

For production and unemployment, behavior of expected and observed series is similar; in case of employment, expectations exhibit lower averages and medians, across all ownership sectors, than realizations; in case of general business conditions, the opposite is true. Public enterprises exhibit lower average and higher variability of entropy, as measured by standard deviation, than private enterprises, with the sole exception of standard deviation for observed general business conditions.

5. EMPIRICAL RESULTS: DISSIMILARITY OF STRUCTURES

Figures 5-8 provide graphical summary of results obtained for differences between measures of dissimilarity of structures $P(S_q^n : S_p^n)$ calculated for 3-month horizon and 4-month horizon[2].

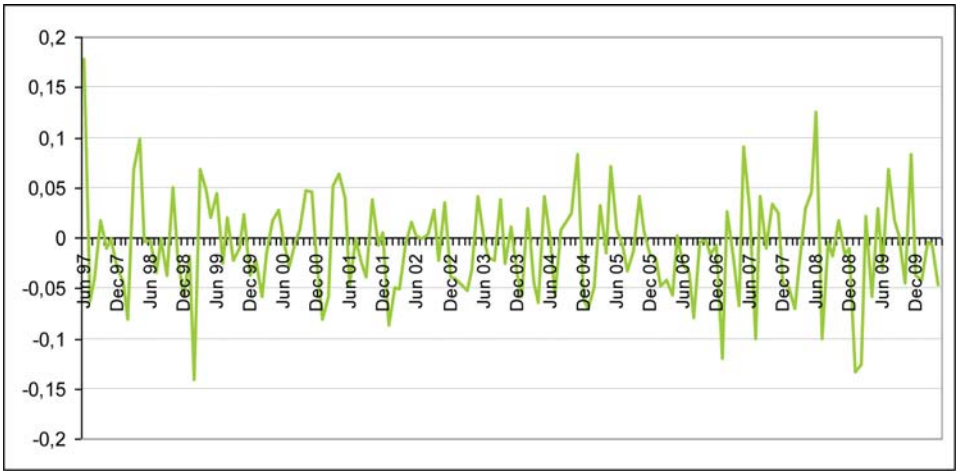

Figure 5. Differences between dissimilarity measures for $k = 3$ and $k = 4$, production (all firms)

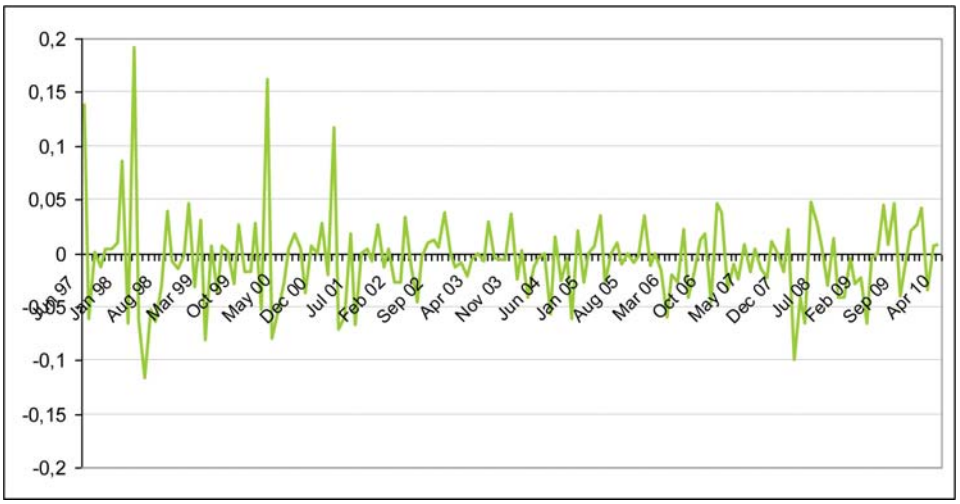

Figure 6. Differences between dissimilarity measures for $k = 3$ and $k = 4$, prices (all firms)

[2] Detailed results (values calculated on the basis of formula (5) for all four variables, three ownership types and separately for 3-month forecast horizon and 4-month forecast horizon) are available from author upon request.

