METODY ILOŚCIOWE
W BADANIACH EKONOMICZNYCH

QUANTITATIVE METHODS IN ECONOMICS

Vol. XIV, No. 2

Warsaw University of Life Sciences – SGGW
Faculty of Applied Informatics and Mathematics
Department of Econometrics and Statistics

METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH

QUANTITATIVE METHODS IN ECONOMICS

Volume XIV, No. 2

Warsaw 2013

ISSN 2082 – 792X

Warsaw 2013, Volume XIV, No. 2

The original version is the paper version

Journal homepage: qme.sggw.pl

Published by Warsaw University of Life Sciences Press

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013

CONTENTS

Lidia Luty – Demographic development of the powiats of the Malopolskie voivodeship 7

Edyta Łaszkiewicz – Sample size and structure for multilevel modelling: Monte Carlo investigation for the balanced design 19

Rafał Łochowski – On an upper gain bound for strategies with constant and proportional number of assets traded 29

Andrzej Łodziński – The method of supporting decisions under risk based on multiobjective optimization 39

Małgorzata Machowska–Szewczyk – Fuzzy classification of symbolic objects 51

Paulina Malaczewska – Useful government expenditure influence on the shadow economy 61

Maciej Malaczewski – Endogenous technological progress and economic growth in a model with natural resources 70

Jerzy Marzec, Andrzej Pisulewski – Technical efficiency measurement of dairy farms in Poland: an application of Bayesian VED model 78

Aldona Migała–Warchoł, Marek Sobolewski – Evaluation of voivodeships diversification in Poland according to transport infrastructure indicators 89

Katarzyna Miszczyńska – Multivariate analysis of healthcare systems in selected European Union countries. Cluster analysis 99

Piotr M. Miszczyński – Measuring the efficiency of local government units management in the central region of Poland in a dynamic perspective 108

Joanna Muszyńska, Iwona Müller–Frączek – The development of agriculture in Poland in the years 2004-2011– the taxonomic and econometric analyses 118

Sylwia Nieszporska – Ordering and classification of the Silesian voivodeship region with respect to a health care system activity 129

Maria Parlinska, Iryna Petrovska – The role of information systems in logistic enterprices 139

Michał Bernard Pietrzak, Justyna Wilk Mariola Chrzanowska – Economic situation of eastern Poland and population migration movement 148

Artur Prędki – Subsampling approach for statistical inference within stochastic DEA models 158

Aneta Ptak–Chmielewska – Semiparametric Cox regression model in estimation of small and micro enterprises' survival in the Malopolska voivodeship 169

Maria Sarama – Comparative analysis of the information society development level in the poviats of the Podkarpackie voivodship 181

Iwona Skrodzka – Spatial diversity of human capital in the European Union 191

Aneta Sobiechowska–Ziegert, Aniela Mikulska – Measure of the level of socio-economic development in provinces 200

Agnieszka Sompolska-Rzechuła, Grzegorz Spychalski – The use of correspondence analysis in the evaluation of the role of fibrous and medicinal plants in plant production in farms 210

Aleksander Strasburger, Olga Zajkowska – Who wants to work longer? 220

Anna Sznajderska – Foreign exchange rates in Central European economies: nonlinearities in adjustment to interest rate differentials 229

Ryszard Szupiluk, Piotr Wojewnik, Tomasz Ząbkowski – Multivariate decompositions for value at risk modeling 240

Andrzej Szuwarzyński – Evaluation of the efficiency of flexicurity implementation in OECD countries 251

Agnieszka Tłuczak – The analysis of the phenomenon of spatial autocorrelation of indices of agricultural output 261

Olga Zajkowska – Gender pay gap in Poland – Blinder–Oaxaca decomposition 272

Wojciech Zatoń – Investors' preferences and payoffs from structured products 279

Tomasz Ząbkowski, Krzysztof Gajowniczek – Forecasting of individual electricity usage using smart meter data 289

Monika Zielińska–Sitkiewicz – Application of multivariate discriminant analysis for assessment of condition of construction companies 298

Wojciech Zieliński – Confidence intervals for fraction in finite populations: minimal sample size 309

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 7 – 18

DEMOGRAPHIC DEVELOPMENT OF THE POWIATS OF THE MAŁOPOLSKIE VOIVODESHIP

Lidia Luty
Department of Mathematical Statistics
University of Agriculture in Cracow
e-mail: rrdutka@cyf-kr.edu.pl

Abstract: The evolution of the demographic phenomena both in spatial and time terms allows to assess the development of the region. The purpose of the article was an attempt to identify similarities in the selected demographic processes in the powiats of the Małopolskie voivodeship in the years 2002-2011. In the first part of the analysis of the phenomenon the powiats are organized using indicator of demographic development estimated in the first and in the last year of the analysis. For separated four groups of powiats, similar in terms of the analysed indicator, representatives were selected using the method of the centre of gravity, for which shows the process of changes of demographic characteristics such as: birth rate per 1000 population; gross reproduction rate; non-productive age population, per 100 persons of working age; the number of infant deaths per 1000 live births in terms of time.

Keywords: indicator of demographic development, classification, the method of the centre of gravity

INTRODUCTION

Forming of the demographic phenomena both in terms of space and time allows to assess the development of the region. The purpose of this article is an attempt to identify similarities in the selected demographic processes in different powiats of the Małopolskie voivodeship in the years 2002-2011. In the first part of the consideration of the phenomenon the powiats are organized using indicator of demographic development estimated in the first and last year of the analysis. For separated groups of powiats similar in terms of selected indicator, representatives were chosen, for which it was shown the process for selected demographics phenomenon in terms of time.

METHOD OF ANALYSIS

Population of *n* objects O_i $(i = 1,\ 2,\ ...,\ n)$ in defined unit of time is characterized by *m* characteristics. Values of characteristics X_j $(j = 1,\ 2,\ ...,\ m)$ corresponding to objects are described by matrix:

$$[x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nm} \end{bmatrix} \qquad (i = 1,\ 2,\ ...,\ n;\ j = 1,\ 2,\ ...,\ m),$$

where x_{ij} - value of the *j*-characteristic for the *i*-object in defined unit of time.

Every object we can equate with a point in the *m*-dimensional Euclidean space, which the individual axes correspond to the variable X_j.

To determine the indicator of the relative demographic development for each object, you should:

- standardize values of characteristics X_j according to the formula:

$$z_{ij} = \begin{cases} \dfrac{x_{ij} - \bar{x}_j}{S_j}, & X_j \in S \\ \dfrac{\bar{x}_j - x_{ij}}{S_j}, & X_j \in D \end{cases} \tag{1}$$

where: x_{ij} – value of the *j*-characteristic for the *i*-object,

$\bar{x}_j$ – the arithmetic mean of the *j*-characteristic,

S_j – the standard deviation of the *j*-characteristic,

S, D – accordingly, a set of stimulant and destimulant[1];

- transform standardized characteristics' values in accordance with the formula:

$$z_{ij}^{*} = z_{ij} - \min_i \{z_{ij}\} \tag{2}$$

- get the value of the indicator of demographic development for each object O_i [Pociecha, 1988]:

[1] The term of stimulant and destimulant was introduced by Z. Hellwig [1968].

$$W_i = \frac{\sum_{j=1}^{m} z_{ij}^*}{\sum_{j=1}^{m} \max_i \left\{ z_{ij}^* \right\}} \tag{3}$$

Indicator W_i takes the values from the range $\langle 0,\ 1 \rangle$, higher values mean a higher level of development. This measure is relative, based on it you can evaluate the level of development of the object in relation to the level of the rest of the analyzed objects and group tested objects.

CLASSIFICATION OF POWIATS IN RELATION TO THE INDICATOR OF DEMOGRAPHIC DEVELOPMENT

Quantitative and qualitative changes in the population in the area, describe number of statistical characteristics. Using the criterion of low correlation between variables forming set of variables to determine the indicator of demographic development in one year, which is the basis for the classification of powiats of the Małopolskie voivodeship, the following variables were selected:

X_1 - natural growth per 1000 population,

X_2 - gross reproduction rate,

X_3 - non-productive age population, per 100 persons of working age,

X_4 - the number of infant deaths per 1000 live births.

Thanks to this method, powiats were organized based on value of the indicator of demographic development. Values of indicator and positions occupied by individual powiats due to the value of this measure in two years is shown in the table 1.

The highest value of the indicator of demographic development in both presented years reached bocheński powiat. In turn, the lowest value W_i in year 2002 reached olkuski powiat, and in the year 2011 chrzanowski powiat. Diversity of values of estimated indicator was higher in 2002. The coefficient of variation of the estimated indicators of development in 2002 was 48,9%, and in 2011 - 36,8%. In 2011 average value of W_i was 0,432 and it was higher than average value of his measure estimated in 2002. Powiats with indicators of demographic development higher than average value W_i form a coherent whole territory (south part of voivodeship) both in 2002 (bocheński, nowosądecki, nowotarski, suski, tatrzański, limanowski, gorlicki, tarnowski, brzeski, wadowicki powiats) and in 2011 (bocheński, nowosądecki, nowotarski, limanowski, gorlicki, suski, brzeski, tatrzański powiats).

Table 1. Values of the indicator of demographic development for powiats

Powiat	Year 2002		Year 2011	
	W_i	position (d_{li})	W_i	position (d_{ki})
bocheński	0,860	1	0,807	1
brzeski	0,363	9	0,471	7
chrzanowski	0,140	18	0,158	19
dąbrowski	0,261	12	0,244	18
gorlicki	0,378	7	0,559	5
krakowski	0,256	13	0,311	14
limanowski	0,406	6	0,581	4
miechowski	0,255	14	0,295	16
myślenicki	0,306	11	0,426	9
nowosądecki	0,478	2	0,679	2
nowotarski	0,442	3	0,608	3
olkuski	0,128	19	0,354	13
oświęcimski	0,211	15	0,252	17
proszowicki	0,167	16	0,415	10
suski	0,424	4	0,478	6
tarnowski	0,376	8	0,393	12
tatrzański	0,414	5	0,470	8
wadowicki	0,343	10	0,304	15
wielicki	0,148	17	0,397	11

Source: own elaboration

Conformity assessment of hierarchy of sorting out of powiats in two classifications we can make estimate of the Spearman's ranks correlation coefficient, using the formula:

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(d_{li} - d_{ki})^2}{n(n^2 - 1)} \tag{4}$$

where: d_{li}, d_{ki} - position of the *i*-object respectively in sort outs of *l* and *k*; *n* – number of objects.

To test the compatibility of the sort out measure, we use statistics:

$$u = r_S\sqrt{n-1} \tag{5}$$

that is, assuming, that the sort outs are only coincide at random, has a normal distribution.

The value of the Spearman's ranks correlation coefficient for presented arrangement of powiats is 0,804, which is statistically significant ($u = 3{,}409 > u_{\alpha=0,05} = 1{,}960$), so there is no major changes in hierarchy of powiats.

Within the arranged set of powiats, respectively in 2002 and in 2011, four disjoint subsets of similar objects were separated in the following ways:

– I a group of objects, for which: $W_i > \overline{W} + S_W$,

– II a group of objects, for which: $\overline{W} < W_i \leq \overline{W} + S_W$,

– III a group of objects, for which: $\overline{W} - S_W < W_i \leq \overline{W}$,

– IV a group composed of objects, for which: $W_i \leq \overline{W} - S_W$,

where: $\overline{W}$ - the arithmetic mean of W_i, S_W - the standard deviation of W_i.

Summary of the results of grouping of powiats against the designated demographic development indicator measure shows graphically on the figure 1.

In the first year of the analysis only one powiat (bocheński) was assigned to the group I, to group II nine powiats, mostly southern and central Małopolska. The third and the fourth in 2002 formed the powiats adjacent to the city of Kraków, and powiats put forth the most of the northwest and dąbrowski powiat.

Table 2. The minimum and maximum characteristics in groups of powiats

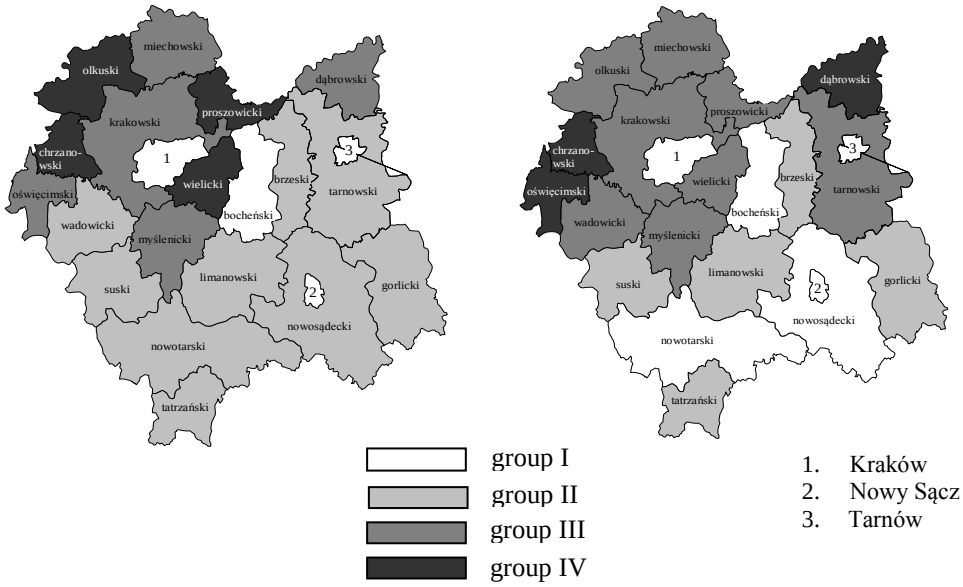
Group	Description	Year 2002				Year 2011			
		X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
I	$\min_i x_{ij}$	1,9	0,764	70,2	6,4	3,5	0,722	58,1	2,6
	$\max_i x_{ij}$	1,9	0,764	70,2	6,4	5,6	0,844	62,0	6,4
II	$\min_i x_{ij}$	1,8	0,625	66,0	1,6	2,3	0,639	59,0	3,0
	$\max_i x_{ij}$	6,4	0,920	77,5	9,3	5,3	0,833	64,2	4,2
III	$\min_i x_{ij}$	-3,3	0,585	61,5	2,1	-4,3	0,554	54,9	2,7
	$\max_i x_{ij}$	3,9	0,725	71,6	7,5	4,7	0,746	63,0	6,2
IV	$\min_i x_{ij}$	-1,7	0,564	59,1	6,0	-1,1	0,570	54,4	5,3
	$\max_i x_{ij}$	0,8	0,723	68,2	12,1	0,3	0,717	56,9	7,5

Source: own elaboration

In 2011, there was a slight realignment, the largest group was the third group. In the group of top classified it was next to bocheński powiat, nowotarski and nowosądecki powiats. In group IV remained chrzanowski powiat and joined oświęcimski and dąbrowski powiats.

The minimum and maximum values of characteristics, on the basis of which it was estimated the economic development level indicators in separated groups of powiats is shown in the Table 2.

Figure 1. Diversity of powiats according to the indicator of level demographic development in 2002 and 2011



Source: own elaboration based on the Table 1

In 2011, comparing to the year 2002 decreased within each group, both the maximum and minimum values for the number of people in the non-productive age, per 100 persons of working age. The minimum natural growth in groups generally increased (the exception is in a group three), the maximum values of that characteristic decreased in the second and third group. Changes in the value of the minimum and maximum number of live-born girls per one woman of child-bearing age have the same direction as the natural growth change (exception - first group). The maximum values of the number of deaths of infants per 1000 live births in all separated groups of powiats in 2011 comparing to 2002 did not increase.

To assess the compliance of designated classification of powiats we apply measures [Podolec, 1978]:

$$S_{lk} = 1 - \frac{z_1}{n(n-1)} \tag{6}$$

$$S_{lk}^{*} = 1 - \frac{2(z_2 - n)}{\sum_{i=1}^{4}(n_{li}^2 - n_{li}) + \sum_{i=1}^{4}(n_{ki}^2 - n_{ki})} \tag{7}$$

where:

n – the number of tested objects,

$\left[z_{ij}\right]=\left[p_{ij}^{l}\right]+\left[p_{ij}^{k}\right]$ – compatibility assignment matrix, where $\left[p_{ij}^{l}\right]$ - ordering classification matrix l, for which $p_{ij}^{l}=1$, where objects O_i and O_j were assigned to the same subset, and $p_{ij}^{l}=0$, where objects O_i and O_j were in different subsets; $\left[p_{ij}^{k}\right]$ - ordering classification matrix *k*, for which $p_{ij}^{k}=1$, where objects O_i and O_j were assigned to the same subset, and $p_{ij}^{l}=0$, where objects O_i and O_j were in different subsets,

z_1 – number of ones in compatibility assignment matrix,

z_2 – number of twos in compatibility assignment matrix,

i – subset's number,

n_{li} – number of objects in i-subset created in classification 1,

n_{ki} – number of objects in i-subset created in classification k.

Measures S_{lk}, S_{lk}^{*} take a value in range $\langle 0,\ 1\rangle$. Value S_{lk} tells you, what is the probability that a randomly chosen pair of objects were compatibly allocated under the classification 1 and k. If $S_{lk}>0{,}8$ it can be concluded that the divisions are compatible. Measure S_{lk}^{*} specifies, what part of the "connections" between objects created by one of the divisions is covered in the second division.

For two classifications of powiats of the Małopolskie voivodeship, respectively in 2002 and 2011, $S_{lk}=0{,}661$ and $S_{lk}^{*}=0{,}376$, this shows very weak compatibility of divisions.

Selection of representatives of groups of powiats in 2002 was conducted by the method of the centre of gravity, as a measure of distance, Euclidean distance was selected [Pluta 1977]. How to select the representatives of the groups of this method depends on the size of groups of objects. Objects forming one-piece groups become automatically representatives. We choose representatives of the multiple-element groups (the number of elements greater than two) after the calculation of the sum of distances of each object from the other group's objects and indicate the representative object, for which the sum of the distances from other objects in the group is the smallest.

Group I represents bocheński powiat, group II – brzeski powiat, group III – krakowski powiat, and group IV – olkuski powiat.

CHARACTERISTIC OF DEMOGRAPHIC INDICATORS FOR SELECTED POWIATS

General trends, that characterize the development of the population in the years 2002-2011 in selected powiats of the Malopolskie voivodeship is shown in table 3. Linear trends presented in selected population of powiats are of good compatibility.

Table. 3. Population trend models in selected powiats in the years 2002-2011

Group	Powiat	The estimated trend model
I	bocheński	$\hat{y}_t = \underset{(415,55)}{97644,93} + \underset{(66,972)}{567,376}t, \quad t = 1, \ 2, \ ..., \ 10$ $V = 0,604\%, \qquad R^2 = 0,9$
II	brzeski	$\hat{y}_t = \underset{(263,675)}{88961,67} + \underset{(42,495)}{287,297}t, \quad t = 1, \ 2, \ ..., \ 10$ $V = 0,426\%, \qquad R^2 = 0,851$
III	krakowski	$\hat{y}_t = \underset{(2007,386)}{235572,9} + \underset{(323,520)}{2205,461}t, \quad t = 1, \ 2, \ ..., \ 10$ $V = 1,186\%, \qquad R^2 = 0,853$
IV	olkuski	$\hat{y}_t = \underset{(228,142)}{114776,1} - \underset{(36,768)}{75,721}t, \quad t = 1, \ 2, \ ..., \ 10$ $V = 0,292\%, \qquad R^2 = 0,346$

Source: own elaboration

The exception is olkuski powiat, in which the number of people in the last analyzed year increased significantly compared to previous years, which largely contributed to the mismatch of trend. If we assess the trend of olkuski powiat without taking into account the year 2011 we would get:

$$\hat{y}_t = \underset{(125,289)}{114980,1} - \underset{(22,264)}{131,483}t, \quad t = 1, \ 2,..., \ 9 \qquad (V = 1,509\%; \ R^2 = 0,832)$$

.

Models for the powiats of the first three groups provide for further increases in population, if you continue the trend so far. This can not be said about the representative of Group IV, olkuski powiat, in which the model predicts a further decrease in the number of population.

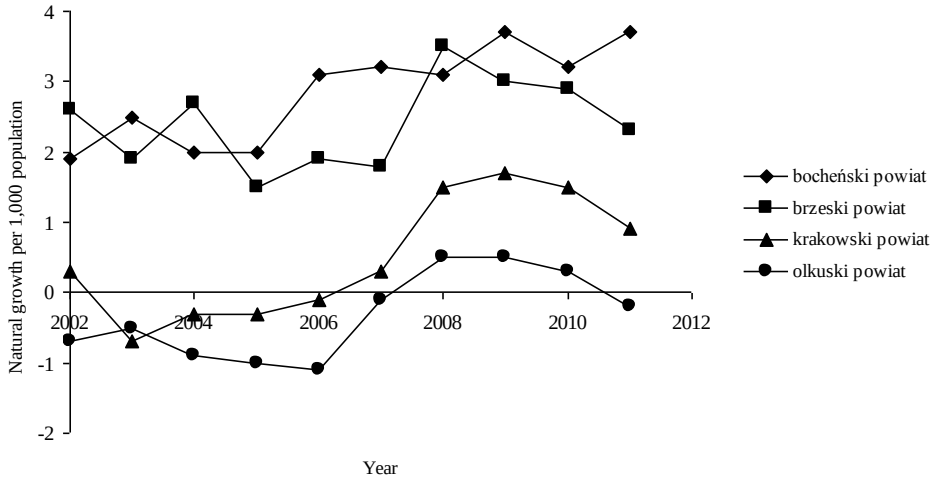
The observed changes in the characteristics, on the basis of which we defined the relative indicators of demographic development for selected powiats in the years 2002-2011 are presented at figures 2-5. On this basis, we conclude that:

– bocheński and brzeski powiats had throughout analyzed period of time a positive natural growth;

– only in bocheński powiat in the last analyzed year, natural growth was higher than in 2010;

– the lowest natural growth indicator in almost all years (an exception is the year 2003) had olkuski powiat;

Figure 2. Natural growth per 1,000 population in the years 2002-2011



Source: own elaboration

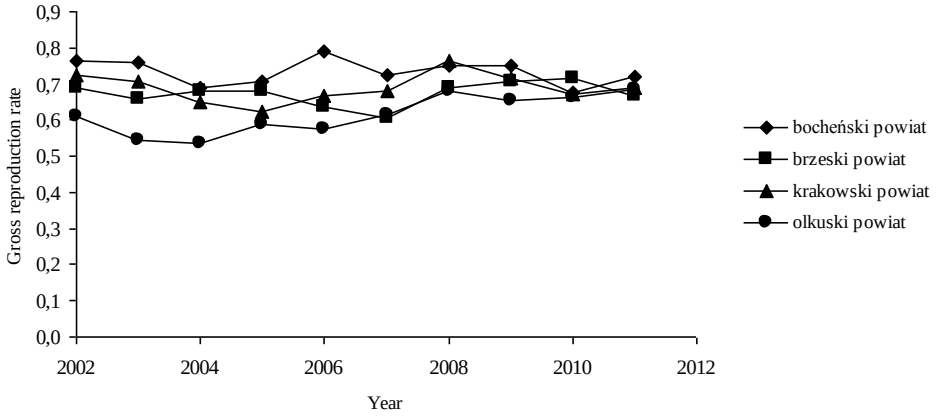
– the number of live-born girls per one woman is currently of childbearing age, showed throughout the analyzed period of time, slight fluctuation;

– differences (in absolute value) in the value of the gross reproduction rate in representative powiats decreased from year to year so that in 2011, were no more than 0,055;

– in 2002, the non-productive age population per 100 persons of working age was the biggest in brzeski powiat (70,2), the lowest in olkuski powiat (60,1);

– average rates of changes in non-productive age population per 100 persons of working age were less than 1, indicating that from year to year in these powiats this indicator decreased by 1% (olkuski powiat), and 2% (bocheński, brzeski, krakowski powiats);

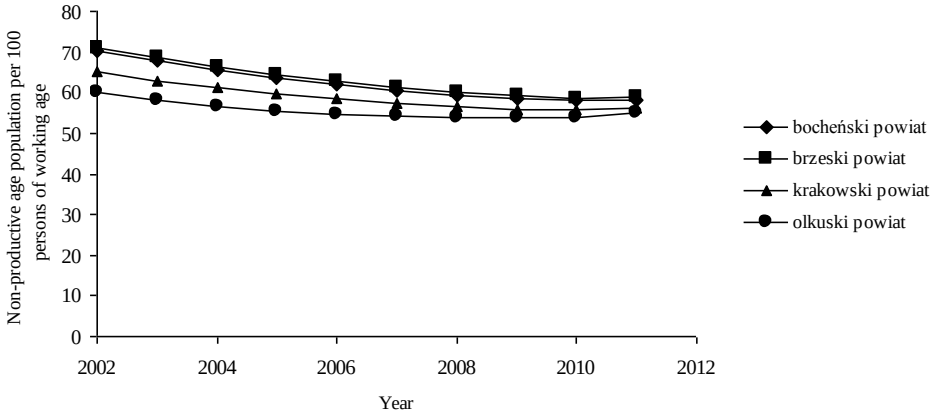
Figure 3. Gross reproduction rate in the years 2002-2011



Source: own elaboration

– in 2011, non-productive age population per 100 persons of working age in all powiats did not exceed 59,0 (brzeski powiat) and was not lower than 54,9 (olkuski powiat);

Figure 4. Non-productive age population, per 100 persons of working age in the years 2002-2011



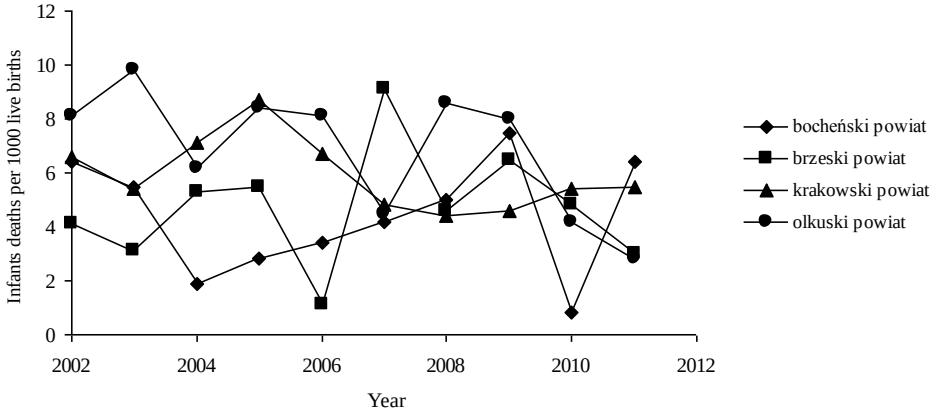
Source: own elaboration

– the number of infant deaths per 1000 live births in powiats in analyzed period of time did not show constant trends;

– in 2002, the number of infant deaths per 1,000 live births was the highest in olkuski powiat (8,1), but from year to year in this powiat decreased on average by 11% and in 2011, has reached a value of 2,8;

– in krakowski and brzeski powiats number of infant deaths per 1000 live births in the last analyzed year was lower than in the first year of analysis; in turn, within bocheński powiat, in those years was the same (6,4).

Figure 5. The number of infant deaths per 1000 live births during the years 2002-2011



Source: own elaboration

SUMMARY

1. The analysis was based on selected demographic indicators, which may decide about the development of the population in the area.
2. Powiats of the Małopolskie voivodeship are diverse in terms of the level of demographic development. We can distinguish four groups of powiats with similar characteristics describing analyzed phenomenon.
3. Assessment of the demographic development is definitely higher in powiats of South and central part of the Małopolskie voivodeship.
4. Linear trends of population in selected powiats generally provide further increase in population (the exception is olkuski powiat).
5. Natural growth in bocheński and brzeski powiats in all analysed years was much larger than in krakowski and olkuski powiats.

6. The differences in the number of live-born girls per one woman who is currently of childbearing and non-productive age population, per 100 persons of working age, in selected powiats were decreasing from year to year.
7. The number of infant deaths per 1000 live births in both powiats, as well as in years was varied.

REFERENCES

Hellwig Z. (1968) Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr, Przegląd Statystyczny nr 4.

Pluta W. (1977) Wielowymiarowa analiza porównawcza w badaniach ekonomicznych, PWE, Warszawa.

Pociecha J., Podolec B., Sokołowski A., Zając K. (1988) Metody taksonomiczne w badaniach społeczno-ekonomicznych, PWN, Warszawa.

Podolec B., Sokołowski A., Woźniak M., Zając K. (1978) Metody badania zgodności poziomu rozwoju demograficznego i gospodarczego, w: Statystyka społeczno-ekonomiczna w Polsce. Stan i perspektywy, Warszawa.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No.2, 2013, pp. 19 – 28

SAMPLE SIZE AND STRUCTURE FOR MULTILEVEL MODELLING: MONTE CARLO INVESTIGATION FOR THE BALANCED DESIGN

Edyta Łaszkiewicz
Department of Spatial Econometrics, University of Lodz
e-mail: elaszkiewicz@uni.lodz.pl

Abstract: The aim of the study is to examine the robustness of the estimates and standard errors in the case of different structure of the sample and its size. The two-level model with a random intercept, slope and fixed effects, estimated using maximum likelihood, was taken into account. We used Monte Carlo simulation, performed on a sample of the equipotent groups.

Keywords: multilevel model, Monte Carlo, sample size

INTRODUCTION

Sufficient sample is one of the most important problem in the multilevel modelling (see e.g. Mass and Hox [2004, 2005] or Snijders [2005] to mention just a few). The most basic design conditions like a number of groups at each level of the analysis and its size determine the ability to obtain accurate (unbiased) estimates of the regression coefficients, standard errors and power of tests[1]. Additionally, Busing [1993] found out the insufficient sample size (10 to 50 groups with 5 or 10 individuals) might be responsible for the model nonconvergence. Despite the asymptotic properties of the multilevel models estimators (like REML or IGLS), due to which larger sample guarantees the bias reduction, in the centre of interest is the downward limit of the sample [Mass and Hox 2005]. Accordingly, the adequate (sufficient) sample size can be define as such the minimum sample, which guarantees the unbiasedness (or more precisely: acceptable low size of the bias). Such definition is consistent with Snijders and Bosker [1993], who use the

[1]Other factors like the estimation method, proportion of singletons, value of the intraclass correlation, collinearity or model complexity, which also might affect the estimates, are not wider describe as they are not take into consideration in this study.

term ‘conditionally optimal’ to characterise the sample size which allows to yield the minimal standard errors for the particular parameters or other constraints. Although the literature about the sufficient sample size is large, there is still no consensus how it should looks like, what is the result of i.e. using different simulation conditions and/or simulation designs. Let review only the guidelines for 2-level models estimated using the balanced sample. We start from the recommendations for the unbiased parameter and standard errors estimates, then concentrate on the suggestions based on the maximization the power of the tests.

Kreft [1996] recommended ‘30/30’ rule which means minimum 30 observations per group and minimum 30 units at each level of the analysis to unbiased estimate all parameters and their standard errors. As pointed by Mass and Hox [2005], such number of groups gives unbiased results except the standard error estimates of the random effects at the level-2. Accordingly, Hox [1998] recommended minimum 20 observations for 50 groups if the cross-level interaction is tested. Although both the number of groups and the number of observations per group are important to obtain the unbiased results, the sensitivity of the fixed and random effects (and their standard errors) estimates to above is different. When the accuracy of the variance components estimates is influenced strongly by the number of groups, fixed effects estimates are less susceptible to the data sparseness. Similar conclusions were drawn by Newsom and Nishishiba [2002] and Clarke and Wheaton [2007], who confirmed that the unbiased estimates of the fixed effects might be received even for the small sample. As the variance components estimates are often in the main centre of the interest in the multilevel models, additional suggestions dealing with the random effects were concerned in detail. Mok [1995] noticed that 5 groups at the second level gives a notably bias of the variance estimates, while Clarke and Wheaton [2007] suggested at least 10 observation per group for at least 100 groups is needed to obtain the unbiased estimate of the intercept variance. If the slope variance is estimated they recommended at least 200 groups with minimum 20 observation per group. Although for the accurate estimates of the variance components (often underestimated) at least 100 units is needed, in practise such sample would be hard to obtain [see Mass and Hox 2004]. According to all of the mentioned guidelines, rather than the large number of observations per unit, the large number of groups seems to be more important to receive the accurate estimates.

Sufficient sample size is considered also due to the accuracy of the standard errors estimates but such investigations are in the minority [Mass and Hox 2005]. In the simulation research the most common way to validate standard errors estimates is by checking the accuracy of the significance test or the coverage of the confidence interval (generated by using standard normal distribution and gamma distribution)[2]. Accordingly, Browne and Draper [2000] showed, using IGLS and

[2] Although the assumption about the normality is not optimal, especially if the confidence intervals of the random effects are considered (because of the lack of the confidence

RIGLS estimators, that for at least 48 groups the coverage of the nominal 95% intervals is unbiased (for the fixed effects estimates), when the intervals for the covariance matrix parameters are substantially biased (below 95%). Similarly, Mass and Hox [2005] found out that negative influence of as small as 30 number of groups is small for the standard errors of the fixed effect coefficients (6.0% and 6.4% for the intercept and regression coefficient) and higher for the standard errors of the variance components (around 9% for the level-2 intercept and slope variances). Additionally, in a large (5760 conditions) Monte Carlo experiments Bell et al. [2010] found out that for each type of the predictor variable, treated as the fixed effect, estimated confidence interval coverage is rather constant and higher than for the level-2 estimates, what is consistent with the previous reviewed researches. Finally, according to Snijders [2005] group size is less important for the power of the tests than the number of groups, what is similar to the results for the estimates. The only limitation of the small group size for the power of testing are the random slope variances. As the power of the tests is the result of the standard error size, consistency of the conclusions seems to be natural.

There is no agreement about the negative influence of the data sparseness on the convergence. Although Bell et. al. [2010], Mass and Hox [2004] found out that there is no problem with the model convergence using ML and RIGLS estimator, according to Busing's [1993] findings such problem might occurs if the sample is too small. In practice the generalisation of the presented rules is always limited to the specific cases, e.g. the type of the estimated effect (random, fixed, interaction, cross-level, etc.) or the estimation method.

In the literature, to set the optimal/sufficient sample size, in the multilevel modelling, the simulation method has been chosen more frequently. Another way is to use the approximate formula, relating effect size and standard errors to statistical power of the significance test [Snijders and Bosker 1993]. As was showed by Snijders [2005], the way of computing the sufficient sample size depends on the parameter estimates which the researcher is interested in. Also Moerbeek et al. [2001] presented formulas for calculating the optimal design (the sample size) for the 2-level models with detailed evaluation using *D*-optimality and *L*-optimality criteria. Although the approximate formula seems to be faster in using, its limitation (like the lack of the generalisation) makes Monte Carlo simulation more flexible tool for evaluation the sufficiency of the sample size.

The motivation for this paper is to evaluate by the Monte Carlo simulation the influence of the sample size and its structure on the estimates biasness. The fixed and random parameter estimates and their standard errors are examined in the 2-level model estimated by maximum likelihood (ML). The rest of the paper is divided into the simulation method description and the results discussion.

---

symmetry), in most of the simulation studies such method of evaluation of the standard errors estimates are using [see e.g. Busing 1993, Van der Leeden et al. 1997].

SIMULATION DESIGN

The 2-level model (for the continuous outcome variable Y_{ij}) with two explanatory variables $X_{1,ij}$, $X_{2,ij}$ on the level-1 was examined. The random (or stochastic) part of the model contains: residual error terms at the level-2: $\mu_{0,j} \sim N(0, \sigma^2_{\mu_0})$, $\mu_{1,j} \sim N(0, \sigma^2_{\mu_1})$ and individual-level (level-1) residuals $\varepsilon_{ij} \sim (0,1)$. The fixed (or determinist) part contains β_0, β_1, β_2 coefficients. This model can be written as [Goldstein 2010]:

$$Y_{ij} = \beta_{0,j} + \beta_{1,j} X_{1,ij} + \beta_2 X_{2,ij} + \varepsilon_{ij}, \quad (1)$$
$$\beta_{0,j} = \beta_0 + \mu_{0,j},$$
$$\beta_{1,j} = \beta_1 + \mu_{1,j},$$

where: $i = 1, \dots, M$ and $j = 1, \dots, J$. We assume the structure of the variance-covariance matrix as in the standard multilevel models: $\forall i \neq i' \ \mathrm{cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = 0$, $E(\mu_{1,j}) = E(\mu_{0,j}) = 0$, $j \neq j' \ \mathrm{cov}(\mu_{0,j'}, \mu_{0,j}) = \mathrm{cov}(\mu_{1,j'}, \mu_{1,j}) = 0$, $\mathrm{cov}(\mu_{0,j}, \varepsilon_{ij}) = \mathrm{cov}(\mu_{1,j}, \varepsilon_{ij}) = 0$. The values of the predictors were drawn independently from the normal distribution with variance 1. Model (1) was estimated via ML.

Three conditions were varied in the simulation: (1) number of groups J={5, 10, 20, 30, 50, 70, 90}, (2) number of observations per group M={5, 10, 20}, (3) values of the parameters (in Table 1). As the value of the intraclass correlation (ICC) influence the results the two different values of ICC were tested. The ICC was calculated as follows: $(\sigma^2_{\mu_0} + \sigma^2_{\mu_1})/(\sigma^2_{\mu_0} + \sigma^2_{\mu_1} + \sigma^2_{\varepsilon})$.

Table 1.Target values of parameters

variant/parameter	β_0	β_1	β_2	$\sigma^2_{\mu_0}$	$\sigma^2_{\mu_1}$	ICC
1	0.60	0.50	0.30	0.50	0.40	0.47
2	0.20	0.30	0.80	0.20	0.30	0.33
3	0.30	0.70	0.80	0.20	0.30	0.33

Source: own calculation

The large variation of the groups number was evaluated because this factor might affects the estimate much more than the group size. For each of the 63 conditions 1000 datasets were simulated using user-written syntax in STATA[3] based on the xtmixed command which allows for the multilevel model estimation.

The accuracy of the estimates was indicated using two measures commonly used in the evaluation of the simulation results:

- Relative bias of an estimator $\widehat{\theta_l}$ for parameter θ_l, defined as:

[3] Monte Carlo simulation syntax is available at: https://sites.google.com/site/elaszkiewicz.

$$B(\widehat{\theta_l}) = \frac{\overline{\widehat{\theta_l}} - \theta_l}{\theta_l} \cdot 100\%, \quad (2)$$

where $\overline{\widehat{\theta_l}}$ is the arithmetic mean calculated from *K*=1000 simulation runs of $\widehat{\theta_{lk}}$. According to Hoogland and Boomsm (1998) unbiased estimates are those for which the relative bias is less than 5%. The relative biases were calculated to evaluate only the parameter estimates.

- Rate of the coverage, calculated as:

$$C\left(se(\widehat{\theta_l})\right) = \frac{\sum C\left(se(\widehat{\theta_{lk}})\right)}{K} \cdot 100\% \quad (3)$$

$$C\left(se(\widehat{\theta_{lk}})\right) = \begin{cases} 1 \; if \; \theta_l \in CI \\ 0 \; if \; \theta_l \notin CI \end{cases},$$

where $se(\widehat{\theta_{lk}})$ is the estimated standard error of the $\widehat{\theta_{lk}}$ at *k*-th run, CI is the 95% confidence interval established separately for the fixed effects as: $\widehat{\theta_{lk}} \pm u_\alpha \cdot se(\widehat{\theta_{lk}})$ and for the random effects as: $\exp(\ln(\widehat{\theta_{lk}}) \pm u_\alpha \cdot \frac{1}{\widehat{\theta_{lk}}} \cdot se(\widehat{\theta_{lk}})$. The indicator was used to check the bias of the standard error estimates.

Additionally, to compare different conditions ANOVA (for the parameter estimates) and logistic regression (for the confidence interval evaluation) were used.

RESULTS AND DISCUSSION

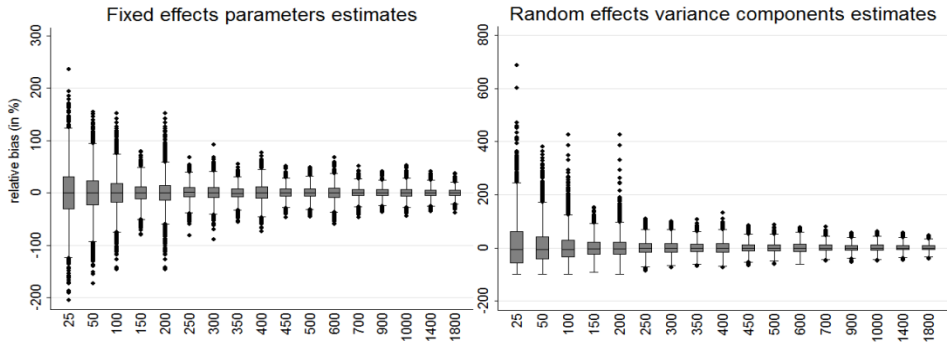
The convergence of model was achieved almost in each case, even for the smallest sample size. However, for the sample of 5 groups with 5 observations per group it was more frequently impossible to estimate standard errors for the random effects variance components due to the singular variance-covariance matrix of the random effects [see e.g. Henderson 1986].

Parameter estimates

The average relative bias for the fixed effect estimates (0.01%) was lower from the random effect estimates bias, which was 1.07%. Although, there was no significant differences in the relative bias across the fixed parameter estimates, the biases of the $\widehat{\sigma}^2_{\mu_0}$ and $\widehat{\sigma}^2_{\mu_1}$ were significantly different and higher for the first one. Additionally, there was no significant differences between the relative bias of the fixed parameter estimates when three variants of the target values of the parameters were compared. However, the influence of the ICC on the random effect estimates was revealed. For the higher value of the ICC, the lower relative bias of the random effect estimates was achieved. This is consistent with e.g. Newson and Nishishiba [2002], who showed that the ICC value determines the accuracy of the estimates.

The unbiased estimates were achieved for the fixed effect estimates for each of the simulated sample size (Figure 1). Even for the sample as small as 25 observations the relative biases were less than 1% for all of the fixed parameters estimates. In the case of the random effects estimates only for the sample of 25 observations the results were biased. The relative bias for the random intercept variance estimates was 16% and for the random slope variance estimates almost 10%. The relative biases less than 1% for the variance components estimates were achieved for the sample size equal to 100 or higher. Additionally, as the sample size increases, the variance of the parameters estimates has decreased strongly.

Figure 1. Effect of group size on the relative bias of the parameter estimate



Source: own calculation

Table 2. Relative biases (in %) and significance of the group size effect

parameter /group size	5	10	20	*p*-value*
β_0	-0.15	0.10	-0.10	0.89
β_1	-0.22	0.01	0.43	0.27
β_2	-0.20	0.08	0.16	0.22
$\sigma^2_{\mu_0}$	2.70	0.92	0.55	0.00
$\sigma^2_{\mu_1}$	1.97	0.53	-0.25	0.00

* *p*-value for the effect of group size on the relative bias of the parameter estimate

Source: own calculation

Although the unbiased results might be obtained even if 5 observation per unit occurs, the sensitivity of the fixed and random effects estimates for the group size was different (presented in Table 2). Only for the variance components estimates ($\sigma^2_{\mu_0}, \sigma^2_{\mu_1}$) increase of the group size affects significantly the value of the relative bias of the estimates. Such results are similar to Newsom and Nishishiba [2002], Clarke and Wheaton [2007].

According to Table 3, all of the fixed effects estimates are unbiased for the sample consisting of 5 groups. In the opposite, the random effects estimates are biased in such a case. It means that for the unbiased estimates for all of the

parameters the sample of at least 10 groups with 5 observations per group is needed. This is less than in Kreft's '30/30' rule or as Hox [1998] suggested. The differences in the recommendations are the result of not taking into account the unbiasness of the standard errors estimates.

Table3. Relative biases (in %) and significance of the number of groups effect

parameter / nr of groups	5	10	20	30	50	70	90	*p*-value*
β_0	-0.15	-0.01	-0.08	-0.10	0.67	-0.75	0.08	0.79
β_1	0.80	0.23	-0.29	-0.11	-0.10	-0.20	0.18	0.65
β_2	-0.27	0.06	0.04	0.15	0.12	-0.03	0.02	0.89
$\sigma^2_{\mu_0}$	7.90	1.48	0.45	0.14	-0.14	0.26	-0.38	0.00
$\sigma^2_{\mu_1}$	4.73	0.63	0.21	-0.18	-0.46	0.16	0.15	0.00

* *p*-value for the effect of number of groups on the relative bias of the parameter estimate

Source: own calculation

The results showed that unbiasness of the random effects estimates depends more on the number of groups in the sample, than the group size. This conclusion is consistent with Snijders and Bosker[1994].

Standard errors

The coverage of the 95% confidence interval (CI) was similar for the fixed effect parameters (93.47%) and for the random effect estimates (94.78%). The results of the logistic regression showed that the rate of the coverage rate for the CI for the random effects depends on the ICC but the fixed effects seems to be not affected by the level of the ICC.

Table4. Influence of the ICC value on the coverage of the 95% confidence interval

parameter / nr of groups*	5	10	20	30	50	70	90
$\sigma^2_{\mu_0}$	0.00	0.00	0.23	0.79	0.95	0.90	0.15
$\sigma^2_{\mu_1}$	0.00	0.16	0.39	0.53	0.47	0.20	0.51

* *p*-value from the logistic regression, where the value of ICC was independent variable

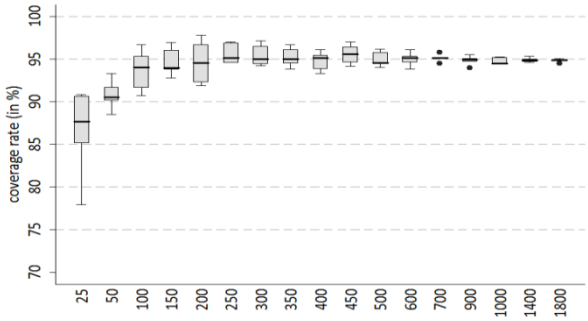
Source: own calculation

Inspired by Mass and Hox [2005], who proved that for the ICC=0.1, 0.2, 0.3, the influence of the ICC value on the coverage rate (for the random effects estimates) occurs only for the extremely small sample size (like 10 groups with 5observations), we checked if the influence of the ICC varies across the number of groups. Our results (presented in Table 4) are more detailed than Mass and Hox [2005] as for each variant of the number of groups separate regression has been done. For at least 10 groups in the sample the impact of the ICC value on the coverage of the 95% CI for random intercepts variance was statistically significant.

In comparison, the CI for random slopes variance was affected by the ICC value only in the sample of 5 groups.

Next, the sample size effect was tested. As expected, with increasing of the sample size results the rate of the coverage grows (Figure 2). For the samples of less than 100 observations at least for one of the parameters the coverage rate of the CI was lower than 90%. It means that for the unbiased standard errors estimates such sample size is the minimum.

Figure 2. Coverage rate of the 95% confidence interval by the group size



Source: own calculation

Although for the smallest sample size there are significant differences in the coverage rate of the CI for the fixed (89.70%) and random (81.52%) effects parameters, the differences decrease as the sample size increase. For the sample of 1800 observations each of the parameter achieved the coverage rate around 95%.

Table 5. Coverage rate of the 95% CI (in %) by group size with significance

parameter /group size	5	10	20	*p*-value*
β_0	93.22	93.15	93.03	0.45
β_1	92.67	93.11	92.90	0.54
β_2	93.91	94.51	94.69	0.00
$\sigma^2_{\mu_0}$	92.31	95.35	95.34	0.00
$\sigma^2_{\mu_1}$	93.80	95.94	95.93	0.00

* *p*-value from the logistic regression, where groups size was independent variable

Source: own calculation

Similar to the relative bias behaviour, the coverage rate increases when the size of the groups grows (Table 5). The positive, significant effect of rising of the group size was noticed for the variance components (boththe intercepts and the slopes) and for the one of the fixed effect parameter. However, difference between coverage rate of the CI for fixed and random parameters was small.Additionally, the significance of the number of groups effect was tested (Table 6). For each of the parameter the coverage rate of the 95% CI depends on the number of groups in

the sample. In the sample of 5 groups 7.09-12.64% of the cases were outside of the 95% CI. As the number of groups increased, the coverage rate also increased.

Table 6. Coverage rate of the 95% CI (in %) by number of groups with significance

parameter / nr of groups	5	10	20	30	50	70	90	*p*-value*
β_0	89.29	92.12	93.22	93.91	94.32	94.60	94.48	0.00
β_1	88.01	91.28	93.36	93.92	94.44	94.69	94.56	0.00
β_2	92.91	93.50	94.82	94.69	94.70	94.74	95.22	0.00
$\sigma^2_{\mu_0}$	87.36	93.48	95.94	96.16	96.06	95.70	95.64	0.00
$\sigma^2_{\mu_1}$	92.21	94.79	96.16	96.49	95.98	95.56	95.36	0.00

* *p*-value from the logistic regression, where number of groups was independent variable

Source: own calculation

Surprisingly, for each variant of the number of groups the coverage rate was higher for the variance components estimates. In the case of the 30 groups the non-coverage rate for the fixed effects parameters was 6% (the same as in Mass and Hox [2005]), however for the random effects parameters we obtained 3%, when in Mass and Hox [2005] study the non-coverage rate was 8%. The differences in the results are the effect of the way how the CIs were build. Mass and Hox [2005] used standard normal distribution to establish the CIs for the variance components parameters, when we used Wald-type CIs, which were better performed.

CONCLUSIONS

Our results are consistent with the previous investigation. The unbiased estimates of the fixed effects parameters might be obtained even for the extremely small samples. The structure of the sample (number of groups and group size) do not affect negatively the fixed effects estimates. For the unbiased estimates of the variance components both design conditions are important, but only for too small (less than 10) number of groups results were biased. The evaluation of the standard errors estimates once again proved the major role of the number of groups to guarantee the satisfactory coverage rate of the CIs. Also, we showed that for the random effects the Wald-type confidence interval are better than based on the standard normal distribution. Our results might be generalized but only to the presented conditions. Additional researches are required to examine more advanced multilevel model specification, e.g. cross-classified or multiple membership.

REFERENCES

Bell B. A., Morgan G. B., Kromrey J. D., Ferron, J. M. (2010) The impact of small cluster size on multilevel models: a Monte Carlo examination of two-level models with binary

and continuous predictors, JSM Proceedings, Survey Research Methods Section, pp. 4057-4067.

Browne W. J., Draper D. (2000) Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models, Computational Statistics, 15, pp. 391-420.

Busing F. (1993) Distribution characteristics of variance estimates in two-level models, Unpublished manuscript, Leiden University.

Clarke P., Wheaton B. (2007) Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods, Sociological Methods & Research, 35, pp. 311-351.

Goldstein H. (2010) Multilevel statistical models (4th ed.), New York: Hodder Arnold

Henderson C. R. (1986) Estimation of singular covariance matrices of random effects, Journal of Dairy Science 69.9, pp. 2379-2385.

Hoogland J., Boomsma, A. (1998) Robustness studies in covariance structure modeling: An overview and a meta-analysis, Sociological Methods and Research, 26(3), pp. 329-367.

Hox J. J. (1998) Multilevel modeling: When and why, [in:] Balderjahn I., Mathar R., Schader M., Classification, data analysis, and data highways, New York: Springer Verlag, pp. 147-154.

Kreft I. G. G. (1996) Are multilevel techniques necessary? An overview, including simulation studies, Unpublished manuscript, California State University at Los Angeles.

Maas C. J. M., Hox J. J. (2004) Robustness issues in multilevel regression analysis, Statistica Neerlandica, 58, pp. 127-137.

Maas C. J. M., Hox, J. J. (2005) Sufficient sample sizes for multilevel modeling, Methodology, 1, pp. 86-92.

Mass C. J., Hox J. J. (2002) Robustness of multilevel parameter estimates against small sample sizes. Unpublished Paper, Utrecht University.

Moerbeek M., Van Breukelen G. J., Berger M. P. (2001) Optimal experimental designs for multilevel logistic models, Journal of the Royal Statistical Society, Vol.50(1), pp. 17-30.

Mok M. (1995) Sample size requirements for 2-level designs in educational research, Unpublished manuscript, Macquarie University.

Newsom J. T., Nishishiba M. (2002) Nonconvergence and sample bias in hierarchical linear modeling of dyadic data. Unpublished Manuscript, Portland State University.

Snijders T. A. B. (2005) Power and Sample Size in Multilevel Linear Models', [in:] Everitt B.S., Howell D.C. (eds.) Encyclopedia of Statistics in Behavioral Science, Vol. 3, Wiley, pp. 1570–1573.

Snijders T. A. B., Bosker R. J. (1993) Standard Errors and Sample Sizes for Two-Level Research, Journal of Educational Statistics, Vol. 18, No. 3, pp. 237-259.

Van der Leeden R., Busing F. (1994) First iteration versus IGLS RIGLS estimates in two-level models: A Monte Carlo study with ML3, Unpublished manuscript, Leiden University.

Van der Leeden R., Busing F., Meijer E. (1997) Applications of bootstrap methods for two-level models, Paper presented at the Multilevel Conference, Amsterdam.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 29 – 38

ON AN UPPER GAIN BOUND FOR STRATEGIES WITH CONSTANT AND PROPORTIONAL NUMBER OF ASSETS TRADED[1]

Rafał Łochowski
Department of Mathematics and Mathematical Economics
Warsaw School of Economics
e-mail: rlocho@sgh.waw.pl
and
Department of Core Mathematics and Social Sciences
Prince Mohammad Bin Fahd University, Saudi Arabia
e-mail: rlochowski@pmu.edu.sa

Abstract: We introduce general formulas for the upper bound of gain obtained from any finite-time trading strategy in discrete and continuous time models. We consider strategies with constant number of assets traded and strategies with proportional number of assets traded. Unfortunately, the estimates obtained in the discrete case become trivial in the continuous case, hence we introduce transaction costs. This leads to the interesting estimates in terms of the so called truncated variation of the price series. We apply the obtained estimates in specific cases of financial time series.

Keywords: trading strategy, transaction costs, truncated variation, AR(1) process, Wiener process, Ornstein-Uhlenbeck process, random walk, the Black-Scholes model

INTRODUCTION

In [Łochowski 2010] we considered the following investment problem: let $P(n)$ and $Q(n), n=0,1,2,\ldots$ be two non-stationary time series representing the evolution of the prices of futures contracts for two commodities P and Q. Assuming that the prices of P and Q are cointegrated, such that for some positive

[1]Research financed by the National Science Centre in Poland under decision no. DEC-2011/01/B/ST1/05089.

α and β the process αP-βQ is stationary, we considered the following long-run investment strategy: buy the combination αP-βQ when its value fails below some threshold $-a$ and sell it when the value of the combination exceeds threshold a. Buying the combination physically means entering into α long positions in commodity P contracts and entering into β short positions in commodity Q contracts. Similarly, selling the combination physically means entering into α short positions in commodity P contracts and entering into β long positions in commodity Q contracts.

Naturally, similar problem may be considered for larger number of cointegrated commodity contract prices.

The natural question arises whether the strategy described gives the best possible gains or, at least, to compare it with some upper bound for the best possible gain. To do so, in this article we introduce very general formulas for the upper bound for gain obtained from any finite-time trading strategy in discrete and continuous time. We consider two types of strategies:

1. Strategies with constant number of contracts or assets traded. In these models one always buys the same number of contracts or assets.
2. Strategies with proportional number of contracts or assets traded. In these models one always invests all money earned in the previous trading.

The bounds obtained are closely related to the path variation of the price time series (which, on the other hand, is closely related to volatility). Unfortunately, the bounds obtained in the discrete case become trivial in the continuous case, hence we introduce (constant or proportional) transaction costs. This leads to interesting bounds in terms of the so called truncated variation of the price series.

We apply the obtained bounds in specific cases. In the models with constant number of contracts traded we assume the AR(1) structure of the cointegrated price series and in the models with proportional number of assets traded we assume exponential random walk structure of the price series. The bounds obtained for the maximal gain in both cases reveal quite strong boundedness properties – they have finite moments of all orders.

UPPER BOUND FOR GAIN IN MODELS WITH CONSTANT NUMBER OF CONTRACTS TRADED

Discrete case

Let $R(n)$ denote the value of the linear combination αP-βQ of α long positions in commodity P contracts and β short positions in commodity Q contracts (or the value of linear combination of greater number of contracts, under the condition that it is stationary) at the moment $n=0,1,2,\ldots$ Buying this combination

at moments $0 \le b_1 < b_2 < \ldots < b_n < T$ and selling it (i.e. closing all long and short positions) at moments $0 < s_1 < s_2 < \ldots < s_n \le T$ such that $b_1 < s_1 < b_2 < s_2 < \ldots$ we obtain the following gain

$$G = R(s_1) - R(b_1) + R(s_2) - R(b_2) + \ldots + R(s_n) - R(b_n) \tag{1}$$

(note that G may be negative). The immediate upper bound for G reads as

$$G \le sup_n sup_{0 \le t_0 < t_1 < \ldots < t_n \le T} \sum_{i=1}^{n} |R(t_i) - R(t_{i-1})|. \tag{2}$$

The right-hand side of Eq. (2) is simply the total path variation of the time series $R(n)$ and we will denote it as

$$TV(R,[0;T]) = sup_n sup_{0 \le t_0 < t_1 < \ldots < t_n \le T} \sum_{i=1}^{n} |R(t_i) - R(t_{i-1})|. \tag{3}$$

Due to the triangle inequality

$$|a-c| \le |a-b| + |b-c|$$

we simply obtain

$$|R(t_i) - R(t_{i-1})| \le |R(t_i) - R(t_i - 1)| + |R(t_i - 1) - R(t_i - 2)| + \ldots + |R(t_i - (t_i - t_{i-1}) + 1) - R(t_{i-1})|$$

$$= \sum_{i=t_{i-1}+1}^{t_i} |R(i) - R(i-1)|.$$

Hence

$$G \le sup_n sup_{0 \le t_0 < t_1 < \ldots < t_n \le T} \sum_{i=1}^{n} |R(t_i) - R(t_{i-1})| \le \sum_{i=1}^{T} |R(i) - R(i-1)|. \tag{4}$$

(On the other hand, the opposite equality:

$$sup_n sup_{0 \le t_0 < t_1 < \ldots < t_n \le T} \sum_{i=1}^{n} |R(t_i) - R(t_{i-1})| \ge \sum_{i=1}^{T} |R(i) - R(i-1)|$$

is also true and we have $TV(R,[0;T]) = \sum_{i=1}^{T} |R(i) - R(i-1)|$.)Knowing the specific structure of the series $R(n)$ we may calculate the distribution of the random variable $\sum_{i=1}^{T} |R(i) - R(i-1)|$ or e.g. certain characteristics of this distribution (like the expected value).

Remark. Reasoning similarly it is easy to obtain more accurate bound for G -the positive path variation of the time series $R(n)$ - which may be calculated with the following formula

$$UTV(R,[0;T])=\sum_{i=1}^{T} max\{R(i)-R(i-1),0\}.$$

To illustrate the possible application of the obtained bound in a specific case, let usassume as in [Łochowski 2010] that $R(n)$ is a stationary, mean zeroAR(1) process, such that for some $\gamma\in(-1;1)$ and the sequence $Z(0),Z(1),Z(2),\ldots$ of independent random variables with normal $N(0,\sigma^2)$ distribution we have

$$R(n+1)=\gamma R(n)+Z(n). \tag{5}$$

Knowing that $R(0)$ and $Z(0),Z(1),Z(2),\ldots$ are independent, we obtain that (cf. [Łochowski 2010]) $R(n), n=0,1,\ldots$ has normal distribution $N(0,\sigma^2/(1-\gamma^2))$. Hence

$$\begin{aligned} R(i)-R(i-1)&=(\gamma-1)R(i-1)+Z(n)\sim N\left(0,\frac{(1-\gamma)^2\sigma^2}{1-\gamma^2}+\sigma^2\right)\\ &\sim N\left(0,\frac{2\sigma^2}{1+\gamma}\right). \end{aligned} \tag{6}$$

From (6) we easily obtain the expected value of the variable $TV(R,[0;T])$.

$$\begin{aligned} \mathrm{E}G&\le \mathrm{E}\sum_{i=1}^{T}|R(i)-R(i-1)|=T\mathrm{E}|R(1)-R(0)|\\ &=T\frac{\sqrt{2}\sigma}{\sqrt{1+\gamma}}\mathrm{E}|Y|=T\frac{\sqrt{2}\sigma}{\sqrt{1+\gamma}}\frac{\sqrt{2}}{\sqrt{\pi}}=\frac{2}{\sqrt{\pi}}\frac{\sigma}{\sqrt{1+\gamma}}T, \end{aligned}$$

where Y is a standard normal random variable and $\pi\approx 3.1415926$.

Continuous case with constant transaction costs

Now let us turn to the situation when the price process is observed in continuous time and we may sell or buy the combination of the contracts at any time between the moments 0 and T. As it was already noticed in [Łochowski 2010], for $\gamma\in(0;1)$ the continuous counterpart of the AR(1) process given by the recursion (5) is the Ornstein-Uhlenbeck process given by the following sde (stochastic differential equation):

$$dR(t)=ln(\gamma)R(t)dt+\sigma\sqrt{\frac{2ln(\gamma)}{\gamma^2-1}}dW(t), \tag{7}$$

where $W(t), t\geq 0,$ is a standard Brownian motion.

It is well known that the total variation of any process being the solution of any sde driven by a standard Brownian motion, $dR(t)=\mu(t,R(t))dt+\sigma(t,R(t))dW(t),$ satisfying some mild regularity conditions (e.g. the continuity of the functions μ,σ and $\sigma\neq 0$) has infinite total variation, given by the right-hand side of Eq. (2) and Eq. (3) (see [Revuz and Yor 2005, Chapt. IV Proposition 1.2]),

$$TV(R,[0;T])=+\infty.$$

Thus, our estimate (2) becomes trivial. Notice however, that it is no longer the case, when we introduce constant transaction costs.

Let $c/2>0$ be the value of a constant commission, paid for every transaction (regardless of the transaction value). In this setting, the right-hand side of Eq. (1) shall be replaced with

$$\begin{aligned} G&=R(s_1)-c/2-R(b_1)-c/2+R(s_2)-c/2-R(b_2)-c/2+\ldots+R(s_n)-c/2-R(b_n)-c/2 \\ &=R(s_1)-R(b_1)-c+R(s_2)-R(b_2)-c+\ldots+R(s_n)-R(b_n)-c \end{aligned} \tag{8}$$

and the estimate (1) becomes

$$\begin{aligned} G&\leq sup_n sup_{0\leq t_0<t_1<\ldots<t_n\leq T}\sum_{i=1}^{n}\left(\left|R(t_i)-R(t_{i-1})\right|-c\right) \\ &\leq sup_n sup_{0\leq t_0<t_1<\ldots<t_n\leq T}\sum_{i=1}^{n}max\left(\left|R(t_i)-R(t_{i-1})\right|-c,0\right). \end{aligned} \tag{9}$$

The last estimate we will call truncated variation of the process $R(t), t\geq 0,$ and we will denote it as

$$TV^c(R,[0;T])=sup_n sup_{0\leq t_0<t_1<\ldots<t_n\leq T}\sum_{i=1}^{n}max\left(\left|R(t_i)-R(t_{i-1})\right|-c,0\right) \tag{10}$$

Remark. Again, more accurate bound for G in continuous time setting with constant transaction costs is the upward truncated variation of the process $R(t), t\geq 0,$ which is defined with the following formula

$$UTV^c(R,[0;T])=sup_n sup_{0\leq t_0<t_1<\ldots<t_n\leq T}\sum_{i=1}^{n}max\left\{R(t_i)-R(t_{i-1})-c,0\right\}.$$

It is possible to prove that the truncated variation is always finite for any process $R(t), t\geq 0,$ with continuous (cf. [Łochowski 2011]), càdlàg (cf. [Łochowski 2012]) or even regulated paths (cf. [Ghomrasni and Łochowski 2013]). By a càdlàg path we mean a path which is right continuous, $lim_{t\downarrow t_0}R(t)=R(t_0),$ and its left limits, $lim_{t\uparrow t_0}R(t)=R(t_0-),$ exist but may not coincide with the right limit

limits.A regulated path is a path with left $lim_{t\uparrow t_0} R(t)=R(t_0-)$ and right limits $lim_{t\downarrow t_0} R(t)=R(t_0+)$, which may not coincide with the value $R(t_0)$.

The properties of the truncated variation as the function of parameters c and T are well known for broad class of stochastic processes (see [Łochowski 2012], [Bednorz and Łochowski 2012]). In particular, for the process given by Eq. (7) we have the following estimate of the exponential moments of $TV^c(R,[0;T])$, stemming from [Bednorz and Łochowski 2012, Theorem 2]:

$$\mathrm{E}exp\left(\lambda TV^c(R,[0;T])\right)\le 2exp\left(\lambda^2 T\alpha(\gamma,\sigma)+\lambda Tc^{-1}\beta(\gamma,\sigma)+\lambda\theta(T,\gamma,\sigma)\right)\times$$
$$\times\left(1+8\lambda\eta(T,\gamma,\sigma)exp\left(\lambda^2\eta(T,\gamma,\sigma)\right)\right).$$

Here, $\alpha(\gamma,\sigma)$ and $\beta(\gamma,\sigma)$ are constants depending on γ and σ only and $\theta(T,\gamma,\sigma)$ and $\eta(T,\gamma,\sigma)$ are constants depending on T,γ and σ.

UPPER BOUND FOR GAIN IN MODELS WITH PROPORTIONAL NUMBER OF ASSETS TRADED

Discrete case

Now let us turn to another situation, when we buy assets (or portfolio of assets) but we exclude the possibility of a short sale. Let $S(n)=S(0)exp(V(n)), n=0,1,2,\ldots$ denotes the price of the asset at the moment n. Again, we assume that we buy this asset at moments $0\le b_1<b_2<\ldots<b_n<T$ and sell it at moments $0<s_1<s_2<\ldots<s_n\le T$ such that $b_1<s_1<b_2<s_2<\ldots$ but contrary to the previous strategy, where we were buying constant amount of the combination of contracts, we invest all money available. To make it clear, we will calculate the return from this strategy. The return from buying at the moment b_1 and selling at the moment s_1 reads as

$$\frac{S(s_1)}{S(b_1)}-1.$$

Similarly, the return from buying at the moment b_2 and selling at the moment s_2 reads as

$$\frac{S(s_2)}{S(b_2)}-1.$$

When we invest all the money obtained from selling the asset at the moment s_1 to buy the asset at the moment b_2, the return from all four operations reads as

$$\frac{S(s_1)}{S(b_1)}\frac{S(s_2)}{S(b_2)}-1.$$

Similarly, when we always invest all the money earned in the previous trading to buy the asset again, the return from buying the asset at moments $0\le b_1<b_2<\ldots<b_n<T$ and selling it at moments $0<s_1<s_2<\ldots<s_n\le T$ reads as

$$\begin{aligned}R&=\frac{S(s_1)}{S(b_1)}\frac{S(s_2)}{S(b_2)}\cdots\frac{S(s_n)}{S(b_n)}-1\\&=exp\left(V(s_1)-V(b_1)+V(s_2)-V(b_2)+\ldots+V(s_n)-V(b_n)\right)-1.\end{aligned}$$

Reasoning similarly as in the preceding section we obtain the following upper bound

$$\begin{aligned}R&=\frac{S(s_1)}{S(b_1)}\frac{S(s_2)}{S(b_2)}\cdots\frac{S(s_n)}{S(b_n)}-1\\&=exp\left(V(s_1)-V(b_1)+V(s_2)-V(b_2)+\ldots+V(s_n)-V(b_n)\right)-1\\&\le exp\left(TV(V,[0;T])\right)-1=exp\left(\sum_{i=1}^{T}|V(i)-V(i\text{-}1)|\right)-1.\end{aligned}\tag{11}$$

Remark. Similarly as for models with constant transaction costs, more accurate bound for G is expressed with the exponent of the positive path variation of the time series $R(n)$ and may be calculated with the following formula

$$exp\left(UTV(R,[0;T])\right)-1=exp\left(\sum_{i=1}^{T}max\{R(i)-R(i\text{-}1),0\}\right)-1.$$

Assuming the specific structure of the series $V(n), n=0,1,\ldots,$ we may again calculate the distribution or characteristics of the upper bound obtained. The simple yet widely used model assumes that the series $V(n), n=0,1,\ldots,$ is a random walk, i.e.

$$V(n)=X(1)+X(2)+\ldots+X(n),\tag{12}$$

where $X(n), n=0,1,\ldots,$ are i.i.d. (independent and identically distributed). In particular, assuming that $X(1),X(2),\ldots\sim N(\mu,\sigma^2),$ we may calculate

$$\begin{aligned}&\mathrm{E}R\le \mathrm{E}exp\left(\sum_{i=1}^{T}|V(i)-V(i\text{-}1)|\right)-1\le \mathrm{E}exp\left(\sum_{i=1}^{T}|X(i)|\right)-1\\&=\left(\mathrm{E}exp\left(|X(1)|\right)\right)^T-1=\left(\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}exp\left(|x|-\frac{(x-\mu)^2}{2\sigma^2}\right)dx\right)^T-1.\end{aligned}$$

Continuous case with proportional transaction costs

Similarly as before, let us now turn to the situation when the price process is observed in continuous time and we may sell or buy the assets at any time between the moments 0 and T. The continuous counterpart of the discrete random walk (12) with normally distributed increments $X(1), X(2), \ldots \sim N(\mu, \sigma^2)$ is the classical Black-Scholes model of the evolution of stock prices, given by the following sde:

$$dS(t) = \left(\mu - \frac{\sigma^2}{2}\right) S(t)dt + \sigma S(t)dW(t), \tag{13}$$

where $W(t), t \geq 0$, denotes, as before, the standard Brownian motion. Under the assumption that $W(t), t \geq 0$, is independent from $S(0)$ the solution of Eq. (13) reads as:

$$S(t) = S(0)\exp(\mu t + \sigma W(t)) \tag{14}$$

and the process $V(t), t \geq 0$, may be written as

$$V(t) = \mu t + \sigma W(t). \tag{15}$$

(From the properties of the standard Brownian motion we immediately obtain that

$$X(i) = V(i) - V(i-1) = \mu + \sigma(W(i) - W(i-1)) \sim N(\mu, \sigma^2)$$

are for $i = 1, 2, \ldots$ i.i.d. random variables.)

Again, in the continuous case, for $V(t)$ given e.g. by Eq. (15), the upper bound given by Eq. (11) becomes trivial, since

$$TV(V, [0;T]) = +\infty.$$

Thus, similarly as in the previous section, let us introduce transaction costs. In the present case the transaction costs shall not be constant but rather proportional to the transaction value. Let $\delta \in (0;1)$ denote the ratio of every transaction value paid as a commission. Now, the return from buying at the moment b_1 and selling at the moment s_1 reads as

$$\frac{S(s_1)(1-\delta)}{S(b_1)(1+\delta)} - 1.$$

Similarly, the return from buying the asset at moments $0\le b_1<b_2<\ldots<b_n<T$ and selling it at moments $0<s_1<s_2<\ldots<s_n\le T$ reads as

$$R=\frac{S(s_1)}{S(b_1)}\frac{1-\delta}{1+\delta}\frac{S(s_2)}{S(b_2)}\frac{1-\delta}{1+\delta}\cdots\frac{S(s_n)}{S(b_n)}\frac{1-\delta}{1+\delta}-1$$

$$=exp\left(V(s_1)-V(b_1)-ln\frac{1+\delta}{1-\delta}+V(s_2)-V(b_2)-ln\frac{1+\delta}{1-\delta}+\ldots+V(s_n)-V(b_n)-ln\frac{1+\delta}{1-\delta}\right)-1.$$

Denoting $c=ln\frac{1+\delta}{1-\delta}>0$ we obtain the following estimate

$$R=\frac{S(s_1)}{S(b_1)}\frac{1-\delta}{1+\delta}\frac{S(s_2)}{S(b_2)}\frac{1-\delta}{1+\delta}\cdots\frac{S(s_n)}{S(b_n)}\frac{1-\delta}{1+\delta}-1$$

$$=exp\left(V(s_1)-V(b_1)-c+V(s_2)-V(b_2)-c+\ldots+V(s_n)-V(b_n)-c\right)-1 \text{ Remark}$$

$$\le exp\left(TV^c(V,[0;T])\right)-1<+\infty.$$

Again, more accurate bound for G in continuous time setting with proportional transaction costs is expressed with the exponent of the upward truncated variation of the process $V(t), t\ge 0$,

$$R\le exp\left(UTV^c(V,[0;T])\right)-1$$

$$=exp\left(sup_n sup_{0\le t_0<t_1<\ldots<t_n\le T}\sum_{i=1}^{n} max\{R(t_i)-R(t_{i-1})-c,0\}\right)-1.$$

By results of [Łochowski 2011, Sect. 2] it follows that for V being the Wiener process with drift, given by Eq. (15), for any real p we have $\mathrm{E}R^p<+\infty$, and from the results of [Bednorz and Łochowski 2012, Theorem 2]) we get more precise estimate of the form

$$\mathrm{E}R^p\le \mathrm{E}exp\left(pTV^c(V,[0;T])\right)\le 2exp\left(p^2T\alpha(\sigma)+pTc^{-1}\beta(\sigma)+p|\mu|\right).$$

Here $\alpha(\sigma)$ and $\beta(\sigma)$ are constants depending on σ only.

REFERENCES

Bednorz W., Łochowski R. (2012) Integrability and Concentration of Sample Paths' Truncated Variation, Bernoulli, accepted, preprint available on the journal web page.

Ghomrasni R., Łochowski R. (2013) The Play Operator, the Truncated Variation and the Generalisation of the Jordan Decomposition, preprint arXiv:1311:6405, accepted for publication in Mathematical Methods in Applied Sciences.

Łochowski R. (2010) On upper gain bound for trading strategy based on cointegration, Quantitative Methods in Economics 11, Tom I, str. 110 – 117.

Łochowski R. (2011) Truncated variation, upward truncated variation and downward truncated variation of Brownian motion with drift – their characteristics and their applications, Stochastic Processes and Their Applications 121, Tom II, str. 378 – 393.

Łochowski R. (2013) On the Generalisation of the Hahn-Jordan Decomposition for Real Càdlàg Functions, Colloquium Mathematicum 132, Tom I, str. 121-138.

Revuz D., Yor M. (2005) Continuous Martingales and Brownian Motion, Springer, Berlin.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 39 – 50

THE METHOD OF SUPPORTING DECISIONS UNDER RISK BASED ON MULTIOBJECTIVE OPTIMIZATION

Andrzej Łodziński
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
e–mail: andrzej_lodzinski@sggw.pl

Abstract: The method of supporting decisions under risk was presented in this paper. Making decision under risk takes place when a result of a given decision is not explicit and depends on the condition of the environment. A decision-making process based on multiobjective optimization has been presented in this paper. Methods of multiobjective optimization do not give one unique solution, but a whole set of them. A decision making relies on interactive conducting of the decision making process. Selection of given decision is made by way of solving a problem with parameters defining aspirations of a decision maker and the evaluation of obtained results. A decision maker defines a parameter, for which a solution is indicated. Then he or she evaluates the received solution by either accepting or rejecting it. In the second case a decision maker provides a new parameter value and a problem is solved again for the new parameter.

Keywords: multiobjective optimization, symmetrically efficient solution, scalarization function, supporting a decision-making process

INTRODUCTION

The method of supporting decisions under risk based on multiobjective optimization has been presented in this paper. Making decisions under risk takes place when the results of activities undertaken by a decision maker are uncertain due to the likelihood of occurring unexpected circumstances or factors that disturb these unexpected circumstances or disturbing factors - conditions of the environment - are called the scenarios. The higher dispersal of results, the higher level of uncertainty. Simultaneously, each scenario explicitly defines completion of results for individual decisions.

For example, the enterprise, which launches its new product on the market must take into account numerous uncertainties, starting from costs of developmental works, sales volume or possible competition reactions. In the case of a decision made under risk it is possible to estimate probabilities, with which these uncertain results take place, e.g. an enterprise can rely on forecasts of experts for market research.

Making decisions under risk is modeled with the help of a special task of multiobjective optimization. It is a task with non–descendingly ordered functions. Methods of multiobjective optimization do not give one unique solution, but a whole set of them. The method of supporting decision is based on interactive conducting of the process of making decisions that is, selection of given decision is made by way of solving a problem with controlling parameters defining aspirations of a decision maker and the evaluation of obtained results. A decision maker defines a parameter, for which a solution is indicated. Then he or she evaluates the received solution by either accepting or rejecting it. In the second case a decision maker provides a new parameter value and a problem is solved again for the new parameter.

MODELLING OF A DECISION SITUATION UNDER RISK

The problem of selecting a decision under risk should be modelled by implementing these scenarios to a problem of selecting decisions, which represent possible conditions of the environment. Probability distribution is provided for the scenarios. If we assume that the probability of occurring individual scenarios are rational numbers, then it is possible to lead to the situation by multiple repetition of appropriate scenarios, in which probability of occurring each scenario is the same. The number of occurring of a definite scenario refers to the probability that is assigned to it. Definite scenarios $S_i, i = 1,...,m$ correspond to realizations of mark functions $f_i(x), i = 1,...,m$. A higher value of a mark function is preferred for each scenario.

We consider a situation, in which for each decision $x \in X_0$ there is one of m possible results $f_1(x),..., f_m(x)$. Probabilities of these results are the same and amount to $p = \frac{1}{m}$.

The problem of making decisions under risk is modeled as a special task of multiobjective optimization:

$$\max_x \{(f_1(x),..., f_m(x)) : \quad x \in X_0\} \tag{1}$$

where: $x \in X_0$ – decision that belongs to the set of admissible decisions, $X_0 \subset R^n$,
$S_i, i = 1,...,m$ - scenarios (environment conditions),
$f = (f_1,..., f_m)$ - vector function, which allocates for each vector of decision variables $x \in X_0$ mark vector $y = f(x)$; individual coordinates $y_i = f_i(x),\ i = 1,...,m$ – represent scalar mark functions - result of a decision x when a scenario $S_i, i = 1,...,m$ takes place,
X_0 - set of feasible decisions.

It is a task for multiobjective optimization put into equally probable scenarios. The results are equally probable - each coordinate of a mark function has the same significance.

Vector function $y = f(x)$ allocates for each vector of decision variables x mark vector $y \in Y_0$, which measures decision quality x from the point of view of the defined set of quality indicators $y_1,..., y_m$. The image of feasible set X_0 for functions f constitutes the set of achievable mark vectors Y_0.

The task (1) is based on finding such admissible decision $x \in X_0$, for whose m marks it assumes the best values. This task is considered in relation to the marks, that is the following task is examined:

$$\max_x \{(y_1,..., y_m): \quad y \in Y_0\} \tag{2}$$

where: $x \in X$ – vector of decision variables,
$y = (y_1,..., y_m)$ - vector quality indicator, individual coordinates $y_i = f_i(x),\ i = 1,...,m$ represent individual scalar criteria,
Y_0 - set of admissible mark vectors.

Mark vector $y = (y_1,..., y_m)$ in the multiobjective problem (2) represents a decision result x in the form of a vector with m equally probable $p = \frac{1}{m}$ coordinates $y_i, i = 1,.., m$.

SYMMETRICALLY EFFICIENT SOLUTION

Making decisions under risk is modeled as a special task of multiobjective optimization with the relation of reference that meets anonymity property. The results, which differ by ordering of coordinates, are not differentiated. The solution to the problem of the selection decision is the decision of symmetrically efficient. It is an efficient decision, which meets an additional domain - anonymity property of preference relation.

Nondominated results (Pareto-optimal) are defined in the following way:

$$\hat{Y}_0 = \{\hat{y} \in Y_0 : (\hat{y} + \tilde{D}) \cap Y_0 = \varnothing)\} \tag{3}$$

where: $\tilde{D} = D \setminus \{0\}$ – positive cone without a top. The following can be assumed as a positive cone $\tilde{D} = R_+^m$.

Appropriate feasible decisions are defined within a decision area. A decision $\hat{x} \in X_0$ becomes an efficient decision (Pareto-optimal), if a mark vector corresponding to it $\hat{y} = f(\hat{x})$ is an nondominated vector [12].

In the multiobjective problem (1), which serves to make decisions under risk with a given set of mark functions, value distribution obtained by these functions is only important, whereas it is not important which value a given function has assumed. The results, which differ by ordering, are not differentiated. This requirement is formed as anonymity (neutrality) domain of preference relation.

Then this relation is called an anonymous relation, when for each mark vector $y = (y_1, y_2, ..., y_m) \in R^m$ for any permutation P of a set $\{1, ..., m\}$ the following domain takes place:

$$(y_{P(1)}, y_{P(2)}, ..., y_{P(m)}) \approx (y_1, y_2, ..., y_m). \tag{4}$$

Mark vectors that have the same coordinates, but in a different order, are identified. Preference relations that meet additional anonymity condition are called anonymous preference relation.

Nondominated vector that meets anonymity property is called a symmetrically nondominated vector. The set of symmetrically nondominated vectors is marked with $\hat{Y}_{0S}$. Symmetrically efficient decisions are defined within a decision area. A decision $\hat{x} \in X_0$ becomes a symmetrically efficient decision, if a

mark vector corresponding to it $\hat{y} = f(\hat{x})$ is a symmetrically nondominated vector. The set of symmetrically efficient decisions is marked with $\hat{X}_{0S}$.

A relation of symmetric dominance can be expresses as an inequality relation for mark vectors, whose coordinates are ordered in the non-descending order. This relation can be presented with the use of transformation $T: R^k \to R^m$ that non-descendingly orders coordinates of mark orders that is a vector $T(y)$ is a vector with non-descendingly ordered coordinates of a vector y that is $T(y) = (T_1(y), T_2(y), ..., T_m(y))$, where $T_1(y) \le T_2(y) \le ... \le T_m(y)$ and there is permutation P of the set $\{1,...,m\}$ so that $T_i(y) = y_{P(i)}$ for $i = 1,..,m$. The relation of symmetric dominance $\geq_a$ is an ordinary vector dominance for non-descendingly ordered vectors [8].

Mark vector y^1 dominates symmetrically prefers a vector y^2 if the following condition is met:

$$y^1 \geq_a y^2 \Leftrightarrow T(y^1) \geq T(y^2) \tag{5}$$

Solving a decision problem involves the determination of symmetrically efficient decision corresponding to preferences of a decision maker.

SCALARIZATION OF A PROBLEM

Special multiobjective task is solved for indicating a solution that is symmetrically efficient for a multiobjective task (1). It is a task with coordinates of a mark vector that are ordered non-descendingly that is the following task:

$$\max_y \{(T_1(y), T_2(y), ..., T_m(y)) : \quad y \in Y_0\} \tag{6}$$

where: $y = (y_1, y_2, ..., y_m)$ – mark vector,

$T(y) = (T_1(y), T_2(y), ..., T_m(y))$, where $T_1(y) \le T_2(y) \le .... \le T_m(y)$ - mark vector that is ordered non-descendingly,

Y_0 - set of achievable mark vectors.

An efficient solving of a task of multiobjective optimization (6) is a symmetrically efficient solution of a multiobjective task (1).

In order to provide a solution of a multiobjective task (6) scalarization of this task is solved that has a scalarizing function $s: Y \times \Omega \to R^1$:

$$\max_x \{s(y, \bar{y}) : x \in X_o\} \tag{7}$$

where: $y = (y_1, y_2, ..., y_m)$ – mark vector,
$\bar{y} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_m)$ - controlling parameter.

It is a task of one-criterion optimization of specially created scalarizing function of two variables - mark vector $y \in Y$ and a controlling parameter $\bar{y} \in \Omega \subset R^m$ with the actual value that is function $s : Y \times \Omega \to R^1$. The parameter $\bar{y} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_m)$ is at disposal of a decision maker, which enables him or her reviewing the set of symmetrically efficient solutions.

Optimum solution of a task (7) should be a solution of a multiobjective task. Scalarizing function should meet some properties - completeness property and sufficiency property. Sufficiency property means that for each controlling parameter $\bar{y}$, solving a scalarizing task means a solution that is symmetrically efficient that is $\hat{y} \in \hat{Y}_{0S}$. Completeness property means that thanks to adequate changes of a parameter $\bar{y}$ any result can be obtained $\hat{y} \in \hat{Y}_{0S}$. Such function fully characterizes symmetrically efficient solutions. Each maximum of this function is a symmetrically efficient solution. Each symmetrically efficient solution can be obtained by assuming appropriate values of controlling parameters $\bar{y}$.

Completeness and insufficient parameterization of the set of symmetrically nondominated vectors $\hat{Y}_{0S}$ can be obtained by applying the reference point method to a task (6). This method is used as controlling parameters of aspiration levels. Aspiration levels are such values of mark functions, which are satisfactory for a decision maker.

Scalarizing function in the method of the point of reference has the following form:

$$s(y, \bar{y}) = \min_{1 \le i \le m} (T_i(y) - T_i(\bar{y})_i) + \varepsilon \cdot \sum_{i=1}^{m} (T_i(y) - T_i(\bar{y})_i) \tag{8}$$

where: $y = (y_1, y_2, ..., y_m)$ – mark vector,
$T(y) = (T_1(y), T_2(y), ..., T_m(y))$,
where $T_1(y) \le T_2(y) \le .... \le T_m(y)$
- mark vector is ordered non-descendingly,
$\bar{y} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_m)$ - vector of aspiration levels,
$T(\bar{y}) = (T_1(\bar{y}), T_2(\bar{y}), ..., T_m(\bar{y}))$,

where $T_1(\bar{y}) \le T_2(\bar{y}) \le ... \le T_m(\bar{y})$
- mark vector is ordered non-descendingly,
ε - arbitrarily small, positive regularizing parameter.

Such scalarizing function is called the function of achievement. This function measures closeness of a given solution to the aspiration level. The aim is to find a solution that is as close as possible to achieve definite requirements - aspiration levels.

Optimum values of this function can be used not only to calculate a symmetrically nondominated vector, but also to evaluate achievability of a given aspiration point $\bar{y}$. If a maximum of the achievement function $s(y, \bar{y})$ is negative then the aspiration point $\bar{y}$ is not achievable, however, the maximum point $\hat{y}$ of this function is symmetrically nondominated vector in some sense equally closest to the point $\bar{y}$. If a maximum of the achievement function $s(y, \bar{y})$ is equal to zero then the aspiration point $\bar{y}$ is achievable and is a symmetrically nondominated vector. If a maximum of the achievement function $s(y, \bar{y})$ is positive then the aspiration point $\bar{y}$ is achievable, however, the maximum point $\hat{y}$ of this function is symmetrically nondominated vector in some sense equally improved to the point $\bar{y}$ [12].

Maximization of such function due to y means a symmetrically efficient solution $\hat{y}$ and a decision that generates a symmetrically efficient decision $\hat{x}$. An indicated symmetrically efficient solution $\hat{x}$ depends on values of aspiration levels $\bar{y}$.

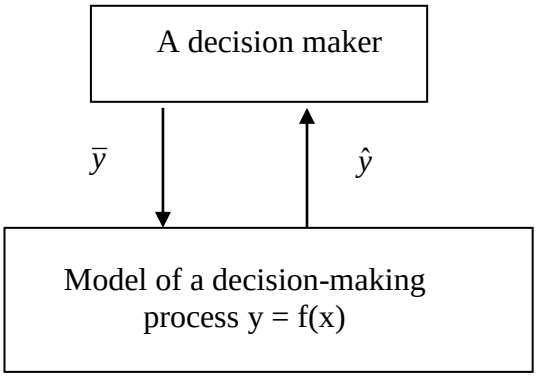
METHOD OF SELECTING SYMMETRICALLY EFFICIENT DECISIONS

The solution of a task of multiobjective optimization is the whole set of solutions, so a decision maker should select a decision with the help of the interactive computer system. Such system enables a controlled preview of the set of solutions. On the basis of the parameter values provided by a decision maker the task is solved and the system presents a solution for analysis that corresponds to current values of these parameters.

The tool for previewing the set of solutions is the function (8). Maximum of this function depends on the parameter $\bar{y}, i = 1,,...,m$, which is used by a decision maker to select a solution. A decision maker, when solving a problem, defines

aspiration levels $\bar{y}, i = 1,,..., m$ as desirable values of individual marks. A decision maker expresses his or her preferences in the reference point method by defining such a value for each mark function, which will be fully satisfactory for him or her. These values constitute the aspiration level for a given mark function. The controlling parameter in the form of aspiration levels represents actual values that are easily understood by a decision maker and characterize his or her preferences. Aspiration levels are expressed in the terms of values of individual mark functions. The method of supporting selection of a decision is presented on the Figure 1.

Figure 1. The method of supporting a decision-making process



Source: own work

Such manner of making decisions does not impose on a decision maker any rigid way of analyzing a decision problem and enables the possibility of modifying his or her preferences while analyzing a problem. A user has a master role in this method of making decisions.

EXAMPLE

The problem of selecting the best investment out of 10 investments is presented in order to illustrate the method of supporting a decision under risk. Probabilities of payments for individual investments are the following:

Investment 1:

payment [thousand PLN]	7	8	10	14	15
probability	0,3	0,2	0,2	0,2	0,1

Investment 2:

payment [thousand PLN]	7	8	9	10	13	14	15
probability	0,1	0,1	0,1	0,1	0,3	0,3	0,2

Investment 3:

payment [thousand PLN]	8	9	10	13	14	15
probability	0,1	0,2	0,1	0,2	0,2	0,2

Investment 4:

payment [thousand PLN]	8	9	10	11	13	14	15
probability	0,1	0,2	0,1	0,1	0,1	0,3	0,1

Investment 5:

payment [thousand PLN]	6	7	9	10	11	13	14	15
probability	0,1	0,1	0,1	0,1	0,1	0,1	0,3	0,10

Investment 6:

payment [thousand PLN]	8	9	13	14
probability	0,3	0,2	0,2	0,3

Investment 7:

payment [thousand PLN]	6	7	8	11	13	14
probability	0,1	0,2	0,2	0,3	0,1	0,1

Investment 8:

payment [thousand PLN]	6	7	9	11	13	15
probability	0,3	0,1	0,1	0,2	0,2	0,1

Investment 9:

payment [thousand PLN]	8	9	11	13	14
probability	0,1	0,3	0,1	0,3	0,2

Investment 10:

payment [thousand PLN]	7	8	9	10	11	14	15
probability	0,1	0,1	0,1	0,1	0,3	0,2	0,1

Each investment requires to invest 10000 PLN.

In order to characterize possible investments we set their values on the set that contains all possible values, which they can assume with non-zero probability:

payment in [thousand PLN]	6	7	8	9	10	11	13	14	15
Investment 1	0	0,3	0,2	0	0,2	0	0	0,2	0,1
Investment 2	0	0,1	0,1	0,1	0,1	0	0,3	0,3	0,2
Investment 3	0	0	0,1	0,2	0,1	0	0,2	0,2	0,2
Investment 4	0	0	0,1	0,2	0,1	0,1	0,1	0,3	0,1
Investment 5	0,1	0,1	0	0,1	0,1	0,1	0,1	0,1	0,1
Investment 6	0	0	0,3	0,2	0	0	0,2	0,3	0
Investment 7	0,1	0,2	0,2	0	0	0,3	0,1	0,1	0
Investment 8	0	0,1	0	0,1	0	0,2	0,2	0	0,1
Investment 9	0	0	0,1	0,3	0	0,1	0,3	0,2	0
Investment 10	0	0,1	0,1	0,1	0,1	0,3	0	0,2	0,1

Source: own calculations

While repeating appropriate scenarios we lead to the situation, where the probability of each scenario is the same and equals $p = \frac{1}{10}$. Then we get the situation, in which the results of each investment decision $i = 1,2,...,10$ are the following mark vectors with equally probable coordinates:

i	1	2	3	4	5	6	7	8	9	10
y^1	7	7	7	8	8	10	10	14	14	15
y^2	7	8	9	10	13	14	14	14	15	15
y^3	8	9	9	10	13	13	14	14	15	15
y^4	8	9	9	10	11	13	14	14	14	15
y^5	6	7	9	10	11	13	14	14	14	15
y^6	8	8	8	9	9	13	13	14	14	14
y^7	6	7	7	8	8	11	11	11	13	14
y^8	6	6	6	7	9	11	11	13	13	15
y^9	8	9	9	9	11	13	13	13	14	14
y^{10}	7	8	9	10	11	11	11	14	14	15

Source: own calculations

The set of symmetrically nondominated vectors is the following $\hat{Y}_{os} = \{y^2, y^3\}$. Two decisions: investment 2 and investment 3 are the symmetrically efficient decisions. When making a decision it is necessary to select between them and reject other investments, irrespective of individual preferences. Investment 2 and investment 3 are incomparable in relation to anonymous preference relations. Selection between them depends on individual preferences of a decision maker.

These two variants are selected by the reference point method with non-descendingly ordered coordinates with adequately determined aspiration level.

Assuming the biggest possible aspiration marks as aspiration levels, e.g. for the vector of aspiration $T(\bar{y}) = (8,\ 9,\ 9,\ 10,\ 13,\ 14,\ 14,\ 14,\ 15,\ 15)$ we get as a solution – investment 2.

Decreasing beginning aspirations, e.g. for the vector of aspiration $T(\bar{y}) = (7{,}5,\ 8{,}5\ 9,\ 10,\ 13,\ 14,\ 14,\ 14,\ 15,\ 15)$ we get as a solution - investment 3.

This method enables to select any symmetrically efficient solution that corresponds to preferences of a decision maker.

CONCLUSION

The method of supporting decisions under risk was presented in this paper. Decision making process takes place by solving a task of multiobjective optimization. This method is characterized by using aspiration points and optimality of the achievement function in order to organize interaction with a user.

The method provides the whole set of solutions that are anonymously efficient and enables a decision maker to choose freely. However, such manner of conduct does not substitute a decision maker in his or her decision making process. The whole process of making decisions is controlled by a user.

REFERENCES

Jaworski P., Micał J. (2005) Mathematical modeling in finance and insurance.(in polish) Poltext, Warszawa.

Lewandowski A. and Wierzbicki A. eds. (1989) Aspiration Based Decision Support Systems. Lecture Notes in Economics and Mathematical Systems. Vol. 331, Springer-Verlag, Berlin-Heidelberg.

Łodziński A (1991) The use of reference objectives for selecting polyoptimal control in multistage process. System Analysis Modelling Simulation. Vol. 8. Akademie Verlag Berlin.

Keeney L., Raiffa H. (1993) Decisions with Multiple Objectives. Preferences and Value Tradeoffs.

Luce D., Raiffa H. (1966) Games and decisions. .(in polish) PWN, Warszawa.

Łucki Z. (2001) Mathematical techniques for managing. (in polish) Publishing house AGH, Kraków.

Ogryczak W. (1997) Multicriteria optimization of linear an discrete. Publishing house UW, Warszawa.

Ogryczak, W. (2002) Multicriteria optimization an Decisions under Risk, Control and Cybernetics.

Sikora W. (2008) Operations research. PWE, Warszawa.

Trzaskalik T. (2006) Multicriteria methods in the Polish financial. PWE, Warszawa.

Tyszka T. (2010) Decisions. Psychological and economic perspective. Publishing house Scholar, Warszawa.

Wierzbicki A., Makowski N., Wessels J. (2000) Model_Based Decision Support Methology with Environmental Applications. IIASA Kluwer, Laxenburg Dordrecht.

Wierzbicki A., Granat J. P. (2003) Optimization in supporting decisions. (typescript) Warszawa.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 51 – 60

FUZZY CLASSIFICATION OF SYMBOLIC OBJECTS

Małgorzata Machowska–Szewczyk
Department of Artificial Intelligence Methods and Applied Mathematics
West Pomeranian University of Technology in Szczecin
e-mail: mmachowska@wi.zut.edu.pl

Abstract: The aim of this work is to present fuzzy clustering algorithm for objects, which can be described by mixed feature-type symbolic data and fuzzy data. The main idea is the transformation of mixed feature-type symbolic data and fuzzy data into histogram-valued symbolic data. Fuzzy classification is very useful in case, when classes are difficult separated, mixed objects can be classified into class with the fixed degree of membership.

Keywords: fuzzy classification, symbolic data, histogram-valued symbolic data

INTRODUCTION

Clustering algorithms of symbolic objects (e.g. only numeric values or interval-valued data) most often assume that the variables used for description are of the same type. However, there is a lot of real objects requiring the symbolic features for description which can be both numeric-valued, interval-valued, a set of categories-valued and an ordered list-valued with weights.
The aim of this study is presenting the proposal of a generalization of the classification methods of symbolic objects, characterized by mixed type features, proposed in the work de Carvalho and de Souza [2010], relating to the case of fuzzy classification and the possibility of including fuzzy features to describe objects. These methods are based on iterative clustering methodology with adaptation of the Euclidean distance. Distances are changed in each iteration of the algorithm, and can either be the same for all classes, or different for particular groups. In the first step the transformation of symbolic values of various types is made to histogram-valued symbolic data. The modification proposed by the author allows to carry out the classification in both the classical sense (then the

classification method of de Carvalho and de Souza is used) as well as in terms of fuzzy classification. The fuzzy classification is very useful in a situation of classes separated with difficulty, the so called mixed objects can be classified into classes with a certain degree of membership. The classical classification forces the assigning of an object only to one class, therefore the objects whose similarity to several classes at the same time is quite high are not recognized, and the quality of the classification obtained is then low. The proposed algorithm, therefore, contributes to an additional opportunity for mixed-value symbolic data analysis.

TRANSFORMATION INTO HISTOGRAM – VALUED SYMBOLIC DATA

Each object i from the set $\Omega = \{1,...,n\}$, described by the p-values of symbolic variables $\{X_1,...,X_p\}$, is identified with the vector of mixed-value symbolic data $\mathbf{x}_i = (x_i^1, x_i^2,...,x_i^p)$, $i = 1,...,n$. This means that the symbolic variable X_j can assume for a given unit i the value x_i^j in the form of [Bock, Diday 2000]:

- set–valued, if given an item i, $X_j(i) = x_i^j \subset A_j$, where $A_j = \{t_1^j, t_2^j,...,t_{H_j}^j\}$ is set of categories;
- ordered list–valued, if given an item i, x_i^j is set-list of ordered list of categories $A_j = [t_1^j, t_2^j,...,t_{H_j}^j]$;
- interval–valued, if given an item i, $X_j(i) = x_i^j = [a_i^j; b_i^j] \subset [a;b]$, where $[a;b] \in \wp$ and $\wp$ is set closed intervals defined from R;
- histogram–valued, if given an item i, $X_j(i) = x_i^j = \left(S^j(i), \mathbf{q}^j(i)\right)$, where $\mathbf{q}^j(i) = (q_{i1}^j, q_{i2}^j,...,q_{iH_j}^j)$ is the vector of weights defined in $S^j(i)$, such that a weight q_{im}^j corresponds to the category m from $S^j(i)$ and $S^j(i)$ is support of measure $\mathbf{q}^j(i)$.

The aim of standard clustering algorithm [Diday i Simon 1976] is to find a partition $P = \left(C_1, C_2,...,C_K\right)$ of the set Ω into a fixed number K of classes and their corresponding patterns $\mathbf{G} = \left(\mathbf{g_1},...,\mathbf{g}_K\right)$ by the local minimization of criterion function W. This criterion assesses the fitting between classes and their respective representatives.

To overcome the difficulty, which is the representation of objects using ordered or non-ordered symbolic data of various types, the pre-processing is made, whose purpose is to obtain a suitable homogenization of symbolic data. It consists

in the transformation of mixed-value symbolic data to histogram-valued symbolic data.

If X_j is a set-valued variable, its transformation into a symbolic histogram-valued variable $\tilde{X}_j$ is achieved as follows: $\tilde{X}_j(i) = \tilde{x}_i^j = (A_j, \mathbf{q}^j(i))$, where $A_j = \{t_1^j, t_2^j, ..., t_{H_j}^j\}$ is a domain of variable X_j and the support of the weight vector $\mathbf{q}^j(i) = (q_1^j(i), q_2^j(i), ..., q_{H_j}^j(i))$. The weight $q_h^j(i)\,(h = 1,...,H_j)$ of the category $t_h^j \in A_j$ is defined as [de Carvalho 1995]:

$$q_h^j(i) = \begin{cases} \dfrac{1}{c(x_i^j)} & \text{if } t_h^j \in x_i^j \\ 0 & \text{if } t_h^j \notin x_i^j \end{cases}, \tag{1}$$

where $c(x_i^j)$ is the cardinality of a finite set of category $c(x_i^j)$.

If X_j is an ordered list-valued variable, then it is transformed into a histogram-valued symbolic variable $\tilde{X}_j$ as follows: $\tilde{X}_j(i) = \tilde{x}_i^j = \left(A_j, \mathbf{Q}^j(i)\right)$, where $A_j = [t_1^j, t_2^j, ..., t_{H_j}^j]$ is support of the vector of cumulative weights $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i))$. The cumulative weights $Q_h^j(i)\,(h = 1,...,H_j)$ of category t_h^j from the list A_j are defined as [de Carvalho 1995]:

$$Q_h^j(i) = \sum_{r=1}^{h} q_r^j(i) \text{ where } q_r^j(i) = \begin{cases} \dfrac{1}{l(x_i^j)} & \text{if } t_r^j \text{ is from the list } x_i^j \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

$l(x_i^j)$ is the length of an ordered list of category x_i^j.

In the case of interval-valued variable X_j it is transformed to histogram-valued symbolic variable $\tilde{X}_j$ as follows: $\tilde{X}_j(i) = \tilde{x}_i^j = (\tilde{A}_j, \mathbf{Q}^j(i))$, where $\tilde{A}_j = \{I_1^j, I_2^j, ..., I_{H_j}^j\}$ is the list of elementary intervals, constituting support of the cumulative weight vector $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i))$. The cumulative weight $Q_h^j(i)\,(h = 1,...,H_j)$ of the elementary interval I_h^j is defined as [de Carvalho 1995]:

$$Q_h^j(i) = \sum_{r=1}^{h} q_r^j(i) \text{ where } q_r^j(i) = \frac{l(I_r^j \cap x_i^j)}{l(x_i^j)}, \tag{3}$$

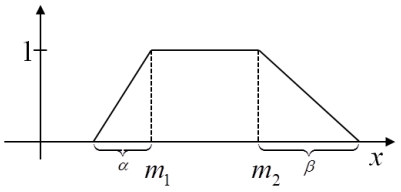
$l(I)$ is the length of the closed interval I.

It can be shown that: $0 \le q_h^j(i) \le 1 \, (h = 1,...,H_j)$ and $\sum_{h=1}^{H_j} q_h^j(i) = 1$. In addition, that: $q_1^j(i) = Q_1^j(i)$ i $q_h^j(i) = Q_h^j(i) - Q_{h-1}^j(i) \, (h = 2,...,H_j)$.

Limits of elementary intervals $I_h^j \, (h = 1,...,H_j)$ are derived from the ordered limits of n+1 intervals $\left\{x_1^j, x_2^j, ..., x_n^j, [a;b]\right\}$ and the number of elementary intervals is at most $2n$. The elementary intervals have the following properties [de Carvalho 1995]:

1. $\sum_{h=1}^{H_j} I_h^j = [a;b]$,
2. $I_h^j \cap I_{h'}^j = \varnothing$ if $h \ne h'$,
3. $\forall h \exists i \in \Omega$ that $I_h^j \cap x_i^j \ne \varnothing$,
4. $\forall i \exists S_i^j \subset \{1,...,H_j\} : \bigcup_{h \in S_i^j} I_h^j = x_i^j$.

Figure 1. Parameterization of *TFN*



Source: own elaboration

The trapezoidal fuzzy numbers *TFN* (see Fig. 1) are in real applications often, represented as *L-R* fuzzy numbers. Let *L* (*R*) be decreasing, shape function from R^+ to $[0,1]$, with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, $L(x) > 0$ for all $x < 1$, $L(x) = 0$ ($L(x) > 0$ for all x and $L(+\infty) = 0$). A fuzzy number A with its membership function μ_A [Zimmermann 1991]:

$$\mu_A(x) = \begin{cases} L\left(\dfrac{m_1 - x}{\alpha}\right) & \text{for } x < m_1 \\ 1 & \text{for } m_1 \le x \le m_2 \\ R\left(\dfrac{x - m_2}{\beta}\right) & \text{for } x > m_2 \end{cases} \tag{4}$$

is called an *L-R* type *TFN*. Symbolically, A can be denoted by $A = (m_1, m_2, \alpha, \beta)_{LR}$, where $\alpha > 0, \beta > 0$ are called left and right spreads, respectively. Using this

parametric representation can be presented four kinds of *TFN*s with real numbers, interval, triangular and trapezoidal fuzzy numbers.

If X_j is variable of trapezoidal fuzzy value $x_i^j = (m_1^j(i), m_2^j(i), \alpha^j(i), \beta^j(i))_{LR}$, its transformation into symbolic histogram-valued variable $\tilde{X}_j$ is accomplished in the following way (author's proposal): $\tilde{X}_j(i) = \tilde{x}_i^j = (A_j, \mathbf{Q}^j(i))$, where $\tilde{A}_j = \{I_1^j, I_2^j, ..., I_{H_j}^j\}$ is the list of interval fuzzy numbers constructed on elementary intervals, constituting support of the cumulative weight vector $\mathbf{Q}^j(i) = (Q_1^j(i), Q_2^j(i), ..., Q_{H_j}^j(i))$. Cumulative weight $Q_h^j(i)\,(h = 1,...,H_j)$ interval fuzzy number I_h^j is defined as:

$$Q_h^j(i) = \sum_{r=1}^{h} q_r^j(i), \text{ where } q_r^j(i) = \frac{l(I_r^j \cap x_i^j)}{l(x_i^j)}, \tag{5}$$

$l(I)$ is a area under a membership function of fuzzy number I.

It can be show, that: $0 \le q_h^j(i) \le 1\,(h = 1,...,H_j)$ and $\sum_{h=1}^{H_j} q_h^j(i) = 1$. Moreover, again $q_1^j(i) = Q_1^j(i)$ i $q_h^j(i) = Q_h^j(i) - Q_{h-1}^j(i)\,(h = 2,...,H_j)$.

The boundaries of fuzzy numbers $I_h^j\,(h = 1,...,H_j)$ are obtained from the ordered boundaries of supports and cores of all considered fuzzy numbers $\{x_1^j, x_2^j, ..., x_n^j\}$.

After the pre-processing step every object i (i=1,…,n) is represented by a histogram-valued symbolic data vector $\tilde{\mathbf{x}}_i = \left(\tilde{x}_i^1, ..., \tilde{x}_i^p\right)$, and $\tilde{x}_i^j = \left(D_j, \mathbf{u}^j(i)\right)$, where D_j (a domain of variable $\tilde{X}_j$) depending on the type of the primary variable is the set of categories, an ordered list of categories or a list of elementary intervals, a list of fuzzy numbers with supports resulting from the elementary intervals, $\mathbf{u}^j(i) = (u_1^j(i), ..., u_{H_j}^j(i))$ is a vector of weights or of cumulative weights. The pattern of class $C_k\left(k = 1,...,K\right)$ is also represented by a histogram-valued symbolic data vector $\mathbf{g}_k = (g_k^1, ..., g_k^p)$, $g_k^j = (D_j, \mathbf{v}^j(k))\ (j = 1,...,p)$ with a vector of weights or of cumulative weights $\mathbf{v}^j(k) = (v_1^j(k), ..., v_{H_j}^j(k))$, where D_j is the set of categories, list of categories or a list of elementary intervals. It is noteworthy that for each variable $\tilde{X}_j\ (j = 1,...,p)$ the support is the same for all units and patterns.

According to the general scheme, the iterative classification algorithm [Diday i Simon 1976] is searching for a set Ω partition $P^* = (C_1^*, C_1^*, ..., C_K^*)$ into a fixed number K of classes, corresponding to K patterns $\mathbf{G}^* = (\mathbf{g}_1^*, ..., \mathbf{g}_K^*)$

representing the classes in P^*, and K of weight vectors parametrizing the squares of adaptive Euclidean distances, for which the criterion function value is minimum:

$$W(\mathbf{G},\mathbf{D},P)=\sum_{k=1}^{K}\sum_{i\in C_k} d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda}_k). \tag{6}$$

In the formula (4) the following is considered:

- squares of adaptive Euclidean distances parameterized by the same vector of weights $\boldsymbol{\lambda}_k=\boldsymbol{\lambda}(k=1,...,K)$, where $\boldsymbol{\lambda}=(\lambda^1,...,\lambda^p)$ changes at each iteration but is the same for all classes:

$$d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda})=\sum_{j=1}^{p}\lambda^j\phi^2\left(\mathbf{u}^j(i),\mathbf{v}^j(k)\right)=\sum_{j=1}^{p}\lambda^j\sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(k)\right)^2, \tag{7}$$

- squares of adaptive Euclidean distances parameterized by the weight vectors $\boldsymbol{\lambda}_k=(\lambda_k^1,...,\lambda_k^p),(k=1,...,K)$, that change with each iteration and are different for particular classes:

$$d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda}_k)=\sum_{j=1}^{p}\lambda_k^j\phi^2\left(\mathbf{u}^j(i),\mathbf{v}^j(k)\right)=\sum_{j=1}^{p}\lambda_k^j\sum_{h=1}^{H_j}\left(u_h^j(i)-v_h^j(k)\right)^2. \tag{8}$$

In the first case the weight vector is estimated globally for all classes at once, while in the second case the weights are estimated locally for each class.