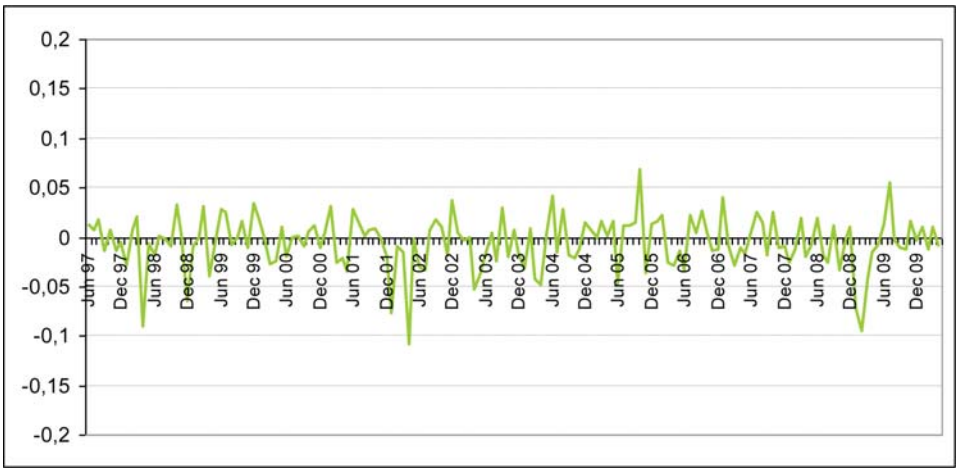

Figure 7. Differences between dissimilarity measures for $k = 3$ and $k = 4$, employment (all firms)

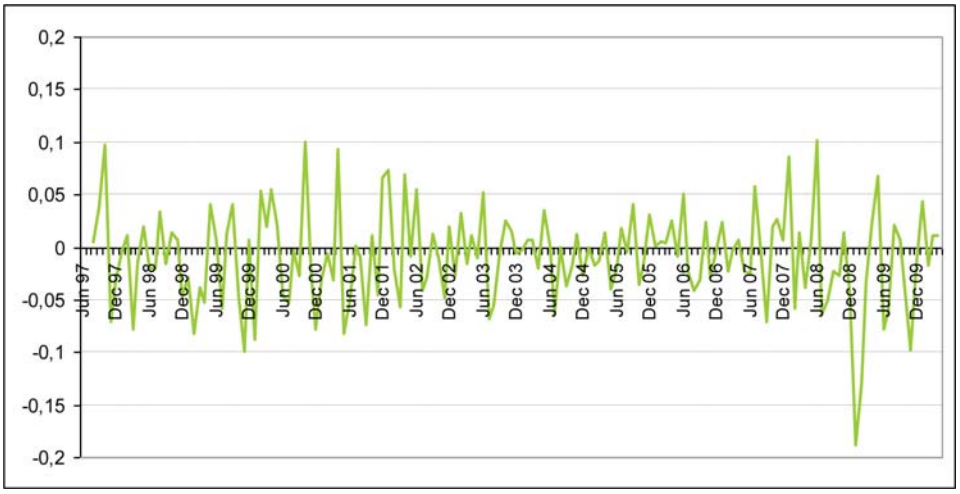

Figure 8. Differences between dissimilarity measures for $k = 3$ and $k = 4$, general business conditions (all firms)

Source: own calculations on the basis of RIED data

In Figures 5-8, negative numbers signify that value of $P(S_q^n : S_p^n)$ for $k = 3$ is smaller than for $k = 4$, that is, structures of realizations and expectations for $k = 3$ are more similar than structures of realizations and expectations for $k = 4$. This can be interpreted as 4-month expectations being less congruent with later observed realization than expectations interpreted in 3-month horizon. This result can be at least partly explained by the well-known observation that uncertainty increases with the forecast horizon, and the tendency of business enterprises to evaluate their performance in quarterly (that is, 3 month) terms.

Table 3 presents summary statistics for differences between dissimilarity measures for $k = 3$ and $k = 4$.

Table 3

Summary statistics for dissimilarity measures: difference between $k = 3$ and $k = 4$

	production	**prices**	**employment**	**business**
min	-0.1420	-0.1160	-0.1080	-0.1880
max	0.1790	0.1920	0.0680	0.1010
avg	-0.0095	-0.0056	-0.0055	-0.0095
median	-0.0100	-0.0050	-0.0030	-0.0060
std dev	0.0480	0.0412	0.0263	0.0446

Source: own calculations on the basis of RIED data

The graphs do not provide basis for determining which forecast horizon is supported by measures of dissimilarity of structure. Average values of dissimilarity statistics (see Table 3) are slightly negative for all variables which suggests that predictions made at the 3-month forecast horizon are more similar to the realizations that those made 4 months in advance. Median values confirm this result. However, the differences between the 3-month and 4-month horizons are very small and should not be considered as a proof of superiority of the shorter forecast horizon, especially in the light of relatively high standard deviations. It is also worth noting that maximum values of dissimilarity measures comparing 3-month and 4-month forecasts for production and prices are twice as high as those obtained for employment and general business conditions. It follows that in production and prices there are (few) periods in which the 4-month forecast horizon is considerably more similar to realizations, although the average remains negative, suggesting that on average the 3-month forecast horizon is closer to the observed changes in analyzed variables.

6. CONCLUDING COMMENTS

On the basis of empirical analysis of the business tendency survey data, the following conclusions have been reached through application of measures of entropy:

1. In case of production, distribution of increase / no change / decrease fractions is relatively uniform, leading to high entropy and providing little information.
2. Entropy of prices is relatively low; since value of entropy allows to evaluate degree of concentration, in case of prices fractions of survey answers seems to be particularly centered on one of the three options provided in the questionnaire. In theory, answers might be centered on either of the three options (increase / no change / decrease) and vary from one questionnaire to another. In practice, however, they are heavily biased towards the "no change" category. In the analyzed period (that is, March 1997 – July 2010), no change in prices is always expected by the majority of

the respondents – that is, "no change" fraction constantly remains the largest among the three options. This result does not hold in case of production, employment, or general business conditions expectations, in confirmation of the results of entropy analysis.

3. Entropy of general business conditions exhibits the highest variability which may be interpreted as volatile changes in information content of surveys from one month to another. In contrast, entropy of production is the least variable.
4. Generally, public enterprises exhibit lower entropy (as measured by average) and higher variability (as measured by standard deviation) than private enterprises; that is, for public enterprises concentration of answers to the survey questions is higher and also more variable.

To evaluate whether structures for 3-month or 4-month forecast horizons are more similar to observed realizations, measures of dissimilarity of structures were calculated. Unfortunately the results do not provide clear answer to this question. A slight tendency towards the 3-month forecast horizon is noted on the basis of negative (but very small in absolute terms) average values of dissimilarity measures across all four variables.

To summarize, results obtained on the basis of entropy and dissimilarity measures provide new insights into behavior of expectations and realizations expressed in business tendency surveys. As this is the first attempt to empirically address the question of information content of the RIED survey data, more work is clearly needed. One of the issues that merit further analysis is whether current situation of an enterprise systematically influences its expectations, and consequently degree of concentration of answers on a particular option. To evaluate usefulness of entropy measures in analyzing questionnaire data, predictive value of a *priori* information should be studied, and predictive properties of various statistic tools compared for different distributions of answers.

Author is employed at the Institute of Econometrics, Warsaw School of Economics.

REFERENCES

[1] Chomątowski S., Sokołowski A. (1978), *Taksonomia struktur*, Przegląd Statystyczny 2:217-226.
[2] Doszyń M. (2002) *Skłonności a entropia*, Przegląd Statystyczny 49:73-78.
[3] Georgescu-Roegen N. (1971) *The Entropy Law and the Economic Process*, Harvard University Press, Cambridge.
[4] Kempa W. (2002), *Zastosowanie entropii empirycznej w badaniu związku korelacyjnego dwóch cech* Przegląd Statystyczny 49:163-173.
[5] Kowalczyk H. (2010), *O eksperckich ocenach niepewności w ankietach makroekonomicznych*, Bank i Kredyt 5:101-122.
[6] Przybyszewski R., Wędrowska E. (2005), *Aksjomatyczna teoria entropii*, Przegląd Statystyczny 52:85-101.
[7] RIED (2010), *Business survey. June 2010*, Warsaw School of Economics, Warsaw.
[8] Rényi A. (1961), *On measures of entropy and information,* Proceedings of the 4^{th} Berkeley Symposium on Mathematics, Statistics and Probability, pp. 547-561.