FUZZY CLUSTERING ALGORITHM FOR SYMBOLIC OBJECTS

The generalization of de Carvalho and de Souza procedure [2010] proposed by the author in this paper for the case of the fuzzy classification will permit, in a situation of classes separated with difficulty, to use the partial membership of classes of objects whose similarity to several classes at the same time is high. Given the degree of membership to particular classes, one can define a function representing the classification criterion as follows:

$$\tilde{W}(\mathbf{G},\mathbf{D},\mu)=\sum_{k=1}^{K}\sum_{i=1}^{n}[\mu_k(i)]^r\, d(\tilde{\mathbf{x}}_i,\mathbf{g}_k/\boldsymbol{\lambda}_k)\rightarrow\min, \tag{9}$$

assuming that $r>1$ is the degree of fuzziness, whereas $\mu_k(i)$ is the degree of the object i membership of class C_k and $\sum_{k=1}^{K}\mu_k(i)=1$.

Assuming that the weights are the same in each class or different, one can use the method of Lagrange multipliers and solve the corresponding systems of equations, to determine the degree of individual objects membership of the classes as follows, respectively:

Using the method of Lagrange multipliers can solve the corresponding systems of equations and determine the degree of individual objects membership of the classes as formula (10) when the weights are the same in each class or formula (11), when the weights are different:

$$\mu_k(i) = \frac{\left[\sum_{j=1}^{p} \lambda^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]^{-1/(r-1)}}{\sum_{q=1}^{K}\left[\sum_{j=1}^{p} \lambda^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]^{-1/(r-1)}}, \tag{10}$$

$$\mu_k(i) = \frac{\left[\sum_{j=1}^{p} \lambda_k^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]^{-1/(r-1)}}{\sum_{q=1}^{K}\left[\sum_{j=1}^{p} \lambda_q^j \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]^{-1/(r-1)}}. \tag{11}$$

Next, proceeding in the analogous manner, one can designate vector of class patterns that minimizes the classification criterion:

$$v_h^j(k) = \frac{\sum_{i=1}^{n} [\mu_k(i)]^r u_h^j(i)}{\sum_{i=1}^{n} [\mu_k(i)]^r}. \tag{12}$$

Similarly, the best weights can be determined for which the criterion function reaches a local minimum, and $\lambda^j > 0$ and $\prod_{j=1}^{p} \lambda^j = \eta$, where $\eta \in \mathsf{R}$ is constant:

$$\lambda^j = \frac{\left\{\eta \prod_{l=1}^{p}\left(\sum_{k=1}^{K}\left[\sum_{i=1}^{n} [\mu_k(i)]^r \sum_{h=1}^{H_l} \left(u_h^l(i) - v_h^l(k)\right)^2\right]\right)\right\}^{\frac{1}{p}}}{\sum_{k=1}^{K}\left[\sum_{i=1}^{n} [\mu_k(i)]^r \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2\right]}. \tag{13}$$

If in the criterion function *W* the squared Euclidean distance is considered, parameterized by weights, which may be different for particular classes, and change with each iteration, then assuming that $\lambda_k^j > 0$ and $\prod_{j=1}^{p} \lambda_k^j = \chi$, where $\chi \in \mathsf{R}$ is constant, to determine the weights that minimize the criterion *W* one can use the method of Lagrange multipliers and some elements of algebra and obtain the formula:

$$\lambda_k^j = \frac{\left\{\chi \prod_{l=1}^{p}\left(\sum_{i=1}^{n} [\mu_k(i)]^r \sum_{h=1}^{H_l} \left(u_h^l(i) - v_h^l(k)\right)^2\right)\right\}^{\frac{1}{p}}}{\sum_{i=1}^{n} [\mu_k(i)]^r \sum_{h=1}^{H_j} \left(u_h^j(i) - v_h^j(k)\right)^2}. \tag{14}$$

The particular steps in the algorithm for fuzzy classification of symbolic data with different types of features are as follows:

1. For $i = 1,...,n$ and $j = 1,...,p$ calculate $\tilde{x}_i^j = (D_j, \mathbf{u}^j(i))$, using equality (1), (2), (3), (5) depending on the type of symbolic variable.
2. Assume $t = 0$.
3. Fix the degree of fuzziness $r > 1$, the initial fuzzy partition $\mu^{(t)} = \left\{\mu_1^{(t)},...,\mu_K^{(t)}\right\}$ and the number $\varepsilon > 0$.

4. Calculate the vector of class patterns $\mathbf{g}_1^{(t)},...,\mathbf{g}_K^{(t)}$, where $\mathbf{g}_k^{(t)} = \left(\left(g_k^1\right)^{(t)},\left(g_k^2\right)^{(t)},...,\left(g_k^p\right)^{(t)}\right)$, $k \in \{1,...,K\}$, $\left(g_k^j\right)^{(t)} = \left(D_j,(\mathbf{v}_j(k))^{(t)}\right)$, $j \in \{1,...,p\}$, is given by (10).
5. Determine the weight vector values for particular variable and classes using the formulas (13) or (14).
6. Calculate the new value of membership function $\mu^{(t+1)} = \left\{\mu_1^{(t+1)},...,\mu_K^{(t+1)}\right\}$ using formulas (10) or (11).
7. If $\left\|\mu^{(t+1)} - \mu^{(t)}\right\| > \varepsilon$ then assume $t := t+1$ and go back to step 4. Otherwise STOP.

EMPIRICAL EXAMPLE

A set of objects consists of 10 brands of cars from four companies: Skoda, Fiat, Citroen and Renault. Each brand is characterized by six features: company, engine capacity, price, available colour, comfort, fuel consumption. The features: company, engine capacity, colour take symbolic values, the price is an real valued, comfort and safety are fuzzy data. The data set is shown in Table 1.

Table 1. Data set of cars

Lp.	Brands	Company	Engine capacity	Price	Colour[1]	Comfort	Fuel consumption
1	Fabia	Skoda	1,4	43,35	B, Br, C, Cz, Cz1, F, M, M1, N, N1, P, S, S1, S2, Z, Ż	[6.5;6.7; 1.5; 0.7]	[5.4; 6.4; 0.7; 1.6]
2	Oktavia	Skoda	1,4	54,5	B, Br, C, Cz, M, M1, N, N1, P, S, S1, S2,	[7.35;7.35; 0.65; 0.65]	[5.9;6.9; 0.8; 1.6]
3	Superb	Skoda	1,8	88,3	B, Br, C, Cz, M, M1, N, N1, P, S, S1, S2,	[8.3;9.3; 0;0]	[7.5;8.7; 0.9;1.9]

[1] In the colour feature the following notation is adopted: B-white, B1-pearl white, Br-burgundy, C-red, Cz-black, Cz1-black pearl, F-violet, Gr-graphite, M- sea green, M1-light sea green, N-blue, N1-light blue, P- pistachio-green, Ps- sand-coloured, S-silver, S1-light gray, S2-gray, Z-green, Zł-golden, Ż-yellow.

Lp.	Brands	Company	Engine capacity	Price	Colour[1]	Comfort	Fuel consumption
4	Panda	Fiat	1,2	26,99	B, C, , Cz, F, M1, N, N1, Ps, S2, Zł, Ż	[6.0;7.0; 0.1;0.4]	[4.9;4.9; 0.9; 1.5]
5	Bravo	Fiat	1,6	66,99	B, B1, C, Cz, N, N1, S1, S2	[7.5;7.5; 0.4; 0.7]	[4.9;4.9; 0.8; 1.4]
6	C3 Picasso	Citroen	1,4	56,6	B1,C, Cz, N, Ps,S2, Z	[6,5;7,5; 0.2;0.2]	[6.1;7.1; 1.1; 1.6]
7	C1	Citroen	1	43,1	B, C, Cz, N, P, Ps,S1, S2	[6.8;6.8; 1;1]	[4.5;4.5; 0.6;1]
8	C5	Citroen	1,6	101,7	B, B1, Br, Cz1, Gr, M, S, S1, S2	[8.5;9.5; 0;0]	[6.6;7.6; 1.1; 1.2]
9	Thalia	Renault	1,2	29,9	B, C, Cz1,N1, Ps, S, S1, S2	[6.8;6.8; 0.7; 0.8]	[5.9;5.9; 1.1; 1.7]
10	Megane	Renault	1,6	54,45	B, Cz, N, S, S1, S2,	[7.2;8.2; 0; 0.1]	[6.3;7.3; 0.8; 1.8]

Source: the author's own elaboration on the basis of www.skoda-auto.pl; www.fiat.pl; www.renault.pl; www.citroen.pl; opinie.auto.com.pl

Table 2. Membership for cars calculated by fuzzy clustering algorithm

Lp.	1	2	3	4	5	6	7	8	9	10
$\mu_1(i)$	0,747	0,694	0,186	0,769	0,640	0,721	0,706	0,177	0,761	0,523
$\mu_2(i)$	0,253	0,306	0,814	0,231	0,360	0,279	0,294	0,823	0,239	0,477

Source: the author's own elaboration

Table 2 shows the values of the membership function of two classes for 10 objects, obtained as a result of the application of fuzzy classification algorithm for various types of symbolic and fuzzy variables using different weights for particular classes, assuming that $r = 2$, $K = 2$ i $\varepsilon = 0{,}0001$.

Analyzing the results in Table 2, two determined classes can be isolated: C1={Fabia, Oktawia, Panda, Bravo, C3 Picasso, C1, Thalia, Megane}, C2={Superb, C5}. Renault Megane is a mixed object whose belonging to both classes is considerably high. It is also possible to notice that Fiat Bravo and Skoda Octavia have a fairly high degree of belonging to the second class, which means that there is a relatively high degree of similarity of these vehicles to the objects belonging to class 2.

CONCLUDING REMARKS

The presented iterative algorithm of the classical and fuzzy classification permits to cluster objects featured by mixed-value symbolic data. The algorithm

using distances with different weights for particular classes is able to identify the classes of different shapes and sizes, which is a definite advantage. The disadvantage is that they are dependent on the initial partition. The experimental evaluations for the interval-valued data have showed the superiority of classification algorithm applying the same weights, in terms of class recognition quality (assessed using the corrected Rand index) in the configuration of data with a priori almost equal dispersion of classes, and the superiority of the algorithm using different weights for particular classes where dispersion of classes is preset in advance as a different one. The proposed fuzzy classification methods for symbolic data with different types of features are a generalization of the methods presented in the work of de Carvalho and de Souza [2010], and therefore they have the same advantages and disadvantages. They allow, however, to assign to the individual objects the degrees of membership of different classes in the range from 0 to 1. This is of particular importance when the classes are separated with difficulty and the classical clustering forces the assignment of a given object to one class only. Therefore, in this case, the fuzzy classification may give better results, identifying the mixed objects whose similarity to several classes at the same time is high.

REFERENCES

Bock H. H., Diday E. (2000) Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag, Berlin, Heidelberg.

De Carvalho F.A.T. (1995) Histograms in symbolic data analysis. Annals of Operations Research 55, 229–322.

De Carvalho F.A.T., de Souza R. (2010) Unsupervised pattern recognition models for mixed feature-type symbolic data, Pattern Recognition Letters 31, 430–443.

Diday E., Simon J.C. (1976) Clustering analysis. In: Fu, K.S. (Ed.), Digital Pattern Clasification. Springer, Berlin, 47–94.

Zimmermann H.J. (1991) Fuzzy Set Theory and Its Applications, Kluwer, Dordrecht.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 61 – 69

USEFUL GOVERNMENT EXPENDITURE INFLUENCE ON THE SHADOW ECONOMY

Paulina Malaczewska
Department of Econometrics, University of Lodz
e-mail: pmalaczewska@uni.lodz.pl

Abstract: This paper contains the attempt to describe the phenomenon of shadow economy as a zero-sum non-cooperative, normal form game between households and the government. In the model government spending can be treated as a government consumption or as an expenses that contribute to increased social welfare and for the provision of public goods and services. We conduct sensitivity analysis of Nash equilibrium in models with two different types of government expenditure and examine whether proposed models indicate a various mechanisms and determinants of the undeclared economic activity.

Keywords: shadow economy, useful government expenditure, game theory, Nash equilibrium

INTRODUCTION. SHADOW ECONOMY

According to the latest estimates of the shadow economy in 31 European Countries informal sector is from 7,1% (Switzerland) to 31,2% (Bulgaria) of official GDP[1]. Therefore the examination of driving forces of the shadow economy seems to be necessary and extremely important.

In economic literature tax burden and social security contributions are mentioned among the most common determinants of the shadow economy, e.g. [Kozyra–Cybulska et al. 2010], [Schneider 2006], [Patera et al. 2007]. Also [Smuga et al. 2005] highlight complex, confusing and inflexible regulations and poor detection of undeclared activities. The results of 22 different empirical studies are summarized in paper [Schneider, Williams 2013]. Authors note that there are

[1] Schneider F. (2013) Size and Development of the Shadow Economies of Portugal and 35 other OECD Countries from 2003 to 2013: Some New Facts, working paper, p. 5.

several factors which explain 78-98% of the variance of the shadow economy. Taxation and social security contributions, quality of public institutions, public services, regulations of labor market, transfer payments and tax morale are mentioned among those driving forces of the shadow economy. [Kabaj 2009] suggests that one determinant of the shadow economy is high unemployment rate. This is consistent with increase in size of the shadow economy which happened in most OECD countries in 2009 during the economic crisis. [Rosser et al. 2000] show positive relationship between income inequality and the size of the shadow economy in transition economies.

Another determinant is mentioned by [Schneider, Dreher 2006]. According to their research the level of corruption has a significant influence on the shadow economy. This influence is ambiguous and depends on level of economic development. In high-developed countries the shadow economy and corruption are mutually substitutable, while in the case of developing countries - complementary.

Other researches show high correlation between size of the shadow economy and fiscal illusion [Buehn et al. 2012], minimal wage [Maloney, Mendez 2004] or rule of law and quality of institutions [Aruoba 2010].

The purpose of this paper is to analyze in a simple theoretical model impact of useful government expenditures on size of the shadow economy. This paper is organized as follows. Section 1 contains description of the model of shadow economy. This model is an extended version of model presented in [Malaczewska 2013] and was enriched with useful government expenditures. In section 2 the solution of the model is provided and detailed sensitivity analysis is conducted. The last section concludes with a discussion of possible extensions and directions for the future research.

MODEL OF SHADOW ECONOMY

The following model is an extension of the basic model of the shadow economy presented in [Malaczewska 2013], which has been enriched with the idea of useful government spending[2].

We consider a model with two different economic entities: households and government. Households have time endowment $\bar{L}$, which can be divided into activity in the shadow economy (L_s) or into activity in formal sector ($\bar{L} - L_s$). As a result, shadow economy is created by households. However, choice of the size of informal sector made by households depends on economic conditions which are determined by the government. Government has two instruments of economic policy - the average tax rate (τ) and the effectiveness of government's control institutions (by determining the amount of expenses for their activities, denoted by

[2] The concept of wasteful government spending has been applied among others in paper Marattin L., Palestini A. (2010) Edgeworth Dependence and Government Spending Multipliers: a Theoretical Analysis, working paper.

W_K). If households are caught on working in the informal sector, they must pay a fine (P) proportional to the amount of income received in the informal sector. Revenues from taxes and fines government spend on the control institutions or government spending (G). Government spending can be treated as socially useful expenses (useful government expenditures) or as government consumption (from household point of view - wasteful government expenditures). We assume balanced budget, so government revenues are equal to expenses. To sum up, government's budget constraint can be written as:

$$w_r(\bar{L} - L_s)\tau + P = W_K + G \tag{1}$$

where $\tau \in (0,1)$ and w_r (w_s) denotes the average wage per hour of work in formal sector (informal sector). We assume that expected revenues from fines P are given by the equation:

$$\mathrm{P} = \left[\beta\, \mathrm{A}_1\left(1 - \mathrm{e}^{-\phi\, \mathrm{W_k}}\right) + (1-\beta)\mathrm{A}_2 \frac{\mathrm{L_s}}{\bar{\mathrm{L}} - \mathrm{L_s}}\right] a\mathrm{w_s} \tag{2}$$

where $\phi \in \mathrm{R}_+$, $\beta, \mathrm{A}_1, \mathrm{A}_2 \in [0,1]$, and $a > 0$ denotes ratio or multiplicity of wage obtained in shadow economy, which determines the amount of fine.

Household income comes from activity in shadow economy ($\mathrm{w_s L_s}$) and from wages paid in formal sector, but decreased by taxation ($\mathrm{w_r}(\bar{\mathrm{L}} - \mathrm{L_s})(1-\tau)$). All revenues households can spend on consumption C or to pay penalties P. No borrowing is allowed. Summarizing, households' budget constraint is as follows:

$$\mathrm{w_r}(\bar{\mathrm{L}} - \mathrm{L_s})(1-\tau) + \mathrm{w_s L_s} = \mathrm{C} + \mathrm{P} \tag{3}$$

Both the government and households maximize their own utility function. We assume that utility function of government depends on two factors – government spending G and level of social support of government. The government, as an entity elected for the term, must take care about popularity among households, to ensure the possibility of reelection. Thus function of social support of government S was created, which depends on tax burden τ (negative relation) and size of formal sector (positive relation):

$$\mathrm{S} = \mathrm{D}_1 \sqrt{1-\tau} + \mathrm{D}_2(\bar{L} - \mathrm{L_s})^2 \tag{4}$$

where $\mathrm{D}_1, \mathrm{D}_2 \in \mathrm{R}_+$.

The utility function of government is assumed to have the normal properties of being concave with respect to law of diminishing marginal utility and can be written as:

$$\mathrm{U_g} = \frac{(\mathrm{G} + \gamma\, \mathrm{S})^{1-\alpha} - 1}{1-\alpha} \tag{5}$$

where $\gamma \in \mathrm{R}_+, \alpha \in (0, \infty)$.

Similarly, the utility of households is a function of two factors: consumption C and useful government expenditures (ψG):

$$\mathrm{U_h} = \frac{(\mathrm{C} + \psi\, \mathrm{G})^{1-\delta} - 1}{1-\delta} \tag{6}$$

where $\delta \in (0,\infty), \psi \in [0,1)$. Parameter ψ represents how much taxes and penalties paid by households are returned to them in the form of socially useful government spending. When $\psi = 0$ we consider basic model presented in [Malaczewska 2013] without useful government expenditures. In this case, all government expenditures G constitute government consumption and are only used to meet the government needs. When $\psi \neq 0$ some part of government spending contributes to the welfare of society and is used to provide public goods and services. In this paper, we analyze the case when $\psi \neq 0$. We will examine whether extension of the analysis of the shadow economy of useful government expenditures will change significantly equilibrium state of the model and its sensitivity. Then both cases will be compared and appropriate conclusions drawn.

Based on a previous discussion, government maximizes its utility function

$$\mathrm{U_g} = \frac{\left(\mathrm{G}+\gamma\left[\mathrm{D_1}\sqrt{1-\tau}+\mathrm{D_2}(\bar{L}-\mathrm{L_s})^2\right]\right)^{1-\alpha}-1}{1-\alpha} \tag{7}$$

subject to constraint

$$w_r(\bar{L}-L_s)\tau + \left[\beta\,\mathrm{A_1}\left(1-\mathrm{e}^{-\phi\,\mathrm{W_k}}\right)+(1-\beta)\mathrm{A_2}\frac{\mathrm{L_s}}{\bar{\mathrm{L}}-\mathrm{L_s}}\right]\mathrm{aw_s} = W_K + G \tag{8}$$

by choosing tax rate τ and amount of expenses for government's control institutions W_K. Similarly, households choose size of the shadow economy L_s which maximizes their utility function

$$\mathrm{U_h} = \frac{(\mathrm{C}+\psi\,\mathrm{G})^{1-\delta}-1}{1-\delta} \tag{9}$$

subject to the constraint

$$\mathrm{w_r}(\bar{\mathrm{L}}-\mathrm{L_s})(1-\tau)+\mathrm{w_s}\mathrm{L_s} = \mathrm{C} + \left[\beta\,\mathrm{A_1}\left(1-\mathrm{e}^{-\phi\,\mathrm{W_k}}\right)+(1-\beta)\mathrm{A_2}\frac{\mathrm{L_s}}{\bar{\mathrm{L}}-\mathrm{L_s}}\right]\mathrm{aw_s} \tag{10}$$

where $\tau \in (0,1),\ \psi \in [0,1),\ \beta, \mathrm{A_1}, \mathrm{A_2} \in [0,1],\ \phi, \mathrm{D_1}, \mathrm{D_2}, \gamma, \mathrm{a}, \alpha, \delta \in \mathrm{R_+}$.

SOLUTION AND SENSITIVITY ANALYSIS OF THE MODEL

The model has been solved using the method of Lagrange multipliers. The first order conditions can be written as[3]:

$$\begin{cases} \frac{\partial\,\mathrm{U_h}}{\partial\,\mathrm{L_s}} = 0 \Rightarrow\ -\mathrm{w_r}[1-\tau(1-\psi)]+\mathrm{w_s} = (1-\beta)\mathrm{A_2}\mathrm{aw_s}\frac{\bar{\mathrm{L}}}{(\bar{\mathrm{L}}-\mathrm{L_s})^2}(1-\psi) \\ \frac{\partial\,\mathrm{U_g}}{\partial\,\tau} = 0 \Rightarrow \mathrm{w_r}(\bar{\mathrm{L}}-\mathrm{L_s}) = \frac{\gamma\,\mathrm{D_1}}{2\sqrt{1-\tau}} \\ \frac{\partial\,\mathrm{U_g}}{\partial\,\mathrm{W_K}} = 0 \Rightarrow \phi\,\beta\,\mathrm{A_1}\,\mathrm{aw_s}\,\mathrm{e}^{-\phi\,\mathrm{W_k}} = 1 \end{cases} \tag{11}$$

After several calculations stationary point of the Lagrange function is obtained. Sufficient conditions are fulfilled, so this point is both optimal solution and Nash equilibrium of considered model. The analytical solution of the model, when $\psi = 0$ is as follows:

[3] All calculations are available from the author on request.

$$\begin{cases} W_K^* = \frac{1}{\phi}(\ln\phi + \ln\beta + \ln A_1 + \ln a + \ln w_s) \\ \tau^* = 1 - \frac{w_s}{w_r + (1-\beta)A_2 a w_s \bar{L}\frac{4w_r^2}{\gamma^2 D_1^2}} \\ L_s^* = \bar{L} - \frac{\gamma D_1}{2w_r}\sqrt{\frac{w_r + (1-\beta)A_2 a w_s \bar{L}\frac{4w_r^2}{\gamma^2 D_1^2}}{w_s}} \end{cases} \quad (12)$$

while the analytical solution of the model, when $\psi \neq 0$ can be written as:

$$\begin{cases} W_K^* = \frac{1}{\phi}(\ln\phi + \ln\beta + \ln A_1 + \ln a + \ln w_s) \\ \tau^* = 1 - \frac{-w_r \psi + w_s}{(1-\psi)\left[w_r + (1-\beta)A_2 a w_s \bar{L}\frac{4w_r^2}{\gamma^2 D_1^2}\right]} \\ L_s^* = \bar{L} - \sqrt{\frac{(1-\beta)(1-\psi)A_2 a w_s \bar{L} + \frac{1}{4}\frac{1}{w_r}(1-\psi)\gamma^2 D_1^2}{-w_r \psi + w_s}} \end{cases} \quad (13)$$

Following tables 1 and 2 contains results of sensitivity analysis of Nash equilibrium due to changes in parameters values in both cases. Positive (negative) value of the first derivative of any decision variable with respect to given parameter informs us that in a case of two exact economies that are different only in a size of given parameter, the one with greater level of that parameter also has greater (lower) level of decision variable of interest.

Table 1. Sensitivity analysis of Nash equilibrium due to changes in parameters values, when $\psi = 0$

Variable/parameter	β	a	w_S	w_r
W_K^*	$\frac{\partial W_k^*}{\partial \beta} > 0$	$\frac{\partial W_k^*}{\partial a} > 0$	$\frac{\partial W_k^*}{\partial w_s} > 0$	no relation
τ^*	$\frac{\partial \tau^*}{\partial \beta} < 0$	$\frac{\partial \tau^*}{\partial a} > 0$	$\frac{\partial \tau^*}{\partial w_s} < 0$	$\frac{\partial \tau^*}{\partial w_r} > 0$
L_s^*	$\frac{\partial L_s^*}{\partial \beta} > 0$	$\frac{\partial L_s^*}{\partial a} < 0$	$\frac{\partial L_s^*}{\partial w_s} > 0$	$\frac{\partial L_s^*}{\partial w_r} > 0$

Source: own calculations

Table 2. Sensitivity analysis of Nash equilibrium due to changes in parameters values, when $\psi \neq 0$

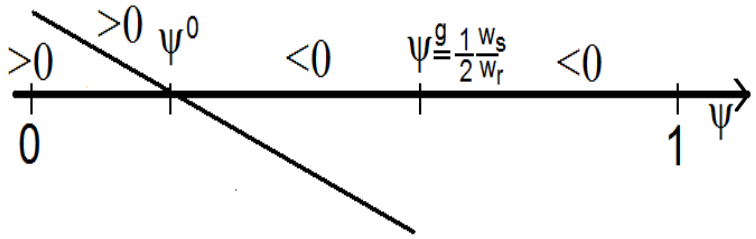
Variable/parameter	β	a	w_S	w_r
W_K^*	$\frac{\partial W_k^*}{\partial \beta} > 0$	$\frac{\partial W_k^*}{\partial a} > 0$	$\frac{\partial W_k^*}{\partial w_s} > 0$	no relation
τ^*	$\frac{\partial \tau^*}{\partial \beta} < 0$	$\frac{\partial \tau^*}{\partial a} > 0$	$\frac{\partial \tau^*}{\partial w_s} < 0$	$\frac{\partial \tau^*}{\partial w_r} > 0$
L_s^*	$\frac{\partial L_s^*}{\partial \beta} > 0$	$\frac{\partial L_s^*}{\partial a} < 0$	$\frac{\partial L_s^*}{\partial w_s} > 0$	$\frac{\partial L_s^*}{\partial w_r} = ?$

Source: own calculations

Analysis of tables 1 and 2 leads us to following conclusions:

- If the probability of detection of activity in the shadow economy dependent on W_K will grow (greater level of β), then it can be expected that consequently the government will raise the level of expenditures on control institutions (hence $\frac{\partial W_K}{\partial \beta} > 0$).
- The increase in w_s makes the work in the shadow economy more attractive. As a result, greater number of households work in shadow economy and the size of informal sector is growing.
- Increase in parameter a leads to increase in optimal level of W_K^*. Increasing a is equivalent to increasing the penalties for activities in the shadow economy. In this case, it is profitable for government to raise expenditures W_K, because it will be compensated by increase of revenues from penalties.
- The rest of the results of the sensitivity analysis, except $\frac{\partial L_s^*}{\partial w_r}$, is consistent with standard economic theory.
- The sign of the partial derivative $\frac{\partial L_s^*}{\partial w_r}$ depends on the value of the parameter ψ, which is depicted on figure 1.

Figure 1. The sign of the derivative $\frac{\partial L_s^*}{\partial w_r}$ depending on the value of the parameter ψ



Source: based on own calculation

By ψ^0 on Figure 1 we denote value of the parameter ψ for which the size of the shadow economy is insensitive to the wage rate in formal sector ($\frac{\partial L_S^*}{\partial w_r} = 0$), and by ψ^g – value of the parameter ψ from which we can unambiguously determine the sign of the partial derivative $\frac{\partial L_S^*}{\partial w_r}$. Level of ψ^0 is given by equation:

$$\psi^0 = \frac{\frac{w_S}{4w_r^2}\gamma^2 D_1^2}{\frac{1}{2w_r}\gamma^2 D_1^2 + w_S(1-\beta)A_2 a\bar{L}} \tag{14}$$

For ψ lower than ψ^0 (in particular for $\psi = 0$) we have $\frac{\partial L_S^*}{\partial w_r} > 0$, so with greater level of w_r greater levels of L_S are associated. When ψ is greater than ψ^0 (in particular for $\psi > \psi^g$) we have $\frac{\partial L_S^*}{\partial w_r} < 0$. Only in this case increase in level of wages in formal sector leads to decrease in the size of the shadow economy.

In model, when $\psi \neq 0$, we can also determine partial derivatives $\frac{\partial L_S^*}{\partial \psi}$ i $\frac{\partial \tau^*}{\partial \psi}$. The results are ambiguous and depend on the relationship between earnings in the informal sector and the shadow economy:

- if $w_s > w_r > w_r(1-\tau) + \tau\, w_r\psi$, then $\frac{\partial L_S^*}{\partial \psi} > 0$ i $\frac{\partial \tau^*}{\partial \psi} < 0$

This result is surprising, because the increasing share of government spending on public goods and services which meet the needs of households (increase of ψ) contributes to the growth of the shadow economy and to the decline in the tax burden. Probably, wages from the shadow economy are so large, that even improved care about needs of society and the tax cuts will not lead to a decrease in the shadow economy, but on the contrary – will contribute to its growth. This requires further studies.

- if $w_r > w_s > w_r(1-\tau) + \tau\, w_r\psi$, then $\frac{\partial L_S^*}{\partial \psi} < 0$ i $\frac{\partial \tau^*}{\partial \psi} > 0$

This result is consistent with our economic knowledge. Households encouraged by useful government expenditures (increase of ψ) leave shadow economy and return to formal sector despite tax increases. Apparently in this case households notice the benefits of paying taxes – bigger part of tax revenue returns as public goods and services provided by the government.

SUMMARY AND CONCLUSIONS

In this paper model describing the shadow economy as a result of interaction between households and government has been created. Each of economic entities maximize their own utility function – households by choosing optimal size of the shadow economy, and government by choosing tax rate and level of expenditures on control institutions. Additionally, model has been extended by useful government expenditures to analyze their influence on the size of shadow

economy. Model with useful government expenditures has been described, solved and compared with basic model. Analysis of the model leads to following conclusions:

1. Extension of the model of useful government expenditures does not change the effect of increase of parameter β, a and w_s on the equilibrium values L_S^*, τ^* and W_K^*.
2. Increasing wages in the formal sector always lead to an increase in taxation.
3. Increasing wages in the formal sector have ambiguous impact on the size of the shadow economy depends on the value of parameter ψ. When the share of useful government expenditures is low (low ψ or $\psi = 0$), then increase in wages in the formal sector leads to increase of the size of the shadow economy and taxation. Therefore, if the government intends to use this instrument (increasing wages in formal sector, e.g. by increasing minimum wage[4]) in order to reduce the shadow economy, it brings the opposite result. On the other hand, when the share of useful government expenditures is high (high ψ), then increase wages in the formal sector lead to decrease the size of the shadow economy, despite an increase in taxation.
4. Also ambiguous relationship is obtained for the effect of changing parameter ψ on the size of the shadow economy and tax burden. It appears that if wage in the shadow economy is significantly larger than the gross wage in the formal sector, then increase of the share of useful government spending leads to the increase of the shadow economy, despite the lower taxes. In this case, increasing ψ will not decrease the size of shadow economy. On the other hand, when wage in the shadow economy is smaller than gross wage in formal sector (but higher than net wage), the increase of ψ will lead to a decline in the shadow economy, even while the taxation has been increased.

Presented model and analysis are not faultless and require some improvements. First of all, it is necessary to extend the model to another economic entity – firms and, as a result, create labor market. Also other theoretical determinants of the shadow economy should be included in the model, such as corruption, fiscal illusion etc. Moreover, the assumption of balanced budget is doubtful and counterfactual. These observations will be developed in further research.

REFERENCES

Aruoba S. B. (2010) Informal Sector, Government Policy and Institutions, Working paper.

[4] The relationship between the size of the shadow economy and the level of minimum wage is analyzed in paper Maloney W., Mendez J. (2004) Measuring the impact of minimum wages. Evidence from Latin America, Law and Employment: Lessons from Latin America and the Caribbean. University of Chicago Press.

Buehn A., Dell'Anno R., Schneider F. (2012) Fiscal illusion and the shadow economy: Two sides of the same coin?, MPRA Paper No. 42531.

Dreher, A., Schneider F. (2006) Corruption and the shadow economy : an empirical analysis, CESifo working papers, No. 1653.

Kabaj M. (2009) Praca nierejestrowana we współczesnej literaturze ekonomicznej, Polityka Społeczna nr 10.

Kozyra-Cybulska M., Molenda A., Wojnar E., Zielański M (2010) Badanie warunków i jakości życia oraz zachowań ekonomicznych w gospodarstwach domowych, Działalność nierejestrowana, PTS, Rzeszów.

Malaczewska P. (2013) Analiza zjawiska szarej strefy jako gry niekooperacyjnej, Matematyka i Informatyka na Usługach Ekonomii, Zeszyty Naukowe, UEP, w druku.

Maloney W., Mendez J. (2004) Measuring the impact of minimum wages. Evidence from Latin America, Law and Employment: Lessons from Latin America and the Caribbean. University of Chicago Press.

Marattin L., Palestini A. (2010) Edgeworth Dependence and Government Spending Multipliers: a Theoretical Analysis, working paper.

Patera K. et al. (2007) Projekt Przyczyny pracy nierejestrowanej, jej skala, charakter i skutki społeczne, CBOS, IPiSS, Warszawa.

Rosser Jr, J. B., Rosser M. V., Ahmed E. (2000) Income inequality and the informal economy in transition economies, Journal of Comparative Economics, 28.1.

Schneider F. (2006) Shadow Economies and Corruption All Over the World: What Do We Really Know?, Economics Discussion Papers, No 2007-9, Kiel Institute for the World Economy.

Schneider F. (2013) Size and Development of the Shadow Economies of Portugal and 35 other OECD Countries from 2003 to 2013: Some New Facts, working paper.

Schneider F., Williams C. (2013) The Shadow Economy, The Institute of Economic Affairs, London.

Smuga T. et al. (2005) Metodologia badań szarej strefy na rynku usług turystycznych, Instytut Koniunktury i Cen Handlu Zagranicznego, Warszawa.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No.2, 2013, pp. 70 – 77

ENDOGENOUS TECHNOLOGICAL PROGRESS AND ECONOMIC GROWTH IN A MODEL WITH NATURAL RESOURCES

Maciej Malaczewski
Department of Econometrics, University of Lodz
e-mail: mmalaczewski@uni.lodz.pl

Abstract: Idea of technological progress based on two different types of research that generate radical and incremental innovations, became new approach in endogenous growth modeling. This approach seems to be useful in modeling relationships between technological progress, natural resources, environmental quality and economic growth.

The purpose of this paper is to answer questions about relationships between long-term economic growth, technological progress and use of natural resources. The main object is an impact of natural resources use on growth rate and a role of endogenous technological progress.

Keywords: economic growth, technological progress, natural resources, technological opportunities

INTRODUCTION

Formal description of innovation process is not easy, it is even much more complicated to include it in a standard economic growth model, especially when natural resources use also has to be included. There are at least three dominating approaches to technological progress modeling: human capital approach (considered extensively, for example, in [Lucas 1988]), research and development sector with rising number of patents (like in [Romer 1990]), and, recently, technological opportunities approach, introduced by [Olsson 2000]. The latter concept is mostly based on [Kuhn's 2012] theory of scientific revolutions.

The purpose of this paper is an attempt to answer question about theoretical dependences among long-run economic growth, technological progress and natural resources use, with implication of technological opportunities approach. Main subject of this research is influence of natural resources on rate of economic growth and a role of endogenous technological progress in that dependence.

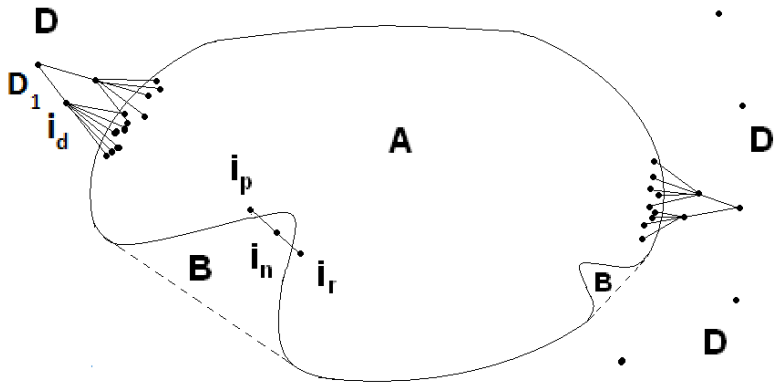
The structure of this paper is as follows. In part one we describe Olsson's idea of endogenous technical change, based on technological opportunities. Construction of an economic growth model based on this idea is described in section two. After that, in section three, we present solution of the model and perform its analysis. Whole paper is ended with a short summary.

Author acknowledge support from KBN in 2010 – 2013 years, grant Nr NN 112 553138.

IDEA OF TECHNOLOGICAL OPPORTUNITIES

Concept of technological opportunities was introduced by [Olsson 2000, 2005], but even earlier some authors mentioned that new knowledge is a result of combining few old existing theories. Figure 1 shows basics of this idea.

Figure 1. Idea of technological opportunities



Source: [Olsson 2000]

Description of Olsson's idea is as follows. Let $A \subset R^k$ be a set of all existing, known and widespread ideas. In this set can be found all basic ideas (like adding and subtracting natural numbers), but also more complex ones (like e.g. quantum mechanics). New ideas appear in three different ways: as a scientific discovery, as a radical innovation or as an incremental innovation.

Incremental innovation is a result of regular research work, which is based on use of existing ideas and combine them, which leads to form new idea. Simple example of incremental innovation might be a new application for smartphones which is created with a use of existing hardware (phone), its software and some algorithmic schemes. These are not revolutionary ideas, because they need existing ones, they do not lead to significant expansion of knowledge.

New idea i_n arises as a convex combination of early existing ideas i_p and i_r, so:

$$i_n = \alpha \cdot i_p + (1 - \alpha) \cdot i_r, \quad (1)$$

where $\alpha \in (0; 1)$. It is possible to match in convex combination more than two existing ideas. Convexity of this linear combination comes as a result of mixing ideas, if one of ideas i_p, i_r is used more (with greater share) that result would be different[1]. Convexity of linear combinations implies that new ideas, emerge as a result of incremental innovations, are in a convex hull of A (which we denote as a $conv(A)$). Set $B = conv(A) - A$ is a set of technological opportunities – it contains all of ideas, that are reachable with existing state of knowledge, but still not invented. Whenever new idea is invented as an incremental innovation set of technological opportunities become smaller and set of all ideas become greater. It is easy to see, that if set B would not grow eventually all of ideas possible to invent will be invented. So there must be some way to increase size of set of technological possibilities. This is based on another two types of scientific innovations.

The most important are scientific discoveries, which are more accidental, non intentional side-effects of research than effects of directed research. Appearance of scientific discoveries always creates new paradigm, they are anomalies with respect to current set of knowledge. These discoveries (D) are outside of set A, in some distance from it. With existing level of knowledge it is not clear what are the causes of these discoveries, but they become inspiration to directed and intentional research.

Existence of ideas outside of set A implies possibilities of combine ideas from set A and scientific discovery (for example D_1). That combination, which is an effect of intentional research, creates new knowledge (i_d) outside of set A, leads to its enlargement and enlargement of convex hull, and leads to enlargement of set B. New research might be continued until scientific discovery would be included into set A and would become a part of general knowledge. Scientific discovery leads to new paradigm, which opens new technological opportunities. That leads to creation of new ideas.

In the next section we describe a model of economic growth based on idea of technological opportunities.

[1] It is hard to imagine what would arise as a result of mixing idea of car tires (with a share, let's say, 0.98) and idea of Phillips Curve (with a share 0.02). Therefore concept of Olsson seems to be more adequate inside of a single scientific discipline or on a cross-section of similar disciplines.

MODEL

Concept of technological opportunities, described in details in section one, seems quite interesting but not easy to apply in a growth theory. There are not many papers which include that idea in standard economic growth model. One, that need to be mentioned, is a paper by Growiec and Schumacher [2013]. In that paper standard economic growth model with endogenous technological progress is presented. This progress may only take place if there are technological opportunities, and those opportunities may be created only with a use of existing knowledge. Appearance of new knowledge raises TFP, which leads to increase in production. Paper considers both centralized economy and social planner case, results are similar in both cases. It seems to be much harder to apply technological opportunities approach to modeling of natural resources use. Probably the only paper considering this problem was [Lundström 2003].

Our model is based on [Olsson 2000, 2005], [Lundström 2003] and [Growiec, Schumacher 2013] models. We use standard optimal control approach to long-run economic growth modeling. We consider closed economy. Households contains L citizens, their number increase with a rate of growth equal to n:

$$\dot{L} = nL \tag{2}$$

We assume, for simplicity, that at the beginning the number of citizens is equal to 1. Labor supply is shared between three different activities: production, basic research and applied research.

Applied research creates new ideas. Evolution of level of knowledge A is of following form:

$$\dot{A} = \delta(u_A L)^{\gamma} B^{\mu} \tag{3}$$

where u_A is a share of time devoted to applied research, B is a level of technological opportunity, $\delta, \gamma, \mu > 0$ are constant parameters.

Basic research creates new technological opportunities:

$$\dot{B} = \zeta(u_B L)^{\gamma} A^{\mu} - \delta(u_A L)^{\gamma} B^{\mu} \tag{4}$$

where u_B is a share of time devoted by households to basic research, $\zeta > 0$ is a parameter. Creation of new technological opportunities depends on existing level of knowledge, but amount of technological opportunities decreases by an amount of new, created knowledge.

Level of production is given by production function of standard, Cobb-Douglas form:

$$Y = A^{\sigma} M^{\alpha} (u_Y L)^{1-\alpha} \tag{5}$$

where $\sigma > 0, \alpha \in (0,1)$, M is effective physical capital. By effective physical capital we understand physical capital that may be powered with produced energy:

$$M = \min\{aK, bE\} \tag{6}$$

where K is physical capital stock and E is a flow of produced energy. Energy production function uses two factors – existing stock of capital and flow of natural resources:

$$E = A^{\kappa} K^{\beta} R^{1-\beta} \tag{7}$$

where $\kappa > 0, \beta \in (0,1)$, R is natural resources flow used to energy production. We assume that there is always enough of physical capital, the problem lies only in energy production, so

$$M = bE. \tag{8}$$

Evolution of physical capital is in a standard form:

$$\dot{K} = Y - C - dK \tag{9}$$

where C is a level of consumption and d is depreciation rate.

Economy is endowed with supply S of natural resources, which is extracted successively:

$$\dot{S} = -R \tag{10}$$

Households maximize their lifetime utility from present moment to infinity given by following utility function:

$$L_0 \int_0^{+\infty} e^{-(\rho-n)t} \frac{1}{1-\theta} \left(c^{1-\theta} - 1\right) dt \to max \tag{11}$$

where $\rho > 0$ is a discount rate, $L_0 = 1$ and $-\theta$ is equal to elasticity of marginal utility. Small letter c denotes consumption per capita.

We express the model in variables in per capita terms and denote them with small letters, for example, $k = \frac{K}{L}$. Our model is now of the following form:

$$\int_0^{+\infty} e^{-(\rho-n)t} \frac{1}{1-\theta} \left(c^{1-\theta} - 1\right) dt \to max \tag{12}$$

$$\dot{k} = A^{\sigma+\kappa\alpha} b^{\kappa\alpha} k^{\beta\alpha} r^{(1-\beta)\alpha} (1 - u_A - u_B)^{1-\alpha} - c - (d+n)k \tag{13}$$

$$\dot{A} = \delta u_A^{\gamma} e^{\gamma n t} B^{\mu} \tag{14}$$

$$\dot{B} = \zeta u_B^{\gamma} e^{\gamma n t} A^{\mu} - \delta u_A^{\gamma} e^{\gamma n t} B^{\mu} \tag{15}$$

$$\dot{s} = -r - ns \tag{16}$$

Households choose their level of per capita consumption c, use of natural resource per capita r and share of time devoted between work, applied research and basic research u_Y, u_A, u_B. Obviously

$$u_Y = 1 - u_A - u_B. \tag{17}$$

In the next section we derive solution of this model and draw some conclusions.

SOLUTION AND DISCUSSION

We maximize present-value Hamiltonian, given by:

$$H(u_A, u_B, r, c, A, B, s, k, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = e^{-(\rho-n)t}\frac{1}{1-\theta}\left(c^{1-\theta} - 1\right) + \lambda_1\left(\delta u_A^{\gamma} e^{\gamma n t} B^{\mu}\right) + \lambda_2\left(\zeta u_B^{\gamma} e^{\gamma\, nt} A^{\mu} - \delta u_A^{\gamma} e^{\gamma n t} B^{\mu}\right) + \lambda_3(-r - ns) + \lambda_4\left(A^{\sigma+\kappa\alpha} b^{\kappa\alpha} k^{\alpha\beta} r^{\alpha(1-\beta)}(1 - u_A - u_B)^{1-\alpha} - c - (d+n)k\right). \quad (18)$$

Transversality conditions are as follows:

$$\lim_{t\to+\infty} \lambda_1 A = 0, \quad (19)$$

$$\lim_{t\to+\infty} \lambda_2 B = 0, \quad (20)$$

$$\lim_{t\to+\infty} \lambda_3 s = 0, \quad (21)$$

$$\lim_{t\to+\infty} \lambda_4 k = 0. \quad (22)$$

First order conditions are of following form:

$$e^{-(\rho-n)\,t} c^{-\theta} = \lambda_4, \quad (23)$$

$$\lambda_4 \alpha(1-\beta) A^{\sigma+\kappa\alpha} b^{\kappa\alpha} k^{\alpha\beta} r^{\alpha(1-\beta)-1}(1 - u_A - u_B)^{1-\alpha} = \lambda_3, \quad (24)$$

$$(\lambda_1 - \lambda_2)\delta\gamma u_A^{\gamma-1} e^{\gamma n t} B^{\mu} = \lambda_4(1-\alpha) A^{\sigma+\kappa\alpha} b^{\kappa\alpha} k^{\alpha\beta} r^{\alpha(1-\beta)}(1 - u_A - u_B)^{(1-\alpha)-1} \quad (25)$$

$$\lambda_2 \zeta\eta u_B^{\gamma-1} e^{\gamma n t} A^{\mu} = \lambda_4(1-\alpha) A^{\sigma+\kappa\alpha} b^{\kappa\alpha} k^{\alpha\beta} r^{\alpha(1-\beta)}(1 - u_A - u_B)^{-\alpha}, \quad (26)$$

$$-\dot{\lambda}_1 = \lambda_2 \zeta\mu u_B^{\gamma} e^{\gamma n t} A^{\mu-1} + \lambda_4(\sigma + \kappa\alpha) A^{\sigma+\kappa\alpha-1} b^{\kappa\alpha} k^{\alpha\beta} r^{\alpha(1-\beta)}(1 - u_A - u_B)^{1-\alpha} \quad (27)$$

$$-\dot{\lambda}_2 = \lambda_1 \delta\mu u_A^{\gamma} e^{\gamma n t} B^{\mu-1} - \lambda_2 \delta\mu u_A^{\gamma} e^{\gamma n t} B^{\mu-1} \quad (28)$$

$$-\dot{\lambda}_3 = -\lambda_3 n \quad (29)$$

$$-\dot{\lambda}_4 = \lambda_4\left(\alpha\beta A^{\sigma+\kappa\alpha} b^{\kappa\alpha} k^{\alpha\beta-1} r^{\alpha(1-\beta)}(1 - u_A - u_B)^{1-\alpha} - (d+n)\right). \quad (30)$$

We define steady-state as a state when all variables grow at constant rates. With this assumption, we use standard procedure to obtain solution. Growth rates of variables are as follows:

$$g_y = g_c = g_k = \frac{(\sigma+\kappa\alpha)\frac{\gamma n}{1-\mu} - \alpha(1-\beta)\rho}{1-\alpha+\alpha\theta(1-\beta)} > 0 \quad (31)$$

$$g_s = g_r = (1-\theta) g_y - \rho < 0 \quad (32)$$

$$g_A = g_B = \frac{\gamma n}{1-\mu} \quad (33)$$

All four transversality conditions come down to a single inequality:

$$-\rho + n + (1-\theta) g_y < 0 \quad (34)$$

which is fulfilled[2].

Formulas for u_Y, u_A and u_B are also possible to derive, but they are too complex to present them here[3]. All rates of growth depend only on parameters. Table 1 presented below contains signs of first partial derivatives of rates of growth with respect to chosen parameters. During process of obtaining those signs we assumed for simplicity that $\theta > 1$.

Table 1. Signs of first partial derivatives (with assumption $\theta > 1$)

	$x = \rho$	$x = \sigma$	$x = \kappa$	$x = \gamma$	$x = n$	$x = \mu$	$x = \theta$
$\frac{\partial g_y}{\partial x} = \frac{\partial g_k}{\partial x} = \frac{\partial g_c}{\partial x}$	< 0	> 0	> 0	> 0	> 0	> 0	< 0
$\frac{\partial g_r}{\partial x} = \frac{\partial g_s}{\partial x}$	< 0	< 0	< 0	< 0	< 0	< 0	< 0
$\frac{\partial g_A}{\partial x} = \frac{\partial g_B}{\partial x}$	$= 0$	$= 0$	$= 0$	> 0	> 0	> 0	$= 0$

Source: own calculations

The most important conclusions drawn from Table 1 are as follows.

- Higher discount rate leads to lower rate of economic growth. This result might be understood with a following logic. In two identical economies, which are different only in a size of ρ, the one with greater discount rate has lower rate of growth of production per capita and natural resources are used in a more intensive way. To maximize utility this economy switch production from future to present moments, which decreases rate of capital accumulation. Size of ρ has no effect on rate of growth of technological progress.
- Increase in σ or κ leads to increase in g_y and more intense use of natural resources (lower g_r). It also has no effect on rate of technological progress.
- Whenever γ, n or μ is higher it is connected to higher rate of growth of A and production per capita, higher intensity of use of natural resources and higher rate of growth of technological opportunities.
- θ represents tendency of consumers to smooth (less volatile) path of consumption in time. Higher θ implies lower g_y, in the limit, when $\theta \to +\infty$, g_y reaches zero.

Performed analysis leads to conclusion, that technological progress in general leads to more intensive extraction of natural resources – higher rate of growth of A is related to higher (in absolute value) g_r. This interesting impact of technological

[2] $-\rho + n + (1 - \theta)g_y$ is equal to rate of use of natural resources (g_R), so it has to be negative.

[3] Available upon request from the author.

progress on natural resource use is due to substitutability of natural resources and physical capital – it is optimal to extract natural resources as soon as possible to produce enough energy to power more physical capital and use it in production. This obviously increases level of production and level of investments, which leads to higher stock of K. Higher level of physical capital substitute natural resources in production and allow to entirely exploit them sooner.

SUMMARY

Concept of technological opportunities, introduced by [Olsson 2000], opens many interesting directions of research in theory of economic growth. Obviously, this idea leads to mathematically more complex endogenous growth models, much harder in analysis, but with interesting consequences. The first attempt of formulating Olsson's theory in economic growth model, [Lundström 2003], cannot be treated as a good example, because in modeled economy natural resources were not a production factor, but source of all income. On the other hand, [Growiec, Schumacher 2013] model includes concept of technological opportunities, but without natural resources.

An attempt to modeling an economy with natural resources in an economic growth model taken in this paper should not be treated as final one. Substitutability between physical capital and natural resources is a flaw of proposed model. In further research this substitutability should be replaced by complementarity between physical capital and natural resources or (as in some papers) by treating natural resource as a factor of production of physical capital.

REFERENCES

Growiec J., Schumacher I. (2013) Technological opportunity, long-run growth, and convergence, Oxford Economic Papers, 65(2), pp. 323-351.

Kuhn, T. S. (2012) The structure of scientific revolutions, University of Chicago press.

Lucas R. (1988) On the Mechanics of Economic Development, Journal of Monetary Economics, vol. 22, no.1 (July), pp. 3 - 42.

Lundström S. (2003) Technological Opportunities and Growth in the Natural Resource Sector, Working Papers in Economics, (116).

Olsson O. (2000) Knowledge as a set in idea space: An epistemological view on growth. Journal of Economic Growth, 5(3), pp. 253-275.

Olsson O. (2005) Technological opportunity and growth. Journal of Economic Growth, 10(1), pp. 31-53

Romer P. (1990) Endogenous technological change, Journal of Political Economy, vol. 98, no 5, pp. S71 - S102.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 78 – 88

TECHNICAL EFFICIENCY MEASUREMENT OF DAIRY FARMS IN POLAND: AN APPLICATION OF BAYESIAN VED MODEL[1]

Jerzy Marzec
Department of Econometrics and Operations Research
Cracow University of Economics
e-mail: marzecj@uek.krakow.pl
Andrzej Pisulewski
Ph.D. student at the Faculty of Economics and International Relations
Cracow University of Economics
e-mail: andrzej.pisulewski@gmail.com

Abstract: The purpose of this paper is to measure the technical efficiency of Polish dairy farms using a Bayesian Varying Efficiency Distribution (VED) model. In particular, the paper presents the design and assumptions of frontier stochastic production function for panel data. Furthermore, it specifies the microeconomic production function based on panel data, derived from the Polish FADN (Farm Accountancy Data Network). The main part of the paper presents key findings which form the basis of understanding the technological characteristics and average efficiency of Polish dairy farms. Moreover, the exogenous variables affecting the level of average farm efficiency are identified. They are the source of significant differences in levels of efficiency of dairy farmers surveyed.

Keywords: stochastic frontier models, Bayesian VED model, technical efficiency, dairy farms

INTRODUCTION

In 2010, Poland was the 12th-largest milk producer in the world and the 4th in the European Union [Statistical Yearbook of Agriculture 2012]. Although the number of dairy farms in Poland decreased from 874,000 in 2002 to 424,000 in 2010 [Raport z wyników, Powszechny Spis Rolny 2010], the production of milk increased from 11.5 billion liters in

[1] This research was supported by a grant from the National Science Centre (NCN, Poland, decision no. 2013/09/N/HS4/03833), which was awarded to the second author.

2004 to 12.1 billion liters in 2011 [Statistical Yearbook of Agriculture 2012]. However, the share of milk production in gross agricultural output (in current prices) decreased in 2011 compared with 2005: from 17.1% to 14.9% [Statistical Yearbook of Agriculture 2012]. These changes in the Polish dairy sector motivate our study, which analyses the technical efficiency of Polish farms, post-accession to the EU.

In previous studies of technical efficiency on different types of farms, inefficiency proved to be an inherent element in farming. The consequence of inefficiency is higher production costs, which of course negatively affect competitiveness. Therefore, the study of the causes of inefficiency has proven to be an important one [Alvarez and Arias 2004]. We can distinguish the following determinants of inefficiency among those commonly analysed: subsidies, the size of the land, the farm's economic size and its degree of specialization.

Since the milk market in the European Union is strictly regulated, the influence of Common Agricultural Policy (CAP) subsidies on the performance of dairy farms is relevant to a discussion of technical efficiency. The various CAP initiatives influence the farmer's optimal decisions through different mechanisms. Therefore, the impact of subsidies on the farms economic performance is an interesting question for policy makers who want to evaluate the effects of their decisions [Zhu et al. 2008].

Moreover, the empirical studies of technical efficiency of farms in Central and Eastern European (CEE) countries have caused disagreements about the relationship between farm size and efficiency. The commonly used measure of farm size is land area, but as this can be inappropriate for intensive livestock production, the weighting approach (e.g. European Size Unit – ESU) seems more appropriate. However, it is rarely used in efficiency studies [Gorton and Davidova 2004].

Another factor influencing the inefficiency of farms, which many authors have investigated in the studies on farming efficiency in CEE countries, is the degree of specialization in farm production.

The aim of this study is to perform a quantitative analysis of technical efficiency on Polish dairy farms and to identify the determinants of inefficiency using Bayesian VED model.

PREVIOUS STUDIES CONCERNING TECHNICAL EFFICIENCY OF LIVESTOCK AND DAIRY FARMS IN POLAND

The review of papers analysing efficiency of agriculture in Central and Eastern Europe is found in Gorton and Davidova [2004]. Papers on the efficiency of Polish farms include Munroe [2001] and Latruffe et al. [2004]. Research which in particular assesses the technical efficiency of Polish dairy farms can be found in Brümmer et al. [2002]. Analysis which presents the influence of CAP subsidies on dairy farms is found in Zhu et al.. [2008], Latruffe et al. [2012]. The relationship between technical efficiency and farm size is investigated in many papers [see review Alvarez and Arias, 2004]. The results of this relationship in Polish agriculture are presented by Van Zyl et al. [1996], Munroe [2001] and Latruffe et al. [2005]. Studies on how specialization influences technical efficiency in CEE countries are found in Brümmer [2001], Mathijs and Vranken [2001], Bojenc and Latruffe [2009].

In Polish scientific papers on efficiency analysis in agriculture, the dominant methodology of research was Data Envelopment Analysis (DEA) see, e.g., Świtłyk [1999, 2011], Rusielik [2002], Ziółkowska [2008], Rusielik and Świtłyk [2009], Kagan et al. [2010], Czyżewski and Smędzik [2010], Smędzik [2010, 2012], Bezat [2011]. The studies in which parametric approach i.e. stochastic frontiers models was used are for example: Kulawik [2008], Czekaj [2008], Czekaj et al. [2009], Rusielik and Świtłyk [2012]. This paper, on the other hand, utilizes the Bayesian approach to technical efficiency measurement to evaluate the efficiency of Polish dairy farms. The advantages of this approach are often discussed in the literature.

MODEL SPECIFICATION

There is a long history of economists quantifying inefficiency measures in production. Farrell [1957] was the first to measure productive efficiency. Presently, there are two main approaches which identify inefficiency as a deviation from a production or cost frontier: the parametric stochastic approach (Stochastic Frontier Analysis - SFA) and the nonparametric deterministic approach (DEA) [Prędki 2003]. Stochastic frontier models were simultaneously introduced by Aigner, Lovell and Schmidt [Aigner at al. 1977], as well as Meeusen and van den Broeck [1977]. In brief, SFA is based on an econometric model, which uses a conventional production function with two independent random disturbances, a symmetric around zero pure stochastic noise and a nonnegative error term representing inefficiency.

The model for a farm i (i=1,…,N) in period t (t=1,…,T) is written as follows:

$$\ln y_{it} = h(\ln x_{it}, \beta) + v_{it} + z_i \tag{1}$$

where:

y_{it} is the observed output quantity,

h is the production function,

x_{it} is the vector of the input quantities used by the farm,

β is the vector of parameters to be estimated,

v_{it} is a random error term representing random shocks ($v_{it} \sim N(0, \sigma_v^2)$).

Technical inefficiency relative to the stochastic production frontier is represented by the one-sided error component $z_i \geq 0$. Several distributions have been proposed for z_i, with the most common being the half normal, truncated normal or gamma distribution. Conventional assumption is that z_i and v_{it} are distributed independently of each other. Technical efficiency will be measured as $r_{ti} = \exp(-z_i)$, which is easily quantifiable (0,1]. A higher value for z_i equates to an increase in technical inefficiency. If z_i is zero, the farm is perfectly efficient. In many situations, the researcher is interested in making the inefficiency (a individual specific effect) depends on certain farm characteristics. It can be quite reasonable to assume that groups of similar farms, e.g. defined through their size or other factors, have similar efficiencies. Nevertheless the inefficiency distribution varies between groups.

This paper uses the Varying Efficiency Distribution model (VED) proposed by Koop, Osiewalski and Steel [Koop et al. 1997], which is more flexible than traditional frontier models. One of the advantages of this model is that it allows efficiency to vary while retaining certain individual characteristics of farms. The authors mentioned above propose that z_i represents an exponential distribution with a mean (and standard deviation) λ_i. The mean of z_i can depend on some (m-1) dummy exogenous variables s_{ij} ($j = 2,...,m$) explaining possible systematic differences in efficiency levels. The parameterization for the average efficiency takes the form

$$\lambda_i = \prod_{j=1}^{m} \phi_j^{-s_{ij}} \tag{2}$$

where $\phi_j > 0$ is the unknown parameters and, by construction, $s_{i1} \equiv 1$. m = 1 is an important special case and is called the Common Efficiency Distribution (CED) model. The parameter vector ϕ indicates how the mean of the inefficiency distribution changes with the farm characteristics in s. In Bayesian analysis the parameters treated as random variables. This means that the inefficiency of farms are a priori linked through the ϕ.

The estimation of the model in equations (1) and (2) is possible using the maximum likelihood method if the parameters are assumed to be nonrandom constants, of course. However, in practice, this method is hampered by computational difficulties. Most often some non-Bayesian method used a two-step approach. Firstly, the model is estimated without the determinants of efficiency. Afterwards, at the second stage, the efficiency estimates obtained at the first stage were regressed on these farm characteristics. Therefore, this paper employs the Bayesian approach and in particular the Gibbs sampling algorithm for performing Monte Carlo integration (see [Osiewalski, Steel 1998], [Marzec, Osiewalski 2008]). In this context, it is worth mentioning that the continuous variables for s_{ij} ($j > 1$) can be used to explain inefficiency. But it causes some numerical difficulties because it requires a hybrid algorithm that combines Metropolis-Hastings and Gibbs sampling.

It is commonplace in frontier literature to impose regularity conditions drawn from economic theory. This is because the imposition of regularity conditions is relatively simple when employing Bayesian techniques compared to classical estimation. This paper makes use of a translog production function, but with no monotonicity conditions because it was satisfied for the entire sample data.

DATA AND RESULTS

The data used in the present study is taken from Polish Farm Accountancy Data Network[2]. The data covers farms whose main source of revenue in the analysed period came from milk production. The estimation of the Bayesian frontier model in this study involves a balanced panel data from 1,212 Polish dairy farms over the period 2004–2011.

The variables, names and symbols, used to construct output and inputs in the model, are according to European Commission document no. RI/CC 882 Rev.9 "Definitions of Variables Used in FADN Standard Results" dated November 2011. The output (Q)

[2] The authors are very grateful to dr inż. Dariusz Osuch from IERiGŻ, Warsaw, for providing access to the data.

includes production (SE131) and subsidies calculated according to the Polish methodology (SE605PL). Based on the literature and the data available, our empirical model includes the following 4 inputs: fixed capital (K), measured in Polish currency (PLN), labour (L), intermediate consumption, including materials and energy (M), measured in PLN and total utilised agricultural area (A), measured in hectares (SE025). The variable K is the sum of the value of buildings (SE450), machinery (SE455) and breeding livestock (SE460). The variable L is the total labour input, expressed in hours (SE011). Intermediate consumption is the sum of total specific costs (SE281) and total farming overheads (SE336). The construction of output and inputs is slightly different from those proposed in the literature, such as Bezat–Jarzębowska et al. [2012].

The production technology of dairy farms in this study is assumed to be specified by the translog function defined as follows:

$$\begin{aligned} h(x_{it};\beta) &= \beta_0 + \beta_1 \ln K_{it} + \beta_2 \ln L_{it} + \beta_3 \ln M_{it} + \beta_4 \ln A_{it} + \beta_5 \ln K_{it} \ln L_{it} \\ &+ \beta_6 \ln K_{it} \ln M_{it} + \beta_7 \ln K_{it} \ln A_{it} + \beta_8 \ln L_{it} \ln M_{it} + \beta_9 \ln L_{it} \ln A_{it} \\ &+ \beta_{10} \ln M_{it} \ln A_{it} + \beta_{11} \ln^2 K_{it} + \beta_{12} \ln^2 L_{it} + \beta_{13} \ln^2 M_{it} + \beta_{14} \ln^2 A_{it} + \beta_{15} t, \end{aligned} \tag{3}$$

where

t = time trend and symbols K, M, L and A represent the input set, which have been explained above.

The translog belongs to the class of so-called flexible functional forms. In contrast to a Cobb–Douglas production function, where returns to scale are global, translog functional form allows the estimated returns to scale to be different for each observation.

In the present study, there is an additional analysis, the aim of which is to identify exogenous determinants on dairy farm efficiency. The assumed factors determining inefficiency are: farm size measured by land size (classified by UAA – Utilised Agricultural Area) and the economic size (SE005). Both are expressed on a binary scale, i.e., if SE005 > 3 (on a 6-point ordinal scale), the dummy variable takes the value of 1, and zero otherwise. In addition, during the eight-year period cited above, farms earned as much as 39% of their primary income from agricultural activities other than the production of milk. Thus, the next additional determinant is the dummy variable indicating strict specialization. Indicated by an abbreviation, "specialization" is set to one when milk production is the main source of farm income in each of the eight years, or set to zero otherwise. Furthermore, as eighty-two percent (82%) of dairy farmers do not receive subsidies, this fact is reflected in this analysis: the variable "coupled subsidies" equals 1 if total subsidies on livestock (SE615) are greater than zero. These four dummy variables reflect the potential variation in farm efficiency.

Table 1, below, shows the mean values of the samples of individual variables. The average annual milk production is 187,000 PLN. Other characteristics show some features of the variables, including the skewed distribution in the population.

Table 1. Basic characteristics of sample farms: descriptive statistics

Variable	Mean	StDev	Percentile			Min	Max
			25th	50th	75th		
Output (‘000 PLN)	187	187	83	134	224	7	3486
Capital (‘000 PLN)	407	368	189	301	497	17	3750
Labour (in hours)	4 446	1 530	3 652	4 378	4 994	484	29 572
Intermediate consumption (‘000 PLN)	89	96	39	62	105	3	2176
Utilised agricultural area (in ha)	29	29	16	22	35	3	699

Source: own calculations based on data from Polish Farm Accountancy Data Network
Note: min and max denote minimum and maximum, respectively

Table 2 shows posterior mean parameter estimates of the Bayesian estimation of the stochastic frontier model.

Table 2. Posterior results of estimation the translog production frontiers (posterior means and standard deviation)

Variable	Mean	SD	Variable	Mean	SD
const	5.872	1.549	lnL lnM	0.002	0.026
lnK	-0.221	0.200	lnL lnA	-0.027	0.028
lnL	-0.913	0.297	lnM lnA	-0.042	0.019
lnM	0.786	0.217	$\ln^2$K	0.018	0.010
lnA	0.814	0.269	$\ln^2$L	0.043	0.021
lnK lnL	0.030	0.022	$\ln^2$M	0.009	0.012
lnK lnM	-0.025	0.018	ln2A	0.011	0.014
lnK lnA	0.002	0.017	T	0.025	0.001

Source: own calculations

A test comparing the translog versus Cobb-Douglas (C-D) specifications, with the restriction $\beta_5=\ldots=\beta_{15}=0$, revealed that this restriction (which has the Wald-test-statistic p-value of 1.2×10^{-5}) is rejected by the data. Therefore, the translog function seems a better representation of production technology on dairy farms.

Intermediate consumption (M) has the greatest impact on the volume of production. A 1% increase in the quantity of this factor results in an increase in production of about 0.56% (± 0.01%), ceteris paribus. A 1% increase in the utilised agricultural area (A) results in an increase in production quantity of about 0.22% (± 0.01%), ceteris paribus. The elasticity of buildings, machinery and breeding livestock (C) is 0.21%, so the impact of this factor is slightly smaller than utilized agricultural area factor. The smallest change (0.11%) in the quantity of production is effected by 1% growth in hours spent on farming.

Table 3. Posterior means and standard deviations of elasticities for a sample mean (variables on a logarithmic scale)

Variables	Average value	Mean	SD[3]
Capital (C)	305,300 PLN	0.21*	0.01
Labour (L)	4,246 h	0.11*	0.01
Intermediate consumption (M)	64,700 PLN	0.56*	0.01
Utilised agricultural area (A)	23 ha	0.22*	0.01
Returns to scale (RTS)	-	1.10*	0.01

Source: own calculations; Note: * significance at 1% levels

The coefficient on the time-trend variables in equation (1) is interpreted as a measure of pure technical change. The estimate of the parameter suggests that farms achieve an increase in production due to technical change. The growth rate in production over the past eight years has been 2.5% per year.

Another important issue is measuring economies of scale. A typical Polish producer of milk is characterized by increasing returns to scale, which is about 1.1 (±0.01). Almost all of the farms are characterized by increasing returns to scale, which does not exceed 1.3. Thus, a proportionate increase in input observably led to a more than proportionate increase in the production function. The opposite was true in only about 0.43% of the cases observed. Detailed information about the sample is presented in Table 4.

Table 4. Frequency distribution posterior means for RTS

Interval	[0.8; 0.9)	[0.9; 1.0)	[1.0; 1.1)	[1.1; 1.2)	[1.2; 1.3)
Frequency	1	41	4238	5393	23
Structure	≈0.0%	0.4%	43.7%	55.6%	0.2%

Source: own calculations

One of the main objectives of this study is to assess the technical efficiency of the dairy farms surveyed. The average efficiency of dairy farms is 0.86, which means that observed production amounts to 86% of potential output, i.e. the maximum output from the given inputs. The median efficiency score is 0.88 and the standard deviation is 0.05, reflecting also the low dispersion of efficiency scores along the sample. The efficiency level for the least efficient farms is 0.56 (±0.04). More than 35% of dairy farms should have efficiency levels in the range [0.8. 0.9], and almost 41% of farms have efficiency scores greater than 0.9.

In this study, there are four dummy variables to reflect the variation in farm efficiency. Table 5 also reports the posterior parameter estimates for the explanatory variables included in equation (2). A negative sign indicates that this variable has a negative impact on technical efficiency. Only two variables, the specialization and the land size, seem statistically significant. The Wald test indicates the null hypothesis that other dummy variables – "coupled subsidies" and "economic size" – are none-significant, i.e. not rejected (p-value equals 0.83).

[3] Identical values of standard deviation are due to rounding only.

Table 5. Sources of the technical inefficiency - posterior results

Variable (0-1)	Average value of variable	Parameter	Mean	SD
Specialization (s_{i2})	61%	$\ln(\phi_2)$	-0.23*	0.06
Coupled subsidies (s_{i3})	82%	$\ln(\phi_3)$	0.05	0.08
Land size (s_{i4})	24%	$\ln(\phi_4)$	-0.37*	0.08
Economic size (s_{i5})	68%	$\ln(\phi_5)$	-0.01	0.08

Source: own calculations; Note: * significance at 1% levels

SUMMARY AND DISCUSSION

The parameters of the translog production function are in line with the results of Brümmer et al. [2002], who also showed that the intermediate consumption factor had the highest elasticity and the labour factor, the lowest. Similar results for the Cobb – Douglas production function were obtained by Latruffe et al. [2004]. However, the parameters of the translog production function contradicted the results of Bezat–Jarzębowska et al. [2012] for the C–D production function. Furthermore, the results proved that elasticities vary over farms, revealing the Cobb–Douglas specifications to be inadequate.

The rate of technical change in the present study is higher than in Brümmer et al. [2002], who reported a technical regress (nearly 9% p.a.) in the period 1991 – 1994 for dairy farms. The negative rate of technical change was also found over the period 1996 – 2000 by Latruffe et al. [2008].

In the present study the majority of dairy farms were operating under increasing returns to scale. This result is in line with Latruffe et al. [2005] who reported that in 1996, 37% livestock farms had increasing RTS, but in 2000 this number jumped to 64%. The values obtained for RTS are confirmed by the results presented by Bezat–Jarzębowska et al. [2012] for the C-D function. However, Bezat–Jarzębowska et al. [2012] reported also decreasing RTS for a CES production function specification for two-and three factors of production.

The average technical efficiency level (0.86) in the covered period 2004 – 2011 is consistent with the results reported by Latruffe et al. [2004] who reported an average technical efficiency of 0.88 for a livestock farms panel in 2000, while in research conducted by Brümmer et al. [2002] for a sample of dairy farms in the Poznań region, average efficiency over the period 1991 – 1994 was equal to 75%. The mean total technical efficiency obtained by Latruffe at al [2005] using DEA method for a sample of livestock farms in 1996 was 0.85, which decreased to 0.71 in 2000. Because the results vary and our study was done using panel methods of estimation while previous studies used a cross-sectional data, it is not possible to clearly state if the average efficiency level increased after the accession the EU.

The results indicate that more diversified dairy farms are more technically efficient. This finding is confirmed by studies of farms in Slovenia conducted by Brümmer [2001]. However, this finding contrasts with the results reported by Mathijs and Vranken [2001] for dairy farms in Hungary, and Bojenc and Latruffe [2009] for Slovenian farms.

The analysis of the relationship between size and technical efficiency show that large farms (above 20 ha) are less technically efficient than smaller farms. These results are

in line with Van Zyl et al. [1996], Munroe [2001] and Latruffe et al. [2005]. However it contradicts the conclusions drawn by Davidova et al. [2002]. Moreover, these results are not consistent with the previous study for the same sample of dairy farms by Marzec and Pisulewski [2013], but that research was conducted using a simple method.

REFERENCES

Aigner D., Lovell C.A.K., Schmidt P. (1977) Formulation and Estimation of Stochastic Frontier Production Function Models, Journal of Econometrics 6, pp. 21–37.

Alvarez A., Arias C. (2004) Technical Efficiency and Farm Size: a Conditional Analysis, Agricultural Economics 30, pp. 241–250.

Bezat–Jarzębowska A., Rembisz W., Sielska A. (2012) Wybrane postacie analityczne funkcji produkcji w ocenie relacji czynnik – czynnik oraz czynnik – produkt dla gospodarstw rolnych FADN, [w:] „Studia i Monografie" 154, IERiGŻ-PIB, Warszawa.

Bojenc S., Latruffe L. (2009) Determinants of Technical Efficiency of Slovenian Farms, Post – Communist Economies 21, pp. 117–124.

Brümmer B., Glauben T., Thijssen G. (2002) Decomposition of Productivity Growth Using Distance Function: The Case of Dairy Farms in Three European Countries, American Journal of Agricultural Economics 84 (3), pp. 628–644.

Brümmer B. (2001) Estimating Confidence Intervals for Technical Efficiency: the Case of Private Farms in Slovenia, European Review of Agricultural Economics 28, pp. 285–306.

Czekaj T. (2008) Techniczna efektywność gospodarstw rolnych a skłonność do korzystania ze wsparcia inwestycji środkami publicznymi, Zagadnienia Ekonomiki Rolnej nr 3 (316), pp. 31 – 44.

Czekaj T., Ziółkowska J., Kulawik J. (2009) Analiza efektywności ekonomicznej i produktywności [w:] J. Kulawik (red.) Analiza efektywności ekonomicznej i finansowej przedsiębiorstw rolnych powstałych na bazie majątku WRSP, IERiGŻ-PIB, Warszawa, pp. 150 – 256.

Czyżewski A., Smędzik K. (2010) Efektywność techniczna i środowiskowa gospodarstw rolnych w Polsce według ich typów i klas wielkości w latach 2006-2008, Roczniki Nauk Rolniczych, Seria G, T. 97, z. 3, pp. 61 – 71

Davidova S., Gorton M., Ratinger T., Zawalinska K., Iraizoz B., Kovacs B., Mizo T. (2002) An Analysis of Competitiveness at the Farm Level in the CEEs, Joint Research Project IDARA, Working Paper 2/11.

Farell J. (1957) The Measurement of Productive Efficiency, Journal of the Royal Statistical Society, Series A, Vol. 120 (3), pp. 253–290.

Gorton M., Davidova S. (2004) Farm Productivity and Efficiency in the CEE Applicant Countries: A Synthesis of Results, Agricultural Economics 30, pp. 1–16.

Kagan A., Góral J., Kulawik J. (2010) Efektywność techniczna przy zastosowaniu metody DEA [w:] Sytuacja produkcyjna, efektywność finansowa i techniczna gospodarstw powstałych w oparciu o mienie byłych państwowych przedsiębiorstw gospodarki rolnej, IERiGŻ, Warszawa, pp. 180 - 239.

Koop G., Osiewalski J., Steel M.F.J. (1997) Bayesian Efficiency Analysis Through Individual Effects: Hospital Cost Frontiers, Journal of Econometrics 76, pp. 77–105.

Koop G., Osiewalski J., Steel M.F.J. (1999) The Components of Output Growth: A Stochastic Frontier Analysis, Oxford Bulletin of Economics and Statistics 61 (4), pp. 455 – 487.

Kulawik J. (red.) (2008) Analiza efektywności ekonomicznej i finansowej przedsiębiorstw rolnych powstałych na bazie majątku WRSP, IERiGŻ-PIB, Warszawa.

Latruffe L., Balcombe K., Davidova S. Zawalinska K. (2004) Determinants of Technical Efficiency of Crop and Livestock Farms in Poland, Applied Economics 36, pp. 1255-1263.

Latruffe L., Balcombe K., Davidova S., Zawalinska K. (2005) Technical and Scale Efficiency of Crop and Livestock Farms in Poland: Does Specialization Matter?, Agricultural Economics 32, pp. 281–296.

Latruffe L., Balcombe K., Davidova S. (2008) Productivity Change in Polish Agriculture: an Application of a Bootstrap Procedure to Malmquist Indices, Post-Communist Economies 20 (4), pp. 449–460.

Latruffe L., Bravo – Ureta B. E., Moreira V. H., Desjeux Y., Dupraz P. (2012) Productivity and Subsidies in the European Union: An Analysis for Dairy Farms Using Input Distance Frontiers, Paper prepared for presentation at the International Association of Agricultural Economists (IAAE) Triennial Conference, Foz do Iguaçu, Brazil.

Marzec J., Osiewalski J. (2008) Bayesian Inference on Technology and Cost Efficiency of Bank Branches, Bank i Kredyt 9, pp. 29–43.

Marzec J., Pisulewski A. (2013) Ekonometryczna analiza efektywności technicznej farm mlecznych w Polsce na podstawie danych z lat 2004 – 2011, Roczniki Kolegium Analiz Ekonomicznych nr 30 (red. M. Bernardelli, B. Witkowski), pp. 255–271.

Mathijs E., Vranken L. (2001) Human Capital, Gender and Organisation in Transition Agriculture: Measuring and Explaining Technical Efficiency of Bulgarian and Hungarian Farms, Post – Communist Economies 13, pp. 171–187.

Meeusen W., van den Broeck J. (1977) Efficiency Estimation from Cobb–Douglas Production Function with Composed Error, International Economic review 18, pp. 435–444.

Munroe D. K. (2001) Economic Efficiency in Polish Peasant Farming: An International Perspective, Regional Studies 35, pp. 461–471.

Osiewalski J., Steel M.F.J. (1998) Numerical Tools for the Bayesian Analysis of Stochastic Frontier Models, Journal of Productivity Analysis 10, pp. 103–117.

Prędki A. (2003) Analiza efektywności za pomocą metody DEA: podstawy formalne i ilustracja ekonomiczna, Przegląd Statystyczny (Statistical Review), 50 (1), pp. 87–100.

Raport z wyników, Powszechny Spis Rolny 2010, (2011) Główny Urząd Statystyczny, Warszawa.

Rusielik R. (2002) Pomiar efektywności produkcji mleka z wykorzystaniem metody DEA, Parce Naukowe Akademii Ekonomicznej we Wrocławiu nr 941, t. 2, pp. 286 – 292.

Rusielik R., Świtłyk M. (2009) Zmiany efektywności technicznej rolnictwa w Polsce w latach 1998 – 2006, Roczniki Nauk Rolniczych, Seria G, T. 96, z. 3, pp. 20 – 27.

Rusielik M., Świtłyk M. (2012) Efektywność techniczna produkcji mleka w wybranych europejskich gospodarstwach w latach 2008 – 2010, Roczniki Nauk Rolniczych, Seria G, T. 99, z. 1, pp. 88 – 99.

Smędzik K. (2012) Czynniki wpływające na efektywność techniczną gospodarstw rolnych osób fizycznych, wyspecjalizowanych w produkcji zwierzęcej (na przykładzie gospodarstw Polskiego FADN z powiatu gostyńskiego), Journal of Agribusiness and Rural Development, nr 3 (25), pp. 241-250.

Smędzik K. (2010) Problem skali produkcji w różnych typach indywidualnych gospodarstw rolnych w Polsce z zastosowaniem modeli DEA, Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu, nr XII/3, pp. 343-348.

Statistical Yearbook of Agriculture 2012, Central Statistical Office, Warsaw.

Świtłyk M. (1999) Zastosowanie metody DEA do analizy efektywności gospodarstw rolnych, Zagadnienia Ekonomiki Rolnej nr 6, pp. 28 – 41.

Świtłyk M. (2011) Efektywność polskiego rolnictwa w latach 1998 – 2009, Zagadnienia Ekonomiki Rolnej nr 4, pp. 59 – 75.

van Zyl J., Miller B. R., Parker A. (1996) Agrarian Structure in Poland. The Myth of Large-Farm Superiority, Policy Research Working Paper No. 1596, The World Bank, Washington, DC.

Zhu X., Demeter R. M., Oude Lansink A. (2008) Competitiveness of Dairy Farms in Three Countries: the Role of CAP Subsidies, Paper presented at the 12th European Association of Agricultural Economists (EAAE) Congress, Gent, Belgium, 27-30 August.

Ziółkowska J. (2008) Efektywność techniczna w gospodarstwach wielkotowarowych, „Studia i Monografie" nr 140, Instytut Ekonomiki Rolnictwa i Gospodarki Żywnościowej, Warszawa.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 89 – 98

EVALUATION OF VOIVODESHIPS DIVERSIFICATION IN POLAND ACCORDING TO TRANSPORT INFRASTRUCTURE INDICATORS

Aldona Migała–Warchoł, Marek Sobolewski
Department of Quantitative Methods, Rzeszów University of Technology
e-mail: amigala@prz.edu.pl, msobolew@prz.edu.pl

Abstract: The aim of this paper is to determine the level of diversification of Polish voivodeships in terms of the selected indicators of transport infrastructure. The data collected from the Local Data Bank of Central Statistical Office will be used in the paper. By means of some methods of linear ordering and cluster analysis the ordering and classification of the Polish voivodeships will be carried out. The obtained results will allow to make an evaluation of Polish voivodeships in terms of the level of development of transport infrastructure.

Keywords: transport, multivariate statistical analysis, synthetic measures, cluster analysis

INTRODUCTION

The concept of infrastructure, despite the fact that it has been operating in the Polish language for many years, there is still no generally accepted definition, and thus it is not clearly understood. The term *infrastructure* derives from English and it means "the foundation of base i.e. the necessary basis for the economy."

In Polish literature the concept of infrastructure in the most general terms is defined as the basic facilities and institutions necessary for the proper functioning of the economy. These devices include man-made, permanently located, line and point objects for public use, which are the foundation of socio-economic development, in view of their functions to move people and goods (transport), news (communications), electricity (energy) and water (water management) [Rydzkowski 2011].

The purpose of infrastructure is to provide the basic conditions of development of the socio-economic system as a whole and the rest elements

of the economy. Accordingly, the transport infrastructure includes man-made, permanently sited basic facilities of roads (linear infrastructure) and transport points (point infrastructure). It affects the economy and society by creating favorable conditions for the movement of people and goods (freight) in the direct and indirect form. From an economic point of view the most important feature of the transport infrastructure is the public nature of its services. There are of course exceptions, i.e. the transport infrastructure services as private goods. Transport infrastructure is formed mainly by three basic groups:

• roads of all modes of transport;
• transport points (airports, ports, etc.);
• ancillary equipment used for the direct control of the roads and transport points [Gołembska 2008].

The aim of this paper is to present a comprehensive analysis of the level of transport infrastructure in Polish voivodeships in 2011[1]. In the first section the criteria for the selection of variables describing the transport infrastructure were described in detail. Pre-selected set of variables was verified statistically, then the synthesis of the information contained therein, by means of taxonomic methods - linear ordering and clustering was done. The obtained results were subjected to the interpretation, and then there were examined the correlations of the level of development of transport infrastructure with the selected macroeconomic indicators.

SUBSTANTIVE SELECTION CRITERIA OF DIAGNOSTIC VARIABLES

The selection of diagnostic variables that describe directly immeasurable social and economic phenomenon, is a point of reference adopted by the researcher. In case of the voivodeships ranking according to the level of development of transport infrastructure an important element is a matter of analyses recipients – either it is an analysis prepared for the individual needs of the residents, or the possibility of development of large manufacturing companies [*Atrakcyjność Inwestycyjna…* 2012].

In the first case, the more important is the functioning of urban and public transport and the quality of local roads, whereas from the point of view of large companies, it is important a network of national roads and the location of the voivodeship with respect to the state border, sea ports or airports. Evaluation of diagnostic variables will be different from the point of view of the citizens who take long trips (and they care about good communication between regions and

[1]It is worth to appeal to a broader context and remind that according to comparative international studies, Poland is at a distant 74. place due to the operation of railway transport, and in the case of road infrastructure is the 134. place (!) [*The Global Competitiveness Report…* 2011].

states) than the use of the transport network within the region. In this paper, an attempt was done to get such diagnostic variables, which would describe the level of development of transport infrastructure in the most comprehensive way, and therefore there were both variables of "global" character (a network of expressways and highways) and "local" one, such as the length of cycle paths or the use of urban transport.

Pre-selection, based on a review of the literature [Atrakcyjność inwestycyjna ... 2012; Wierzbińska, Chudy 2011], led to the creation of a list of 17 diagnostic variables that are listed in the table. The values of some of them are taken directly from the publication of the CSO, some of them are the result of simple calculations of raw data allowing to determine the intensity ratios of certain phenomena. For each variable it was indicated, whether in terms of rankings created, they will be treated as a stimulant or destimulant. The vast majority of the features are stimulants. There is presented the list of diagnostic variables together with the selected descriptive statistics: median, minimum, maximum and coefficient of variation in the first table.

Table 1. List of diagnostic variables together with the selected descriptive statistics

Diagnostic features	**Me**	**min**	**max**	***V***
Traffic at airports [thous. people] ↑	7,6	0,0	88,0	136%
Transportation of cargo at airports [t] ↑	106	0	50 951	227%
Travel time to the nearest sea port [h][a)] ↓	4:59	1:00	8:59	52%
Bridges and overpasses fixed at 100 km of roads ↑	6,6	3,8	20,5	55%
Tunnels and subways, the 1000 km of roads ↑	0,8	0,4	5,6	100%
Hard-surface roads [in km per 100 km^2] ↑	89,8	53,3	179,6	37%
Roads with improved surface [in % of all roads] ↑	91,5	85,3	95,5	3%
Motorways and highways [in km per 100 km^2] ↑	0,67	0,00	1,98	79%
Bicycle paths for 10 thous. km^2 [km] ↑	179,3	83,3	454,1	52%
The urban population served by public transport [%] ↑	75,6	43,9	93,6	19%
Places in urban vehicles per thous. inhabitants ↑	89,8	67,1	147,5	22%
Cars for 1 thous. inhabitants ↑	462,5	409,5	530,8	8%
Lorries for 1 thous. inhabitants ↑	75,2	66,7	108,0	15%
Railway lines per 100 km^2 [km] ↑	6,5	3,8	17,4	46%
Number of passengers of the regional transport per 1 inhabitant ↑	3,1	0,8	20,5	111%
Interregional transport passengers per 1 inhabitant ↑	1,3	0,8	2,4	35%
The share of electrified lines [%] ↑	55,0	29,1	91,1	36%

↑ – stimulant ↓ – destimulant

[a)] Information on road transport was taken from the "map of road connections" available on the website of the General Directorate for National Roads and Motorways (www.gddkia.gov.pl)
Source: own studies

Diagnostic variables concern the existing infrastructure and its actual use in the field of aviation, maritime, road and rail transport. In view of the crucial importance of road transport, this mode of transport is described by the most variables, and the information about the possible use of maritime transport was described by contrast by only one variable which contains information about the average time of travel from the voivodeship to the nearest major sea port (Szczecin or Gdańsk).

It should be noted that some of the features that contain information about the state of the infrastructure - such as rail - can lead to erroneous conclusions. For example, the region with the highest density of railway network is in Śląskie Voivodeship but the share of rail passenger transport is much greater in the Pomorskie and Mazowieckie Voivodeships.

STATISTICAL METHODS

Statistical verification of the diagnostic value of the pre-selected diagnostic variables consisted of two phases:

- analysis of variance (based on classical coefficient of variation);
- correlation analysis (it was used the Potential Information Analysis – PIA).

In the first case, for each diagnostic variable there was determined the coefficient of variation, but for further analysis the characteristics for which the value exceeded 10% were automatically enrolled. When the coefficient of variation did not exceed 10%, it was made another analysis of the substantive value of the given variable, in terms of accuracy of the description of transport for regions and then it was taken the decision whether to exclude or allow further analysis.

PIA method consists in searching the variables which are the most strongly correlated with the others and removing from the analysis those for which the correlation coefficient exceeds a predetermined threshold (a detailed description of these and other alternative methods of reducing the set of diagnostic variables can be found in the literature [Grabiński et al. 1989]). For the study it was adopted a fairly strict level of the correlation coefficient $R = 0.80$. In the case of variables with a high degree of correlation, the decision on their exclusion from the analysis was preceded by a reassessment of their substantive meaning.

The voivodeships ranking was made by means of unsupervised linear ordering methods, and to standardize the data it was applied the zero unitarisation method [Kukuła 2000]. In order to group the voivodeships according to indicators of the transport development it was used the hierarchical clustering procedure – the Ward's method. Both methods are among the most popular procedures for taxonomic studies used in the socio-economic conditions.

Both while doing the ranking, as well as cluster analysis, there was not used the weighing of diagnostic features. Grouping procedure, using the Ward's method, was based on the feature values that were subjected to classic standardization. The

differentiation between objects and clusters, according to the specific of the Ward's method was measured by the square of the Euclidean distance.

To examine the compatibility of the results with other selected indicators of the level of socio-economic development, the correlation analysis methods were used. To avoid distortions associated with the presence of outliers, it was applied nonparametric Spearman rank correlation coefficient. There were also provided the value of test probability *p* which allows to assess whether the tested relationship is not just about "accidental" relationship of these two traits.

The calculations were performed by using *STATISTICA.* There were applied the standard procedures of the program, supplemented by the authors' extensions created in *STATISTICA Visual Basic* language with which it is achieved the complete taxonomic analysis report directly in *WORD.* Minimized in this way the calculation time was used to deepen the interpretation of the results.

VERIFICATION OF DIAGNOSTIC VARIABLES

The values of the variation coefficient *(V)* is given in Table 1. For two characteristics: the share of roads with improved surface and the ratio of passenger cars per 1 thousand inhabitants low coefficient of variation (3 and 8%) constitute the rationale for excluding them from the further analysis.

As it is debatable to what extent the level of motorization (especially when it comes to the number of cars) is associated with the development of transport infrastructure, it was decided to omit this feature in further considerations. The more complex and interesting issue is related with the fraction of roads with improved surfaces in the total length of roads in the region. Although this feature is characterized by low volatility, but it should be noted that this rate of structure can be defined as a fraction of roads with "unimproved" surface, which will change the average level, but not the standard deviation. This reflects some kind of methodological "weakness" of uncritical use of the variation coefficient to reduce the set of diagnostic variables. As for the indicator of the "unimproved" roads the variation coefficient is more than 30%, thus taking into account the substantive importance of this indicator, particularly in terms of accessibility of rural areas, it was decided to leave it for further analysis.

In the second step, for the reduced by one characteristic a set of diagnostic variables there was conducted an analysis of information potential. There were distinguished the characteristics the most correlated with the others (so-called central variables) and it was discussed the exclusion of those which were correlated with them to an extent exceeding a certain threshold value, which is assumed to be $R^* = 0.8$ (so-called satellite variables).

The analysis led to the finding of two central characteristics and two satellite ones that should be excluded from further consideration. The first pair are the rate of passenger air transport (central variable) and the transport of goods at airports (satellite variable). Since these two variables describe the same mode

of transport, it seems that without worry to omit some relevant information the satellite variable (transport of goods by air) can be excluded from a further analysis.

We have a different situation in the case of the second pair of variables: density of the rail network per 100 km^2 (central variable) and the ratio of the density of expressways and motorways (satellite variable). As they describe the state of the infrastructure of two different modes of transport, it was decided to leave both of them for further analysis.

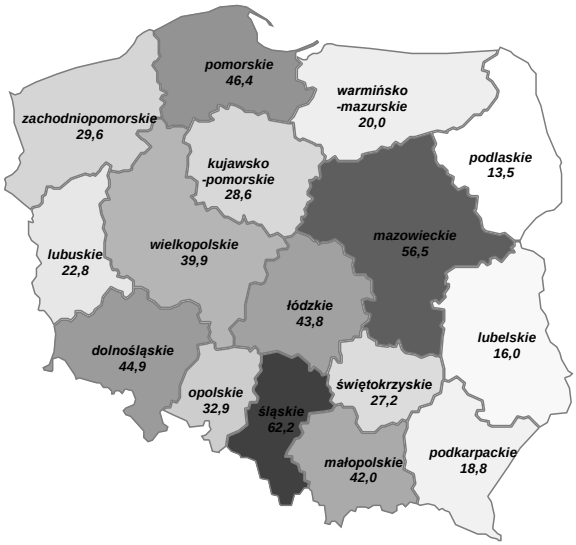
VOIVODESHIPS RANKING ACCORDING TO THE DEVELOPMENT OF TRANSPORT INFRASTRUCTURE

Summing up the considerations discussed in the previous paragraph, from the initial list of variables (Table 1), after the statistical verification there were removed only two features: the motorization indicator and cargo air transport. Thus, the taxonomic analyses will be conducted based on the values of 15 diagnostic variables.

By using the method of linear ordering it was determined the synthetic indicator, which made it possible to prioritize regions because of the level of transport infrastructure (Figure 1).

Śląskie Voivodeship has the highest value of synthetic measure of transport development. The second place in the ranking took Mazowieckie Voivodeship. Another region (Pomorze, Dolny Śląsk) have the value of the synthetic measure much lower than the two leaders indicated above. By far the worst assessment of development and the use of transport infrastructure in the region is characterized by Eastern Poland (Warmia and Mazury, Podkarpackie, Lubelskie and at the very end – Podlaskie Voivodeship).

Figure 1. Ranking of voivodeships by the development of transport infrastructure in 2011



Source: own studies

From a practical point of view, it is interesting how much the level of development of transport is associated with other indicators of socio-economic development. As it results from conducted analysis of correlation (Table 2), a voivodeship with a high degree of development of transport infrastructure has a lower unemployment rate, higher average salary and industrial production index. These correlations are statistically significant and have a relatively high strength. Only the annual growth rate of GDP is correlated to a weaker extent with the level of transport development.

Table 2. Relationship of synthetic measure of transport infrastructure with some selected indicators of socio-economic development

Socio-economic development indicator	**Synthetic measure of transport infrastructure**
Unemployment rate	R = -0,70 (p = 0,0027**)
Average salary (gross)	R = 0,79 (p = 0,0003***)
Sold industry production per one inhabitant	R = 0,87 (p = 0,0000***)
GDP dynamics	R = 0,48 (p = 0,0596)

R – Spearman's rank correlation coefficient, p – test probability value

Source: own studies

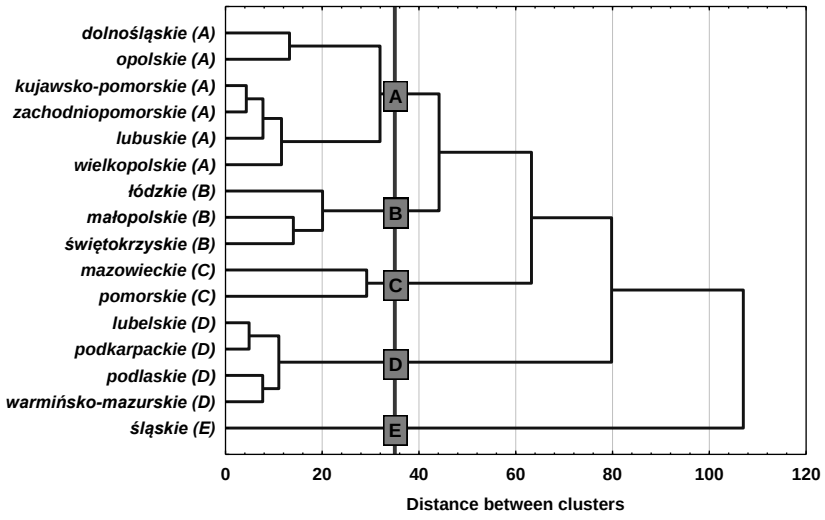
Of course, the analysis of the relation of transport infrastructure development measure with other indicators should be broadened, taking also into

account the dynamics of all the features in the longer term. On the basis of connections to a single, 2011 year, it is impossible to predict the direction of the reason-result relation between the transport level and other indicators of economic development.

VOIVODESHIPS GROUPING ACCORDING TO THE LEVEL OF TRANSPORT INFRASTRUCTURE DEVELOPMENT

As the result of the clustering by the Ward's method the dendrogram was developed which presents the various stages of the agglomeration (Fig. 2). Based on the analysis of the clustering process, taking into account the substantive issues, it was decided to divide the Polish voivodeships into five clusters. The elements of particular groups (designated by the letters A-E) can be read from the dendrogram.

Figure 2. Results of cluster analysis by the Ward's method



Source: own studies

Table 3 shows the group average indicators of all the diagnostic features created for each cluster. The values included in the table are quotients of group average and the average for all the regions. If the ratio is greater than 1, one can say that the group is characterized by relatively preferred values of the feature (contrary to destimuli). The values of the group average lower than 1 indicate a weaker position of the region (or vice versa for destimulants).

Table 3. Relation of group means of individual diagnostic variables to the total mean

Diagnostic variables	Indicators of group mean values				
	A	B	C	D	E
People checked in at airports	0,57↓↓	0,94	3,84↑↑	0,16↓↓	1,47↑↑
Time travel to the seaport	0,75↑	1,23↓	0,60↑↑	1,31↓	1,38↓↓
Bridges and overpasses	1,08	1,05	0,72↓	0,77↓	1,81↑↑
Tunnels and subways	0,73↓	0,87↓	1,49↑↑	0,45↓↓	4,24↑↑
Hard-surface roads	0,86↓	1,36↑	0,88↓	0,76↓	1,92↑↑
Unimproved surface roads	1,08	1,30↓	0,82↑	0,78↑	0,86↑
Express roads and highways	1,40↑↑	0,80↓	0,81↓	0,12↓↓	3,08↑↑
Bicycle paths	1,19	0,64↓↓	1,27↑	0,50↓↓	2,38↑↑
Population of cities with public transport	0,85↓	1,16↑	1,13↑	0,97	1,29↑
Places in urban vehicles	1,02	1,11↑	1,28↑	0,80↓	0,84↓
Lorries per 1 thousand inhabitants	0,99	1,09	1,22↑	0,87↓	0,88↓
Density of railways	1,06	0,93	0,83↓	0,67↓	2,50↑↑
Passenger traffic in the region	0,89	0,34↓↓	3,62↑↑	0,43↓↓	0,67↓
Passenger traffic between regions	1,06	1,22↑	1,19↑	0,70↓	0,78↓
The share of electrified lines	0,92	1,45↑↑	1,06	0,64↓	1,44↑↑

Average deviation of the group mean from the total mean:
↑ - favorable
↓ - unfavorable
(the groups which are distinguished by low or high indicators of group means are marked with different shades of gray)

Source: own studies

The table is constructed in such a way as to maximally facilitate the results interpretation. The exact values of group means are also illustrated by the marks.

CONCLUSIONS

In the paper it was made an analysis of the spatial differentiation of Polish voivodeships in respect to quality of transport infrastructure. The leader turned out to be Śląskie Voivodeship, which was slightly ahead of Mazowieckie Voivodeship. The voivodeships with the lowest development level of transport infrastructure were Podlaskie and then Lubelskie and Podkarpackie Voivodeships. The lowest level of transport infrastructure are characterized by far the voivodeships of the Eastern Poland.

A natural extension of the presented in the analyses will be taking into account the time aspect. Based on these same diagnostic variables there will be created the time-space ranking of the level of infrastructure development, which will help to determine the pace of development of the various regions. Particularly

valuable is linking the changes in the level of development of transport infrastructure with changes of the selected macroeconomic indicators.

REFERENCES

Investment attractiveness of Polish regions and sub-regions 2012 (2012), the Institute for Market Economy, Gdańsk.

Coyle J. (2002), Zarządzanie logistyczne, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Gołembska E. (2008), Kompendium wiedzy o logistyce, PWN, Warszawa.

Grabiński T. (1992), Metody taksonometrii, AE Kraków.

Grabiński T., Wydymus S., Zeliaś A. (1989), Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych, PWN, Warszawa.

Kukuła K. (2000), Metoda unitaryzacji zerowanej, PWN, Warszawa.

Nowak E. (1990), Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych, Państwowe Wydaw. Ekonomiczne, Warszawa.

Ostasiewicz W. (1998), Statystyczne metody analizy danych, AE Wrocław.

Rydzkowski W. (2011), Usługi logistyczne, Instytut Logistyki i Magazynowania Poznań.

Stanisz A. (2007), Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, StatSoft, Kraków.

The Global Competitiveness Report 2011-2012 (2011), World Economic Forum, Geneva, Switzerland.

Transport. Results of operations in 2011 (2012), Central Statistical Office, Warsaw.

Wierzbińska M., Chudy-Laskowska K. (2011), Zróżnicowanie województw pod względem infrastruktury transportowej w Polsce – wyniki badań, Zeszyty Naukowe Politechniki Rzeszowskiej, nr 282, Ekonomia i Nauki Humanistyczne, z. 18.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 99 – 107

MULTIVARIATE ANALYSIS OF HEALTHCARE SYSTEMS IN SELECTED EUROPEAN UNION COUNTRIES. CLUSTER ANALYSIS

Katarzyna Miszczyńska
Department of Public Finance, University of Lodz
e-mail: katarzyna.miszczynska@uni.lodz.pl

Abstract: The European Union as a whole tends to be a highly diversified entity. Therefore it should not be surprising that the main goal of EU authorities is the limitation of inequalities among member countries. The aim of this study is to identify some groups of countries from the EU that show similar characteristics concerning the functioning of healthcare sectors. The selected countries will be analysed through the prism of characteristics connected with the financial and organizational issues. The study will be backed up by a chosen multivariate statistical analysis — cluster analysis.

Keywords: multivariate statistical analysis, cluster analysis, EU healthcare systems

INTRODUCTION

The European Union is undoubtedly an extremely diversified entity. Its countries differ in many ways such as language, culture and organization. Each country's economy is organized in a slightly different way. Therefore, it should not be surprising that the main goal of the EU authorities is the reduction of inequalities among member countries. However, identification of possible directions of changes can be done only with a prior recognition of those inequalities. The matter of health care has been a subject of research in many research centres all over Europe [Willan and Kowgier 2008; Journard et al. 2010; Hass-Symotiuk 2011; Nojszewska 2011; Knapp et al. 2012; Vanberkel et al. 2012; Jena and Philipson 2013]. The problem of the organization of healthcare systems in European countries is of crucial importance. Despite the fact that all the healthcare systems derive from four basic models, they slightly differ from country to country. The aim of this study is to identify some groups of the EU countries from the EU

that behave similarly or show similar characteristics concerning the performance of healthcare sectors. Selected countries will be analysed through the prism of characteristics connected with the financial and organizational issues. The study in question will be backed up by a multivariate statistical analysis - cluster analysis, which finds similarities of objects (in this study – countries) described as data and puts them into groups. This method has been widely used in the field of healthcare services in many different research centres in the world [Chan et al. 2006; Roy et al. 2009; Liu and Liu 2011]. This analysis will lead to the possibility of creating a classification of the EU countries based on healthcare issues which may constitute valuable information for all managers of healthcare units, not to mention the countries' authorities.

MULTIVARIATE STATISTICAL INFERENCE METHODS – CLUSTER ANALYSIS

Introduction to the method

Numerical techniques for deriving classifications originated largely in natural sciences such as biology in an effort to rid taxonomy of its traditionally subjective nature. Their aim was to provide stable and objective classifications; *objective* in a sense that the analysis of the same set of organisms by the same sequence of numerical methods produces the same classification, *stable* in a sense that the classification remains the same under a wide variety of additions of organisms or of new characteristics describing them. Those numerical methods have been named in many different names depending on the area of application. In that sense numerical taxonomy is generally used in biology; Q analysis is applied in psychology; unsupervised pattern recognition in the artificial intelligence literature; and segmentation in market research. However, nowadays cluster analysis is the most commonly used term of procedures which seek to uncover groups in data. Cluster analysis tends to be the most familiar of all approaches to exploratory multivariate analysis, although it is not always thought of as a multivariate technique parallel to, for example, a principal components analysis. Cluster analysis mimics one of the human mind's fundamental ways of dealing with complicated variability: categorizing, or putting things into groups [Drennan 2010].

Clustering is a statistical tool for those who need to arrange large quantities of multivariate data into natural groups. Methods for clustering items (either observations or variables) depend upon how similar or dissimilar they are to each other. Similar items are treated as a homogeneous group, whereas dissimilar items form additional groups. The great majority of the output of this analysis is visual. The results are displayed as scatterplots, trees, dendrograms, silhouette plots, and heat maps [Izenman 2008].

Cluster analysis is not a statistical test, but a collection of different algorithms that group the objects in focus. Cluster analysis methods are used when there are not any a priori hypotheses, but the research is in exploratory phase [Stanisz 2007].

What is a cluster? Unfortunately there is no universally accepted definition. A cluster is a group of items in which each item is "close" (in some appropriate sense) to a central item of a cluster and that members of different clusters are "far away" from each other [Izenman 2008].

The clustering methods

The clustering solutions are found by applying an algorithm which determines the rules by which observations are aggregated. Algorithms can be classified into two basic groups: hierarchical methods and nonhierarchical methods.

Hierarchical algorithms are the most common methods of cluster analysis. In such algorithms, observations are added to each other one by one in a treelike fashion. As a result of these methods of aggregation the dendrogram is created. These methods do not require a prior application of the number of clusters. [Gatignon 2010]

Hierarchical clustering techniques may be subdivided into agglomerative methods, which proceed by a series of successive fusions of the "n" individuals into groups, and divisive methods, which separate the "n" individuals successively into finer groupings. [Everitt et al. 2011]

The group of hierarchical methods includes e.g. [Everitt et al. 2011]:

- single linkage method - also known as the nearest-neighbour technique; the distance between groups is defined as that of the closest pair of individuals, where only pairs consisting of one individual from each group are considered,
- complete linkage method - also known as furthest neighbour is opposite to single linkage, in a sense that distance between groups is now defined as that of the most distant pair of individuals,
- unweighted pair-group method using arithmetic averages (UPGMA) - the distance between two clusters is the average of the distance between all pairs of individuals that are made up of one individual from each group,
- ward method - the fusion of two clusters is based on the size of an error sum-of-squares criterion. The objective at each stage is to minimize the increase in the total within-cluster error sum of squares, E, given by [Everitt et al. 2011]:

$$E = \sum_{m=1}^{g} E_m \tag{1}$$

where,

$$E_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p_k} \left(x_{ml,k} - \bar{x}_{m.k}\right)^2 \tag{2}$$

$\bar{x}_{m.k} = \left(\frac{1}{n_m}\right) \sum_{l=1}^{n_m} x_{ml,k}$ (the mean of the m_{th} cluster for the k_{th} variable),

$x_{ml,k}$ - being the score on the k_{th} variable (k= 1,…, p) for the l_{th} object (l=1,…,nm) in the m_{th} cluster (m=1,…,g)

Nonhierarchical methods are fast but require providing a prior number of clusters. The choice of number of clusters has a large impact on the quality of the resulting segmentation. Application of too many clusters can cause the situation in which the clusters will be internally homogeneous, however their interpretation will be difficult. On the other hand, the smaller the number of clusters is, the less uniform is the concentration. [Everitt et al. 2011]

Basic steps in cluster analysis

Cluster analysis should begin with the standardization of selected variables. The standardization is carried out using the following formula:

$$\bar{z}_{l,k} = \frac{x_{lk} - \bar{x}_{lk}}{S_k} \tag{3}$$

where,

l – object,

k – variable,

$\bar{x}_{lk}$ –arithmetic average,

S_k – the standard deviation of the variable sample.

Standardization is made to make an objective assessment of similarity, apart from a given scale, in which the individual variables are expressed [Adamowicz and Janulewicz 2012].

The next step of cluster analysis is similarity measurement. Since similarity is fundamental to the definition of a cluster, a measure of the similarity between two patterns drawn from the same feature space is essential to most clustering procedures. Because of the variety of feature types and scales, the distance measure (or measures) must be chosen carefully. It is most common to calculate the dissimilarity between two patterns using a distance measure defined on the feature space [Abonyi and Feil 2007].

The most common metric for continuous features is the Euclidean distance shown in the formula below:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{4}$$

where,

x = $(x_1,\ldots,x_p)$

y = $(y_1,\ldots,y_p)$

The last step to be made, in order to conduct cluster analysis, is the choice of hierarchical or nonhierarchical classification methods presented earlier in the paper.

RESULTS OF THE RESERACH

Empirical data

The conducted research was based on the data from two main resources: the Organisation for Economic Co-operation and Development database and Eurostat database. Due to poor availability of data, the empirical data used in the study apply to the healthcare system in the year 2010. The main objects of the study are countries. In order to carry out this study the cluster analysis method was chosen because it allowed to classify the countries in question according to the chosen criteria.

In order to examine the similarities of the selected countries, and thus to create clusters, there were chosen 21 European Union member countries. At the beginning there was made an assumption to conduct the research on all EU countries, however due to scarcity of data, the sample was limited.