[9] Shannon C. E. (1948)*A mathematical theory of communication*, The Bell System Technical Journal 27:379-423, 623-656.
[10] Theil H. (1967), *Economics and Information Theory*, North-Holland Publishing Company, Amsterdam.
[11] Tomczyk E. (2005) *Are expectations of Polish industrial enterprises rational? Evidence from business tendency surveys*, in:Adamowicz E., Klimkowska J. (eds.) *Economic Tendency Surveys and Cyclical Indicators. Polish contribution to the 27th CIRET Conference,* Warsaw School of Economics, Warszawa.
[12] WCED (1987), *Report of the World Commission on Environment and Development: Our Common Future*, NGO Committee on Education (http://www.un-documents.net/wced-ocf.htm).
[13] Wędrowska E. (2009), *Oczekiwana ilość informacji o zmianie struktur jako miara niepodobieństwa struktur*, paper presented at the XI*th* Conference "Dynamic Econometric Models", September 2009, Toruń.

APPLICATION OF MEASURES OF ENTROPY, INFORMATION CONTENT AND DISSIMILARITY OF STRUCTURES TO BUSINESS TENDENCY SURVEY DATA

Summary

This paper evaluates similarities between *a priori* information supplied by business tendency surveys (that is, expectations), and *a posteriori* information (that is, realizations). A *priori* structure is defined by fractions of respondents expressing expectations, and *a posteriori* structure – by fractions of respondents declaring observed changes in economic variables (realizations). On the basis of empirical analysis of the business tendency survey data on production, prices, employment and general business conditions, the following conclusions have been reached. Production time series exhibits the highest entropy, and prices data – the lowest. Since value of entropy allows to evaluate degree of concentration, in case of prices fractions of survey answers seems to be particularly centered on one of the three options provided in the questionnaire (that is, increase – no change – decrease). Entropy of general business conditions exhibits the highest variability which may be interpreted as volatile changes in information content of surveys from one month to another; in contrast, entropy of production is the least variable. It is also found that public enterprises exhibit lower entropy (as measured by average) and higher variability (as measured by standard deviation) than private enterprises.

Key words: tendency surveys, expectations, entropy, dissimilarity of structures

ZASTOSOWANIE MIAR ENTROPII, ZAWARTOŚCI INFORMACYJNEJ I NIEPODOBIEŃSTWA STRUKTUR DO DANYCH TESTU KONIUNKTURY

Streszczenie

Artykuł bada podobieństwo między informacją *a priori* dostarczaną przez respondentów testu koniunktury (oczekiwaniami) a informacją *a posteriori* (zaobserwowanymi realizacjami). Struktura *a priori*

definiowana jest poprzez odsetki respondentów wyrażających swoje oczekiwania, a struktura *a posteriori* – przez odsetki respondentów stwierdzających zaobserwowane zmiany. Na podstawie empirycznej analizy danych testu koniunktury na temat produkcji, cen, zatrudnienia oraz ogólnej sytuacji gospodarczej, sformułowano następujące wnioski. Produkcja cechuje się najwyższą entropią, a ceny – najniższą. Ponieważ poziom entropii może być interpretowany jako stopień koncentracji, w przypadku cen odsetki odpowiedzi na pytania testu koniunktury wydaje się być szczególnie mocno skoncentrowany na jednej z trzech opcji (wzrost – brak zmiany – spadek). Entropia ogólnej sytuacji gospodarczej wykazuje największą zmienność, co można zinterpretować jako przejaw dynamicznych zmian zawartości informacyjnej ankiety w poszczególnych miesiącach; entropia produkcji jest najbardziej stabilna. Co więcej, przedsiębiorstwa sektora publicznego cechuje średnio niższa entropia i wyższa jej zmienność (mierzona odchyleniem standardowym) niż przedsiębiorstwa prywatne.

Słowa kluczowe: badania ankietowe, oczekiwania, entropia, niepodobieństwo struktur