The countries were analysed according to three groups of data: finance, resources and health status. Into those three groups six different variables were assigned. The group gathering data connected with financial issues of countries in question consists of two variables:

- Total health expenditure per capita [in purchasing power parity in US $] – defined as the sum of expenditure on activities that – through application of medical, paramedical, and nursing knowledge and technology – have the goals of e.g.: - promoting health and preventing disease, treatment and reducing premature mortality, providing and administering public health [OECD 2012a]
- Public health expenditure as a percentage of total health expenditure – public health expenditures are health expenditures incurred by public funds. Public funds are state, regional and local government bodies and social security schemes. [OECD 2012a]

To the group of data connected with health care resources there were assigned the following variables:

- Total hospital beds [per 1000 inhabitants] - all hospital beds which are regularly maintained and staffed and immediately available for the care of admitted patients [OECD 2012b]
- Practising physicians [per 1000 inhabitants] – physicians who provide services directly to patients, including e.g.: people who have completed studies in medicine at university level and who are licensed to practice; interns and resident physicians; foreign physicians licensed to practice and actively practising in the country. From this group the following are excluded, e.g.: students who have not yet graduated; dentists and dental surgeons; physicians working in administration, research and in other posts that exclude direct contact with patients [Eurostat 2013]

The last group of data consists of:

- Perceived health status - percentage of the population, aged 15 years old and over who report their health to be 'good' or 'better'. Perceived health status reflects people's overall perception of their health, including both physical and psychological dimensions [OECD 2011]
- Life expectancy at birth - the average number of years that a person at that age can be expected to live, assuming that age-specific mortality levels remain constant. [OECD 2012c]

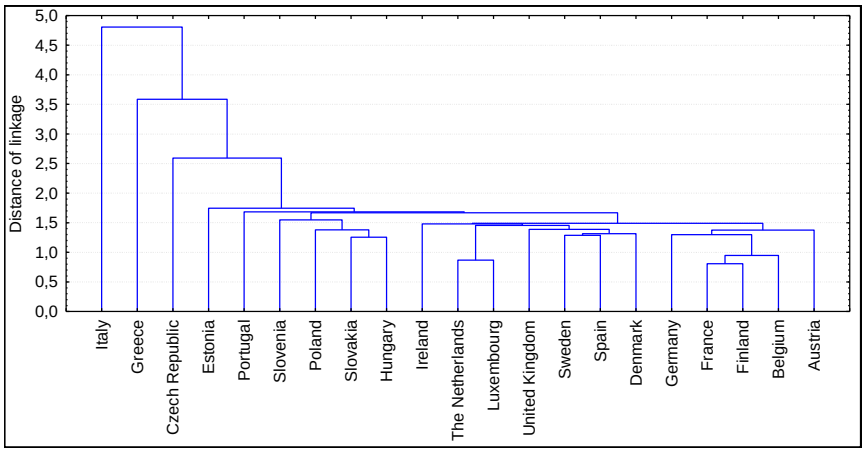
In order to conduct the analysis Statistica PL package was used.

Results

The main goal of the study was to identify some groups of countries from the EU that behave similarly or show similar characteristics concerning the performance of healthcare sectors. The study was conducted by one of agglomeration method of creating clusters – Ward method. However, in order to check the correctness of reasoning and compare the results obtained by different methods, the results of Ward's method were compared with three other agglomeration methods: single linkage, complete linkage and unweighted pair-group method using arithmetic averages (UPGMA). The research was carried out with an assumption of Euclidean distance, as a method of distance calculation.

At first there was applied a single linkage method. As a result there was received a dendrogram presented by Figure 1. This method showed a clear division between countries of eastern and western Europe. There is no clear distinction for many clusters. The first cluster is created by Hungary, Poland and Slovenia and the second one by Austria, Belgium, Finland, France, Spain, Sweden, UK, Luxembourg, the Netherlands and Ireland.

Figure 1. Dendrogram: single linkage method, Euclidean distance

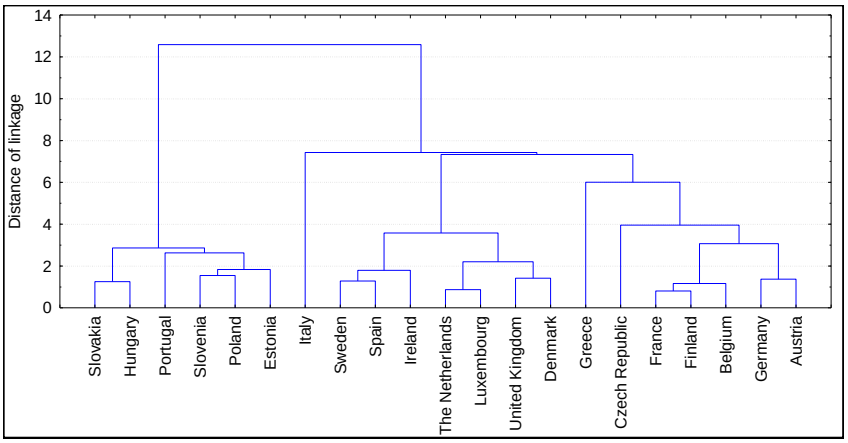


Source: own application in Statistica PL

According to the results of complete linkage method and UPGMA method (unweighted pair-group method using arithmetic averages) the composition of their clusters tends to be similar. In the complete linkage method three clusters were created, however those groups were not fairly homogenous. Poland remained in a group with Slovenia, Slovakia and Hungary. UPGMA method brought more specified results of four clusters. Contrary to the previous method, Austria and Germany created there a separate group. Similarly to the complete linkage method, Poland remained in the same cluster. In both methods Italy, Greece and Czech Republic constituted single-element clusters.

According to the research conducted by Ward's method three clusters can be easily seen. This method gave the most obvious results (see Figure 2), what is justified because this method is considered to be the most effective. The first cluster consists of Austria, Germany, Belgium, France, Finland, and optionally Czech Republic. The second group is created by Denmark, UK, the Netherlands, Sweden, Ireland, Spain and Luxembourg. The last group consists of Poland Estonia, Portugal, Slovenia, Hungary and Slovakia. In this type of segmentation method there is also observed a division into middle - eastern and western Europe countries. Only Portugal makes an exception.

Figure 2. Dendrogram: Ward's method, Euclidean distance



Source: own application in Statistica PL

SUMMARY

Hierarchical methods form the backbone of cluster analysis in practice. They are widely available in software packages and easy to use, although clustering large data sets is time-consuming. Choices that the investigator needs to make refer to the measure of proximity, the clustering method and, often, the number of clusters. The main problem in practice is that no particular clustering method

can be recommended, since methods with favourable mathematical properties (such as single linkage) often do not seem to produce empirically interpretable results. Furthermore, application of the results involves choosing a partition, but hardly do we know the best way of doing that. When a particular partition is required and there is no underlying hierarchy, the nonhierarchical algorithms may be more appropriate [Everitt 2011].

According to the study in question it should be marked that all the clustering methods chosen produced similar but not identical solutions. All the methods applied presented a clear division into countries of eastern/middle-east and western Europe, which is justified from the economic point of view. Italy, Greece and Czech Republic were identified as detached items, creating at the same time, single-itemed clusters. Those three countries were mostly included into the clusters at the last moment, which proved their dissimilarity. In all the cases there is a clear division into groups of eastern / middle-east and western European countries. This shows some still existing differences in healthcare systems in the countries in question. There is still a gap in finance, resources and health status among the EU countries. Poland is always placed in a group with Estonia, Slovenia, Hungary and Slovakia. This fact should be treated as a sign that despite all conducted reforms there is still a substantial gap between Poland and other western countries.

To sum up, it should be said that the greatest limitation of research results was poor availability of data in question. That is why this analysis should be considered as a starting point for further, more detailed analysis.

REFERENCES

Abonyi J., Feil B. (2007) Cluster Analysis for Data Mining and System Indentification, Birkhauser Verlag, Berlin, pp. 1-46.

Adamowicz M., Janulewicz P. (2012), Wykorzystanie metod wielowymiarowych w określeniu pozycji konkurencyjnej gminy na przykładzie województwa lubelskiego, in: Metody ilościowe w badaniach ekonomicznych, Tom XIII/1, pp. 17 – 28.

Chan M.F., Chung L.Y.F., Lee A.S.C., Wong W.K., Lee G.S.C., Lau C.Y., Lau W.Z., Hung T.T., Liu M.L., Ng J.W.S. (2006) Investigating spiritual care perceptions and practice patterns in Hong Kong nurses: Results of a cluster analysis, Nurse Education Today No. 26, Elsevier, pp. 139-150.

Drennan R. D. (2010) Statistics for Archeologists. A common Scene Approach, Springer US, pp. 309-320.

Eustotat (2013) Eurostat – Health care Staff. Physicians, http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/Annexes/hlth_res_esms_an1.pdf, date of access: 1.06.2013.

Everitt B.S., Lanau S., Leese M., Stahl D. (2011) Cluster Analysis, 5th Edition, Wiley, United Kingdom, pp. 1-110.

Gatignon H. (2010) Statistical Analysis of Management Data, Springer US, pp. 295-322

Hass-Symotiuk M (red.) (2011) *System pomiaru i oceny dokonań szpita*la, Wolters Kluwer, Warszawa.

Izenman A. J. (2008) Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning, Springer, New York, pp. 407-462.

Jena A.B., Philipson T.J. (2013) Endogenous Cost-Effectiveness Analysis and Health care technology Adoption, Journal of Health Economics, vol. 32, pp. 172-180.

Journard I., Andre Ch., Nicq Ch. (2010) Health Care Systems: Efficiency and Institutions, Economic Department Working Papers No. 769, ECO/WKP vol. 25.

Knapp M., Pan Y.J., McCrone P. (2012) Cost-Effectiveness Comparison between Antidepressant Treatments in Depression: Evidence from Database Analyses and Prospective Studies, Journal of Affective Disorders, vol. 139, pp. 113-125.

Liu Ch.Y., Liu J.Sh. (2011) Mining the optimal clustering of people's characteristics of health care choices, Expert Sustems with Applications 38, Elsevier, pp. 1400-1404

Nojszewska E. (2011) *System ochrony zdrowia w Polsce*, Wolters Kluwer, Warszawa.

OECD (2011), Perceived health status, in Health at a Glance 2011: OECD Indicators, OECD Publishing, p. 40.

OECD (2012a) OCED Health Data 2012 – Definitions, Sources and Methods. Health expenditure and financing, OECD, pp. 1-20.

OECD (2012b) OCED Health Data 2012 – Definitions, Sources and Methods. Total hospital beds, OECD, pp. 1-8.

OECD (2012c) OCED Health Data 2012 – Definitions, Sources and Methods. Life expectancy at birth and various ages, OECD, pp. 1-7.

Roy S., Pharm B., Madhavan S.S. (2009) Cluster analysis of state Medicaid prescription drug benefit programs based on potential determinants of per capita drug expenditure, Research in Social and Administrative Pahrmacy 5, Elsevier, pp. 51-62.

Stanisz A. (2007) Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, Tom 3. Analizy wielowymiarowe, Statsoft, Kraków, pp. 113-164.

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N. (2012) Efficiency evaluation for pooling resources in health care, OR Spectrum vol.34, pp. 371-390.

Willan A.R., Kowgier M.E. (2008) Cost-Effectiveness Analysis of a Multinational RCT with a Binary Measure of Effectiveness and an Interacting Covariate, Journal of Health Economics, vol. 17, pp. 777-791.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 108 – 117

MEASURING THE EFFICIENCY OF LOCAL GOVERNMENT UNITS MANAGEMENT IN THE CENTRAL REGION OF POLAND IN A DYNAMIC PERSPECTIVE

Piotr M. Miszczyński
Department of Operations Research, University of Lodz
e-mail: piotr.miszczynski@uni.lodz.pl

Abstract: Local government units in Poland are obliged to improve the quality of life of their inhabitants concerning rules of sustainable development. The study described in the article is intended to measure the relative efficiency of management including examination of the relationship between various inputs and outputs in local government units. The analysis in the paper shows efficiency differences between local governments in comparison with sub-region leaders (group frontiers) and region leaders (meta-frontiers). The division into sub-regions and regions was made according to NUTS classification. The measurement of inputs and outputs in local government management units was based on indicators of sustainable development from SAS (Local Government Research System) database. Apart from static comparison (for a particular year) the main purpose of the article is to show changes of efficiency in 5 years horizon with application of dynamic meta-frontier approach. The outcome of the analysis made it possible to indicate some reference points (benchmarks), which may contribute to improve the efficiency of management in the local government units in Poland under research. The concepts delivered in the paper are employed for the purpose of assessing growth performance of local governments using a data set covering 128 cities in period 2006-2010.

Keywords: meta-frontier approach, Malmquist index, catch-up index, relative efficiency measurement, management in local governments, operations research, sustainable development

INTRODUCTION

Effective delivery of public services by local governments is a very important factor in improving the quality of life in the community. The local

governance is assessed by indicators of sustainable development. The analysis in the paper is based on measuring the efficiency of municipalities in Poland. Thanks to the advantages of Data Envelopment Analysis (DEA) the performance of local governments is assessed in a multivariate way. Many factors (indicators of sustainable development), beside standard univariate interpretation, can also be treated as a whole, as measures of costs and benefits borne by the community to achieve the goal of sustainable development and to improve the quality of life of the inhabitants of the local community. Particular indicators of sustainable development are recognized as inputs or outputs for local government activities. DEA, as a result, gives a relative efficiency measure which is one, multivariate indicator of efficiency in a particular local government. Nevertheless, in the paper there is considered a novel in DEA approach, which is holistic analysis of group of local governments in particular sub-regions to recognize sub-region leaders. Thanks to the concept of Rambaldi, Rao and Dolan [Rambaldi et al. 2007] there is measured the efficiency gap between local governments in the sub-region in comparison to region leaders. Finally, having efficiency measures for all considered units it is possible to get average technology gap between sub-region and region, both in static (Technology Gap Ratio measure) and dynamic (catch-up index measure) points of views.

EFFICIENCY IN LOCAL GOVERNMENT

Conceptual basis of the analysis in the paper is taken from SAS[1] (Local Government Research System) database which is also used in the empirical part of research. SAS is based on Potkanski and Rogala's model [Rogala et al. 2008] of hierarchical relationship between quality of life, sustainable development and public services provided by local governments (municipalities). Quality of life of inhabitants is the main goal of local governance. Social, economic and environmental-spatial dimensions are achieved together through the concept of sustainable development, which is required by Polish Constitution[2]. Instrumentation of sustainable development is a public service provided by the municipality, which manages the economic, social and environmental spheres in the local area.

To differentiate the levels of local governments there is used a NUTS system introduced by the European Union and applied for statistical offices in the United Europe.[3] In this paper there are considered NTS1 level (central region of Poland only), NTS2 level (lodzkie voievodeship-the province of Lodz and

[1] Documentation and database available in internet: www.sas24.org (only Polish language)
[2] Polish Constitution - „Konstytucja Rzeczpospolitej Polskiej", Chapter I, art. 5
[3] For more information about NUTS nomenclature see European Commision Web site: http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

mazowieckie voievodeship/province as two sub-regions of central region), and NTS5 level (all gminas in considered central region of Poland).

METHODOLOGY

To measure efficiency of local governance there is used a Data Envelopment Analysis (DEA) method. It is a non-parametric mathematical modeling method which measures "technical efficiency" as ratio of weighted sum of inputs divided by weighted sum of outputs.[4] Using DEA terminology, local government is named DMU – Decision Making Unit. Development level of local government (DMU) is considered as technology used by all other local governments (other DMUs) in particular sub-regions and the region considered in the analysis.
To obtain the efficiency measure results there was used a basic DEA output oriented BCC model[5] for every DMU (for every object *o*) (1-4):

$$\theta_o \rightarrow \max \tag{1}$$

$$\sum_{j=1}^{n} x_{ij}\lambda_{oj} \leq x_{no} \qquad \text{for } i = 1, \ldots, m \tag{2}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_{oj} \geq \theta_o y_{ro} \qquad \text{for } r = 1, \ldots, s \tag{3}$$

$$\sum_{j=1}^{n} \lambda_{oj} = 1 \tag{4}$$

$$\theta_o, \lambda_{o1}, \lambda_{o2}, \ldots, \lambda_{on} \geq 0 \tag{5}$$

where:

- θ_o - output level coefficient (technical efficiency measure) for considered object o, ($\theta_o = 1$ means that object o is effective, is a frontier)
- y_{rj} – *r*-th ouput of object *j*,
- x_{ij} – *i*-th input of object *j*,
- y_{ro} – *r*-th ouput of object *o*,
- x_{io} – *i*-th input of object *o*,
- *m* – number of inputs,
- *s* – number of outputs,
- *n* – number of objects,
- λ_{oj} – optimal technology coefficient for object *j* (variable in the model, value bigger than 0 says that effective object *j* is benchmark for ineffective object *o*)

The model above brings static (one year) comparison between DMUs (local governments). To achieve a dynamic perspective there is used a Malmquist index which shows efficiency change from year to year for particular DMUs.

---

[4] For theoretical background see Charnes et al. (1996)

[5] BCC model assumes variable return of scale (vrs), constrain (4); for more theoretical background of choice vrs model see O'Donnell et al. (2008)

To assess changes in efficiency of DMU in time there is used the Malmquist index:

$$M^{t,t+1}(y^t, x^t, y^{t+1}, x^{t+1}) = \sqrt{\frac{\frac{1}{D^t(y^{t+1},x^{t+1})}}{\frac{1}{D^t(y^t,x^t)}} \times \frac{\frac{1}{D^{t+1}(y^{t+1},x^{t+1})}}{\frac{1}{D^{t+1}(y^t,x^t)}}} \quad (6)$$

and

$$D^t(y^{t+1}, x^{t+1}) = \frac{1}{\theta} \quad (7)$$

where:
$D^t(y^{t+1}, x^{t+1})$ is a measure of distance between object in period t+1 and technology (all considered DMUs input-output mix) in period t.

Malmquist index can be decomposed with use of algebraic transformation into technical change (TC) and technical efficiency change (TEC).

$$TEC^{t,t+1}(y^t, x^t, y^{t+1}, x^{t+1}) = \frac{\frac{1}{D^{t+1}(y^{t+1},x^{t+1})}}{\frac{1}{D^t(y^t,x^t)}} \quad (8)$$

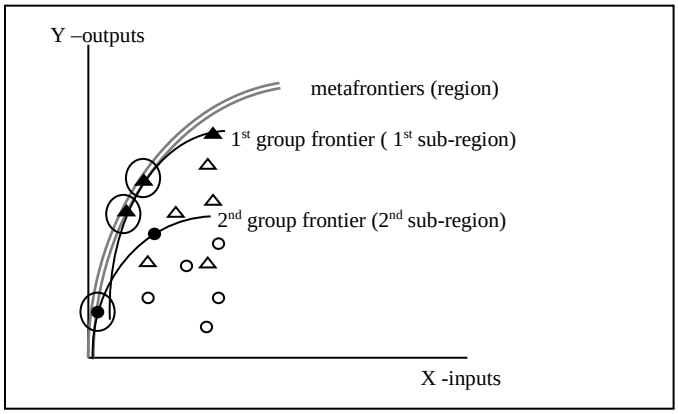
$$TC^{t,t+1}(y^t, x^t, y^{t+1}, x^{t+1}) = \sqrt{\frac{\frac{1}{D^t(y^{t+1},x^{t+1})}}{\frac{1}{D^{t+1}(y^{t+1},x^{t+1})}} \times \frac{\frac{1}{D^t(y^t,x^t)}}{\frac{1}{D^{t+1}(y^t,x^t)}}} \quad (9)$$

where:
TEC – efficiency change which does not consider change in technology in time (frontiers change)
TC – efficiency that considers changes in technology development between periods t and t+1

In the paper there is considered efficiency comparison across regions. It is made by measuring efficiency relative to a metafrontier, which is a boundary of an unrestricted technology set (all region, not only a sub-region). There are also group frontiers to be the boundaries of restricted technology sets in sub-regions (there is assumed that restrictions result from lack of economic infrastructure or other characteristics of the DMU environment in particular sub-region in comparison to other sub-regions). Metafrontiers (region frontiers) envelop the group frontiers (sub-region frontiers). In the paper there are measured two types of efficiency: [O'Donnell et al. 2008]

- distance to the group frontier (common TC) and
- distance between the group frontier and the metafrontier (TGR).

Figure 1. Group frontiers with relation to metafrontiers



Source: self-prepared

The sub-region specific group technologies are sub-sets of the region metatechnology. For all groups ($k = 1,\ldots,K$) the distance in time period t with respect to the k-group frontier is greater than or equal to the distance in time period t with respect to the metafrontier [Rambaldi et al. 2007, p.10].

$$D^t_{meta}(y^t, x^t) \geq D^t_k(y^t, x^t) \tag{10}$$

For the output-oriented model there can be obtained Technology Gap Ratio (TGR) at time *t*:

$$TGR^t_k(y^t, x^t) = \frac{D^t_{meta}(y^t,x^t)}{D^t_k(y^t,x^t)} \tag{11}$$

$TGR^t_k < 1$ shows that between k-sub-region and region frontiers there is a technology gap.

Technology gap ratio can be considered dynamically by the technology gap ratio growth index.

$$TGR\_GR^{t,t+1}_k(y^t, x^t, y^{t+1}, x^{t+1}) = \frac{TGR^{t+1}_k(y^{t+1},x^{t+1})}{TGR^t_k(y^t,x^t)} \tag{12}$$

After a few algebraic manipulations of Rambaldi, Rao and Dolan [Rambaldi et al. 2007, pp. 15-18] there are distinguished two types of technology gap ratio growth indexes concerning decomposition: one for technical efficiency change, another for technical change:

$$TGR\_GR^{t,t+1}_k(y^t, x^t, y^{t+1}, x^{t+1}) = \frac{TEC^{t,t+1}_{meta}(y^t,x^t,y^{t+1},x^{t+1})}{TEC^{t,t+1}_k(y^t,x^t,y^{t+1},x^{t+1})} \tag{13}$$

$$[TGR\_GR^{t,t+1}_k(y^t, x^t, y^{t+1}, x^{t+1})]^{-1} = \frac{TC^{t,t+1}_{meta}(y^t,x^t,y^{t+1},x^{t+1})}{TC^{t,t+1}_k(y^t,x^t,y^{t+1},x^{t+1})} \tag{14}$$

where:

(13)– technology gap ratio growth rate which does not consider change in technology (frontiers change) in time, can be interpreted as the relative

technological progress or regress of the local government (DMU) in sub-region *k* with respect to change in the metatechnology (region technology) change. When $\mathrm{TGR\_GR}_{k}^{t,t+1} < 1$ then the gap between the sub-region frontier and the metafrontier is decreasing (particular sub-region is experiencing technological progress at a faster rate than in the whole region).

(14)– inverse technology gap ratio growth rate TC – efficiency that considers changes in technology development between periods *t* and *t*+1and removes the issue of choosing a relevant benchmark time period by averaging the input-output mixes (technologies) of two different time periods. Interpretation of the ratio is analogical to (13), when $[\mathrm{TGR\_GR}_{k}^{t,t+1}]^{-1} > 1$ then the gap between the sub-region frontier and the region metafrontier is decreasing.

The metafrontier approach brings two types of Malmquist indexes:

- considering the region specific technology (k-sub-region frontier) (15), and
- considering the metatechnology (metafrontier) (16).

$$\mathrm{M}_{k}^{t,t+1}(y^{t}, x^{t}, y^{t+1}, x^{t+1}) = \sqrt{\frac{\frac{1}{D_{k}^{t}(y^{t+1}, x^{t+1})}}{\frac{1}{D_{k}^{t}(y^{t}, x^{t})}} \times \frac{\frac{1}{D_{k}^{t+1}(y^{t+1}, x^{t+1})}}{\frac{1}{D_{k}^{t+1}(y^{t}, x^{t})}}} \tag{15}$$

$$\mathrm{M}_{meta}^{t,t+1}(y^{t}, x^{t}, y^{t+1}, x^{t+1}) = \sqrt{\frac{\frac{1}{D_{meta}^{t}(y^{t+1}, x^{t+1})}}{\frac{1}{D_{meta}^{t}(y^{t}, x^{t})}} \times \frac{\frac{1}{D_{meta}^{t+1}(y^{t+1}, x^{t+1})}}{\frac{1}{D_{meta}^{t+1}(y^{t}, x^{t})}}} \tag{16}$$

There are also given respectively to (8) and (9): $\mathrm{TEC}_{meta}^{t,t+1}$, $\mathrm{TC}_{meta}^{t,t+1}$, $\mathrm{TEC}_{k}^{t,t+1}$, $\mathrm{TC}_{k}^{t,t+1}$.

As a summary of algebraic manipulation of Rambaldi, Rao and Dolan [Rambaldi et al. 2007, pp. 15-19] there is achieved an equation as follows (17):

$$\mathrm{catch-up}_{k}^{t,t+1} = \frac{\mathrm{M}_{k}^{t,t+1}}{\mathrm{M}_{meta}^{t,t+1}} \tag{17}$$

The $\mathrm{catch-up}_{k}^{t,t+1} > 1$ means that particular *k*-th sub-region shows catch- up with the whole region (metafrontier) technology over the periods *t* to *t*+1.

Thanks to the DEA BCC output oriented model, Malmquist indexes and metafrontier approach Rambaldi, Rao and Dolan [Rambaldi et al. 2007] assume that three types of efficiency changes can be identified and recognized between *t* and *t*+1:

- changes in the input mix,
- changes in technology at the sub-regional level (shifts of the sub-regional group frontiers), and
- changes in the metatechnology (shifts of the metafrontier).

DATA

The data used in the study come from the SAS (Local Government Research System) database[6]. The basis of SAS brings together in one place data on all municipalities in Poland and the study of quality of life at the local level. Most of the data in the database are especially designed indicators according to the assumptions of authors of the system (database). The indicators are based on data provided by the Polish Central Statistical Office and the Polish Ministry of Finance.

As the input and output to test the efficiency of DEA there were used especially designed sustainable development indicators. The selection of indicators was based on the work methodology of SAS database, which is divided into three main types: economic, environmental - spatial, and social. The indicators of each type are divided into 10 areas. In total, that gives the 30 types of indicators. In the analysis shown in this paper for each area there was chosen a representative indicator. The selection of inputs and outputs (here the indicators of sustainable development) was fairly limited due to the availability of empirical data. The indicators which are classified as inputs as well as those considered to be the outputs do not involve a full picture of the situation of the activities in local governments. In this research selected variables (sustainable development indicators) are treated as symptoms [Guzik, 2009, p.64] of all the studied topic.

EMPIRICAL CASE

Technical efficiency estimated in DEA method is often used for improving management in DMU (in this case local governments). Additional information to design programs gives estimation of the technology gap between group frontiers and the metafrontier. It gives a wider view on the environment of DMU not only at local, but also at sub-regional and regional levels in this case. In the considered case there were measured year-to-year dynamics of considered indicators in 5 year scope (2006-2010).

In Poland there are 908 cities and in the Central Region there are 128 cities. All are managed by local governments. According to three- stage territorial division of Poland and with respect to NUTS[7] United Europe classification there are distinguished sub-regions which gather cities (local governments, DMUs) in groups. In the paper cities are grouped according to NTS-2 level, provinces (or voivodships). In the analysis there are considered two of them: province of Lodz called lodzkie voivodship which gathers 43 cities and mazowieckie voivodship

[6] Available in internet: www.sas24.org (only Polish language)

[7] More datails at EU WebPage
http://europa.eu/legislation_summaries/regional_policy/management/g24218_en.htm
[15.08.2013]

which gathers 85 cities. Efficiency of 43 cities in lodzkie voivodship (grouped as members of first sub-region) and 85 cities in mazowieckie voivodship grouped as members of second sub-region) are compared to regional frontiers (metafrontiers) of Polish Central Region (which gathers two considered sub-regions into one region NTS-1 level).

The analysis brought about some results for all particular DMUs (cities). For the sake of a wider comparison there were made sub-region results aggregates which are averages of indicators for DMUs for particular sub-regions. Some examples of such aggregated indicators for two considered voivodships (sub-regions) are presented below.

Table 1. Average results for DMUs in two considered voivodships in 2006-2010

lodzkie voivodship									
[%]	TEC_k	TC_k	M_k	TEC _meta	TC _meta	M _meta	TECm /TECk	TCm /TCk	catch_up
2007	98,48	105,54	**103,42**	96,70	106,93	**102,46**	98,07	101,20	**100,98**
2008	106,03	103,95	**109,41**	106,90	104,73	**110,67**	100,73	100,63	**98,87**
2009	97,57	100,42	**97,40**	94,63	100,03	**93,72**	97,18	99,50	**105,24**
2010	90,65	121,52	**106,47**	91,47	121,23	**109,21**	102,69	101,17	**97,65**
mazowieckie voivodship									
[%]	TEC_k	TC_k	M_k	TEC _meta	TC _meta	M _meta	TECm /TECk	TCm /TCk	catch_up
2007	100,67	103,24	**102,91**	101,17	102,88	**102,59**	100,39	99,56	**100,33**
2008	100,86	108,21	**107,19**	102,92	109,36	**110,57**	102,17	101,04	**97,03**
2009	99,44	97,33	**95,35**	100,26	95,83	**94,58**	100,81	98,42	**100,92**
2010	98,42	113,67	**109,29**	95,90	117,43	**109,52**	97,43	103,13	**99,82**

Source: self-prepared with use of computation in EMS and MS Excel

The Table above provides some useful information to consider. For the clarity of observations, it has to be mentioned that Mamquist indexes (with respect to group frontiers and metafrontiers) for average efficiency growth in both voivodships show the efficiency growth in all the periods concerned except for 2009, where the indexes show a decline. Concerning the gap between group and meta frontiers, mazowieckie voivodship in 2007 and 2009 shows a catch-up value closer to 1, which means that more DMUs from this sub-region were metafrontiers in that time. The biggest cath-up for lodzkie voivodship in 2009 can result from

a narrowing gap between this group frontiers and metafrontiers, which can be perceived by means of a technology gap ratio growth rate and inversed technology gap ratio growth rate values below 1 in that period.

CONCLUSIONS AND FUTURE RESEARCH

DEA method gives possibility to measure technical efficiency of several units in a particular group and within a particular period of time. Malmquist approach in DEA enables the comparison of technical efficiency change in several time periods. The metafrontier approach makes it possible to compare different groups of units. Finally, Rambaldi, Rao and Dolan's [Rambaldi et al. 2007] proposal gives connections of these three separate approaches. The DEA approach, enriched by the metafrontier idea accompanied by the Malmquist index method, gives new possibilities of analysis at local, sub-regional, regional, but also (which was shown by Rambaldi, Rao and Dolan) country and world level comparisons. It is especially useful to consider it dynamically through several time periods. The case presented above is an introduction to wider analysis at country level. Promising results of the region analysis give a chance for deeper conclusions from more than two - province comparisons in the future. There was an intention of the author to focus on possibilities created by the DEA method and its metafrontier approach for deeper regional and sub-regional analysis within a particular country in a dynamic perspective. The next step leads to comparative analysis of existing approaches published by other authors.

REFERENCES

Charnes A. (1996) Cooper W., Lewin A.Y., Seiford L.M., Data Envelopment Analysis: Theory, Metodology, and Application, Kluwer Academic Publishers

Coelli, T. (1998) Prasada Rao, D, and S., Battese, G. E, An introduction to efficiency and productivity analysis, Boston, Kluwer Academic Publishers, 1998.

Farrell, M. J. (1957) The Measurement of Productive Efficiency, Journal of the Royal Statistical Society Series A (General) vol.120, no.3, pp. 253-290

Guzik B. (2009) Podstawowe modele DEA w badaniu efektywności gospodarczej i społecznej, wyd. Uniwersytetu Ekonomicznego, Poznań

O'Donnell C. J., Rao, D.S.P., Battese G. E.(2008) Metafrontier frameworks for the study of firm-level efficiencies and technology ratios, Empirical Economics, Springer-Verlag, 34: p.231–255

Rambaldi, A.N., Rao, D.S.P. and Dolan, D. (2007) Measuring Productivity Growth Performance using Meta-Frontiers with applications to Regional Productivity Growth analysis in a Global Context. In: O'Donnell, C.J., Australian Meeting of the Econometric Society ESAM07, Brisbane

Rogala P. (2008) Wskaźniki zrównoważonego rozwoju i badania jakości życia, Jelenia Góra – Poznań

Scheel H. (2000) EMS – Efficiency Measurement System Version 1.3.0, manual for computer program for DEA calculation

United Europe Web Page [15.08.2013] http://europa.eu/legislation_summaries/regional_policy/management/g24218_en.htm

Eurostat Web Page [15.08.2013], http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

Zibaei M. (2012) Technical Efficiency Analysis of Fisheries: Toward an Optimal Fleet Capacity, Sustainable Agriculture Research Vol.1, No.1

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 118 – 128

THE DEVELOPMENT OF AGRICULTURE IN POLAND IN THE YEARS 2004-2011 – THE TAXONOMIC AND ECONOMETRIC ANALYSES

Joanna Muszyńska, Iwona Müller–Frączek
Department of Econometrics and Statistics
Nicolas Copernicus University in Torun
e-mail: Joanna.Muszynska@umk.pl, Muller@econ.umk.pl

Abstract: The aim of the paper was to assess the regional differentiation of the level of agriculture and its changes over time. Based on the synthetic measure of development the rankings of regions (provinces) were created. The objects were also classified and divided into groups of a similar level of agriculture. In addition, in order to identify the long-term tendency in this sector of economy, the process of β-convergence of the level of agriculture has been studied. For the verification of hypotheses dynamic panel models were applied. All computations were performed in the Gretl, based on CSO data.

Keywords: regional differentiation, synthetic measure of development, β-convergence

INTRODUCTION

Polish membership to the European Union, and thus the possibility to use EU funds have a significant impact on the development of Polish agriculture. Actions taken by farmers, due in part to the use of the Structural Funds have changed the nature of many farms. They cease to be only the source of income for the farmer and his family. More and more often, the farms become the enterprises that compete on the market of food producers. Implementation of new technologies, increasing the scale of production and the specialization are aimed to create a financial surplus to enable the further development of the farm.

Issues concerning the development of Polish agriculture and its regional differentiation, especially in the context of Polish accession to the European Union, have been widely discussed in the literature. These topics can be found in the

works, among others, of Binderman [2010, 2012], Muszyńska [2009, 2010], and Szewczyk [2012].

The article presents the analysis of the level of agricultural development in Poland in the years 2004-2011. The study refers to private farms with area exceeding 1ha. The average farm in the province was adopted as the research unit.

For the purpose of the analysis, the synthetic measure of development was created. Its construction was based on the different aspects of agriculture. The economic size of the farm was one of the above mentioned characteristics. It is measured in PLN and determines production capacity of the farm, expressed as its potential income. Regional coefficients of standard output (SO), applied in the calculations, allowed to reflect local conditions, different for the four statistical regions in Poland.

The aim of the paper was to assess the regional differentiation of the level of agriculture and its changes over time. Based on the synthetic measure of development, constructed for each of the years of study, the rankings of regions (provinces) were created. The objects were also classified and divided into groups of a similar level of agriculture. In addition, the process of β-convergence of the level of agriculture has been studied. The validation of the hypothesis of the absolute β-convergence has allowed identifying the long-term tendency in this sector of economy. Based on the analysis of the conditional β-convergence, the article indicates the main determinants of development.

During the study, the authors applied taxonomic methods, widely discussed in the literature, inter alia, by Jajuga [1993] and Kolenda [2006]. In order to verify the hypotheses, which have been posed in the analysis, dynamic panel data models were used. The models were constructed and estimated according to the methods described in the literature, among others, by Baltagi [2005] and Dańska–Borsiak [2011]. All computations were performed in GRETL, using data available in the public statistics.

THE TAXONOMIC ANALYSIS

The empirical study was based on data derived from the Local Data Bank and the statistical yearbooks, published by CSO. Availability of statistical data limited the scope of the analysis, both in space (provinces) and in time (years 2004-2011). It also enabled to take into account only some of the aspects of agricultural development[1]. The average in the province, private farm with an area exceeding 1 ha was the research unit. Diagnostic variables, used in the

[1] Due to the lack of data, some of the characteristics were not taken into account, e.g. education level of farm owner, number of employees, the degree of mechanization of farms and others.

analysis were: economic size of farm[2], agricultural land area of farm, level of investment in agriculture and fixed assets value.

The a.m. variables reflected the most important determinants of the development of agriculture. Production capacity of farm, structure and marketability of crops were described with the economic size of the farm. The other variables presented agricultural land area, the volume of investment in agricultural production and value of fixed assets of the farm.

All of the diagnostic variables were stimulants. They also met the postulate of maximum spatial differentiation. To ensure variable uniformity all of them were standardised before aggregation. Upon the value of the determinants Hellwig's measure of development was created. The measure was constructed in accordance with the formula:

$$d_i = 1 - \frac{c_{i0}}{\bar{c}_0 + 2s_0}, \tag{1}$$

where:

c_{i0} – Euclidean distance[3] of the object *i* to the pattern[4],

$\bar{c}_0$ – average distance of the objects to the pattern,

s_0 – standard deviation of the distance.

Table 1 presents the value of the synthetic measure of development of agriculture (d_i), during the years 2004-2011.

As it can be seen in Table 1, the values of the synthetic measure of development (d_i) for the best agricultural provinces were several times greater than the values of the weakest regions. This fact confirms the strong regional differentiation of the level of agriculture in Poland. Simultaneously, the low level of diversification of agriculture over time can be observed. During the analysed period, d_i for most provinces remained at a similar level. For seven of sixteen regions, the synthetic measure of development did not exceed the value 0,5. In the whole period under investigation it remained at a low, almost constant level.

---

[2] Economic size was calculated based on the regional coefficients of standard output (SO) and the data on major crops and acreage of basic animal husbandry. Details can be found in the paper of Müller-Frączek I., Muszyńska J. (2013) Regionalne zróżnicowanie wielkości ekonomicznej indywidualnych gospodarstw rolnych w Polsce, The Annals of The Polish Association of Agricultural and Agribusiness Economists, volume XV, no. 4.

[3] Euclidean distance was calculated according to the formula: $c_{i0} = \sqrt{\sum_{j=1}^{m} (z_{ij} - z_{0j})^2}$

where: z_{ij} – the standardised value of variable *j* for the object *i*, z_{0j} – the standardised value of variable *j* for the pattern.

[4] Pattern – a hypothetical object with the best values of all diagnostic variables (in case of stimulants – maximum values).

Table 1. The value of synthetic measure of development of agriculture

year / province	2004	2005	2006	2007	2008	2009	2010	2011
dolnośląskie	0,52	0,50	0,48	0,50	0,52	0,54	0,52	0,56
kujawsko-pomorskie	0,67	0,62	0,62	0,61	0,70	0,71	0,69	0,68
lubelskie	0,30	0,27	0,26	0,26	0,28	0,29	0,27	0,26
lubuskie	0,51	0,55	0,60	0,49	0,61	0,61	0,67	0,68
łódzkie	0,34	0,33	0,31	0,32	0,33	0,33	0,33	0,32
małopolskie	0,12	0,13	0,13	0,13	0,14	0,13	0,15	0,13
mazowieckie	0,37	0,34	0,34	0,35	0,38	0,37	0,37	0,41
opolskie	0,68	0,61	0,63	0,65	0,62	0,69	0,76	0,84
podkarpackie	0,12	0,14	0,13	0,14	0,14	0,13	0,14	0,14
podlaskie	0,56	0,51	0,52	0,52	0,56	0,59	0,56	0,52
pomorskie	0,69	0,62	0,62	0,64	0,73	0,70	0,66	0,77
śląskie	0,22	0,21	0,24	0,25	0,26	0,26	0,23	0,23
świętokrzyskie	0,25	0,21	0,21	0,20	0,21	0,21	0,20	0,21
warmińsko-mazurskie	0,98	0,87	0,85	0,84	0,94	0,94	0,88	0,80
wielkopolskie	0,74	0,68	0,67	0,70	0,75	0,73	0,72	0,77
zachodniopomorskie	0,74	0,87	0,87	0,89	0,87	0,87	0,86	0,82

Source: own calculations based on CSO data

Based on the values of the synthetic measure of development rankings of the provinces were constructed. The results are shown in Table 2.

Table 2. The rankings of the provinces

year / province	2004	2005	2006	2007	2008	2009	2010	2011
warmińsko-mazurskie	1	2	2	2	1	1	1	3
wielkopolskie	2	3	3	3	3	3	4	5
zachodniopomorskie	3	1	1	1	2	2	2	2
pomorskie	4	4	6	5	4	5	7	4
opolskie	5	6	4	4	6	6	3	1
kujawsko-pomorskie	6	5	5	6	5	4	5	7
podlaskie	7	8	8	7	8	8	8	9
dolnośląskie	8	9	9	8	9	9	9	8
lubuskie	9	7	7	9	7	7	6	6
mazowieckie	10	10	10	10	10	10	10	10
łódzkie	11	11	11	11	11	11	11	11
lubelskie	12	12	12	12	12	12	12	12
świętokrzyskie	13	14	14	14	14	14	14	14
śląskie	14	13	13	13	13	13	13	13
małopolskie	15	16	15	16	15	15	15	16
podkarpackie	16	15	16	15	16	16	16	15

Source: own calculations based on CSO data

Compatibility of the results of the following years was assessed using Kendall's coefficient of concordance *W*, expressed by formula:

$$W = 12 \cdot \frac{N\left(\sum_{i=1}^{N} T_i^2\right) - \left(\sum_{i=1}^{N} T_i\right)^2}{m^2(N^4 - N^2)}, \tag{2}$$

where:

- N – sample size,
- m – number of rankings,
- T_i – the sum of all ranks of the object i .

The Kendall's coefficient yielded an observed W=0,975. Very high and statistically significant value of the coefficient has proved the compatibility of the rankings in the considered period.

The next step of the taxonomic analysis was to classify the regions and divide them into four groups with the same level of agricultural development. The classification was carried out using two methods: the standard deviation and maximum gradient. The results of clustering (see Table 3) were very similar for both methods. In most cases, the region was assigned into the same group or a neighboring group.

We can distinguish three groups of provinces, for which the results of clustering were consistent in the whole period of the study:

- the best agricultural regions (group I) – provinces: warmińsko–mazurskie and zachodniopomorskie,
- average level of agricultural development (group II or I) – provinces: kujawsko–pomorskie, opolskie, pomorskie, wielkopolskie,
- the weakest agricultural regions (group IV) – provinces: małopolskie and podkarpackie.

For the remaining eight regions the results were not so unequivocal.

Table 3. The results of the clustering

year / province	clustering method	2004	2005	2006	2007	2008	2009	2010	2011
dolnośląskie	*	II	II	II	II	II	II	II	II
	**	III		III	III	III		III	
kujawsko-pomorskie	*	II	II	II	II	II	II	II	II
	**								I
lubelskie	*	III	III	III	III	III	III	III	III
	**	IV		IV	IV	IV		IV	IV
lubuskie	*	II	II	II	II	II	II	II	II
	**	III			III	III			I
łódzkie	*	III	III	III	III	III	III	III	III
	**	IV		IV	IV	IV		IV	IV
małopolskie	*	IV	IV	IV	IV	IV	IV	IV	IV
	**								
mazowieckie	*	III	III	III	III	III	III	III	III
	**	IV		IV	IV	IV		IV	
opolskie	*	II	II	II	II	II	II	I	I
	**					III		II	
podkarpackie	*	IV	IV	IV	IV	IV	IV	IV	IV
	**								
podlaskie	*	II	II	II	II	II	II	II	II
	**	III		III	III	III		III	
pomorskie	*	II	II	II	II	II	II	II	I
	**								
śląskie	*	IV	IV	III	III	III	III	IV	IV
	**		III	IV	IV	IV			
świętokrzyskie	*	III	IV	IV	IV	IV	IV	IV	IV
	**	IV	III				III		
warmińsko-mazurskie	*	I	I	I	I	I	I	I	I
	**								
wielkopolskie	*	I	II	II	I	I	II	II	I
	**	II			II	II			
zachodniopomorskie	*	I	I	I	I	I	I	I	I
	**	II							

* – standard deviation method, ** – maximum gradient method

Source: own calculations based on CSO data

THE ECONOMETRIC ANALYSIS

The next stage of research concerned the future of agriculture in Poland. Its aim was to assess the convergence of the level of agricultural development of private farms. The average in the province, private farm with an area exceeding 1 ha remained the research unit. The agricultural development was defined by the synthetic measure d_i, as it was described in the previous section. Analysis was based on a dynamic panel data model[5]:

$$\ln \frac{Y_{i,t}}{Y_{i,t-1}} = \alpha - \beta \ln Y_{i,t-1} + \eta_i + u_{it}, \tag{3}$$

where:

- Y – the level of development,
- i – the number of the region, $i = 1, ..., N$,
- t – number of period $t = 1, ..., T$,
- η_i – group effects,
- u_{it} – error term.

The phenomenon of unconditional β-convergence of the process Y occurs when the parameter β, in equation (3) is a positive value. It proves there is a constant over time, negative correlation between the level of the process and its growth rate. The existence of unconditional β-convergence means that the regions with initially lower level of the investigated process will catch up the better developed provinces. The speed of convergence to equilibrium (the rate of catching up) can be calculated according to the formula:

$$\lambda = -\ln(1-\beta). \tag{4}$$

In order to estimate parameters the dynamic panel data model, described by the equation (3), was transformed to the model:

$$y_{i,t} = \alpha + (1-\beta) y_{i,t-1} + \eta_i + u_{it}, \tag{5}$$

where: $y_{i,t} = \ln Y_{i,t}$.

Based on the values contained in table no 1, the empirical model of unconditional β-convergence was estimated. It took the following form:

$$\hat{y}_{i,t} = -0{,}083 + \underset{(\pm 0{,}061)}{(1-0{,}104)} y_{i,t-1}, \tag{6}$$

Model parameters were estimated using the Blundell and Bond System Generalized Method of Moments Estimator (GMM-sys). The correctness of the estimated model was verified using the Arellano-Bond test for autocorrelation and the Sargan test of over-identifying restrictions. The estimation methods of dynamic panel data

[5] Static approach was unfeasible because of insufficient number of observations.

models and the statistical tests mentioned above are widely described in the literature, inter alia, by Ciołek [2004] and Dańska-Borsiak [2011].

The Sargan test checks if over-identifying restrictions omitted from the estimation process were correct. The null hypothesis of the test states that the applied instruments are correct in the sense of their being uncorrelated with the error terms of the first difference model. The Arellano-Bond test verifies the assumption regarding autocorrelation of the model error term. The model is properly specified (the GMM method provides consistent estimator) if there is no arguments for rejecting the null hypothesis about the absence of the second-order autocorrelation of the first difference model error term. Existence of the first-order autocorrelation is an expected phenomenon, resulting from the model construction.

Table 4. The test results for models described by equations (6) and (9)*

test	model (6)		model (9)	
	value of the test statistics	p-value	value of the test statistics	p-value
AR(1)	-1,741	0,081	-1,881	0,060
AR(2)	0,675	0,500	0,214	0,831
Sargan	14,943	0,958	14,979	0,958
Wald	218,638	0,000	11357,900	0,000

*-verification was conducted at 10% level of significance

Source: own computations performed in GRETL

The tests results are compiled in Table 4. All the tests confirmed the proper specification of the models. For both models, the Sargan test gave no arguments for rejecting the null hypotheses. The instruments applied during the estimation process were not correlated with the error terms of the models. Also the Arellano-Bond test, used to verify the assumption about the absence of the second-order autocorrelation, provided no grounds for rejecting the null hypotheses. That means there was no the second-order autocorrelation of error terms in both models. Significance of the parameter estimates was proved using the Wald test.

A positive value of the coefficient $\beta = 0{,}104$ in the model (6) positively verified the hypothesis regarding the existence of β-convergence process of the level of agricultural development of private farms in Poland. The rate of convergence was estimated at $\lambda = 11\%$ and the time to cover halfway to the common equilibrium point were about 6,3 years[6].

The existence of β-convergence of the level of agricultural development has imposed the question of the conditions of this phenomenon. The next step in the analysis was therefore to test the conditional β-convergence, which takes into account the effect of other factors on the growth rate of the investigated process. This study was designed to not only confirm the impact of factors on convergence,

[6] The time was calculated according to the formula: $t = -\ln(0{,}5) / \lambda$.

in other words, to demonstrate the existence of conditional convergence. Its aim was to assess the strength of this effect. The speed of the conditional convergence was applied as the research tool.

The study of the conditional β-convergence was based on a model:

$$\ln \frac{Y_{i,t}}{Y_{i,t-1}} = \alpha - \beta \ln Y_{i,t-1} + \gamma \ln X_{it} + \eta_i + u_{it}, \tag{7}$$

where X is an explanatory variable (a factor that affects the process of the study).

Same as before, the conditional convergence occurs when the parameter β is positive, so there is a negative correlation between the process and its rate of growth. The rate of convergence can be estimated in accordance with the formula (4). However, this rate is determined by the strong assumption that the conditions affecting the growth rate of the process Y, in other words, the process X are the same for all regions.

For the purpose of the estimation equation (7) is converted to the form:

$$y_{i,t} = \alpha + (1-\beta) y_{i,t-1} + \gamma x_{it} + \eta_i + u_{it}, \tag{8}$$

where: $y_{i,t} = \ln Y_{i,t}$ and $x_{i,t} = \ln X_{i,t}$.

The empirical model of conditional β-convergence[7], with the investments in agriculture as an explanatory variable took the form:

$$\hat{y}_{i,t} = \underset{(\pm 0{,}037)}{-1{,}271 + (1-0{,}147) y_{i,t-1}} + \underset{(\pm 0{,}035)}{0{,}150}\ x_{it}, \tag{9}$$

where $y_{i,t}$ is the logarithm of the measure of development, and $x_{i,t}$ the logarithm of investments in agriculture of the average farm in the region i and year t.

A positive value of the coefficient β=0,147 in model (9) positively verified hypothesis regarding the existence of the conditional β-convergence with the investment in agriculture as a variable determining the phenomenon. The rate of convergence, assuming that the average investments in all the provinces is the same, was estimated at λ=15.9%. In comparison to the unconditional convergence the rate grew by 4,9%. Thus, by changing the level of investment in agriculture, the region would cover half the distance to the point of equilibrium in about 4 years.

The econometric analysis confirmed that it is possible to even out the average level of agricultural development of private farms in all regions in Poland. In addition it has indicated investments as a factor strongly influencing this phenomenon.

[7] For tests results see Table 4.

SUMMARY AND CONCLUSIONS

The study, described in the article, did not cover all aspects of agricultural development. It was an attempt to assess the regional differentiation of this phenomenon. However, in spite of its simplicity, the synthetic measure of development, presented in the paper, seemed to characterise the level of agriculture in Poland properly. The analyses performed on the basis of this measure provided reliable results both in terms of content and statistics.

The survey showed a strong regional diversification in the level of the agricultural development and simultaneously a slight differentiation of this phenomenon in time. The econometric analysis confirmed the possibility of levelling of the agricultural development of private farms. In addition, the study has indicated investment as a key factor in this process.

Because of the short period of the study and incomplete statistical information the analysis did not cover many aspects of agriculture. Therefore, the next step will be to extend the synthetic measure of development using wider range of diagnostic variables.

REFERENCES

Baltagi B.H. (2005) Econometric Analysis of Panel Data, John Wiley & Sons, Ltd., Chichester.

Binderman A. (2010) Wpływ sposobu normalizacji zmiennych na ocenę regionalnego zróżnicowania rolnictwa, Quantitative Methods in Economics, volume XI, no.2.

Binderman A. (2012) Rozwój polskiego rolnictwa w kontekście regionalnego zróżnicowania w latach 1998-2010, Quantitative Methods in Economics, volume XIII, no.3.

Ciołek D. (2004) Konwergencja krajów w okresie transformacji do Unii Europejskiej, the doctors thesis, Uniwersytet Gdański, Gdańsk.

Dańska-Borsiak B. (2011) Dynamiczne modele panelowe w badaniach ekonomicznych, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Goraj L., Bocian M., Cholewa I., Nachtman G., Tarasiuk R. (2012) Współczynniki Standardowej Produkcji "2007" dla celów Wspólnotowej Typologii Gospodarstw Rolnych, IERiGŻ PIB, Warszawa.

Hellwig Z. (1968) Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom rozwoju oraz zasoby i strukturę wykwalifikowanych kadr, Przegląd Statystyczny, no. 4.

Jajuga K. (1993) Statystyczna analiza wielowymiarowa, PWN, Warszawa.

Kolenda M. (2006) Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Muszyńska J. (2009) Regionalne zróżnicowanie rolnictwa w Polsce w roku 2007, The Annals of The Polish Association of Agricultural and Agribusiness Economists, volume XI, no. 4.

Muszyńska J.(2010) Zmiany regionalnego zróżnicowania poziomu rolnictwa w Polsce, The Annals of The Polish Association of Agricultural and Agribusiness Economists, volume XII, no. 1.

Müller-Frączek I., Muszyńska J. (2013) Regionalne zróżnicowanie wielkości ekonomicznej indywidualnych gospodarstw rolnych w Polsce, The Annals of The Polish Association of Agricultural and Agribusiness Economists, volume XV, no. 4.

Szewczyk J. (2012) Miara zróżnicowania wyposażenia gospodarstw rolnych w techniczne środki produkcji, Quantitative Methods in Economics, volume XIII, no.1.

http://www.fadn.pl

http://www.stat.gov.pl/bank danych lokalnych

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 129 – 138

ORDERING AND CLASSIFICATION OF THE SILESIAN VOIVODESHIP REGION WITH RESPECT TO A HEALTH CARE SYSTEM ACTIVITY

Sylwia Nieszporska
Department of Statistics and Econometrics
Czestochowa University of Technology
e-mail: sylniesz@poczta.onet.pl

Abstract: The health care system in Poland or any other country should ensure an equal and unrestricted availability of the health service to its population. This idea, though being noble and harmonious with the country's constitution, seems to differentiate small administrative regions of the Silesian voivodeship. Therefore the presented paper is an attempt to estimate and analyse conditions of the health care system activity in this region. Those regions were systematized based on a structure of health care system. Also, a homogeneous groups of regions were created. The analysis is based on the cluster analysis.

Keywords: numerical taxonomy, synthetic ratio, health care system, state of health, availability of a health service

INTRODUCTION

The general functioning of the health care system in Poland, which was reformed in 1999, consists in the cooperation between its participants and their mutual collaboration. One can enumerate three main goals of this exchange [Kautsch, Whitfield, Klich 2001, p. 31]: service consumers, providers and payers. The role of the medical service consumer is performed by the patient who purchases health care services from a doctor or another health care provider through the payer. The National Health Fund, which is a public insurer and intermediary in the exchange of services between the healthcare provider and the patient, has been the payer since 2003.

Responsibility for the protection of the public health needs and for the financial consequences resulting therefrom does not lie exclusively on the insurer.

A significant role in the financing of health care services is also played by the state and local governments, which are the founding bodies of the majority of state healthcare centres.

Each province in Poland is therefore required to ensure that the health care needs of its inhabitants are met on an appropriate level of expertise and quality. However various economic determinants of the functioning of the provinces and geographic, demographic or economic factors typical of individual provinces, allow one to assume existence of disproportions in the access to a wide range of health care services. This problem was discussed in a study devoted to one of the most interesting regions in Poland, namely Silesia. This region is important not only because of its population density (which in 2008 amounted to 377 persons per one square kilometer and was the highest one in Poland [www.stat.gov.pl (23.02.2009)]) or an interesting regional labour market. The uniqueness of Silesia is mainly connected with its activity in the field of health care.

In 2011, 2716 outpatient clinics [Zdrowie i ochrona zdrowia w 2011, p.190] and 116 hospitals [Zdrowie i ochrona zdrowia w 2011, p. 233] operated in the Silesian province, which were the highest values in the entire country. The Silesian region is also an important center for the education of medical staff at all levels, It can also boast a recognized and well-known heart transplantation centre (Silesian Center for Heart Diseases in Zabrze), oncology centre (with the Oncology Centre in Gliwice, which is the most modern centre in the country) and finally the recognized Burns Treatment Centre in Siemianowice Śląskie [www.silesia-region.pl (23.02.2009)].

The above mentioned advantages of the Silesian region constitute a basis for monitoring the functioning of the health care system within its boundaries. This study may be thus helpful in the assessment of health care activities in this region, compared to other regions in Poland. It can also constitute a basis for the assessment of their effectiveness in particular districts and cities. Therefore the main goal of the presented paper is to check if the access to the health service is the same in different regions of Silesian voivodeship. It is focused only on the infra-structure and inputs that can help in the health service realization. Thus we don't analyse the results of the activity of the health care system, e.g. the health state. In this context, analyses of the former Częstochowa, Katowice and Bielsko–Biała provinces, merged into one region, which is an entirety in administrative terms, but remains divided in historical terms, seem to be of particular importance.

DISTRICTS AND CITIES OF THE SILESIAN PROVINCE

There are 17 districts in the Silesian Province. This province is characterized by a relatively high level of diversity in terms of the number of people living in its districts (the coefficient of variation in 2011 is 43,99%).

The Bieruń–Lędziny district, which in 2011 was home to just over 58 thousand people, is the smallest one. The biggest district in terms of the number of inhabitants was at that time the Cieszyn district, with 176.6 thousand inhabitants [www.stat.gov.pl (24.02.2012)].

One of the characteristics of the Silesian districts is the professional activity of its inhabitants. The largest percentage of employed persons was in the Bieruń–Lędziny district, in which as much as 35.23% of its total population was registered as working people. An alarmingly low percentage of workers (only 14.39%) was recorded in the Rybnik district in 2010 [www.stat.gov.pl (24.02.2012)].

Natural increase of the population was extremely diverse in the Silesian districts in 2010. While the lowest value of this feature amounted at the time to -3.3 births per 1 thousand people in the Zawiercie district, the highest value of 3.5 births was recorded in the Bieruń–Lędziny district [www.stat.gov.pl (24.02.2012)].

Among the many characteristics of the population of the Silesian province, one should mention the number of infant deaths. This feature describes not only the demographic conditions in the region, but it also provides a certain insight into the condition and level of the health care system. Thus, the Będzin district, with 2.9 infant deaths per one thousand live births, could boast the lowest value of the number of deaths in the entire region in 2011[www.stat.gov.pl (24.02.2012)]. In contrast, the highest rate was recorded in the Kłobuck district, with almost 12 deaths per one thousand live births.

The Silesian province has 19 cities with district rights. Among them one can mention such small towns as Świętochłowice (about 53 thousand people in 2011), Piekary Śląskie (about 58 thousand in 2011) or Żory (62 thousand in 2011), and large cities like Katowice (about 309 thousand in 2011), Częstochowa (236 thousand in 2011) or Sosnowiec (about 215 thousand in 2011), which are not the biggest cities in the country, but stand out in their region.

The economic environment in which the Silesian province functioned in 2010 shows that, compared to all the cities of the region, the worst market conditions were prevailing in Świętochłowice, Siemianowice Śląskie and Żory. These cities recorded the lowest percentage of workers in the entire Silesian province, compared to the number of inhabitants.

If one considers the demographic conditions of the functioning of cities with district rights of the Silesian Province, one can say that they are characterized by a similarly large dispersion in relation to the population growth, which was also the case of the districts. The highest value of this feature in 2010 was recorded in Żory, and it amounted to 5.1 births per 1000 people. This value was significantly higher than that recorded in Rybnik, for which the growth rate was 3.1 people per one thousand inhabitants. The lowest value was recorded in Sosnowiec, with the rate of - 3.2 births per 1000 people [www.stat.gov.pl (24.02.2012)].

Compared to the Silesian province districts, the cities of the region were characterized by higher values of the infant deaths, because in Katowice,

Świętochłowice and Rybnik one recorded more than 9 infant deaths per one thousand live births in 2011. The lowest infant death rate among the cities with district rights was recorded in Jastrzębie Zdrój, which had 2.2 infant deaths per one thousand live births [www.stat.gov.pl (24.02.2012)].

RANKING OF DISTRICTS AND CITIES OF THE SILESIAN PROVINCE

Assessment of the functioning of the health care system in the Silesian region, with division into districts and cities with district rights, seems to be plausible, considering the fact that the presently existing provinces (including the Silesian one) are usually a cluster of other provinces, existing before 1999. Is each district and city of the present Silesian province characterized by the same efficiency in the field of health care and can the patients of these regions expect the same access to medical services?

The above question can be answered with the application of numerical taxonomy methods, because it is the numerical taxonomy which provides one with methods to classify and rank the analysed objects.

The notion of taxonomy is derived from Greek, in which *taxis* means ordering, and *nomos* - rule. Basically, taxonomy can be defined as "a scientific discipline, dealing with the rules and procedures of classification (ordering, grouping, discriminating, eliminating, dividing)" [Kolenda 2006, p.17].

The basic concepts in the taxonomy are *object* and *feature*. The districts and cities with district rights (considered as a separate group of objects) were the object of analysis in this paper.

The analysed units were described by several features. These include:

- the number of hospitals per 1 thousand inhabitants (LS)
- the number of hospital beds per 10 thousand inhabitants (LL),
- the number of outpatient care centres, increased by the number of medical practices, per 1 thousand inhabitants (LA)
- the number of drug stores per 1 thousand inhabitants (A)
- the number of day nurseries and nursery units per 1 thousand inhabitants (LZ)
- the number of infant deaths per one thousand live births (ZN)
- the number of places in full-time social welfare centres per 1 thousand inhabitants (PS) and
- the amount of budgetary expenditure of the districts on health care in PLN thousands per 1 thousand inhabitants (W).

The author has chosen the above-mentioned variables, because, due to the availability of data, they describe in relatively thorough (as part of the health care system) and uniform terms the functioning of the health care system within the analysed area [Kozierkiewicz 2000]. All variables are a cumulative value within

each of the 17 districts and 19 cities with district rights of the Silesian province. These features allow one to assess the resources and infrastructure of the health care system in this area, and to connect them with expenditure allocated by the districts for health care protection purposes. Unfortunately, the commonly available data about the districts do not allow one to compare the results and effects of the functioning of the health care system, expressed, for example, by the number of inpatients or outpatients.

The data used in the study cover the year 2007.

In the procedure of classification and ranking of objects, one can distinguish several stages: transformation of features into stimulants, normalization of features, weighting of the feature.

So in the first place it was necessary to transform all the features that were de-stimulants and nominants to stimulants. Almost all analysed features are stimulants, except for the number of infant deaths, which, according to the transformation procedures, was transformed into a stimulant, using the formula:

$$z_i = -x_i, \tag{1}$$

where x_i, are the values of the X feature (de-stimulants), and z_i is its value transformed into a stimulant.

To order the districts and cities of the Silesian province, one has determined for each of them the so-called. synthetic feature, which is a kind of resultant of all analysed features. The synthetic value was calculated based on the following formula:

$$g_i = \sum_{j=1}^{k} x_{ij}^{'} w_j \tag{2}$$

wherex_{ij}'- normalized values for the j-th feature and i-th object, w_j- weights for the j-th feature, k - number of features.

The first stage in determining the value of the synthetic feature was to normalize the features, because the normalization procedures enable one to compare them. In this case, the normalization was achieved by standardization.

In order to determine the degree of importance of each analysed feature, one has given them weights, with the application of a ranking method, based on orthogonal projection. The w_j weights were determined on the basis of the following formula:

$$w_j = \frac{x_{Wj} - x_{Aj}}{d(A,W)} \tag{3}$$

where: x_{Wj} – value of the so-called standard for the j-th feature, x_{Aj}– value of the so-called anti-standard for the j-th feature, $d(A,W)$ - Euclidean distance between x_{Wj} and x_{Aj}.

The literature on the subject describes a few methods of choosing the standard and anti-standard. The selection criterion is the value of the directional variance determined for the value of the synthetic feature [Kolenda 2006, pp. 139-141]. The empirical research conducted by the author shows that the highest value of the directional variance, so the best method allowing one to select the ranking method is the one for which the standard is the value of the third quartile of the normalized features, and the anti-standard - value of the first quartile.

The above procedure was used to order districts and cities with district rights, taking into account eight features analysed in 2008. Table 1 shows the obtained results.

Table 1.Values of the synthetic feature for districts and cities of the Silesian province in 2007

District	**Values of the synthetic feature**	**City**	**Values of the synthetic feature**
Będzin	0,22199	Bielsko-Biała	1,3946
Bielsko-Biała	-0,97432	Bytom	-0,1666
Bieruń-Lędziny	-0,94093	Chorzów	2,1583
Cieszyn	4,82349	Częstochowa	1,3609
Częstochowa	-1,81389	Dąbrowa Górnicza	-1,5072
Gliwice	0,16009	Gliwice	-0,9569
Kłobuck	-0,41443	Jastrzębie Zdrój	-0,5244
Lubliniec	2,81814	Jaworzno	-0,6459
Mikołów	0,27042	Katowice	3,5980
Myszków	-1,12775	Mysłowice	1,1979
Pszczyna	0,71931	Piekary Śląskie	0,1490
Racibórz	-0,53049	Ruda Śląska	-1,8807
Rybnik	-3,38378	Rybnik	-1,7825
Tarnowskie Góry	1,56900	Siemianowice Śląskie	1,5376
Wodzisław	0,08469	Sosnowiec	0,7522
Zawiercie	-0,44409	Świętochłowice	-1,0801
Żywiec	-1,03745	Tychy	-0,7125
		Zabrze	-0,1963
		Żory	-2,6956

Source: own study

In 2007 the Cieszyn district was leading in terms of the functioning of the health care system, with the synthetic indicator exceeding the value of 4, and significantly higher than indicators in other districts. The Lubliniec district, with the value of almost 3, ranked second.

The Rybnik district was the lowest-ranked one in 2007 (the value of the synthetic feature was about -3.4), followed by Częstochowa, Myszków and Żywiec districts, with values of the synthetic feature lower than -1

The ranking of cities is opened by Katowice, for which the value of the synthetic feature, as in the case of the districts, was far higher than that of the other cities, and amounted to 3.6. Chorzów came second in the ranking. The lowest value of the synthetic feature was recorded in Żory (value of around -2.7, Ruda Śląska (value of around -1.9) and Rybnik (value of around -1.8).

CLASSIFICATION OF DISTRICTS AND CITIES OF THE SILESIAN PROVINCE

Classification is a very important concept in taxonomy. This broad concept is connected with the "methodology of sorting a set of objects as well as the process of classification itself" [Zeliaś A (ed.) 1991, s.75]. Therefore it boils down to the division of all analysed objects into groups, so that each group will contain only homogeneous objects. Such determination of typological groups allows one to identify similar objects , to classify them into one group and to assign different objects to other clusters.

To arrange districts and cities with district rights into groups operating and functioning in the health care sector, the author has applied the concentration method, and more precisely the Ward's method. It is a linking method in which one seeks to minimize the sum of squares of any two clusters and to connect objects, so that the value of the intra-group variance of features describing the objects in the created clusters can be as low as possible.

The Ward's method is considered to be an agglomeration combinatorial method, in which each analysed object is initially treated as a single cluster, so that one can reduce the number of such groups in subsequent analyses, combining the previous clusters into the so-called groups of higher rank. The entire process ends with the creation of a single cluster, which includes all the analysed objects.

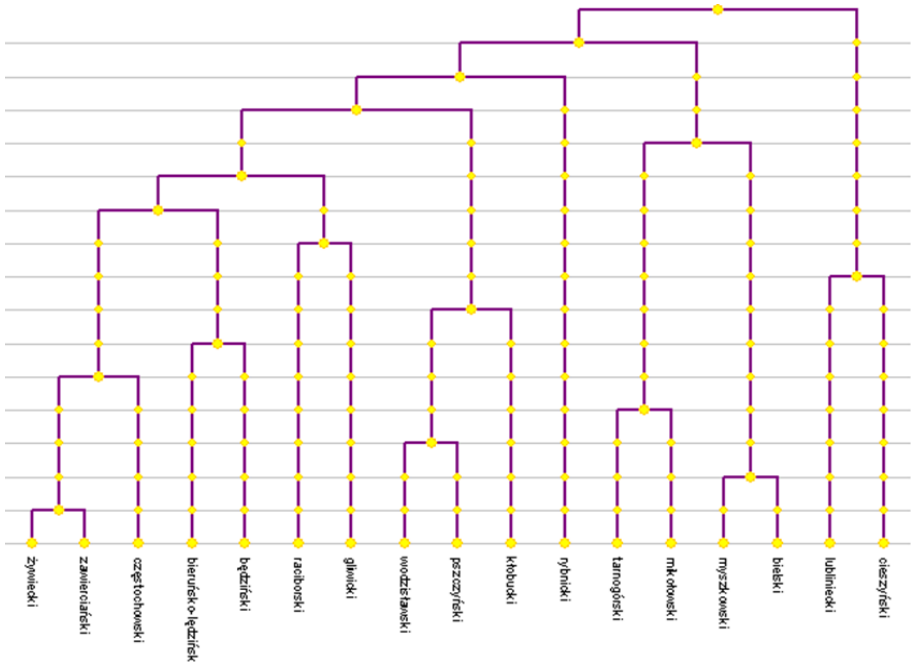
The results of clustering are usually presented in a graphic form with the so-called dendrite tree (dendrogram). Dendrite presents formation of subsequent clusters of increasingly higher orders with specific binding distances, which shows similarities and differences between the analysed objects from the point of view of the considered features.

Therefore districts were grouped first, starting from a one-element cluster, through a cluster which connects most similar districts, and ending with one, that connects all the tested objects.

District analysis results are shown in a dendrogram presented in Figure 1. In a group of 17 districts, one can identify (for example, on the level of link 11) seven groups. The first group consists of the following districts: Żywiec, Zawiercie, Częstochowa, Bieruń–Lędziny and Będzin. They are characterized by

the synthetic feature value oscillating around -1. The second group consists of the Racibórz and Gliwice districts, which seem to be better off than districts of the first group in terms of the analysed features. Third cluster consists of three regions: Wodzisław, Pszczyna and Kłobuck.

Figure 1.Dendrogram for districts of the Silesian province



Source: own study

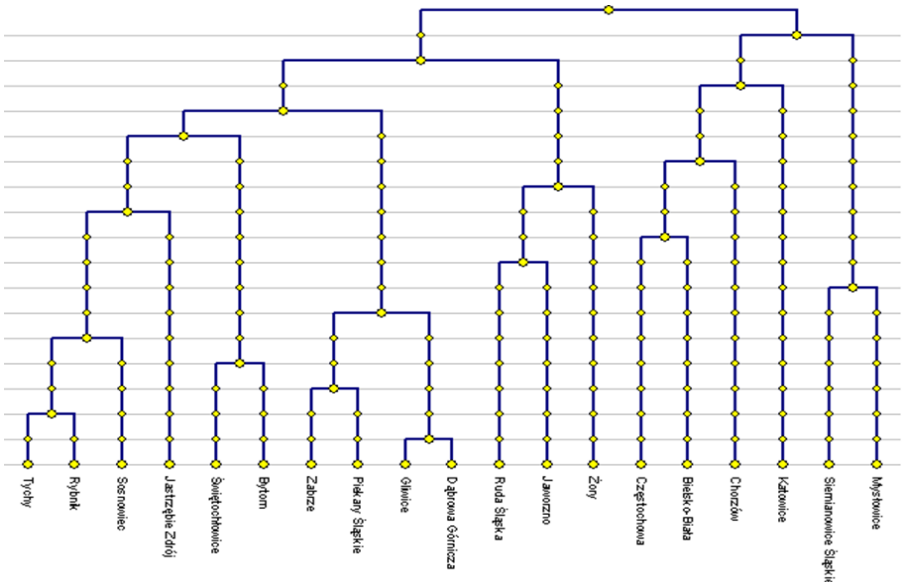
Another group includes only one district (the Rybnik district), differing significantly from the other districts, which is considered to be the worst in that group (Table 1). The next two clusters are formed by Tarnowskie Góry and Mikołów districts, and Bielsko–Biała and Myszków districts, respectively. The best functioning districts of the Silesian province include districts from the seventh group, namely: the Cieszyn and Lubliniec districts.

Cities of the Silesian province were subjected to a similar analysis, and classification and their results are presented in Figure 2. Analysis of the Silesian cities allows one to distinguish (e.g. on level 13) seven clusters of cities.

One cluster includes only one city, namely Katowice, which, according to an earlier stage of the analysis (Table 1), was considered to be the best in this group. A separate group includes Częstochowa and Bielsko–Biała, which seems to be important due to the analysed implications of incorporating these cities into the Silesian province. These cities are joined on higher level of linking by Chorzów, and finally by Katowice.

According to the calculations the results of which are presented in Table 1, the worst city in terms of the functioning of the health care system in Silesia was Żory, around which, in accordance with the classification rules, Jaworzno and Ruda Śląska were also clustered.

Figure 2.Dendrogram for cities with district rights



Source: own study

Similarities in terms of the analysed features were also noticed, in accordance with the conducted study, among Zabrze, Piekary Śląskie, Dąbrowa Górnicza and Gliwice. Separate clusters are formed by Bytom and Świętochłowice, Jastrzębie Zdrój and Sosnowiec, Rybnik and Tychy, as well as Siemianowice Śląskie and Mysłowice.

CONCLUSIONS

Application of quantitative methods in the analyses of not only the health care sector, but also of the wider economy, seems to be an important method of reading and noticing certain regularities and rules. Such analyses are of special importance in a sector in which the human health and life are at stake.

In determining relations and interactions between the objects (which can include health care institutions, branches of the National Health Fund or regions of the country), the numerical taxonomy, using quantitative methods, can be of utmost assistance. Such an approach allows one not only to establish relations between the analysed objects, but also to order and synthesize them, which in turn may become helpful in the description of a given phenomenon.

This paper focuses on the analysis of the Silesian province, taking into account available data on the functioning of the health care system in the region with respect to infrastructure of the system only. Districts and cities with district rights constitute the analysed objects.

The results of ranking of districts and cities with district rights indicate fairly significant differences between the analysed objects in terms of their functioning within the health care system. The Cieszyn district can boast the best conditions for the health care system functioning. The Rybnik district comes last. Katowice is the ranking winner. Żory has come in last. Results for the best districts and cities could be explained by the highest value of the analysed features. Similarly, the worst regions had the lowest value of the features. It is worth noting that Częstochowa and Bielsko–Biała, which prior to 1999 used to be provincial cities, in 2007 were functioning in a similar way, when one takes into account the analysed properties of the health care system. Moreover, Katowice, the best city in the region, can also be included in the cluster of these cities. Thus one can make an assumption that creation of the Silesian province in 1999 did not worsen the situation of the former provincial cities, which were incorporated into the province. The same can be said about the districts of the Silesian province, with the Lubliniec and Cieszyn ones occupying higher places than the former districts of the Katowice province.

Although the study results show an existing spatial dispersion of the analysed features, one should be satisfied with the fact that Silesia was the best Polish province in the years 2003-2005, based on similar studies using similar numerical taxonomy methods [Strzelecka, Nieszporska 2009, pp . 91-102.].

REFERENCES

Kautsch M., Whitfield M. (2001) Zdrowie i opieka zdrowotna – zagadnienia uniwersalne i przypadki szczególne, [w:] Kautsch M., Whitfield M., Klich J.(red.), „Zarzadzanie w opiece zdrowotnej" Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków.

Kolenda M. (2006) Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Kozierkiewicz A. (2000) Znaczenie wybranych wskaźników dla podejmowania decyzji w ochronie zdrowia. Część I, Zdrowie Publiczne 2000, suplement 1.

Podstawowe dane z zakresu ochrony zdrowia w 2006 r., Informacje i opracowania statystyczne, Zakład Wydawnictw Statystycznych, Warszawa 2007.

Strzelecka A., Nieszporska S. (2009) Metody taksonomiczne w ocenie funkcjonowania systemu ochrony zdrowia w Polsce, [w:] „Kierunki rozwoju systemu ochrony zdrowia w Polsce" red. Naukowa Ewelina Nojszewska, Szkoła Główna Handlowa w Warszawie, Warszawa, str. 91-102.

Zdrowie i ochrona zdrowia w 2011 r., Główny Urząd Statystyczny, Warszawa 2012.

Zeliaś A. (red.) (1991) Ekonometria przestrzenna, PWE, Warszawa.

www.stat.gov.pl

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 139 – 147

THE ROLE OF INFORMATION SYSTEMS IN LOGISTIC ENTERPRISES

Maria Parlinska, Iryna Petrovska
Department of Agricultural Economics and International Economic Relations
Warsaw University of Life Sciences – SGGW
e-mail: maria_parlinska@sggw.pl

Abstract: Well-developed information channels play very important role for the institution. In this paper had been researched the Ukrainian wholesale market for purpose to make the comparison between information distribution channels on the chosen agricultural wholesale markets in Poland and Ukraine. Also had been defined the recommendations for future development of information distribution systems for purpose to improve the market activity.

Keywords: wholesale market, information distribution channels, Analytical Hierarchy Process

INTRODUCTION

Role of wholesale markets is very important for the country. Now wholesale markets situated not only on specific location but also in the Internet. Many wholesale markers create their own web-sites where there is a possibility to provide online trading. Owing to well-developed information systems this type of trading works very well and makes the trade easier. Now sellers and buyers can find easily each other. They don't need to spend time for looking for buyers and at the same time they reduce their needs for warehousing that reduces costs.

If company work with specific product (for example flowers), there could appear also problem with time between product tracing from the buyer to seller. That is why it is very important for seller to find the buyer as fast as possible in purpose to have production still fresh.

Flower production is very specific and need special care. Flowers should have special condition as for transportation as for warehousing as well. There is two ways of making them fresh as long as possible: to use chemicals and special

equipment or to bring them to buyers immediately. If company needs warehousing, thereby its costs increase.

It is also important to take into account costs benefits of the information systems for the company. If company has information system, then it doesn't need to spend big money for the staff as IS makes everything by itself. Company concentrates mostly on production work and not on calculations and so on.

DEVELOPMENT OF FLOWER MARKET IN POLAND AND UKRAINE ON THE EXAMPLE OF BRONISZE AND STOLYCHNYI WHOLESALE MARKETS

Brining to the buyers immediately need from the seller so called skills of "fast finder of buyers". As the example, where it is possible to find such things work, we can take the Flower Auctions in Holland. Owing to their information systems, buyers even can sell their production directly from their home. There is no faster way for meeting a buyer with the seller. Owing to well-developed information its became possible for our purpose. Thereby members save their time. They can find each other faster, more transactions can be made and as the result – profit will be higher.

Such kind of online trading should be introduced in such countries as Poland and Ukraine. For Poland it would be easier, because here exist well-developed wholesale markets with good infrastructure. Such markets just need some additional technologies to create flower auctions.

In Ukraine the situation is slightly different. Here even a lot of web-sites need further improvements. Some wholesale markets don't have possibility to display the information from the web-page in English language. But other try to use some IS for purpose to make trade better. Such kind of wholesale markets will be described in this article.

Stolychnyi wholesale market just started to work. There are not enough infrastructure facilities. Building process of flower departments still isn't finished. There are four different pavilions. At the market there is trade of vegetables and fruits, also fresh fish and meat.

Project in the market Stolychnyi is to create the competitive European place for supply and demand meeting. There are presented prices from the last trade. Also here are contacts of members of the market with information about them. What is good point – it makes easier for sellers to find the buyers(they just can use information from page – unfortunately only if they know Ukrainian language).

Table 1. Comparison of distribution channels in Ukraine and Poland wholesale markets

Characteristics	Stolychnyi wholesale market, Ukraine	Bronisze wholesale market, Poland
Web-site possibilities	Just Ukrainian language available	Well-developed
Displaying of prices	No historical data of prices, presents of current prices	Presence of historical data of prices, presents of current prices(only for registered users)
Description of activity	There are presented full description about institution activity	There are presented full description about institution activity
Marketing	The market use radio, newspapers, TV, Internet and other sources to promote their activity. Also there is presented information about project implementation progress	The market use radio, newspapers, TV, Internet and other sources to promote their activity
Contacts	Well-organized	Well-organized
Departments	Some departments aren't finished, description presented	Whole description presented. Each department works successfully
Flower department	In progress(doesn't work yet)	Working well
Description of members	Full description about members only from the pavilions with fruits and vegetables	Full description about members
Direct Selling from cars	Available	Available
Number of flower firms	Information isn't presented	Around 60

Source: own work

INFORMATION SYSTEMS AND TRANSACTION COSTS

It has a sense to introduce the ITC only when the costs of ITC are less than transacting costs.
ITC can help the enterprise to reduce:

- Searching costs (ITC can make searching process faster and better, especially when there are published too much information about the market. This situation is very common in the wholesale markets. It's only effective in the situation when the searching costs are higher than the costs for information gathering by ITC)
- Negotiation costs (ITC can make sharing of information between actors more successful and decrease the information asymmetry).
- Enforcement costs

For this purpose Distance sales service is presented in the market Stolychnyi.
In the wholesale market Stolychnyi is presented distance service which provides possibility for traders to save their time and money. Actors are able to stay at their place and don't need to visit the wholesale market.
There are presented the following advantages of such service:

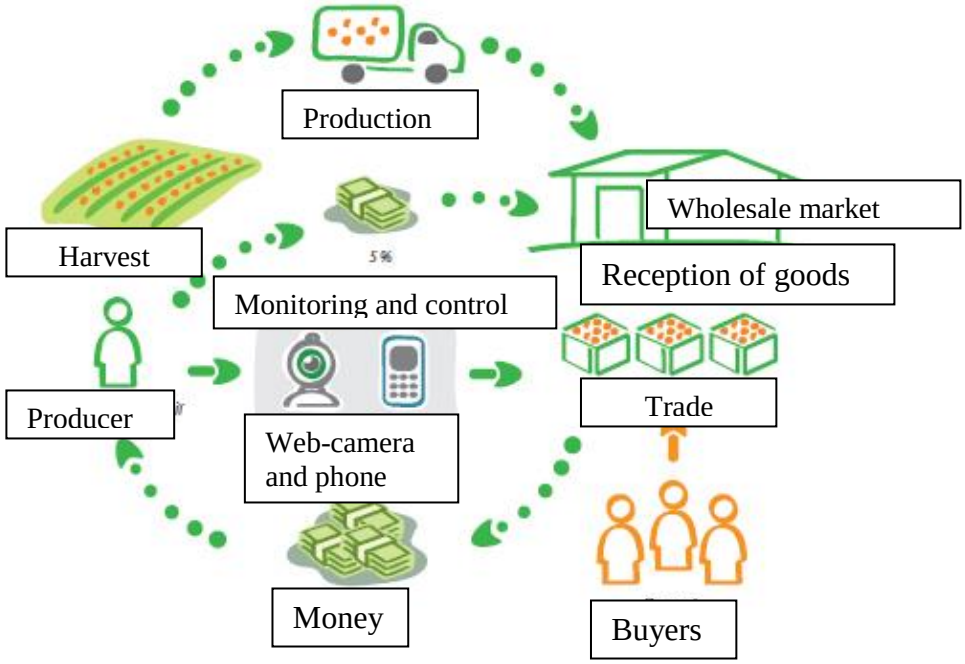
1. The wholesale market price is higher in comparison with price, set when grower sells directly from the field.
2. New distribution channel
3. Costs of this service are not exceed 5% of production costs, what is lower in comparison to costs of other distribution channels
4. Wholesale market is interested to made transactions as fast as possible.
5. Promotion on firm on the biggest trade fairs in Ukraine.
6. Seller has complete information about market situation connected with prices, supply, demand, and quality of production, alternative offers and competitors.
7. Convenience to work through the Internet on PC or cell phone. Control of the process in online regime.
8. Market responsibility of production taken for selling
9. Possibility of warehousing

Wholesale market guarantees:

1. Information about production sold during last 3 days
2. Place for selling goods, professional sellers and loaders
3. Control of realization process in online regime with web-camera
4. Possibility of interactive price bargaining
5. Information about market prices
6. Daily reports about quantity of sold production
7. Information and marketing support

This new service will give possibility for sellers to decrease their transaction costs and as a result –to increase their benefits. The scheme of this process is shown on the graph below.

Figure 1. Information distribution system in the market Stolychnyi



Source: web-page of wholesale market Stolychnyi

Another important service offered by market Stolychnyi is the possibility to buy online on web-page www.zakaz.ua. This web-page is better made in comparison with market one. There is also a possibility to choose English language, observe prices and supply of production. This is online distribution channel which just started to work.

To be able to use all these services the company needs:

- Software (on cell phone, computer and similar access device)
- Access to the Internet

Both of these companies possess such devices.

In the articles there had been made the estimation of quality service of information distribution channels in the wholesale market Stolychnyi.

For this purpose were used statistical method which called Analytical Hierarchy Process. This method was developed by Tomas Saaty and it gives possibility to evaluate qualitative data in quantitative values.

The following hierarchy was used:

- In the first step the goal should be created,
- Establishing of criteria and their ranking,
- Definition of the alternatives.

The scales from 1 to 9 were used for purpose to evaluate criteria and alternatives.
The theoretical sense of the AHP method can be presented as followed:
There is known matrix (n x n) of criteria' preferences

$$A = [a_{ij}]$$

with diagonal with value 1, characterised by

$$a_{ji} = 1 / a_{ij}$$

Vector of priorities:

$$\begin{bmatrix} w_1/w_1 & w_1/w_2 \ldots & w_1/w_n \\ w_2/w_1 & w_2/w_2 \ldots & w_2/w_n \\ w_n/w_1 & w_n/w_2 \ldots & w_n/w_n \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_n \end{bmatrix} = n \begin{bmatrix} w_1 \\ w_2 \\ w_n \end{bmatrix}$$

The principle of determination of the vector of priorities is to consider each lement of the matrix of criteria' preferences

$$a_{ij} = \frac{w_i}{w_j}, w_i, w_j$$

"W" – Weight of the certain estimation relative to i^{th} and j^{th} criteria. The demand of transitivity is fulfilled by cohesion of estimators. It means: if the element I is preferred against j and element j against k, element I is preferred against k.

$$Aw = nw$$

$$Aw = \lambda w$$

$$(A - \lambda I)w = 0$$

$$\det(A - \lambda I) = 0$$

$$\lambda_{\max} = n$$

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

$$CR = \frac{CI}{R}$$

If deviation between λ. max and value n estimated by CR pass the possible limits, the estimation of validity of criteria must be done again.

T. L. Saaty recommends giving to the critical value CR value 0.1. If CR is less or equal 0.1, it can be said that correspondence exists. If CR is more than 0.1, then the analysis must be repeated.
The following criteria of importance were chosen based on the research made by prof. Parlinska:

- Reliability,
- Availability,
- Costs.

There were made ranking of criteria of importance. There were made questionnaires between experts to create it. (Parlińska 2008)

Table 2. Weights of criteria of importance

Source of information(criteria)	Vectors of priorities
Reliability	0,54
Availability	0,297
Costs	0,163
Sum	1

Source: [Parlińska 2008]

Consistency index is equal to 0,01. The sum of these weights must be equal to 1.
According to all these criteria the following distribution channels were evaluated:

- Internet
- Radio
- Press
- TV
- Phone service

Table 3. Ranking of alternatives based on criterion "Reliability"

Source of information (alternatives)	Vectors of priorities
Internet	0,7
Radio	0,001
Press	0,01
TV	0,001
Phone service	0,288

Source: own calculations

Table 4. Ranking of alternatives based on criterion “Availability”

Source of information (alternatives)	Vectors of priorities
Internet	0,6
Radio	0,007
Press	0,002
TV	0,001
Phone service	0,39

Source: own calculations

Table 5. Ranking of alternatives based on criterion “Costs”

Source of information (alternatives)	Vectors of priorities
Internet	0,5
Radio	0
Press	0
TV	0,5
Phone service	

Source: own calculations

Table 6. Final ranking of the information sources

Source of information (alternatives)	Final vector of priorities
Internet	0,6377
Radio	0,002619
Press	0,005994
TV	0,000837
Phone service	0,35285
Sum	1

Source: own calculations

According to the results in table Internet service is the most developed and together with phone service. For clients these sources of information are the most available. The questionnaires were made among people from wholesale market Stolychnyi. The sample contained of 30 responders from different ages.

CONCLUSIONS

In according to the conducted research, it is necessary to remark that Warsaw wholesale market is better developed than Kiev one. But at the same time market Stolychnyi develops from day to day better and better. Building process is going well. Even in this situation trading works there. Only one thing should be changed for purpose to get international level, it is the language of web-site. At least English language must be added.

Possibility of online registration is very important for both countries. For such purpose the good Information System is required. For market Stolychnyi it is a bit early to introduce, but for Bronisze market it is already good time. Bronisze market has well-developed infrastructure and can concentrate its activity on making, for example, online auctions.

If the company will have the opportunity to register online, it will save a lot of time and reduce the costs.

Online trading gives advantages for the company such as finding sellers or buyers, trading directly from the home or working place.

Also it would be useful to create mobile version of web-site. Owing to this, participants will be able to work from each place all over the world. Company will be able to find more potential buyers, check demand and supply, prices and so on. It can make decision making process easier and faster.

REFERENCES

Cordella A. (2006) Transaction costs and information systems: does IT add up? Journal of information technology, 21 (3). pp. 195-202

Parlińska M. (2005) Hierarchy of Decision Making Process. [w:] Monografia: Studia Universitatis Babes-Bolyai; Oeconomica, Printed by Economics Faculty, Babey's Bolays University of Cluj Napoca, Romania, s. 73-78.

Parlinska M (2008) Rola informacji w gospodarce rynkowej na podstawie wybranych rynków rolnych, Wydawnictwo SGGW, Warszawa

Saaty T.L. (2000) The Analytic Hierarchy Process, New York: McGraw Hill. International, Translated to Russian, Portuguese, and Chinese, Revised editions, Pittsburgh: RWS Publications.

http://www.kyivopt.com/

http://www.bronisze.com.pl/

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 148 – 157

ECONOMIC SITUATION OF EASTERN POLAND AND POPULATION MIGRATION MOVEMENT

Michał Bernard Pietrzak[1]
Department of Econometrics and Statistics
Nicolaus Copernicus University in Toruń
e-mail: Michal.Pietrzak@umk.pl
Justyna Wilk
Department of Econometrics and Computer Science
Wrocław University of Economics
e-mail: justyna.wilk@ue.wroc.pl
Mariola Chrzanowska
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
e-mail: mariola_chrzanowska@sggw.pl

Abstract: The aim of this paper is to examine population migration flows in relation to the eastern Poland and its economic situation. The empirical study refers to the period 2008-2010 corresponding to global financial and economic crisis which affected the intensity and directions of internal migrations in Poland. Ratio analysis, as well as taxonomical analysis was applied in the research

Keywords: economic situation, internal migration, synthetic measure, ratio analysis

INTRODUCTION

A host of modern economic theories indicate that regional development is conditioned by the occurrence of regional networks. Among various regional development theories the centre-periphery theory deserves special attention (see: Hryniewicz (2010); Baldwin (2001)). It assumes the dependence of development

[1] The project was co-financed by Nicolaus Copernicus University in Toruń within the UMK research grant no. 1481-E.

between the central and peripheral areas taking into account also the semi-periphery condition. The central region supports the development of peripheral areas. The centre is usually created by a small, though dynamically developing, territorial area where both economic activity and social life is concentrated. Peripheral areas, in turn, are less developed and are exploited by the core. As a result, the development of the peripheral area is totally dependent on the centre (see [Pain 2008], [Kauppila 2011], [Wojnicka et al. 2005]).

Polish regions (NUTS 2) that form the so-called eastern bloc (the Podkarpackie, Lubelskie, Podlaskie and Świętokrzyskie provinces) constitute the periphery of the Polish economic space and are referred to as 'Poland B'. Such a situation results in the main from historical conditions and peripheral location. The social and economic backwardness of eastern Poland necessitates undertaking activity that would stimulate both economic and social development of that region. For that reason the provinces are of special interest to the national developmental policies [Jakubowska 2012], [Małkowski 2011], [Celińska-Janowicz et al. 2010]. Provincial capital cities of eastern Poland function as important regional centers. The major functions they play are administrational, transgenic, and educational[2]. Despite the significant contribution of capital cities to the settlement network, an increase in negative social and economic phenomena can be observed. These phenomena include mass migrations which may result in depopulation of that area[3].

ECONOMIC SITUATION OF EASTERN POLAND

Eastern Poland's macro region comprises the area of approximately 75,000 km^2, which constitutes almost 25% of the country's total area. That area is inhabited by almost 6.8 million people which is about 17.5% of the total population. More than a half of eastern Poland's dwellers live in rural areas. The Podkarpackie, Lubelskie, Podlaskie and Świętokrzyskie provinces belong to the poorest regions within the European Union. Their gross domestic product *per capita* is much below the national average and the unemployment rate is much above the national average. These areas are the least competitive and the problems they are facing include attracting specialists and retaining the best educated inhabitants. The major impediments to the development of the area include poor infrastructure, an ineffective employment structure, an ineffective agricultural

[2] See, for instance, Dziemianowicz, Szlachta and Szmigiel-Rawska (2011), pp. 37-38.

[3] The content of the paper is the continuation of the author's studies of the problem of migrations and the findings of the studies are presented in the works Bal-Domańska, Wilk (2011), Matusik, Pietrzak, Wilk (2012), Pietrzak, Drzewoszewska, Wilk (2012), Pietrzak, Żurek, Matusik, Wilk (2012), Wilk, Pietrzak, Matusik (2013), Pietrzak, Wilk, Matusik (2013), Pietrzak, Wilk (2013), Wilk, Pietrzak (2013).

sector, which dominates in that area, poorly developed services and industry, and poor quality of human resources [Korcelli P. et al. 2008].

A typical feature of the macroregion is the low level of entrepreneurship. The lowest one can be seen in the Podkarpackie while the highest in the Świętokrzyskie province. The low level of the development of the industrial sector can be explained by both weak internal demand and transport network.

Despite the general poor economic situation of the macroregion, if compared with the rest of the country, eastern Poland's area is not economically homogeneous. In order to take a closer look at the economic situation of eastern Poland, the taxonomic development measure (TDM) has been applied (see, for instance [Zeliaś 2000, 2004]). The values of the measure have been designated by means of Hellwig's method (1968). The study comprised Poland's 66 subregions (NUTS 3) and focused on their situation in 2008 which was described by a set of diagnostic characteristics (see Table 1).

Table 1. Values of selected statistics of diagnostic characteristics

No	Variable	Minimum	Maximum	Arithmetic mean	Coefficient of variation (%)
1	Gross value added per employed person (PLN)	48712,00	136953,00	73434,34	22,71
2	Natural person conducting economic activity per 100 working-age persons (number of person)	7,00	19,00	11,22	22,29
3	Share of commercial companies with foreign capital per 100 national economy entities in the private sector (REGON – private sector) (%)	0,28	5,97	1,25	77,22
4	Investment outlays in enterprises per capita (PLN)	1053,00	12024,00	3157,64	60,64
5	Share of persons employed in the service sector (market and non-market) in employed persons (%)	27,01	83,31	46,78	26,60
6	Average monthly gross wages and salaries (PLN)	2401,06	4504,85	2827,96	14,23
7	Share of registered unemployed persons in working-age persons (%)	1,50	13,20	6,27	42,58

Source: own calculations based on data from LDB of CSO of Poland.

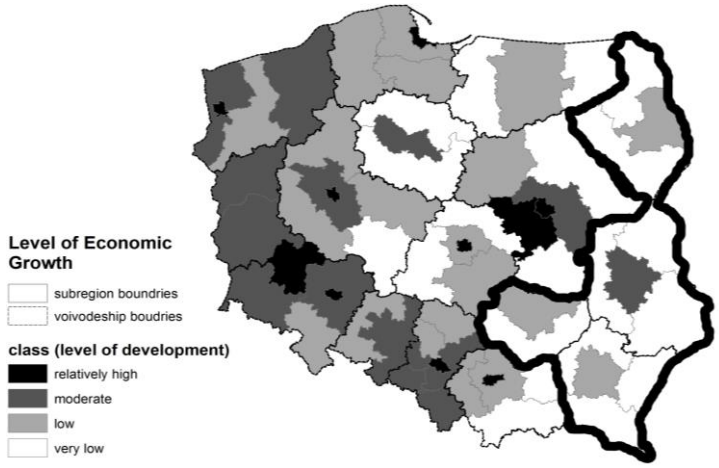
Subregions were grouped in four classes based on the values of the TDM using the 3-means method. The defined classes illustrate relatively high (over 0.4), moderate (0.3, 0.4], low [0.2, 0.3] and very low (under 0.2) levels of economic development (see Figure 1).

The highest value of the measure (0.99) was found for the capital city of Warsaw while the lowest value (0.09) was obtained by the Chełmsko–Zamojski subregion of the Lubelskie province. The highest level of economic development within eastern Poland was observed in the Lubelski subregion. The value of the calculated measure (0.29) allowed the determination of the development level for that area as moderate. A slightly lower value of the TDM was obtained for the Białostocki subregion (0.28), which was classified in a group of subregions characterized by a low level of economic development. The Rzeszowski and Kielecki subregions were also included in that group with the measure value

amounting to 0.23. Other subregions of eastern Poland were included in the group with the lowest level of economic development.

A factor that 'stigmatizes' the area of eastern Poland is the significant migration outflow. As was presented in the Social and Economic Strategy for the Development of Eastern Poland, the area that is most susceptible to the depopulation processes is Lubelszczyzna and the prediction is that as many as 80,000 people will have abandoned it by 2020.

Figure 1. The level of economic development by TDM value classes



Source: elaborated by the authors based on data from LDB of CSO of Poland.

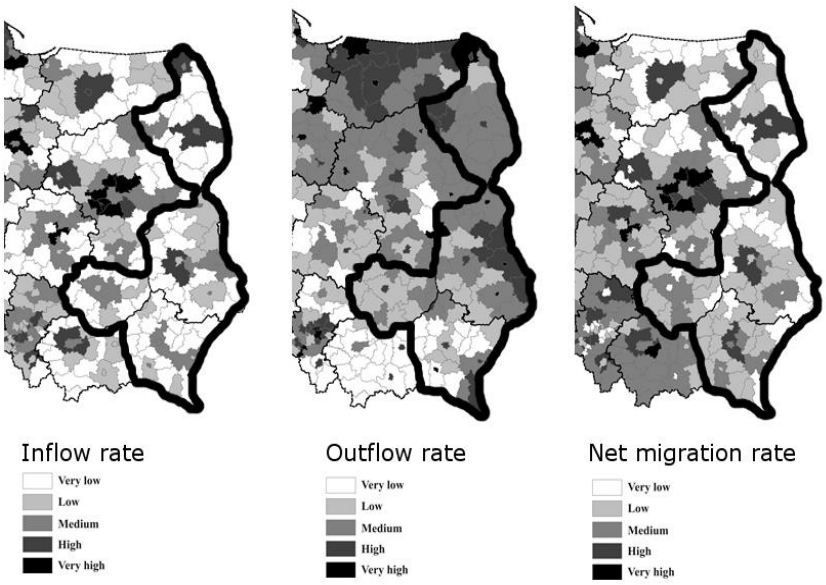
MIGRATION FLOWS OF EASTERN POLAND

Data on internal migration flows in Poland (registered inflows and outflows for permanent residence) were taken from the Local Data Bank of the Central Statistical Office (BDL GUS). The data aggregated for the period 2008-2010 were used to compute the inflow, outflow and migration ratios for districts as well as the net migration ratio for districts situated in the area of eastern Poland[4]. The values of the ratios were divided into five classes following the natural division method. The classes were respectively named as 'very low', 'low', 'medium', 'high', and 'very high' values of the ratio.

[4] The ratios of inflows (outflows) were computed as the share of persons registered (deregistered) in the years 2008-2010 in the average population size for that period. Net migration ratio, in turn, determines the relation of the net migration ratio to the average population size in the period of 2008-2010.

Analysis of interregional (within subregions) migration flows and intraregional (between subregions) migration flows was also made[5]. Values of the interregional flows ratio were divided into three equally numerous classes[6]. Relative to the size of flows, classes were named as 'weak', 'medium' and 'strong'[7]. In the case of the intraregional flows ratio, five groups of flows were marked on the map and these are as follows: 'very strong flows', 'strong flows', '1st class medium flows', '2nd class medium flows' and '3rd class medium flows'[8]. Figure 2 shows the values of the inflows and outflows ratios, and also net migration ratio.

Figure 2. The size of the migration inflow and outflow by districts in the period 2008-2010



Source: elaborated by the authors based on data from LDB of CSO of Poland.

[5] The basis for calculations for the 66 subregions was 66 interregional flows and 4,290 intraregional flows. The values of the ratios were computed as the aggregated value of migrations flows in the years 2008-2010 in relation to the average population size in a destination subregion (the current place of residence).

[6] Classes were created based on specific values of centiles (min, C_{33}), (C_{34}, C_{66}), (C_{67}, max).

[7] The values of interregional flows are not provided for cities with district rights since population flows between districts of those cities are not treated as migration flows.

[8] The classes were distinguished based on the largest values of the ratios with a view to illustrating the directions of the largest migration flows. Specified centiles were taken for the purpose of the creation of the following classes: 'very strong flows' which represent 2% of the most intensive flows within the country and their values belonged to the (C_{98}, max) interval. The 'strong flows' represent (C_{96}, C_{98}), '1st class medium flows' (C_{94}, C_{96}), '2nd class medium flows' (C_{92}, C_{94}), and '3rd class medium flows' (C_{90}, C_{92}).

A high value of the inflow ratio was noted merely within the area of the Białostocki district. This is related to a strong population migration from rural areas primarily caused by economic reasons. In the city of Białystok the value of the ratio was determined on the average level since city dwellers move to the city outskirts, to the adjacent districts. A similar level was noted in the Łomżyński district. Other districts had very low values of the ratio.

The majority of districts of the Podlaskie province noted average levels of the outflows ratio. A high value of the measure can be observed in the city of Białystok and in the Łomżyński and Kolneński districts. The latter is also characterized by a high negative value of the net migration ratio. The same group (with very low values of the ratio) includes the following districts: Grajewski, Sejnerski, Sokólski, Wysokomazowiecki, Siematycki, and Hajnowski.

A high (positive) value of the net migration ratio was noted only in the Białostocki district. The Łomżyński district had a medium value of the ratio. A low level of the net migration ratio was observed in the following districts: Suwalski, Augustowski, Moniecki and in Bielski. Other districts were characterized by very low values of the ratio.

A high level of the inflows ratio was noted in the Lubelski district. However, the city of Lublin undergoes the suburbanization process and in that area the level of the measure is defined as low. The same group (with a low inflows ratio) includes the Łęczycki, Włodawski, Parczewski, Bielski, Rycki and Zamojski districts. In the Lubelskie province, two districts (the Świdnicki and Chełmski districts) reached average levels of the ratio. The value of the analysed measure in the remaining districts (Łukowski, Radzyński, Lubartowski, Puławski, Opolski, Kraśnicki, Janowski, Biłgorajski, and Hrubieszowski) was determined at very low level.

It must be noted that the Lubelskie province is deprived of districts with the outflows ratio on very low level. A low level of the ratio can be observed in the following districts: Lubartowski, Puławski, Kraśnicki, Biłgorajski, Janowski and Krasnystawski. The average value of the measure was ascribed to the Łukowski, Bialski, Opolski, Lubelski, Zamojski and Tomaszowski districts.

High values of the ratio can be noted in the districts situated within the border-line area: Włodawski, Chełmski, Hrubieszowski and Łęczyński. The only district with very high value of the outflows ratio was the Rycki district, which is also characterized by very low value of the net migration ratio. Similar level of the latter ratio was noted in the following districts: Łukowski, Radzyński, Parczewski, Włodawski, Hrubieszowski, Tomaszowski and Opolski.

None of the districts of the Podkarpackie province could boast an inflows ratio at a high level. Medium values of the ratio held for the Przemyski and Rzeszowski districts, and low values can be seen in the Krośnieński, Leski and Tarnobrzeski districts. The remaining districts were characterized by very low value of the inflows ratio.

The Bieszczadzki district is undergoing the depopulation process. The district has very high level of the outflows ratio and, at the same time, very low inflows ratio. As a result the number of dwellers in that area is falling. Similar to other provinces of the eastern part of Poland under scrutiny, the Podkarpackie province is facing the depopulation process of urban areas and moving to peripheral areas. Such cities as Rzeszów, Przemyśl, Tarnobrzeg and Krosno noted very high value of the outflows ratio.

At the same time in those districts (excluding the urban area) the value of the ratio is determined as medium (the Tarnobrzeski and Przemyski districts), very low (the Krośnieński and Tarnobrzeski districts) and low (the Rzeszowski district). The group of districts with low values is formed by the Stalowolski and Przeworski districts. The Leski district had a low value. The level of the ratio for the remaining districts (Mielecki, Kolbuszewski, Dębicki, Jasielski, Sanocki, Ropczycko-Sędziszowski, Strzyżowski) was very low.

As concerns the Świętokrzyskie province, the only province with a high ratio value was the Kielecki district. Medium values were noted in the Buski and Ostrowiecki districts. The level of inflows ratio for other districts is classified as low. The whole of the Świętokrzyskie province is characterised by a significantly high level of the outflows ratio.

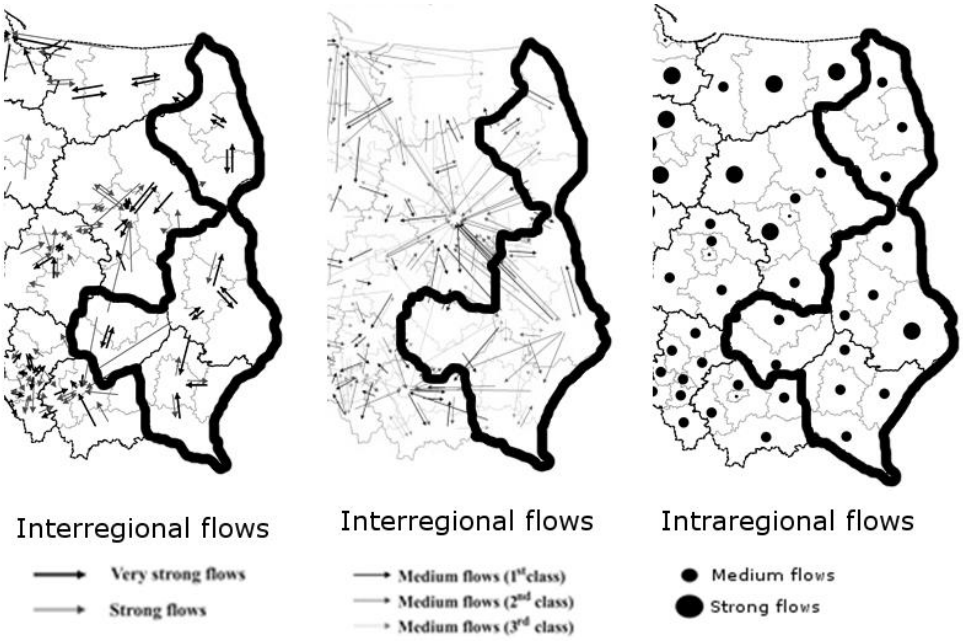
In the city of Kielce that level is classified as high, while in the Opatowski, Sandomierski, Kazimierski and Starachowicki districts the outflows level was determined as low. Very low value of the ratio was identified in the Sandomierski and Skarżyski districts. The whole of the region is characterized by a low net migration ratio. The Kielecki and Buski districts are the only ones where the ratio was on the average level.

Figure 3 presents the directions of the largest migration flows (outflows and inflows) in eastern Poland and the size of flows within subregions in that part of the country. In the Podlaskie province one can see a significantly high level of interregional flows in the Białostocki subregion and a medium level in the remaining subregions. The level of intraregional flows for the subregions of the Podlaskie province was estimated as strong. Strong migration flows were noted in the Chełmsko–Zamojski subregion, while the remaining subregions (Puławski, Lubelski and Bialski) were characterized by a medium level of the phenomenon.

All subregions in the Podkarpackie province had intraregional flows of medium intensity and very strong interregional flows directed to the centre of the province – the Rzeszowski subregion. No major migration flows were noted in the Świętokrzyskie province. Also, slight changes can be observed while analysing intraregional flows. In that region they were classified in the medium-class group.

The major destination for the emigration from the Podlaskie province is the capital city of Warsaw; in the case of the Podkarpackie province the main migration flow is directed towards Cracow. Dwellers from the Lubelskie and Świętokrzyskie provinces migrate to both Warsaw and Cracow.

Figure 3. The size of interregional and intraregional migration flows by subregions in the period 2008-2010



Source: elaborated by the authors based on DL GUS data

CONCLUSIONS

In the contemporary world the following become increasingly important factors of economic growth: entrepreneurship, innovation, and the adaptive capability to changing conditions. The role of migrations in those processes cannot be underestimated. Migrations, however, may lead to a permanent change in the place of residence in another region, or country. If emigration occurs on a mass scale and encompasses the most active, talented and well-educated people, then it may threaten both the development of a region as well as of the whole country (see [Grabowski 2005]).

Poland belongs to a group of EU states with regional bipolarity of demographic processes. On the one hand, one can identify areas with intense migration inflows, while on the other hand, there are areas characterized by mass migration outflows. Migration outflows can be observed mainly in peripheral areas (peripheral refers to both the spatial and socio-economic context), including the macroregion of eastern Poland. However, the region is significantly internally differentiated. The demographically active areas are ones situated in the near vicinity of provincial cities and almost the whole of the Podkarpackie province. The remaining areas of the region are characterized by migration outflows.

REFERENCES

Bal-Domańska B., Wilk J. (2011) Gospodarcze aspekty zrównoważonego rozwoju województw – wielowymiarowa analiza porównawcza, Przegląd Statystyczny nr 3-4, tom 58, 300-322.

Baldwin R. H. (2001) Core-periphery model with forward-looking expectations, Regional Science and Urban Economics, Volume 31, Issue 1, 21-49.

Celińska-Janowicz D., Miszczuk A., Płoszaj A., Smętkowski M. (2010) Aktualne problemy demograficzne regionu Polski wschodniej, Raporty i Analizy EuroReg, Wydawca: Centrum Europejskich Studiów Regionalnych i Lokalnych EUROREG, Warszawa.

Dziemianowicz W. Szlachta J., Szmigiel-Rawska K. (red) (2011) Subregionalne bieguny wzrostu w Polsce, Wydawnictwo Uniwersytetu Warszawskiego, Warszawa.

Grabowski M. (2005) Migracje a rozwój, w: Bos-Karczewska M., Ceglińska A., Duszczyk M., Grabowska-Lusińska I., Maciej Grabowski M., Przybylski W., Szulc M., Migracje szansa, czy zagrożenie. Instytut Badań nad Gospodarką Rynkową, Gdańsk.

Hellwig Z. (1968) Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. Przegląd Statystyczny 15.4.1968.

Hryniewicz J. (2010) Teoria centrum-peryferie w epoce globalizacji, Studia Regionalne i lokalne nr 2 (40) 5-27.

Jakubowska A. (2012) Problem polaryzacji rozwoju obszarów peryferyjnych na przykładzie województwa zachodniopomorskiego, Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu, tom XIV, zeszyt 4, 35-39.

Kauppila P. (2011). Cores and peripheries in a northern periphery: a case study in Finland. Fennia 189: 1, 20-31.

Korcelli P, Degórski M. Komornicki T., Markowski T., Szlachta J., Węcławowicz G. Zaleski J., Zaucha J. (2008) Ekspercki projekt koncepcji zagospodarowania kraju do roku 2033, Opracowanie: Ministerstwo Rozwoju Regionalnego. Warszawa 2008.

Małkowski A. (2011) Regiony przygraniczne jako terytoria peryferyjne na przykładzie wschodniego i zachodniego pogranicza, Problemy regionalizmu i globalizacji Prace Naukowe Uniwersytetu Wrocławskiego nr 221, 364-372.

Matusik S., Pietrzak M., Wilk J. (2012) Ekonomiczne-społeczne uwarunkowania migracji wewnętrznych w Polsce w świetle metody drzew klasyfikacyjnych, Studia Demograficzne, nr 2(162) 3-28.

Pain K. (2008) Examining 'Core–Periphery' Relationships in a Global City-Region: The Case of London and South East England, Regional Studies, Vol. 42.8, 1161-1172.

Pietrzak M.B., Drzewoszewska N., Wilk J. (2012) The analysis of interregional migrations in Poland in the period of 2004-2010 using panel gravity model, Dynamic Econometric Models, Vol. 12, 111-122.

Pietrzak M. B., Żurek M., Matusik S., Wilk J. (2012) Application of Structural Equation Modeling for analysing internal migration phenomena in Poland, Przegląd Statystyczny nr 4, R. LIX, 487-503.

Pietrzak M. B., Wilk J., Matusik S. (2013) Gravity model as a tool for internal migration analysis in Poland in 2004-2010, W: J. Pociecha (red.) Quantitative Methods for

Modelling and Forecasting Economic Processes, Wyd. UE w Krakowie, Kraków (in printing).

Pietrzak M. B., Wilk J. (2013) Obszary metropolitalne Polski południowej a ruch migracyjny ludności, „Ekonomia i Prawo", B. Polszakiewicz, J. Boehlke (red.) Tom XII, nr 3/2013, 498-506.

Wilk J., Pietrzak M. B., Matusik S. (2013) Sytuacja społeczno-gospodarcza jako determinanta migracji wewnętrznych w Polsce, W: K. Jajuga, M. Walesiak (red.) Taksonomia 20-21. Klasyfikacja i analiza danych – teoria i zastosowania, PN UE we Wrocławiu nr 278, 330-342.

Wilk J., Pietrzak M.B. (2013) Analiza migracji wewnętrznych w kontekście aspektów społeczno-gospodarczych – podejście dwuetapowe, W: J. Dziechciarz (red.) Ekonometria 2(40) Wyd. UE we Wrocławiu, 62-73.

Wojnicka E., Tarkowski M., Klimczak P (2005) Przestrzenne i regionalne zróżnicowania ośrodków wzrostu. Polaryzacja a wyrównywanie szans rozwojowych. Przesłanki dla kształtowania polityki regionalnej państwa. Ministerstwo Gospodarki i Pracy, Warszawa.

Strategia rozwoju społeczno-gospodarczego Polski Wschodniej do roku 2020. Opracowanie: Ministerstwo Rozwoju Regionalnego. Warszawa 2008.

Zeliaś A. (2000) (red.) Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym, Wyd. AE w Krakowie, Kraków.

Zeliaś A. (2004) (red.) Poziom życia w Polsce i krajach Unii Europejskiej, PWE, Warszawa.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 158 – 168

SUBSAMPLING APPROACH FOR STATISTICAL INFERENCE WITHIN STOCHASTIC DEA MODELS[1]

Artur Prędki
Department of Econometrics and Operations Research
Cracow University of Economics
e-mail: predkia@uek.krakow.pl

Abstract: In the current literature, many different stochastic extensions of the DEA framework can be found. Such generalized approaches enable one to model the uncertainty inherent to the form of the production possibility set and the value of the technical efficiency measure at any given point of the former. In the paper we provide a detailed discussion of some statistical model and the subsampling algorithm which are used in statistical inference. The methodology is then illustrated with an empirical example using the real-world data from the Polish energy sector.

Keywords: DEA method, technical efficiency, statistical inference, subsampling

INTRODUCTION

Essentially, data envelopment analysis (or DEA, in short) is applied to measure technical efficiency within a group of n production (or, more generally, decision-making) units producing s sorts of outputs (arrayed in vector $\mathbf{y}_j = [y_{1j}, \ldots, y_{sj}]$ for $j = 1, \ldots, n$) out of m sorts of inputs ($\mathbf{x}_j = [x_{1j}, \ldots, x_{mj}]$).

In what follows we assume that the producers are technologically homogenous and the technology itself is represented by the production set T, the specifics of which are discussed in the following section. Additionally, in a deterministic DEA framework, T is specified as a minimal set (in terms of inclusion), containing all the data points $(\mathbf{x}_j, \mathbf{y}_j)$, $j = 1, \ldots, n$. Such an approach

[1] Research conducted under financial support from Cracow University of Economics. The author would like to express his gratitude to Dr Łukasz Kwiatkowski for his assistance with English translation of the paper.

enables one to derive its form explicitly. On the other hand, such an assumption is invalid within stochastic variations of DEA, which is attributable to uncertainty inherent to the data collected[2]. Consequently, the production set is of an unknown form.

Based on T, the so-called Farrell technical efficiency measure (for some feasible production plan $(\mathbf{x}_o, \mathbf{y}_o)$) can be defined. Such a measure underlies studies employing DEA methods and is usually formulated in the input- and output-oriented settings, respectively:

$$\theta\,(\mathbf{x}_o, \mathbf{y}_o) = \min\{\theta \in R: (\theta\mathbf{x}_o, \mathbf{y}_o) \in T\},$$
$$\lambda\,(\mathbf{x}_o, \mathbf{y}_o) = \max\{\lambda \in R: (\mathbf{x}_o, \lambda\mathbf{y}_o) \in T\}.$$

In the research we restrict ourselves to the input orientation solely. In a deterministic DEA framework, owing to a known form of the production set, a unique value of the measure can be calculated by solving appropriate linear programs (see, e.g., [Guzik 2009]). However, it is no longer the case in stochastic DEA methods, where the actual form of T remains unknown.

STATISTICAL MODEL

In the paper we focus on one of the stochastic DEA methods, developed by a team of researchers gathered around Professor Léopold Simar[3]. According to the preceding remarks, the setting requires the form of T to be approximated, so that technical efficiency measure can be further estimated at any given point of T. For the approximation and estimation to be valid one needs to formulate a relevant model. In the present study we resort to the methodology presented by Kneip et al. [2008], recognized for its transparency as compared with earlier suggestions found in the literature (see [Kneip et al. 1998], [Gijbels et al. 1999] and [Park et al. 2000]).

Below we present three most fundamental assumptions that are material to further considerations, starting with the one specifying properties of the production set.

Assumption 1. The units produce *s* sorts of goods out of *m* sorts of inputs and use the same technology, represented by T – a closed and convex production set, satisfying the 'free disposal' (or: inefficiency) condition and the 'no free lunch' condition (i.e. no output can be produced out of zero inputs).

For more details and a review of other possible properties of the production set we refer to, e.g., [Prędki 2012].

The second assumption pertains to the data generating process.

---

[2] The uncertainty may arise as a result of measurement errors or data incompleteness, for instance.

[3] Léopold Simar (born 1946) is Full Professor of Statistics at the Department of Statistics, Université Catholique de Louvain (Belgium).

Assumption 2. The sample $\{(\mathbf{x}_j, \mathbf{y}_j) \in R_{0+}^{m+s}: j = 1, \ldots, n\}$ is drawn from a sequence of independent random vectors (X_j, Y_j) following the same distribution with a continuous (over support T) density function $f_T(\mathbf{x}, \mathbf{y})$: $R_{0+}^{m+s} \to R$.
The presumption above underpins the proof of consistency of the technical efficiency measure estimator (at a given point of T) and implies obvious fact of all the data points $(\mathbf{x}_j, \mathbf{y}_j)$ $(j = 1, \ldots, n)$ are feasible production plans (i.e. fall into T) with probability 1.

The final assumption is meant to lend due 'smoothness' to the efficiency measure as a function of inputs and outputs.

Assumption 3. The function $\theta(\mathbf{x}, \mathbf{y})$ is of class $C^2(T)$ (i.e. the second-order derivatives exist and are continuous).
The above requirement aims at obtaining additional information on asymptotic sample distribution of the estimator of technical efficiency measure (at a given point of T).

EFFICIENCY MEASURE ESTIMATOR

As an approximation of the production set we shall use one of its well-known forms obtained in the deterministic setting. Specifically, T is approximated with a minimal set (in terms of inclusion) $\hat{T}$ that satisfies Assumption 1 and includes all the data points[4]:

$$\hat{T} = \{(\mathbf{x}, \mathbf{y}) \in R_{0+}^{m+s}: \exists \lambda_j \geq 0: \sum_{j=1}^{n} \lambda_j = 1, \mathbf{x} \geq \sum_{j=1}^{n} \lambda_j \mathbf{x}_j, \mathbf{y} \leq \sum_{j=1}^{n} \lambda_j \mathbf{y}_j\}.$$

Replacing T with $\hat{T}$ in the definition of $\theta(\mathbf{x}_o, \mathbf{y}_o)$ one obtains the estimator of input-oriented technical efficiency estimator at $(\mathbf{x}_o, \mathbf{y}_o)$:

$$\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o) = \min\{\theta: (\theta \mathbf{x}_o, \mathbf{y}_o) \in \hat{T}\}.$$

In [Kneip et al. 1998] it has been proved that for any $(\mathbf{x}_o, \mathbf{y}_o)$ lying in the interior of $\hat{T}$ the above formula yields a consistent estimator of the true efficiency measure (with the convergence rate of $n^{-2/(m+s+1)}$), whereas in [Kneip et al. 2008] the asymptotic sample distribution of $\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o) - \theta(\mathbf{x}_o, \mathbf{y}_o)$ has been derived. Regrettably, its form is dependent on $(m + s)(m + s + 1)/2 + 2$ unknown parameters related to the density f_T and the multivariate production frontier. To this day, no consistent method of estimation for these parameters has been designed. This, in turn, prohibits construction of asymptotic confidence intervals and other quantities which could possibly shed some light on the relation between the data uncertainty and the one of efficiency measure estimation.

[4] Inequalities are understood "by coordinates".

BOOTSTRAP

In the situation outlined above, one is prompted to resort to the bootstrap methods, which have gained much favor since the late 1990s. Such an approach, based on simulating additional samples out the original one, allow for approximation of the aforementioned confidence intervals and some measures of statistical dispersion. A necessary condition for any bootstrap procedure to be valid is its consistency, which requires (in the limiting case of $n\to\infty$) the distribution of the bootstrap estimator of an efficiency measure to be corresponding (in a certain manner) with the actual asymptotic distribution of the estimator considered in the previous section. A formal definition of the bootstrap consistency can be found in, e.g., [Simar and Wilson 2000, p. 61]. It seems that early attempts of designing relevant bootstrap procedures (such as the ones presented in [Löthgren and Tambour 1999], [Simar and Wilson 1998, 2000]) have not been quite successful as they have resulted in either methods consistency of which could not be formally verified, or in ones proved positively inconsistent[5]. Moreover, some of these have been founded upon a quite restrictive and non-verifiable assumption of homogeneity of technical efficiency measure. Only recently, in [Kneip et al. 2008], have two consistent bootstrap methods been devised. However, one of them requires solving a huge number of linear programs at each bootstrap iteration, which renders the method (already complex due to its multilevelness and intricate formulae) computationally formidable. Therefore, the other of the two – resting upon the idea of the so-called subsampling – seems a more attractive alternative.

SUBSAMPLING

Since the subsampling approach is the one employed in our empirical study to follow, we present in some more detail the way the method proceeds.

Step 1. Based on the original sample, the value of $\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)$ is calculated.

Step 2. Generate a bootstrap sample $\{(\mathbf{x}_j^*, \mathbf{y}_j^*)\text{: } j = 1, \ldots, k\}$ by randomly drawing (independently, uniformly, and with replacement) $k < n$ observations from the original sample[6].

Step 3. The bootstrap estimate of $\theta(\mathbf{x}_o, \mathbf{y}_o)$ is computed according to the formula:

$$\hat{\theta}^*(\mathbf{x}_o, \mathbf{y}_o) = \min\{\theta\text{: } (\theta\,\mathbf{x}_o, \mathbf{y}_o) \in \hat{T}^*\},$$

where

[5] Only some simulation studies have been carried out, the results of which seem to corroborate consistency of some procedures (under a given Data Generating Process).

[6] While specifying value of $k = k(n)$ in simulation studies one needs to guarantee that $k(n)\to\infty$ and $k(n)/n\to 0$ in the limiting case of $n\to\infty$. Assuming $k = n$ (which corresponds with the so-called naive bootstrap) results in inconsistency of the procedure.

$$\hat{T}^* = \{(\mathbf{x}, \mathbf{y}) \in R_{0+}^{m+s}: \exists \lambda_j \geq 0: \sum_{j=1}^{k} \lambda_j = 1, \mathbf{x} \geq \sum_{j=1}^{k} \lambda_j \mathbf{x}_j^*, \mathbf{y} \leq \sum_{j=1}^{k} \lambda_j \mathbf{y}_j^* \}.$$

Repeat steps 2-3 B times, obtaining a sequence of bootstrap estimates:

$$\{\hat{\theta}_b^*(\mathbf{x}_o, \mathbf{y}_o): b = 1, ..., B\}.$$

The way the algorithm proceeds next depends on the particular aim of the inference. In the current study, for instance, we are to analyze uncertainty related to the efficiency measure calculated in Step 1, which stems from an unknown form of the production set. We start with approximating asymptotic confidence interval for $\theta(\mathbf{x}_o, \mathbf{y}_o)$.

Step 4. For each b = 1, …, B, compute the expression:

$$k^{2/(m+s+1)}\left(\frac{\hat{\theta}_b^*(\mathbf{x}_o, \mathbf{y}_o)}{\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)} - 1\right),$$

and put the values obtained in a non-decreasing order.

Step 5. For a given $\alpha \in (0, 1)$, discard $\alpha/2 \cdot 100\%$ elements in both tails of the sequence. The minimum and maximum of the remaining values are the so-called bootstrap quantiles, denoted by $\delta_{\alpha/2,k}$ and $\delta_{1-\alpha/2,k}$, respectively.

Step 6. Compute the interval:

$$[\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)/(1 + n^{-2/(m+s+1)}\delta_{1-\alpha/2,k}); \hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)/(1 + n^{-2/(m+s+1)}\delta_{\alpha/2,k})],$$

henceforth referred to as the asymptotic bootstrap confidence interval for $\theta(\mathbf{x}_o, \mathbf{y}_o)$.

Some dispersion measures providing one with information about uncertainty related to $\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)$ can be calculated as well, including:

- sample variance of the bootstrap estimates:

$$\hat{\sigma}_o^2 = \left(\frac{k}{n}\right)^{\frac{4}{m+s+1}} B^{-1}\left[\sum_{b=1}^{B} \hat{\theta}_b^*(\mathbf{x}_o, \mathbf{y}_o) - B^{-1}\left(\sum_{b=1}^{B} \hat{\theta}_b^*(\mathbf{x}_o, \mathbf{y}_o)\right)\right]^2,$$

- bootstrap bias of $\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)$:

$$\text{bias}_o = \left(\frac{k}{n}\right)^{\frac{2}{m+s+1}}\left[B^{-1}\left(\sum_{b=1}^{B} \hat{\theta}_b^*(\mathbf{x}_o, \mathbf{y}_o)\right) - \hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)\right].$$

Then, taking the bootstrap bias into account, a revised estimate of technical efficiency measure is readily available:

$$\hat{\hat{\theta}}(\mathbf{x}_o, \mathbf{y}_o) = \hat{\theta}(\mathbf{x}_o, \mathbf{y}_o) - \text{bias}_o.$$

Note, however, that such a correction may result in an increase of the mean square error of the revised estimator (with respect to $\hat{\theta}(\mathbf{x}_o, \mathbf{y}_o)$). In the literature (see, e.g., [Simar and Wilson 2000, p. 63]) it is postulated that the revision should be applied once the following inequality holds:

$$\text{bias}_o^2 > 3\hat{\sigma}_o^2.$$

SPECIFIC PROBLEMS

While using the DEA methods, values of the efficiency measure are calculated for each of the units in the original sample, i.e. for o = 1, …, *n*, as it is the main objective of the approach to compare and rank the units with respect to their efficiency. Analogously, one may wish to obtain (approximated) confidence intervals and the dispersion measures for each unit. However, theoretical results and simulation studies presented in the relevant literature (see [Kneip et al. 2008] and [Simar and Wilson 2011]) consider only a single object $(\mathbf{x}_o, \mathbf{y}_o)$ that is assumed to be independent of the ones in the original sample $\{(\mathbf{x}_j, \mathbf{y}_j) \in R_{0+}^{m+s}: j = 1, \ldots, n\}$[7]. Aiming at employing the subsampling approach in order to compute the confidence intervals and dispersion measures we slightly modify the procedure outlined above. We proceed along some hints found in [Simar and Wilson 2011, p. 40]. Firstly, the subsampling procedure is repeated *n* times for each o = 1, …, *n*. For a given unit, the algorithm proceeds similarly as before with the only modification in Step 2, now consisting in drawing *k* out of $n - 1$ elements (excluding object o). Naturally, in such a case $k < n - 1$, so some information is lost, yet the required independence is guaranteed.

Choosing an appropriate value for *k* is another key issue. Some simulation studies carried out in [Kneip et al. 2008] indicate a critical role of the parameter to final results of the subsampling procedure, thereby proving their pronounced sensitivity. Nevertheless, no universal approach (featuring some desired statistical properties) to selecting the 'right' number for *k* can be pointed. Therefore, we resort to (quite an arbitrary) empirical rule that has been formulated in [Simar and Wilson 2011, pp. 42-43]. The method – presented below – is a particular version of a general procedure designed in [Politis et al. 2001], and has specifically been developed to deal with approximation of confidence intervals.

Step 1. For a given point $(\mathbf{x}_o, \mathbf{y}_o)$, conduct the subsampling scheme for arbitrarily specified values of *k*: $k_1 < k_2 < \ldots < k_W$. As a result one obtains a sequence of corresponding confidence intervals:$\{[L(k_w); R(k_w)]: w = 1, \ldots, W\}$.

Step 2. Choose some arbitrary small natural number *v* (usually, in practice, *v* assumes the value of one, two or three).

Step 3. For any $w \in [v + 1, W - v]$,calculate a sum of standard deviations of the left-hand side endpoints: $L(k_{w-v}),\ldots, L(k_{w+v})$, and of the right-hand side ones: $R(k_{w-v}), \ldots, R(k_{w+v})$.

Step 4. For a given $w \in [v + 1, W - v]$, select the value of $k_w \in \{k_1, k_2, \ldots, k_W\}$ that minimizes the sum computed in Step 3.

[7] In the sense of the random vector (X_o, Y_o) being stochastically independent of each (X_j, Y_j), j = 1, …, *n*.

EMPIRICAL STUDY

We finally proceed to an empirical illustration of the presented methodology, analyzing a real-world data set comprising of 32 Polish power stations and thermal-electric power stations in 1995. Previously, and for the first time, the data have been analyzed in [Osiewalski and Wróbel–Rotter 2002]. With a view to make inference about technical efficiency of these production units, we employ the following variables as the inputs:

x_1 – real capital (gross value of capital assets in thousands of Polish zloty);
x_2 – labor (number of workers);
x_3 – feed energy(in TJ).[8]

The only output of the units is the energy produced, y, measured in TJ.

Utilizing the author's own numerical procedure programmed in GAUSS 12.0, subsampling routines presented in the previous sections (with suitable modifications) have been run.[9] The results are presented in Table 1.

Table 1. Subsampling results

j	$\hat{\theta}_j$	$\hat{\sigma}_j^2$	bias_j	$\hat{\hat{\theta}}_j$	L_j	R_j	k_j
1	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
2	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
3	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
4	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
5	0.67971	0.00810	0.07037	0.60935	0.47925	0.67954	23
6	0.77265	0.00557	0.08513	0.68752	0.60434	0.77246	27
7	0.74739	0.00577	0.07962	0.66777	0.56619	0.74720	27
8	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
9	0.96423	0.00025	0.02402	0.94021	0.93145	0.96399	27
10	0.69205	0.00614	0.07542	0.61663	0.49809	0.69188	24
11	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
12	0.63646	0.00983	0.10151	0.53496	0.43344	0.63630	20
13	0.58258	0.01551	0.14586	0.43671	0.38674	0.58243	16
14	0.97465	0.00011	0.01522	0.95943	0.95105	0.97440	27
15	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
16	0.82456	0.00307	0.07183	0.75273	0.68625	0.82435	27

[8] TJ – terajoule, 1 GWh = 3,6 TJ.

[9] The author is deeply indebted to Professor Anna Pajor for her advice and unwavering support as well as for providing him with additional numerical procedures of her own.

j	$\hat{\theta}_j$	$\hat{\sigma}_j^2$	bias_j	$\hat{\hat{\theta}}_j$	L_j	R_j	k_j
17	0.82574	0.00283	0.05880	0.76694	0.68825	0.82553	27
18	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
19	0.96811	0.00016	0.01122	0.95689	0.93864	0.96786	27
20	0.95963	0.00021	0.02988	0.92974	0.92272	0.95939	27
21	0.65916	0.00217	0.06664	0.59252	0.54118	0.65899	19
22	0.62917	0.00411	0.06698	0.56219	0.47174	0.62901	25
23	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
24	0.86040	0.00338	0.06460	0.79581	0.74578	0.86019	27
25	0.81771	0.00221	0.04918	0.76853	0.67515	0.81751	27
26	0.75999	0.00506	0.05942	0.70057	0.58505	0.75980	27
27	0.75424	0.00629	0.06550	0.68875	0.57647	0.75405	27
28	0.89069	0.00190	0.04488	0.84582	0.79796	0.89047	27
29	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
30	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
31	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5
32	1.00000	0.00000	0.00000	1.00000	0.99975	0.99975	5

Source: own elaboration

In columns 2-4 of Table 1 we report estimates obtained for the measure of efficiency and measures of its dispersion, whereas in columns 5-7 – the revised estimates of the efficiency measure along with approximated asymptotic confidence intervals (with L_j and R_j standing for the left- and the right-hand side endpoints, respectively). The last column displays values of k_j's, i.e. the number of generated bootstrap samples for each data point $(\mathbf{x}_j, \mathbf{y}_j)$ (j = 1, …, n), which have been specified according to the empirical rule discussed in the previous section.[10]

As regards selecting particular values for B and v, we follow the mainstream literature and set $B = 2000$ and $v = 2$, which obviously does not shed the arbitrariness of such a choice. Reflecting on specific choices for k, again, we resorted to general suggestions found in the literature and – avoiding extreme possible values of the parameter – considered $k \in \{3, 4, \ldots, 29\}$. Since $n = 32$ in our study, setting k close to 32 would imply slower convergence due to the algorithm "nearing" the (inconsistent) naive bootstrap (see Footnote 6). On the other hand, low values of the parameter result in a sizeable loss of information and enforce low number of bootstrap subsamples available to be generated.

[10] Let us recall that the bootstrap samples generated for any j do not include the unit itself.

According to values reported in the last column of Table 1, it is clear that following the empirical rule discussed above the most frequent choice for the number of elements in a subsample is 27. With regard to revision of the point estimates of the efficiency measure, only in the case of unit j = 20 the relevant inequality holds so that the correction is then justified.[11] Values of the bias-revised estimator for the other units are presented only for the sake of completeness.

Estimates of the efficiency measures hover between 0.586 and 1.000. The lowest efficiency is indicated for unit $j = 13$, whereas power stations no.: 1-4, 8, 11, 15, 18, 23, 29-32, appear to be the most efficient. Note, however, that the subsampling procedure is not valid for the latter, as it was not possible to obtain either non-zero values of the dispersion measures or reasonable approximations of the confidence intervals. On a technical note, it is caused by a lack of objects with a higher value of the efficiency measure in the original sample. Partly, it is also in accord with the theoretical results discussed above, which hold only for the data points lying in the interior of the production set, thereby lacking in efficiency. Admittedly, one cannot dismiss a possibility that our estimates of technical efficiency are incorrect and that the units may in fact be inefficient. Nevertheless, with the data set in hand it is not possible to diagnose conclusively whether it is actually the case.

For the units that emerged to be inefficient we managed to compute both the bootstrap measures of dispersion as well as approximated confidence intervals. One needs to perceive the results with some caution, however, as the low number of units in the sample undermines the asymptotics.

Table 2 presents relative and absolute measures of dispersion (only for the inefficient stations), sorted increasingly with respect to technical efficiency estimates.[12]

[11] See the end part of the "Subsampling" section.

[12] As an absolute measure of dispersion we consider the diameter of a confidence interval, i.e. $R_j - L_j$, the values of which are reported in the last column of Table 2.

Table 2. Relative and absolute measures of dispersion

j	$\hat{\theta}_j$	$\hat{\sigma}_j / \hat{\theta}_j$	$bias_j / \hat{\theta}_j$	$R_j - L_j$
13	0.58258	0.21380	0.25038	0.195692
22	0.62917	0.101879	0.106458	0.157266
12	0.63646	0.15575	0.15949	0.202866
21	0.65916	0.070723	0.101093	0.117812
5	0.67971	0.13243	0.10353	0.200289
10	0.69205	0.11327	0.10898	0.193788
7	0.74739	0.10166	0.10654	0.181013
27	0.75424	0.10517	0.08684	0.177587
26	0.75999	0.09364	0.07818	0.17475
6	0.77265	0.09656	0.11018	0.168121
25	0.81771	0.05749	0.06015	0.14236
16	0.82456	0.06721	0.08711	0.138101
17	0.82574	0.06441	0.07121	0.137282
24	0.86040	0.06754	0.07508	0.114401
28	0.89069	0.04894	0.05039	0.09251
20	0.95963	0.015051	0.031142	0.03667
9	0.96423	0.01654	0.02491	0.032538
19	0.96811	0.01289	0.01159	0.029225
14	0.97465	0.01083	0.01561	0.023352

Source: own elaboration

Note a clear negative correlation between estimates of efficiency measure and its dispersion. It follows that higher values of $\hat{\theta}_j$ are usually accompanied with tighter confidence intervals. Conversely, for units of lesser efficiency higher uncertainty is typically indicated. A technical reason behind that observation resides in the fact that the generated subsamples contain a large fraction of highly efficient units, with respect to which the highly inefficient ones are compared. It should be pointed, however, that values obtained for the relative dispersion measures do not exceed 25% of the point estimate. Therefore, even for the most inefficient power stations the relative uncertainty is not considerable.

REFERENCES

Gijbels I., Mammen E., Park B.U., Simar L. (1999) On estimation of monotone and concave frontier function, JASA, Vol. 94, pp. 220-228.

Guzik B. (2009) Basic DEA models in analysis of economic and social efficiency, Poznań University of Economics, Poznań.

Kneip A., Park B.U., Simar L. (1998) A Note on the convergence of nonparametric DEA estimators for production efficiency scores, Econometric Theory, Vol. 14, pp. 783-793.

Kneip A., Simar L., Wilson P.W. (2008) Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models, Econometric Theory, Vol. 24, pp. 1663-1697.

Löthgren M., Tambour M. (1999) Testing scale efficiency in DEA models: a bootstrapping approach, Applied Economics, Vol. 31, pp. 1231-1237.

Osiewalski J., Wróbel–Rotter R. (2002) A Bayesian Random Effect Model in Cost Efficiency Analysis (with the Application to Polish Electric Power Stations), Statistical Review, Vol. 49(2), pp. 47-68.

Park B.U., Simar L., Weiner Ch. (2000) The FDH estimator for productivity efficiency scores, Econometric Theory, Vol. 16, pp. 855-877.

Politis D., Romano J., Wolf M. (2001) On the asymptotic theory of subsampling, Statistica Sinica, Vol. 11, pp. 1105-1124.

Prędki A. (2012) The Origins of Production Possibility Sets in the DEA Method [in:] Mathematics and Information Technology at the Service of Economics: Theory - Models, [ed.] W. Jurek, Scientific Bulletin of Poznań University of Economics, no. 241, Poznań, pp. 126-137.

Simar L., Wilson P.W. (1998) Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models, Management Science, Vol. 44, pp. 49-61.

Simar L., Wilson P.W. (2000) A general methodology for bootstrapping In nonparametric frontier models, Journal of Applied Statistics, Vol. 27, pp. 779-802.

Simar L., Wilson P.W. (2011) Inference by the m out of n bootstrap in nonparametric frontier models, Journal of Productivity Analysis, Vol. 36, pp. 33-53.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 169 – 180

SEMIPARAMETRIC COX REGRESSION MODEL IN ESTIMATION OF SMALL AND MICRO ENTERPRISES' SURVIVAL IN THE MALOPOLSKA VOIVODESHIP[1]

Aneta Ptak–Chmielewska
Institute of Statistics and Demography
Warsaw School of Economics
e-mail: aptak@sgh.waw.pl

Abstract: This paper aims at identifying factors (external and internal) affecting the ability of an enterprise to survive on the market. The analysis is based on the results of a retrospective study conducted in 2012 on a sample of enterprises from Malopolska voivodeship. Methods and models of event history analysis, including semiparametric Cox's model were applied to analyse enterprises' survival. The approach based on the event history analysis allows us to include dynamics of the process. The results provided extensive data on how factors such as size, activity sector, market range, legal form and internals conditions like: owner characteristics, investments, profits, reported barriers affect the survival of enterprises.

Keywords: survival analysis, Cox's model, enterprises' survival

BACKGROUND – LITERATURE REVIEW

Enterprises dynamics account for a high percent of total productivity growth, supporting the idea that entrepreneurs are among the driving forces of economic growth and structural change. In the context of enterprises population dynamics, three aspects and areas must be considered: the entry, the exit and survival of entities. This paper aims at identifying factors (external and internal) affecting the ability of an enterprise to survive on the market.

Entry may be motivated by possibilities of higher earnings as self-employed, which supports start-up decision process [Creedy and Johnson 1983, Audretsch

[1] Paper financed from research grant NCN no. 3739/B/H03/2011/40 titled "Business demography. Micro-macro analysis of enterprise life cycle in Poland and UE" lead by Aneta Ptak–Chmielewska.

1995, Geroski 1995, Vivarelli 2004] as well as escape from unemployment as push factor. Barriers of entry like financial constraints on business start-ups [Evans and Jovanovic 1989, Cabral and Mata 2003] also play a significant role. Some characteristics of entrepreneurs may also be important such as self-realization, fulfillment of aspiration, better social status [Creedy and Johnson 1983, Vivarelli 2004]. Some market mechanisms may be assessed by macro-models but human behavioural decisions are not always fully predictable and measurable.

More than 50% of new firms exit the market within the first five years of activity, which is due to just a mistake according to the true Schumpeterian displacement-replacement effect [Geroski and Mazzucato 2001]. Exit is the effect of self-decisions (behavioural) or due to financial difficulties. The big role play loan constraints [Becchetti and Trovato 2002, Hurst and Lusardi 2004]. The internal factor is human capital of workforce and skills of entrepreneur [Lazear 2004, Silva 2006], and also sex of entrepreneur, but this is mixed effect [Cooper et al. 1994], not confirmed. The sector heterogeneity of survival [Marsili 2002] also plays a significant role.

The initial papers regarding the survival of enterprises have appeared in recent years in Poland. In her paper dedicated to survival of enterprises, Markowicz [2012] applies the Cox model for the analysis of enterprises from Szczecin region (only one city region) based on REGON register (not adjusted for non-active enterprises). Dehnel [2010] applies small area estimation techniques for basic business demography ratios for small and micro enterprises in Poland (active before 2001). Nehrebecka and Dzik [2013] apply logistic regression and distinguish between liquidation and bankruptcy as different ways of exiting the market. Some works of Ptak–Chmielewska [2010, 2012c] apply non-parametric (Kaplan-Meier) and semi-parametric (Cox regression) methods for retrospective data for one of regions in Poland.

According to the group of theories connected with management in an enterprise [Poznańska, 2008]: "Liability of Smallness", "Liability of Newness" and "Ecological Economy", the position of a small enterprise (as far as the number of employees and turnover are concerned) on the competitive market is much weaker in comparison with large enterprises. The reasons can be traced to the fact that small enterprises are put at a disadvantage with regard to economy of scale, distribution chains or market research. They employ managers with lower skills. On the other hand, according to the market niche theory, a small size can create many opportunities [Porter 1979]. Such enterprises will retain their size at a given level in order to be able to reach the market niches which are not accessible for large companies. Liability of Newness links the success of an enterprise with the time it functions on the market. The probability of exiting the market is much higher for new enterprises than for those with at least one year history. Ecological economy refers to biological theories [Hannan and Freeman 1989, Freeman et al. 1983] focusing on three groups of enterprises: new, developing and shrinking (end

of life). The ecological theory is focused then on the life cycle of an enterprise: determinants of birth, survival and death.

DATA AND METHODS

Event history analysis (survival analysis) is defined as a set of different statistical techniques used for describing and analysing the life course of an individual: number of events, sequence of events, timing and time spent in different states of the process. The subject of this analysis is a stochastic process with events (states). This process is described by survival time defined as the time from the start of observation till the end of process or the end of survey if it is completed before the end of the process. The subject of this kind of research is a time span between events known as time of the process or episode.

There are many different models used in the survival analysis. Models are differentiated according to assumptions on their functional form of hazard rate and its variability in time. The description of different models – nonparametric, parametric and semi-parametric – may be found in Frątczak [2005, chapters 7, 8, 9]. In practice, the proportional hazards Cox regression model is most frequently used. For this reason this model was presented in more details in this paper. For the Cox regression model the hazard function is given by a formula:

$$h(t \mid x_1,...,x_k) = h_0(t) exp(\alpha_1 x_1 + ... + \alpha_k x_k) \tag{1}$$

where: h_0 – base hazard, parametrically non-specified function of time and $x_1, x_2, \ldots x_k$ – explanatory variables (including time dependent variables).

Cox also proposed a special type of estimation method called pseudo-likelihood [Cox, 1972]. This method divides the likelihood function for the proportional hazards model into two parts: first including only information on parameters and second including information on parameters and hazard function. This division into two components is justified because the former depends only on the sequence of events and does not depend on the exact time of occurrence, while the latter is 0 and is omitted.

The main advantage of Cox's model (and other semiparametric models) is the ability to assess the influence of many variables (including time dependent variables) on the process with no need for base hazard $h_0(t)$ specification. The main disadvantage of Cox's model is hazard proportionality assumption. This assumption imposes a fixed hazard rate for each pair of individuals at any time. Relative hazard (ranking) for individuals is stable. Despite this limitation of Cox's model, it is particularly attractive for researchers as in the case of Blossfeld and Rohwer [2002]:

- unknown shape of hazard in time;
- no theoretical basis for parametrisation;
- no possibility of specifying the functional shape of hazard;

- main interest focused on the influence of explanatory variables on hazard.

The abovementioned advantages of the application of Cox regression model make this model useful in modelling the risk of enterprises' liquidation. The only disadvantage of this model is a proportionality assumption which implies a fixed proportion of hazard for individuals during the observation period. This problem may be solved by including additional time dependent variables to the model such as the interaction between a variable and time. This model is known as non-proportional hazards Cox regression model. The results of Cox model estimation are parameters describing the influence of explanatory variables on the probability of event occurrence and on the base hazard (the same for all individuals, dependent only on time).

The analysis of enterprises' survival requires information on the exact date of the start and exact date of the end of their activity. Data concerning this information that are available in administrative registers such as REGON and cover all enterprises are not up-to-date and thus seem useless in the analysis of enterprises' survival. This requires the application of representative surveys. A panel representative survey on the sample of micro and small enterprises that has been conducted by CSO since 2002 is a good source of data [Warunki… 2010]. However a panel character of the survey leads to a significant information loss in subsequent waves and poses the necessity of missing data imputation. More details on data sources can be found in [Ptak–Chmielewska 2012a]. In 2012 a retrospective survey on the sample of enterprises from the Malopolska voivodeship registered in 2006 was conducted. The survey covered basic information on a given enterprise, activity status (after a five-year period) and type of activity, legal form, number of workers, range of market activity, and changes observed during its five-year activity. Additionally socio-demographic characteristics of the main owner of enterprise and reported barriers were collected.

The calendar time was not analyzed because all enterprises were existing in the same macroeconomic conditions. Time in survival analysis was defined as time in months since the start of the business (different from the date of registration) till the end date of the business: liquidation or suspension. Right censoring was set at the date of the survey 31.12.2011 for enterprises that survived till the end of a five-year period. For these enterprises (censored) time was measured as time in months since the start of the business till its end in 2011 (survey). In overall sample of 1077 enterprises 667 were censored (62%), means still active at the date of survey.

RESULTS OF COX REGRESSION MODEL ESTIMATION

The semiparametric Cox regression model was estimated in three steps. As the first step univariate models were estimated using the variables listed below and insignificant variables were deleted. In the second step, the multivariate model was

estimated using significant variables and checking for correlation between independent variables. In the last step, the proportionality assumption was verified and the model, including interaction over time for variables for which assumption was not satisfied, was estimated.

Variables used in Cox's regression model:

1. legal form –1 if companies (legal persons), including companies without special legal form, and 0 – for sole traders;
2. number of workers –1 if an enterprise had 10-49 workers at the start, and 0 – if an enterprise had 0-9 workers employed at the start;
3. sector of activity-group –1 if sectors were from a low risk group: industry, education, wholesale, hotels and restaurants, repairs, health care, other activity, and 0 – if sectors belonged to a high risk group: construction, retail, financial intermediation, real estate and business activities, transport;
4. change in the sector of activity –1 if during an enterprise activity some changes of the sector took place, and 0 – no changes;
5. geographical area of activity –1 if the area of activity was a cross-national or international market and 0 – if an enterprise was active only on a local or regional market;
6. change of geographical area – 1 if any change in the geographical area took place and 0 – if there was no change;
7. source of financing–1 if funds for starting the business came from a bank loan and 0 if they came from other sources (mainly own funds or family support);
8. export of goods and services –1 if an enterprise has ever exported their goods or services and 0 – if a respondent answered „no";
9. change in the number of workers –1 if a change in the number of workers took place (measured by a positive parameter in linear trend for a five-year period) and 0 – if this trend was negative or there was no growth;
10. owner as the only employee –1 if during the whole period of activity only the owner worked in an enterprise and there was no change and 0 in any other case;
11. profit in the first year –1 if over the first year of activity an enterprise made profits and 0 if it made a loss or suspended its activity (vulnerable or forced);
12. age of the owner –1 if the owner was 35 years old and older or it was a company and 0 if the owner was younger than 35 or the age of the owner was missing;
13. educational level of the owner –1 if the owner had higher education (post-secondary) or it was a company and 0 in other cases;
14. sex of the owner – 1 if the owner was a man or it was a company and 0 if the owner was a woman;
15. source of maintenance – 1 if the employment in this enterprise was the main source of maintenance for the owner or it was a company and 0 in other cases;
16. type of previous job –1 if the previous job of the owner was a job from the low risk group (engineering, white collar, manager, other) or it was a company and 0 if it was a job from the high risk group (farmer, craftsman, seller, student, unemployed);

17. investments –1 if an enterprise invested in the first year of activity and 0 if there were no investments or an enterprise was suspended or no information;
18. barriers –1 if an enterprise (respondent) did not report any barriers in selling their goods or services in the first year of activity and 0 if barriers were reported.

The results of this univariate analysis confirmed the lack of significance of the following variables: source of maintenance and age of the owner. In the multivariate analysis 16 variables were used. The assumption of proportionality was tested and verified with martingale residuals – supremum test (details in Allison 2010, pp. 173 and further). The correlation between explanatory variables was also investigated but there were no significant correlations. In the multivariate model variables such as sector of activity-group, export of goods and services, sexof the owner, education of the owner were not significant.

Enterprises registered and operating as companies had 65% lower risk of liquidation compared to enterprises of sole traders (see table 1). Enterprises with 10-49 workers at the start had 67% lower risk of liquidation compared to enterprises with 0-9 workers at the start. In many cases it turned out that the owner was the only employee working in an enterprise (self-employment).

Table 1. Results for Cox regression model

Variable	Parameter	Std. error	Chi-square	Pr. > chi-sq.	Hazard ratio
Legal form	-1.03945	0.21673	23.0016	<.0001	0.354
Number of workers	-1.12391	0.36727	9.3644	0.0022	0.325
Sector of activity-group	0.12674	0.10842	1.3664	0.2424	1.135
Change of sector of activity	-1.24947	0.45612	7.5039	0.0062	0.287
Geographical area of activity	-0.29889	0.13047	5.2483	0.0220	0.742
Change of geographical area	-0.92062	0.51144	3.2403	0.0718	0.398
Source of financing	-0.75325	0.29621	6.4669	0.0110	0.471
Export of goods and services	-0.52834	0.33484	2.4897	0.1146	0.590
Change in number of workers	-0.22728	0.02763	67.6391	<.0001	0.797
Owner as only employee	-2.28942	0.20973	119.1544	<.0001	0.101
Profit in the first year	-0.68265	0.10342	43.5683	<.0001	0.505
Sex of the owner	-0.04531	0.10337	0.1921	0.6612	0.956
Education of the owner	-0.10960	0.11582	0.8955	0.3440	0.896
Type of the previous job	-0.38199	0.10645	12.8767	0.0003	0.682
Investments	-0.60068	0.10316	33.9060	<.0001	0.548
Barriers	-0.53033	0.10760	24.2919	<.0001	0.588

Source: own calculations based on the results of a retrospective survey on enterprises' survival in Malopolska voivodeship using SAS system

Enterprises with only the owner working and without any changes during the observation period had about 90% lower risk of liquidation in comparison to enterprises that changed the employment (most cases those enterprises employed

people at start and after 1-2 years limited employment).For the variable sector of activity-group the parameter was positive, which seems to be counter-intuitive. This aggregation could be too strong to reveal any differences between risk for different sectors of activity. If an enterprise was flexible and could change the sector of activity according to a market demand it had 60-70% lower risk of liquidation compared to enterprises that did not change their sector of activity. Enterprises operating on cross-national or international markets had 26% lower risk of liquidation compared to enterprises operating only on regional and local markets. Changes in the area of activity (wider markets) were also significant and lowered the risk of liquidation by about 60%.If an enterprise was able to get bank loans to finance the start of the business it had about 53% lower risk of liquidation compared to enterprises that started the business basing only on own funds or family help. Enterprises that exported their goods or services had higher chances for survival (but this influence was not significant). A key risk factor was also the profit gained during the first years of activity. Enterprises that gained profits during the first year of activity had 50% lower risk of liquidation as compared to enterprises that made a loss or suspended their activity. Enterprises with a male owner (or company) and enterprises with an owner with high education had lower risk of liquidation but the effect of these variables was not statistically significant. Higher importance was attributed to enterprise development and barriers on the market. If an enterprise invested during the first year of activity it had about 45% lower risk of liquidation. Enterprises (respondents) that did not report any demand barriers during the first year of activity had 41% lower risk of liquidation compared to enterprises that reported some kinds of barriers.

In the case of variables such as: change in the number of workers, owner as the only employee and profit in the first year and investments or barriers the proportionality assumption was violated (based on results of supremum test). For these variables the interaction over time was included (see table 2).

Table 2. Interaction over time for variables with a violated proportionality assumption

Variable	Parameter	Std. error	Chi-square	Pr. > chi-sq.
Change in the number of workers*time	-0.00311	0.0003667	71.8844	<.0001
Owner as the only employee *time	-0.02507	0.00292	73.6585	<.0001
Profit in the first year*time	-0.02111	0.00169	155.5539	<.0001
Investments*time	-0.01707	0.00175	95.4634	<.0001
Barriers *time	-0.01638	0.00185	78.6185	<.0001

Source: own calculations based on the results of a retrospective survey on enterprises' survival in Malopolska voivodeship using SAS system.

A negative and significant parameter for interaction reveals a substantial and increasing in time effect of a given variable. For example, the effect of a change in the number of workers measured by a positive increase in the number of workers in

time is negative and additionally accelerates with time (negative significant interaction effect). If an increase in the number of workers was observed the risk of liquidation was lower about 22.7% compared to enterprises for which no increase in the number of workers was observed. This effect accelerated with the time of enterprise activity because the estimated parameter for interaction was -0.00311. For example after 12 months of activity the effect of this variable accelerates up to -0.22728 + (-0.00311)*12= - 0.2646.The results for other variables were very similar but the effect of acceleration was weak.

CONCLUSIONS AND DISCUSSION

Discussion of the results may be divided into two areas: internal and external factors driving the process of enterprises' survival.

The first internal factor important for firms survival is the size and age of a company, numerous investigations have found that larger and older firms have lower hazard rates than smaller and younger ones. However, this effect is not uniform. The size-age effects are only significant for single-person firms, and there is no size effect in new branches and subsidiaries of existing firms [Audretsch and Mahmood 1994, Mahmood 2000]. This means that the effect of ''liability of adolescence'' differs across different types and sectors of activity. According to the ecological theory of firms [Hannan 2005], smaller firms have a higher risk of failure - ''liability of smallness'' [Freeman et al. 1983], younger firms have the higher risk of failure - ''liability of newness'' [Stinchcombe 1965]. Enterprises that employed workers at the very beginning of activity were generally more successful and this effect in Cox model was significant. Also a change (growing number) of workforce of the enterprise plays a significant role.

Second, we have those studies that focus on differences in the legal structure of the firm [Mata and Portugal 2002, Esteve et al. 2004, Esteve-Pérez and Mañez-Castillejo 2008]. However, limited liability firms are more likely to exit through voluntary liquidation. In our model (for one of the regions in Poland) the evidence shows that sole traders are the most risky group, and any form of partnership decreases the risk of failure.

Other strategic activities that influence the survival are advertising and exporting [Kimura and Fujii 2003, Esteve et al. 2004, Esteve-Pérez and Mañez-Castillejo 2008]. Firms engaging in these activities seem to have lower hazard rates of exit [Mata and Portugal 2002]. In our sample we considered dummy: exported in first year or not exported, but this effect was not significant. More important was gaining a profit during the first period of activity.

Considering the socio-demographic profile of an entrepreneur as the most deterministic in enterprise failure we describe below the demography of owners. Those are the main internal factors determining the success and failure.

Only 12% of owners were unemployed before starting their own business which means that the influence of the market "push factor" was not significant (see

also [Ptak–Chmielewska 2012b]). In 70% of cases the employment in the own enterprise is the main source of maintenance of the main owner. In our model the effects of sex of the owner, education of the owner were not significant. The type of professional experience of the owner was significant. If the previous job of the owner was a job from the low risk group (engineering, white collar, manager, other) or it was a company the risk of liquidation was lower comparing to the situation where the owner had previously a job from the high risk group (farmer, craftsman, seller, student, unemployed).

The evaluation of the situation and conditions helping or disturbing the enterprises' development was focused on the environment and barriers existing on the market. 40% of enterprises did not invest in their first year of activity. Almost 48% of enterprises form investing group used their own funds, only 10% used loans or subsidies. Half of entrepreneurs did not report any barriers in sales of goods and services and production development. In subsequent years the share of enterprises that did not report barriers decreased slightly. The most important was the first year because this was the main driver of being on the market. 26% of enterprises reported too high competition on the market, other barriers were not significant.

Among external factors, the most important role is played by: the industry, the geographical space, the business cycle [Manjon–Antolı and Arauzo–Carod 2008]. First, the industry-specific characteristics such as technology, entry rates and scale economies seem to explain differences in survival rates across firms. In our research we controlled for sector of activity group. We divided sectors of activity into two groups: low risk group (industry, education, wholesale, hotels and restaurants, repairs, health care, other activity) and high risk group (construction, retail, financial intermediation, real estate and business activities, transport) but the difference was not significant. The change in the sector of activity was significant, confirming that elasticity of an enterprise plays significant role.

The second, the geographical space, according to New Economic Geography factors such as agglomeration economies, affects firm performance [Fujita et al. 1999]. In Poland enterprises registered in urban areas have higher chances for survival [Ptak–Chmielewska 2010]. More important are cross-border activities. According to our results, enterprises running activities on local or regional markets had lower chances for survival compared to enterprises with activities run on cross-national or international markets. Enterprises that succeeded in expanding outside regional and local markets had lower risk of liquidation.

The third, the business cycle, the chances of survival tend to be closely related to the evolution of the business cycle, being higher in the upswings and lower in the downturns. Business cycle upturn, shows negative and statistically significant coefficients in most research results with evidence that firms founded in periods of low unemployment rates have longer time of survival [Audretsch and Mahmood 1995, Mahmood 2000, Disney et al. 2003, Görgand Strobl 2003, Ptak–Chmielewska 2012b]. Macroeconomic conditions at the time of entry into the

market determine the probability of survival during the whole life of the enterprise [Mata et al. 1995]. However, using macroeconomic conditions must be combined with additional control variables.

In conclusion, we found that some internal factors such as previous experience of the owner or motivation play significant role in firms successes. Also, the external situation and barriers on the market matter considerably. Some of our results confirmed basic findings from previous research but some of detailed results gave new insight into the process of start-ups failure or success. The results for one region are generally consistent with results for the country. Expanding the activity into cross-national and international markets plays a significant role and increases the probability of survival. Surviving the first year is crucial, and this is strongly connected to gaining profits, investing in assets and barriers (demand side). Those effects accelerate in time. Research in this area including methods with time varying effect of characteristics was very rare and our work gives more light into this area. The application of event history methods and models rather than traditional methods in the analysis of enterprises' survival opens up opportunities for a better and wider use of the results in supporting enterprises by public policy.

REFERENCES

Allison P.D. (2010) Survival Analysis Using SAS. A Practical Guide Second Edition, SAS Publishing.

Audretsch D.B. (1995) Innovation and industry evolution. The MIT Press, Cambridge, MA

Audretsch D.B., Mahmood T. (1995) New firm survival: new results using a hazard function. Review of Economics and Statistics 77, pp. 97–103.

Becchetti L. Trovato G. (2002) The Determinants of Growth for Small and Medium Sized Firms. The Role of Availability of External Finance, Small Business Economics, 19, pp. 291-306.

Blossfeld H.P., Rohwer G. (2002) Techniques of Event History Modeling. New Approaches to Causal Analysis, Lawrence Elbaum Associates Publishers, London.

Cabral L., Mata J. (2003) On the Evolution of the Firm Size Distribution: Facts and Theory, American Economic Review, 93, pp. 1075–90.

Cooper A.C., Gimeno-Gascon F.J. Woo C.Y. (1994) Initial Human Capital and Financial Capital as Predictors of New Venture Performance, Journal of Business Venturing, 9, pp. 371–96.

Cox D.R. (1972) Regression models and life tables. Journal of the Royal Statistical Society, Series B 34, pp. 187–220.

Creedy J. Johnson P.S. (1983) Firm Formation in Manufacturing Industry, Applied Economics, 15, pp. 177–85.

Dehnel G. (2010) Rozwój mikroprzedsiębiorczości w Polsce w świetle estymacji dla małych domen. UniwersytetEkonomiczny w Poznaniu.

Disney R.Haskel J., Heden Y. (2003) Entry, exit and establishment survival in UK manufacturing. Journal of Industrial Economy 5, pp. 91–112

Ericsson R., Pakes, A. (1995) Markov perfect industry dynamics: A framework for empirical analysis. Review of Economic Studies, nr 62(1), pp. 53–82.

Esteve S., Sanchis A., Sanchis J.A. (2004) The determinants of survival of Spanish manufacturing firms. Review of Industrial Organization 25, pp. 251–273

Esteve-Pérez S., Mañez-Castillejo J.A. (2008) The Resource-Based Theory of the Firm and Firm Survival, Small Business Economics, 30, pp. 231–249.

Evans D.S.,Jovanovic B. (1989) An Estimated Model of Entrepreneurial Choice under Liquidity Constraints, Journal of Political Economics, 97, pp. 808–27.

Frątczak E. (2005) Analiza historii zdarzeń, Oficyna Wydawnicza SGH, Warsaw.

Freeman J., Carroll G., Hannan M. (1983) The liability of newness: age dependence in organizational death rates. American Sociological Review 48, pp. 692–710

Geroski P.A. (1995) What do We know about Entry?, International Journal of Industrial Organization, 13, pp. 421–40.

Geroski P.A. and Mazzucato M. (2001)Modelling the Dynamics of Industry Populations, International Journal of Industrial Organization, 19, pp. 1003–22.

GörgH., Strobl E. (2003) Footlose multinationals? Manchester School 71, pp.1–19

Hannan M.T. (2005) Ecologies of organizations: diversity and identity. Journal of Economy Perspective 19, pp.51–70

Hannan M.T., Freeman, J.H. (1989) Organizational Ecology, Harvard University Press, Cambridge, MA.

Hurst E. Lusardi A. (2004) Liquidity Constraints, Household Wealth and Entrepreneurship, Journal of Political Economy, 112, pp. 319–347.

Johnson P.S. (2005) Targeting Firm Births and Economic Regeneration in a Lagging Region, Small Business Economics, 24, pp. 451–64.

Kimura F., Fujii T. (2003) Globalizing activities and the rate of survival: panel data analysis on Japanese firms. Journal of Japanese Institutional Economics 17, pp.538–560

Lazear E. (2004) Balanced Skills and Entrepreneurship", American Economic Review Papers and Proceedings, 94, pp. 208–11.

Mahmood T. (2000) Survival of newly founded businesses: a log-logistic model approach. Small BusinessEconomics 14, pp. 223–237.

Markowicz I. (2012) Statystyczna analiza żywotności firm, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin.

Marsili O. (2002) Technological Regimes and Sources of Entrepreneurship, Small Business Economics, 19, pp. 217–31.

Mata J., Portugal P. (2002) The survival of new domestic and foreign owned firms. Strategy Management Journal 23, pp. 323–343.

Mata J., Portugal P.and Guimaraes P. (1995) The Survival of New Plants: Start-up Conditions and Post-entry Evolution, International Journal of Industrial Organization, 13, pp. 459–482.

Manjon–Antolı M.C., Arauzo–Carod J-M. (2008) Firm survival: methods and evidence, Empirica 35, pp. 1–24.

Nehrebecka N., Dzik A.M. (2013) Business demography in Poland: microeconomic and macroeconomic determinants of firm survival, University of Warsaw, Faculty of Economic Sciences, Working Papers, No. 8/2013 (93).

Porter M.E. (1979) The structure within industries and companies performance, Review of Economic and Statistics 61, pp. 214–227.

Poznańska K. (2008) Cykle życia przedsiębiorstw a instytucjonalna infrastruktura ich funkcjonowania. In: E. Mączyńska (ed.), Bankructwa przedsiębiorstw. Wybrane aspekty instytucjonalne. Przedsiębiorstwo Współczesne. Kolegium Nauk o Przedsiębiorstwie, SGGW Warsaw.

Ptak–Chmielewska A. (2010) Analiza przeżycia przedsiębiorstw w Polsce na przykładzie wybranego województwa. In: Dittmann P. E. Szabela-Pasierbińska (ed) Prognozowanie w zarządzaniu firmą. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu PN103, Wrocław.

Ptak–Chmielewska A. (2012a) Dostępność i przydatność danych do analizy przeżycia przedsiębiorstw, Wiadomości Statystyczne no. 6/2012, Warsaw

Ptak–Chmielewska A.(2012b) The Relation between Enterprise Population Dynamics and Economic Cycle, International Journal of Business, Humanities and Technology Vol. 2 No. 2, March. http://www.ijbhtnet.com/

Silva O. (2006) The Jack-of-All-Trades Entrepreneur: Innate Talent or Acquired Skill?, IZA discussion paper no. 2264, August, Bonn, IZA.

Stinchcombe F. (1965) Social structure and organizations. In: March JG (ed) Handbook of organizations. Rand McNally, Chicago, pp. 142–193

Vivarelli M. (2004) Are All the Potential Entrepreneurs So Good?, Small Business Economics, 23, pp. 41–49.

Warunki powstania i perspektywy rozwojowe polskich przedsiębiorstw powstałych w latach 2004-2008 (Creation and operation conditions, development prospects of Polish enterprises established in the years 2004-2008), GUS, Warsaw 2010.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 181 – 190

COMPARATIVE ANALYSIS OF THE INFORMATION SOCIETY DEVELOPMENT LEVEL IN THE POVIATS OF THE PODKARPACKIE VOIVODSHIP

Maria Sarama
Department of Quantitative Methods, University of Rzeszow
e-mail: msarama@univ.rzeszow.pl

Abstract: The aim of this paper is to compare the IS development in the poviats of the podkarpackie voivodship. The synthetic development measures were determined according to: the multiplicative aggregation of indices and the TOPSIS method. Two levels of aggregation and weights determined by the AHP method were used. The values of the indicators were calculated based on the results of surveys carried out in the poviats of the podkarpackie voivodship within MNiSW research project "Determination of intra-regional disparities in the information society development".

Keywords: information society, information and communication technologies, synthetic measures, information society development level

INTRODUCTION

In the studies of the information society (IS) one of the main areas of research is comparative analysis of the level of development in different territorial units. In these analyzes quantitative methods are used, i.a. indices, that can be divided into core indicators and composite indices (CI) [ITU 2012, OECD 2011, United Nations et al. 2005]. Many international organizations and research companies for the past thirty years have offered and updated different sets of indicators to measure development of the IS and scope of use of the Information and Communications Technologies (ICT) (some of them are characterized in [Goliński 2011]). Majority of these indicators are indicators for measuring development of the IS in the countries (NUTS1) and are not suitable for smaller territorial units such as NUTS3 or LAU, because of their specific nature. Thus for smaller units, it is necessary to develop other sets of indicators which take into account their specific.

Composite indices have increasingly been accepted as a useful tool for performance comparisons, benchmarking etc. in various fields such as economy, environment and society [OECD 2008, IANIS+ 2007]. Their usefulness depends heavily on the underlying construction scheme, so a problem faced by researchers is to determine the most suitable method. Technically, CI is a mathematical aggregation of a set of sub-indicators for measuring multidimensional concepts that cannot be captured by a single indicator [OECD 2008]. There are many methods developed for constructing CI (see i.a. [OECD 2008, Panek 2009, Młodak 2006, Strahl 2006]). It is worth noting, that in recent years also methods for multiple criteria decision analysis (MCDA), e.g. AHP, ANP, TOPSIS, have been applied to construct CI.

The aim of this paper is to compare the IS development in the poviats of the podkarpackie voivodship.

RESEARCH METHODS AND EMPIRICAL DATA

Composite indices construction involves the definition of study scope, selection of underlying variables (core indicators), data collection and preprocessing, weighting and aggregation of core indicators and post analysis of the derived CI (see i.a. [Panek 2009, OECD 2008, Młodak 2006]).

To compare level of the IS development in the poviats[1], we used (as a data source) results of surveys carried out within MNiSW research project "Determination of intra-regional disparities in the information society development", i.e. data from questionnaires completed by 3670 households and by more than 11 100 residents (aged from 16 to 74 years) of the rural poviats of podkarpackie voivodship.

To measure level of the IS development, we applied 22 core indicators, which are related to five pillars (aspects) of the IS (see table 1). These aspects correspond to three stages in the ITU model of ICT development process towards the information society [ITU 2012] i.e.: ICT readiness (infrastructure, access), ICT use (intensity) and partially ICT impact (outcomes).

These indicators were selected primarily on the basis of their substantive meaning, statistical criteria were also used. Indicators values were not comparable to each other and it was necessary to normalize them. All the core indicators were measured using a ratio scale, so the quotient mapping was applied. As reference values we took value of 100 (for some indicators) or sum of the arithmetic mean and three standard deviations. Adoption of three instead of two standard deviations was due to large differences between values of some indicators. Moreover, to diminish the effect of large number of outliers at the high end of the value scale, values of indicators having high right asymmetry were transformed by square root function.

[1] Poviats are the second-level units of local government and administration in Poland (i.e. local administrative units LAU1, previously called NUTS4).

Table 1. Pillars (sub-indices) and core indicators included in general indices and their weights determined by the AHP method

Pillars (sub-indices) and core indicators	Weight
1. Residents and households readiness for functioning in the information society:	0,10
percentage of households with a desktop computer or a laptop	0,10
percentage of households with the Internet access	0,14
percentage of households with the Internet access having a broadband connection	0,22
average number of computer-related skills held by residents	0,31
average number of skills related to the use of computer networks held by residents	0,23
2. Scope of use of computers by residents:	0,24
percentage of individuals who regularly (i.e. at least once a week) use a computer	0,14
percentage of individuals who regularly use a word processor	0,18
percentage of individuals regularly using a spreadsheet	0,33
percentage of individuals using a database software at least once during the three months	0,35
3. Scope of the Internet use by residents:	0,10
percentage of individuals regularly using the Internet	0,24
percentage of individuals who regularly use e-mail	0,31
percentage of individuals who receive files from the Internet at least once in three months	0,14
percentage of individuals who regularly use instant messaging	0,31
4. Scope of use of e-services offered in the Internet by residents:	0,45
percentage of individuals who make purchases in the Internet at least once in the three months	0,24
percentage of individuals regularly using the Internet to access their bank account	0,36
percentage of individuals who at least once during the three months search the Internet for purchase or sale offers of real estate, cars, etc.	0,09
percentage of individuals who at least once in three months seek, book or buy on the Internet offers such deals	0,22
percentage of individuals regularly receiving information about cultural events from the Internet	0,09
5. Scope of use of e-government services by residents:	0,11
percentage of individuals who at least once in three months contact via the Internet with the public administration (government or local government)	0,24
percentage of individuals submitting tax returns via the Internet	0,44
percentage of individuals using the Internet in dealing with matters relating to personal documents	0,19
percentage of individuals who contact via the Internet with the health services	0,13

Source: own elaboration based on surveys

Two levels of aggregation were applied i.e.: core indicators into sub-indices and sub-indices into general indices. At the first stage of aggregation, we used two different methods: weighted product method (which is one of the "classical" methods for the construction of CI) and the TOPSIS method.

In weighted product method, multiplicative aggregation is applied and composite indices P are calculated as the weighted geometric mean, i.e. according to the formula:

$$P_i = \prod_{j=1}^{m} (x_{ij})^{w_j} \tag{1}$$

where x_{ij} – normalized value of the j-th core indicator for the i-th poviat, w_j – weight assigned to j-th core indicator, m – number of core indicators.

The values of the sub-indices were determined as the weighted geometric mean instead of frequently used the weighted arithmetic mean, because in case of the additive aggregation there is complete substitution of aggregated indicators (which means that low values of some of indicators are "fully compensated" by a sufficiently high values of the other). Whereas, the geometric aggregation is a less compensatory approach, which contributes to take actions to improve underperforming dimensions.

The TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) is based on the concept that the best alternative (or item, e.g. poviat) should have the shortest distance from the ideal one and must have greatest distance from the negative one. A detailed description of this method can be found i.a. in [Rao 2010, Wysocki 2010, Deng et al. 2000]. We defined as ideal poviat – poviat with maximal values of core indicators and as negative-ideal poviat – poviat with minimal values of them, so composite indices T are calculated as:

$$T_i = \frac{\sqrt{\sum_{j=1}^{m} w_j (x_{ij} - \min_i \{x_{ij}\})^2}}{\sqrt{\sum_{j=1}^{m} w_j (x_{ij} - \min_i \{x_{ij}\})^2} + \sqrt{\sum_{j=1}^{m} w_j (x_{ij} - \max_i \{x_{ij}\})^2}} \tag{2}$$

where x_{ij} – normalized value of the j-th core indicator for the i-th poviat, w_j – weight assigned to j-th core indicator, m – number of core indicators.

Both in the weighted product method, as well as, in the TOPSIS method, results can depend strongly on the selected weights. In literature several weighting methods are proposed, e.g. equal weights, weights based on statistical methods and weights based on public/expert opinion (see i.a. [Panek 2009, OECD 2008, Wysocki 2010]). One of methods for calculating the weights in MCDA is the method of AHP (Analytical Hierarchy Process), in which measures of importance of criteria (or weights assigned to indicators) are determined on the basis of comparisons of them to each other in pairs by means of a rating scale introduced T. L. Saaty in the 1970s. A detailed description of this method can be found i.a. in [Rao 2010, Wysocki 2010]. The AHP method was used to determine weights w_j in

formulas (1) and (2). Pair-wise comparison matrix was established on the basis of author's knowledge about the substantive importance of particular indicators.

At the second level of aggregation, sub-indices obtained by both the weighted product method, as well as, in the TOPSIS were aggregated by the multiplicative method with weights determined by the AHP method.

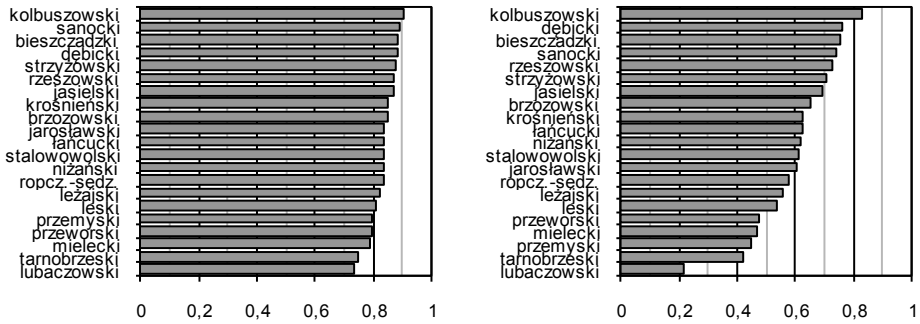
RESEARCH RESULTS

In order to obtain an accurate picture of the spatial differentiation of level of the information society development of in the podkarpackie voivodship, i.e. of the disparities which exist between its poviats, values of the core indicators were calculated on the basis of the collected data. Then, for each poviat, we computed the values of P (by the product method) and T (by the TOPSIS method) sub-indices corresponding to five aspects of the information society development. Obtained results are graphically presented in figures 1–5. In these figures we put two graphs: the left one shows the P values and the right one presents the T values.

The values presented in figures indicate that, for each pillar, poviats rankings created from the sub-indices P and T are very similar. This is also confirmed by the Spearman's rank-order correlation coefficients, which values range from 0,979 (readiness to function in the IS) to 0,994 (use of e-services). A comparison of poviats positions in the rankings that are based on sub-indices P and T shows that the greatest differences between positions in the rankings relate to poviats:

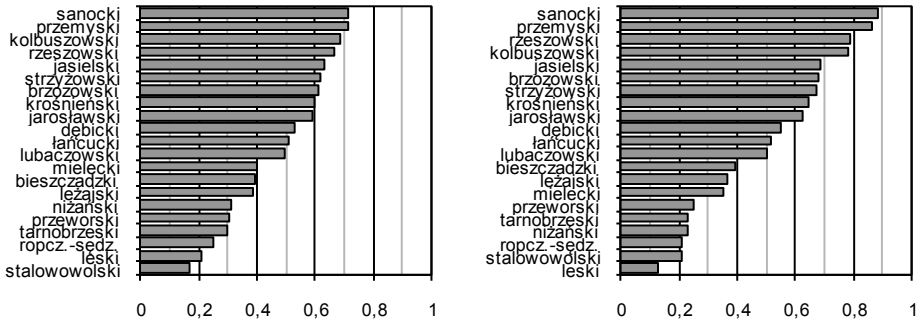
- readiness to function in the IS: jarosławski (3); niżański, przemyski and sanocki (2);
- use of computers: mielecki and niżański (2);
- use of the Internet: brzozowski and niżański (2);
- use of e-services: bieszczadzki and przemyski (2);
- use of e-government services: niżański (3), lubaczowski and przeworski (2).

Figure 1. Residents and households readiness for functioning in the information society (P1 and T1)



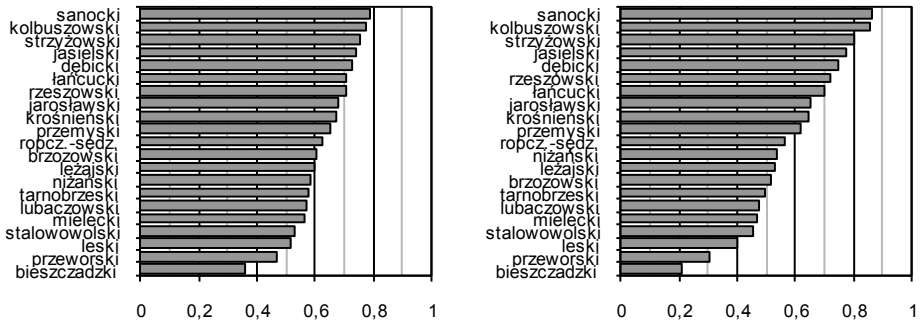
Source: own elaboration based on surveys

Figure 2. Scope of use of computers by residents (P2 and T2)



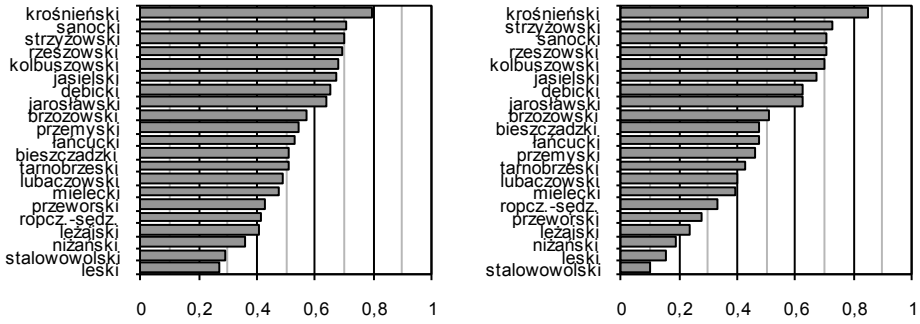
Source: own elaboration based on surveys

Figure 3. Scope of the Internet use by residents (P3 and T3)



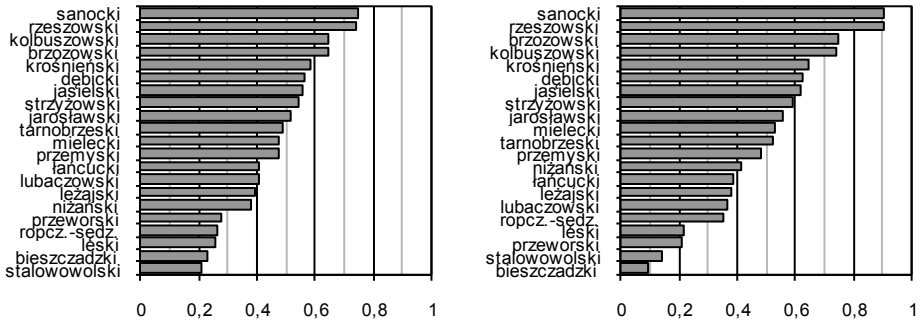
Source: own elaboration based on surveys

Figure 4. Scope of use of e-services offered in the Internet by residents (P4 and T4)



Source: own elaboration based on surveys

Figure 5. Scope of use of e-government services by residents (P5 and T5)



Source: own elaboration based on surveys

Analysis of results, given in figures, shows that the range of variation and differentiation of T sub-indices are greater than of sub-indices P. Coefficients of variation of sub-indices corresponding to aspects of IS development are as follows:

- readiness to function in the IS: for P1 – 5% and for T1 – 23%;
- use of computers: for P2 – 36% and for T2 – 46%;
- use of the Internet: the P3 – 17% and for T3 – 29%;
- use of e-services: for P4 – 26% and for T4 – 43%;
- use of e-government services: for P5 – 34% and for the T5 – 45%.

Sub-indices refer to different aspects of the information society. Therefore, we decided to see how strong correlations exist between their values. The values of Pearson correlation coefficients between sub-indices allow us to draw the following conclusions:

- Sub-indices determined by the multiplicative method: The least correlated with the other sub-indices is P1 (readiness to function in the IS), and the most – sub-indices P3 (use of the Internet) and P4 (use of e-services). At the same time the strongest relationships exist between P2 (use of computers) and P4 – 0,88 and P5 (use of the e-government services) – 0,87, the weakest – between P1 and P5 – 0,46 and P2 – 0,46.
- Sub-indices determined by the TOPSIS method: Also the least correlated with the other sub-indices is T1 (readiness to function in the IS), and the most – sub-indices T3 (use of the Internet) and T4 (use of e-services). At the same time the strongest relationships exist between T2 (use of computers) and T4 – 0,83 and between T3 and T5 (use of e-government services) – 0,80 and the weakest – between T1 and T5 – 0,45 and T2 – 0,46.

From figures it can be seen, that poviats occupy different positions in the rankings based on various sub-indices. High positions in the rankings created on the basis of sub-indices P and T are occupied by poviats: sanocki,

kolbuszowski, rzeszowski, strzyżowski, jasielski. Low places in P rankings are usually occupied by the poviats: leski, stalowowolski, przeworski, ropczycko-sędziszowski, niżański and in T rankings – by poviats: leski, stalowowolski, przeworski, lubaczowski, ropczycko-sędziszowski. Differences between positions occupied in P rankings by poviats: sanocki, jarosławski, jasielski, kolbuszowski, przeworski and in T rankings poviats: jasielski, sanocki, rzeszowski, kolbuszowski, przeworski – are relatively small. Simultaneously, the most diversified in the P rankings are positions occupied by poviats: bieszczadzki, przemyski, tarnobrzeski, stalowowolski, lubaczowski, and in T rankings – poviats: bieszczadzki, przemyski, brzozowski, niżański, stalowowolski.

To obtain an overall assessment of the level of IS development in poviats of the podkarpackie voivodeship, sub-indices corresponding to the aspects of IS development were aggregated into two composite indices IP and IT. IP and IT values were calculated on the basis of sub-indices P and T as the weighted geometric mean with the weights determined by the method of AHP. Similarly as rankings established on the basis of sub-indices, rankings of poviats created from the indices IP and IT are alike, value of the Spearman's rank-order correlation coefficient is equal to 0,98.

Analysis of IP and IT values shows that the range of variation and differentiation of index IT are greater than of index IP, the coefficients of variation are equal to 39% and 25% for index IT and IP respectively. A comparison of poviats positions in the rankings that are based on indices IP and IT shows that the greatest differences between positions in the rankings relate to poviats: ropczycko-sędziszowski (3), kolbuszowski, krośnieński and leżajski (2).

Table 2. Poviats classifications based on the indices IP and IT

Developmental stage	IP	IT
High ($> \bar{I} + s$)	sanocki, krośnieński, rzeszowski, kolbuszowski	sanocki, kolbuszowski, rzeszowski, krośnieński, strzyżowski
Higher than average ($\bar{I} < I \leq \bar{I} + s$)	strzyżowski, jasielski, dębicki, jarosławski, brzozowski, przemyski, łańcucki	jasielski, dębicki, jarosławski, brzozowski, przemyski, łańcucki
Lower than average ($\bar{I} - s < I \leq \bar{I}$)	lubaczowski, mielecki, tarnobrzeski, leżajski, bieszczadzki	mielecki, lubaczowski, tarnobrzeski, bieszczadzki, ropczycko-sędziszowski, leżajski
Low ($I \leq \bar{I} - s$)	przeworski, niżański, ropczycko-sędziszowski, leski, stalowowolski	przeworski, niżański, leski, stalowowolski

Source: own elaboration based on surveys

The values of the indices IP and IT were used to determine the groups of poviats with similar levels of IS development. To determine the limits of the classes we used the arithmetic mean ($\bar{I}$) and standard deviation (s) of indices IP and IT. Table 2 shows received poviats classification.

The results in table 2 indicate that selected groups of poviats do not form distinct clusters on map of the podkarpackie voivodeship (e.g. sanocki and rzeszowski, leski and stalowowolski). There is no center-periphery differentiation, it is sufficient to compare the positions of the two poviats: rzeszowski and sanocki. Also the location of the poviat close to large urban centers (urban poviats) do not always contribute to a high level of IS development in its area (e.g. tarnobrzeski and przemyski).

CONCLUDING REMARKS

No satisfactory and widely accepted definition of information society and the rapid development of information and communication technologies and their increasingly wide applications cause that the substantive meaning of some of the core indicators may change as time goes. Therefore the core indicators, used in this study, were selected so as to concern all stages in the model of ICT development process towards the information society and to be appropriate for measuring and comparing the level of the SI development in territorial units such as LAU1 (NUTS4) now and in the coming years.

The results of the research show that poviats rankings based on indices obtained by the product method and the TOPSIS method are very similar. There is no center-periphery differentiation in the IS development in podkarpackie voivodship and the location of poviat close to large urban centers do not affect the level of IS development in its area.

Having a knowledge of spatial differences and similarities between the IS development in territorial units, allows a more rational allocation resources to support development of the IS and the e-economy. Valuable conclusions can be drawn from the separate analyzes of sub-indices (lower level composite indices) and the relationships between them. Among other things, it is possible to identify the strengths and the weaknesses of each territorial unit in the IS development.

REFERENCES

Goliński M. (2011) Społeczeństwo informacyjne – geneza koncepcji i problematyka pomiaru, Oficyna Wydawnicza SGH, Warszawa.

IANIS+ (2007) Guide to Regional Good Practice Indicators and Benchmarking, Brussels.

Deng H., Yeh C.H., R.J. Willis (2000) Inter-company comparison using modified TOPSIS with objective weights, Computers & Operations Research 27, pp. 963–973.

ITU (2012) Measuring the Information Society 2012, http://www.itu.int/ITU-D/ict/publications/idi/material/2012/MIS2012_without_Annex_4.pdf

Młodak A. (2006) Analiza taksonomiczna w statystyce regionalnej, Difin, Warszawa.

OECD (2008) Handbook on Constructing Composite Indicators: Methodology and User Guide, http://www.oecd.org/dataoecd/37/42/42495745.pdf

OECD (2011) Guide to Measuring the Information Society, OECD Publishing, http://browse.oecdbookshop.org/oecd/pdfs/free/9311021e.pdf

Panek T. (2009) Statystyczne metody wielowymiarowej analizy porównawczej, Oficyna Wydawnicza SGH, Warszawa.

Rao R. V. (2010) Decision Making in the Manufacturing Environment: Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods, Springer-Verlag, London.

Strahl D. (ed.) (2006) Metody oceny rozwoju regionalnego, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

United Nations Partnership on Measuring Information and Communication Technology for Development (2005), Core ICT indicators, http://www.itu.int/en/ITU-D/Statistics/Documents/partnership/CoreICTIndicators.pdf

Wysocki F. (2010) Metody taksonomiczne w rozpoznawaniu typów ekonomicznych rolnictwa i obszarów wiejskich, Wydawnictwo Uniwersytetu Poznańskiego, Poznań.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 191 – 199

SPATIAL DIVERSITY OF HUMAN CAPITAL IN THE EUROPEAN UNION

Iwona Skrodzka
Department of Econometrics and Statistics, University of Bialystok
e-mail: i.skrodzka@uwb.edu.pl

Abstract: The aim of this study is a comparative analysis of European Union countries in terms of human capital. Determination of the stock and prospects of human capital development is an important issue today, both in economic theory and business practice. In this study soft modeling method was used. It allows measurement of unobserved variables.

Keywords: human capital, economic development, soft modeling

INTRODUCTION

Human capital is a significant factor of economic growth [Próchniak 2006, pp. 320-323; Cichy, Malaga 2007, pp. 20-49; Florczak 2007, pp. 126-166]. Therefore, determination of the stock and prospects of human capital development is an important issue today, both in economic theory and business practice.

Human capital can be defined as a stock variable that represents the capacity of an individual, household, nation to generate a sustained flow of earned income [Dagum 2004, p. 1]. Furthermore human capital should be considered as a complex, multifaceted category with various intangible dimensions that are not directly observable and that cannot be measured with precision by a single attribute [Le, Gibson, Oxley 2005, p. 4; Łukasiewicz 2009, p. 96].

The aim of this study is to present spatial diversity of human capital in European Union countries in 2010[1]. In the article human capital is defined as an unobserved variable reflected by such components as: education, knowledge, skills, work experience and health embodied in region society [Domański 1993, p. 19; Marciniak 2000, pp. 157-158].

[1] Data availability influence the year choice.

The method which was used in this research is soft modelling[2]. Soft model enables to investigate the relationships among unobserved variables (latent variables). The values of these variables cannot be directly measured because the lack of a generally accepted definition or the absence of a clear way of measuring them. Soft model consists of two sub-models:

- the internal sub-model – a system of relationships among latent variables, which describes the relationship arising from the theory;
- the external sub-model – defines the latent variables based on observed variables, known as indicators.

Indicators allow indirect observation of latent variables. Latent variables can be define on the basis of deductive or inductive approach. Deductive approach assumes that indicators reflect latent variable. Inductive approach assumes that indicators form latent variable. The choice of approach depend on the theory or intuition of researcher [Rogowski 1990, pp. 25-26].

Thanks to soft model is possible to get synthetic measures of latent variables (as a weighted sum of indicators). One of the most important advantage of soft modelling method is that the construction of synthetic measure base not only on latent variable definition but also on relationships among other categories within model.

The parameters of soft model are estimated using partial least squares method (PLS). Statistical verification is done by Stone-Geisser test and "2s" rule[3].

In the literature description of the method can be found in Wold [1980], its generalization in Rogowski [1990] and examples of application in [Perło 2004, Skrodzka 2012]).

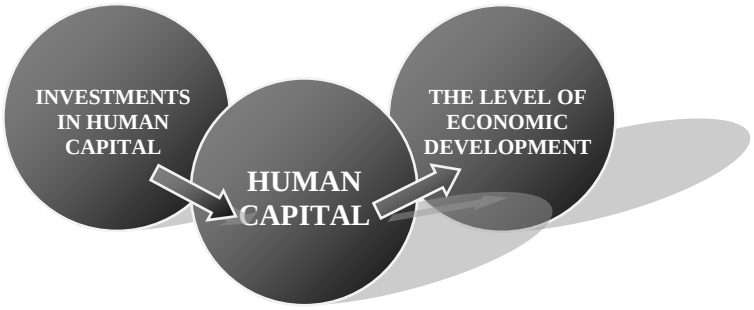
SPECIFICATION OF THE INTERNAL SUB-MODEL

Figure 1 presents the concept of internal sub-model. The concept assumes relationships among three unobserved categories: human capital, investments in human capital and the level of economic development. The first relationship assumes that human capital is a factor of economic development, the second – that human capital can be increased through investments.

[2] Soft modeling is a method proposed by Herman Wold [Wold 1980].

[3] Parameter is statistically significant when value of double error is higher than value of estimator.

Figure 1. The concept of internal sub-model



Source: own elaboration

Estimated model contains two following equations

$$HC_t = \alpha_1 IHC_{t-2} + \alpha_2 IHC_{t-1} + \alpha_3 IHC_t + \alpha_0 + \varepsilon \quad (1)$$

$$LED_t = \beta_1 HC_t + \beta_0 + \xi \quad (2)$$

where
HC – human capital,
IHC – investments in human capital,
LED – the level of economic development,
α_0, α_1, α_2, α_3, β_0, β_1 – structural parameters,
ε, ξ – error terms,
t – 2010.

SPECIFICATION OF THE EXTERNAL SUB-MODEL

Each of unobserved variables is defined by the group of indicators (see Table 1). Deductive approach is used to define above variables. Data use to specify the model were taken from World Bank[4] and Eurostat[5] and they refer to period 2008-2010. Many indicators are covered in databases. The analysis of all indicators would be unclear and difficult to interpret, hence the selection is necessary. The criteria are following:

- universality (commonly respected indicators),
- comparability (indicators as coefficients of intensity),
- variety (coefficient of variation higher than 10%).

[4] http://data.worldbank.org/
[5] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes

Table 1. Indicator of latent variables

Latent variable	Indicator	Meaning	Source[6]
HC	HC01	Persons with tertiary education attainment (%).	E
	HC02	Employees with tertiary education attainment (%).	E
	HC03	Life-long learning of persons aged 25-64 (%).	E
	HC04	Human resources in science and technology (per 100 thous. people).	E
	HC05	Researchers in R&D (per million people).	WB
	HC06	Patent applications (per million people).	WB
	HC07	Percentage of people declaring their health as very good (%)	E
	HC08	Life expectancy at birth (years).	WB
	HC09	Mortality rate, neonatal (per 1000 live births).	E
IHC	IHC01	Total public expenditure on education (% of GDP).	E
	IHC02	Total public expenditure on education (PPS, per capita).	E
	IHC03	Total expenditure on health (% of GDP).	WB
	IHC04	Total expenditure on health (PPS, per capita).	WB
	IHC05	Total expenditure on R&D (% of GDP).	E
	IHC06	Total expenditure on R&D (PPS, per capita).	E
LED	LED01	Gross domestic product (PPS, per capita).	WB
	LED02	Gross value added (euro, per employee).	E
	LED03	The share of agriculture in gross value added (%).	E
	LED04	The share of services in gross value added (%).	E
	LED05	Unemployment rate (%).	WB

Source: own elaboration

Internal model of HC latent variable contains nine indicators. They reflect: education, knowledge, skills and health embodied in the society of the country. One of them (HC09) is destimulant. Internal model of INHC latent variable contains six observed variables which refer to investments in education, health and knowledge. All of them are stimulants. Internal model of LED latent variable contains five indicators. They reflect economic potential of country. Two of them (LED03, LED05) are destimulant.

ESTIMATION RESULTS

Model presented on Figure 1 was estimated using the PLS software created by J. Rogowski[7]. Table 2 contains weight and loadings estimates with regard to

[6] WB – World Bank, E – Eurostat.

[7] PLS software is available at Faculty of Economics and Management University of Bialystok.

external sub-model. All parameters are statistically significant[8]. Moreover, results are consistent with expectations. Stimulants have positive weights and loadings, destimulants have negative ones.

Table 2. Estimates of weights and loadings of the external model

Latent variable	Indicator	Loading	Weight
HC_t	HC01	0,7509	0,1370
	HC02	0,5965	0,0890
	HC03	0,8025	0,1864
	HC04	0,8428	0,1829
	HC05	0,6594	0,1342
	HC06	0,6324	0,1607
	HC07	0,4017	0,1125
	HC08	0,7016	0,2151
	HC09	-0,7811	-0,1973
IHC_t	IHC01	0,6017	0,1647
	IHC02	0,7222	0,1454
	IHC03	0,8462	0,2018
	IHC04	0,8665	0,2196
	IHC05	0,9201	0,2366
	IHC06	0,9285	0,2337
IHC_{t-1}	IHC01	0,5539	0,1570
	IHC02	0,6774	0,1394
	IHC03	0,8550	0,2060
	IHC04	0,8888	0,2248
	IHC05	0,9281	0,2408
	IHC06	0,9359	0,2342
IHC_{t-2}	IHC01	0,5005	0,1425
	IHC02	0,7290	0,1618
	IHC03	0,8686	0,2075
	IHC04	0,8803	0,2201
	IHC05	0,9229	0,2418
	IHC06	0,9274	0,2303
LED_t	LED01	0,9198	0,2241
	LED02	0,9018	0,2271
	LED03	-0,9104	-0,2411
	LED04	0,7658	0,1858
	LED05	-0,5303	-0,0909
	LED06	-0,7746	-0,2312

Source: own calculation

[8] Doubled standard deviation calculated by Tukey cut method were less than the value of the estimator.

Indicators HC04 and HC03 are the most strongly correlated with HC variable. Indicators HC09, HC01 and HC08 have strong influence on HC variable. Indicators HC07 reflects HC variable poorly. To sum up knowledge is the most significant component of human capital in UE-27 countries.

All indicators reflect IHC variable strongly. Indicators connected with R&D sector (IHC05 and IHC06) have the highest influence on IHC variable.

Equations (3) and (4) present estimations of internal relations. Standard deviations calculated basing on Tukey cut method are given in brackets.

$$\underset{}{\hat{HC}_t} = \underset{(0,0196)}{0,5612}IHC_{t-2} + \underset{(0,0317)}{0,2580}IHC_{t-1} + \underset{(0,0456)}{0,0922}IHC_t + \underset{(0,1012)}{5,0458} \tag{3}$$

$$\underset{}{\hat{LED}_t} = \underset{(0,0291)}{0,7678}HC_t - \underset{(0,3532)}{4,6541} \tag{4}$$

Signs of estimators are consistent with expectations. Moreover, all parameters are statistically significant („2s" rule). Coefficient of determination (R^2) have value 0,8 for the equation (3) and value 0,6 for the equation (4). General Stone–Geisser test is equal to 0,36[9]. The model can be verified positively.

Investments in human capital (in 2008, 2009 and 2010) influence on the stock of human capital positively. Investments in 2008 have the highest impact on the stock of human capital, investments in 2010 – the lowest. Furthermore, correlation between human capital and the level of economic development is high and positive. It is possible to claim that countries which invested more in human capital had the higher stock of human capital in 2010. Moreover countries which had the higher stock of human capital, had also the higher level of economic development in 2010.

HUMAN CAPITAL IN EUROPEAN UNION COUNTRIES

Partial Least Square method used to soft model estimation provides calculations of latent variable values. These values can be treated as synthetic measure and used for comparative analysis.

Figure 2 presents diversity of investments in human capital in European Union in 2008. Countries were divided into four groups which were constructed basing on parameters of synthetic measure (z_i): average $(\bar{z})$ and standard deviation (s_z) [Nowak, 1990, pp. 92-93]:

- I group – very high investments in human capital: $z_i \geq \bar{z} + s_z$,
- II group – high investments in human capital: $\bar{z} \leq z_i < \bar{z} + s_z$,

[9] Stone-Geisser test measures prognostic property of soft model. Its values are in the range from -∞ to 1. Positive (negative) value of this test indicates high (poor) quality of model.

- III group – medium and low investments in human capital: $\bar{z} - s_z \leq z_i < \bar{z}$,
- IV group – very low investments in human capital: $z_i < \bar{z} - s_z$.

Denmark, Sweden, Finland, Austria, Luxemburg and Netherlands were the biggest investors in human capital in 2008. The second group is composed of: Belgium, Germany, France, Ireland and United Kingdom. Slovenia, Spain, Portugal, Cyprus, Italy, Greece, Malta, Estonia, Czech Republic and Hungary were classified to third group. The rest of countries, including Poland was classified to the last group with low investments in human capital.

Figure 2. Diversity of investments in human capital in European Union in 2008

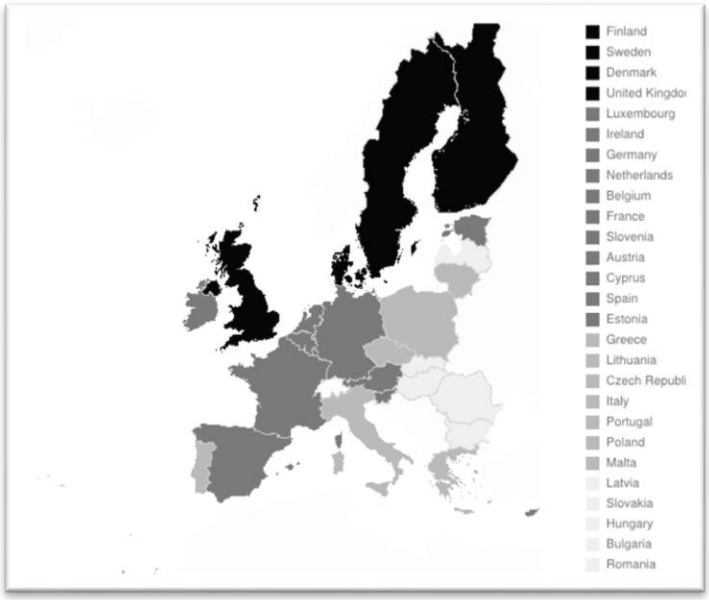
Source: own elaboration

Diversity of human capital in European Union in 2010 is shown in Figure 3. Countries were divided into four groups:

- I group – very high stock of human capital,
- II group – high stock of human capital,
- III group – medium and low stock of human capital,
- IV group – very low stock of human capital.

The highest stock of human capital was concentrated in Finland, Sweden, Denmark and United Kingdom. Luxemburg, Ireland, Germany, Netherlands, Belgium, France, Slovenia, Austria, Cyprus, Spain and Estonia were located in the second group. Greece, Lithuania, Czech Republic, Italy, Portugal, Poland and Malta build third group. The rest of countries was classified to the last group with very low stock of human capital.

Figure 3. Diversity of human capital in European Union in 2010



Source: own elaboration

SUMMARY

The presented soft model has enabled to analyze spatial diversity of human capital and investments in human capital in European Union countries. The rankings of countries were created thanks to estimated values of latent variables. Some conclusions and remarks can be formulated according to the results of this study:

- knowledge is the most significant components of human capital in UE-27,
- expenditures on R&D sector are the most significant form of investing in human capital in UE-27,
- investments in human capital influence on the human capital stock positively in UE-27,
- human capital have positive influence on the level of economic development in UE-27,
- the highest stock of human capital in 2010 was concentrated in Finland, Sweden, Denmark and United Kingdom,
- Denmark, Sweden, Finland, Austria, Luxemburg and Netherlands were the biggest investors in human capital in 2008.

REFERENCES

Cichy K., Malaga K. (2007) Kapitał ludzki w modelach i teorii wzrostu gospodarczego, [in:] M. Herbst (ed.), Kapitał ludzki i kapitał społeczny a rozwój regionalny, Wydawnictwo Naukowe „Scholar", Warszawa.

Dagum C. (2204) Human capital. Encyclopedia of Statistical Sciences, John Wiley & Sons, pp. 1-12.

Domański S. R. (1993) Kapitał ludzki i wzrost gospodarczy, PWN, Warszawa.

Florczak W. (2007) Kapitał ludzki a rozwój gospodarczy, [in]: W. Welfe (ed.) Gospodarka oparta na wiedzy, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Le T., Gibson J., Oxley L. (2005) Measures of Human Capital: A Review of the Literature, Working Paper 05/10, New Zealand Treasury, New Zealand.

Łukasiewicz G. (2009) Kapitał ludzki organizacji. Pomiar i sprawozdawczość, PWN, Warszawa.

Marciniak S. (2000) Innowacje i rozwój gospodarczy, Kolegium Nauk Społecznych i Administracji Politechniki Warszawskiej, Warszawa.

Nowak E. (1990) Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych, PWE, Warszawa.

Perło D. (2004), Źródła finansowania rozwoju regionalnego, Wydawnictwo Wyższej Szkoły Ekonomicznej w Białymstoku, Białystok

Próchniak M. (2006) Czynniki wzrostu gospodarczego – wnioski z badań empirycznych, Ekonomista, no 3.

Rogowski J. (1990) Modele miękkie. Teoria i zastosowanie w badaniach ekonomicznych, Wydawnictwo Filii UW w Białymstoku, Białystok.

Skrodzka I. (2012) Zastosowanie modelowania miękkiego do pomiaru kapitału ludzkiego, rozprawa doktorska, Uniwersytet w Białymstoku, Białystok [maszynopis niepublikowany].

Wold H. (1980) Soft Modelling: Intermediate between Traditional Model Building and Data Analysis, Banach Centre Publication 6, Mathematical Statistics.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 200 – 209

MEASURE OF THE LEVEL OF SOCIO–ECONOMIC DEVELOPMENT IN PROVINCES

Aneta Sobiechowska–Ziegert, Aniela Mikulska
Department of Economic Sciences, Gdansk University of Technology
e-mail: Aneta.Sobiechowska@zie.pg.gda.pl, Aniela.Mikulska@zie.pg.gda.pl

Abstract: The scope of Polish macro-economic data for assessing the level of socio-economic development of the country is largely limited because of the regional variation. Therefore there is a need for cyclic selection and the verification of criteria which allow identifying regions with a similar level of socio-economic development or those that clearly differ from the mean values. The aim of the study is to compare Poland's provinces in terms of their socio-economic development, which helps identify the most similar regions as far as the adopted criteria are concerned.

Keywords: socio-economic development, synthetic variables, classification of regions

INTRODUCTION

Regional differences restrict the use of economic data to assess the level of the social-economic development of selected locations in Poland. Information about diversity can be used to analyse various topics related to social activity including the differences in the intensity of entrepreneurship in the region. The assessment of entrepreneurial activity range can be made through a variety of variables such as structure of employment, labour productivity, level of urbanization and access to infrastructure (including educational) [Strużycki 2004]. The level of socio-economic development of regions determines changes in the SME sector, because changes in the economic system and taking place in the area of efficiency and effectiveness cause the extension of production capacity, and thus they lead to the improvement of the living conditions [Strużycki 2004], which in turn result in the intensification of local entrepreneurial activities. There are many opportunities to acquire and configure variables used to assess the level of socio-economic development. Moreover, there is a real need for cyclic selection,

analysing and verifying the variables (and their groups) which allow for the indication of regions with a similar level of socio-economic development, or those that clearly deviate from the mean. The selection of variables relating to the socio-economic situation seems to be a relatively simple task, like determining the impact direction of a factor on the level of development. However, not only choice but also variable verification by means of statistical measure (e.g. variation or asymmetry coefficient) classified at the appropriate level is important. The issues of availability, integrity and comparability of data to be analysed constitute a major difficulty for researchers . This significantly determined the selection of variables taken into consideration in this study. The ultimate source of information was Local Data Bank of Central Statistical Office (CSO).

The aim of the paper is to identify the most similar regions in terms of the variables chosen for the study. A comparative analysis of provinces in terms of their socio-economic development is provided. A desire to collect information about Polish provinces which could then be used to compare the conditions necessary for their development at the enterprise level inspired the research. Studies in which entrepreneurs usually participate should be supported by an objective assessment of the conditions in the environment in which they take decisions - often of a strategic nature [Bratnicki 2011]. Many publications point to the fact that the period of the transformation of Polish economy has resulted in growing diversity in the level of the socio-economic development of provinces, which has deepened the division of the country into the eastern and western parts [eg Strużycki 2004, Bizon 2011].

VARIABLES SELECTED FOR THE RESEARCH

The regional variation often refers to the basic indicator, which is GDP per capita. In the study presented here this indicator is not used, mainly because of the lack of available data for the year 2010, the period in which the variables were collected, but also because of the fact that the relative value of production does not translate directly and automatically into the standard of living in local communities in provinces [Bizon 2011]. Therefore, the study examines the measures that clearly determine the level of socio-economic development. The research period covers the years 2008-2010 due to the availability of CSO data and the opportunity to make a comparison with available studies.

The variables that meet the criteria of completeness, availability and comparability in the analysed period were pre-qualified for further study and subjected to further selection according to statistical criteria. Table 1 presents the variables describing the level of socio-economic development which were adopted at the initial stage of the analysis. For the sake of further study, only the variables that meet the criteria used in this kind of research were analysed [Zeliaś 2000]. These criteria are:

- sufficient variability measured by variation coefficient with threshold value ε=0,1
- positive asymmetry (in the case of stimulants) or distribution close to symmetric (the skewness index close to zero).

Table 1. Variables subjected to initial verification

Variable	Description
X_1	Average, monthly gross salary in PLN
X_2	Average, monthly disposable income per person in PLN
X_3	Average, monthly disposable income per person in employment in PLN
X_4	Average, monthly disposable income per person in self-employment in PLN
X_5	Registered unemployment rate in %
X_6	Risk of poverty rate (relative poverty line in %)
X_7	Number of physicians per 10000 inhabitants
X_8	Share of expenditure on food and non-alcoholic beverages in the total expenditure in %
X_9	Share of working age population as a % of total population
X_{10}	Average useable floor space in m^2
X_{11}	Registered passenger cars per 1000 population
X_{12}	Structure of employment by sector (agriculture, forestry)
X_{13}	Structure of employment by sector (manufacturing)
X_{14}	Structure of employment by sector (services)
X_{15}	Capital expenditures in the private sector per capita in PLN
X_{16}	Capital expenditures per capita in PLN
X_{17}	Expenditures on R & D per one entity in the business sector in PLN
X_{18}	Expenditures on R & D per one inhabitant in PLN
X_{19}	Hard surface roads in km per 100 km^2
X_{20}	Entities entered in the REGON register per 10 thousand population
X_{21}	Share of the SME sector in sold industrial output total in %
X_{22}	Participation of business in expenditures on R & D in %

Source: Local Data Bank of CSO in Poland

The initial selection of variables allowed for the identification of the following factors relevant to the assessment of differences in the social dimension:

- average, monthly disposable income per person in PLN (X_2) [1],
- registered unemployment rate (X_5),
- risk of poverty rate - relative poverty line (X_6),
- number of physicians per 10000 inhabitants (X_7).

[1] This indicator reflects the purchasing power of households better than the gross salary indicator, and thus allows determining more accurately the perceived level of prosperity, which codetermines the level of socio-economic development.

Taking into consideration the variables concerning income $(X_1 \cdots X_4)$, the amount of average monthly disposable income per one inhabitant seemed the most representative. An attempt was made during the study to determine to what extent the source of income (wage labour, self-employment) was a variable that differentiated the data in the region. This variable, however, was highly correlated with average monthly disposable income per one inhabitant.

Variables approximately describing the living conditions were extended in the initial phase of the study to include other aspects: average useable floor space in m^2, registered passenger cars per 1000 population, share of expenditure on food and non-alcoholic beverages in the total expenditure and share of working age population as a percentage of total population. However, these were eliminated in the next stage of the analysis due to the level of variation coefficient ($\varepsilon \leq 0{,}1$) being lower than it was first assumed. Because of the same reason, the access to the Internet variable was earlier eliminated. Employment structure by sector $(X_{12} \cdots X_{14})$ was not regarded as important because of the very low variation coefficient for the service sector and negative skewness index for the industrial sector.

The criteria that also enabled the researchers to define the level of socio-economic development of the regions were as follows:

- capital expenditures per capita (X_{16}),
- expenditures on R & D per one inhabitant (X_{18}),
- hard surface roads in km per 100 km^2 (X_{19}),
- entities entered in the REGON register per 10 thousand population (X_{20}),
- participation of business in expenditures on R & D (X_{22}).

These indicators correspond to a large extend to the level of infrastructure that determines the success of business ventures. These projects, in turn, are the source and driving force behind socio-economic development in the regions. The rejected variables in this group were as follows:

- capital expenditures in the private sector per capita in PLN (X_{15}) due to the fact, that they do not take into account the expenditures of public funds,
- share of the SME sector in sold industrial output (X_{21}), due to the negative and diverging from zero asymmetry rate,
- expenditures on R & D per one entity in the business sector (X_{17}), due to the correlation with the expenditures on R & D per one inhabitant variable (X_{18}).

Nine diagnostic variables describing the socio-economic situation of the regions were finally selected. These variables are objective, measurable and represent the most important areas of socio-economic development.

RESEARCH METHODOLOGY AND SYNTHETIC VARIABLES DETERMINATION

Variables selected for the study were divided into two groups: social variables and economic variables. They were then subjected to information capacity analysis using Hellwig's method. To do this, correlation coefficients between the variables for each year were determined and then all variables within a group were divided into subgroups containing central and isolated variables, by comparing the correlation coefficients with the assumed threshold of 0,5. This division is presented in Table 2.

Table 2. Division of variables into central and isolated ones

Year	Social variables		Economic variables	
	Central	Isolated	Central	Isolated
2008	X_5	X_2, X_7	X_{16}	X_{19}, X_{22}
2009	X_2	X_5,	X_{16}	X_{19}, X_{22}
2010	X_2	X_5, X_7	X_{16}	X_{19}, X_{22}

Source: own calculations

The final decision on the selection of variables for further analysis was based on the incidence of variables in each subgroup (central or isolated) and fulfilment of all the statistical criteria used, the positive asymmetry in particular. Therefore, for further analysis the following variables were adopted:

- average monthly disposable income per person (X_2),
- registered unemployment rate (X_5)[2],
- capital expenditures per capita (X_{16}),
- hard surface roads in km per 100 km^2 (X_{19}),
- participation of business in expenditures on R & D in (X_{22})

To examine the level of socio-economic development of provinces and to achieve the targets, methods that allow finding similar regions in terms of the level of development (and the selected variables) were used, in particular synthetic variable method. In order to determine a synthetic variable, diagnostic variables were first divided into stimulants and destimulants. It was assumed that among the selected variables there was only one destimulant: registered unemployment rate (X_5). The other variables were considered as stimulants. Variable X_5 was therefore transformed into a stimulant using for this purpose the weighted average rate of unemployment for Poland according to formula (1) [Zeliaś 2000]:

$$x_{ij}^S = 2\overline{x}_j - x_{ij}^D \quad (1)$$

[2] In the group of variables characterizing social development, it was decided to remove X_7 variable because of the growing negative asymmetry rate.

Next, the synthetic variable was calculated as the average of standardized diagnostic variables according to two standardisation variants – alternating and optimal pattern. In the first variant, the average value of diagnostic variable in a given year period was used as a reference point, in the second one - the optimal value in a given year, which is the maximum value in the case of stimulants. In order to compare the results, the synthetic variables were transformed into taxonomic measures according to formula (2):

$$z_i' = \frac{z_i}{\max\limits_i z_i} \tag{2}$$

The formula allows obtaining values in the range $\langle 0,1 \rangle$. The provinces for which taxonomic measures are close to one, will have a better level of socio-economic development in terms of business development. The synthetic variables as well as taxonomic measures are presented in table 3. Next, the provinces were sorted out in terms of socio-economic development. The selected and verified criteria for assessing the level of socio-economic development of regions resulted in achieving groups of regions characterized by a similar level of development. In comparison with individual indicators they better showed regional differences as far as entrepreneurship, innovation, the development of knowledge-based economy, development in the cultural dimension or the purely social one are concerned.

Table 3. Synthetic variables and taxonomic measures according to the adopted standardisation options

Province	2008				2009				2010			
	Variant1		Variant2		Variant1		Variant2		Variant1		Variant2	
	z	z'	z	z'	z	z'	z	z'	z	z'	z	z'
ŁÓDŹ	0,995	0,755	0,620	0,782	0,977	0,670	0,629	0,706	1,015	0,797	0,647	0,810
MAZOVIA	1,242	0,942	0,775	0,976	1,253	0,859	0,807	0,906	1,237	0,971	0,799	1,000
MAŁOPOLSKA	1,182	0,897	0,719	0,906	1,147	0,786	0,720	0,808	1,113	0,874	0,700	0,877
SILESIA	1,318	1,000	0,794	1,000	1,460	1,000	0,891	1,000	1,274	1,000	0,790	0,988
LUBLIN	0,744	0,564	0,475	0,598	0,757	0,518	0,497	0,558	0,808	0,634	0,523	0,655

Province	2008				2009				2010			
	Variant1		Variant2		Variant1		Variant2		Variant1		Variant2	
	z	z'	z	z'	z	z'	z	z'	z	z'	z	z'
PODKARPACIE	1,078	0,818	0,619	0,780	0,995	0,682	0,608	0,682	1,107	0,869	0,665	0,832
PODLASIE	0,946	0,718	0,578	0,729	0,863	0,591	0,557	0,625	0,814	0,639	0,529	0,662
ŚWIĘTOKRZYSKIE	0,891	0,676	0,540	0,680	0,975	0,668	0,608	0,682	0,951	0,746	0,599	0,749
LUBUSKIE	1,012	0,768	0,605	0,762	0,892	0,611	0,566	0,635	0,989	0,776	0,630	0,788
WIELKOPOLSKA	1,084	0,822	0,675	0,850	1,004	0,688	0,652	0,731	1,020	0,801	0,659	0,825
WEST POMERANIA	0,739	0,561	0,477	0,602	0,788	0,540	0,513	0,576	0,814	0,639	0,522	0,653
LOWER SILESIA	1,115	0,846	0,680	0,857	1,073	0,735	0,682	0,765	1,142	0,896	0,713	0,892
OPOLE	0,950	0,721	0,592	0,746	1,107	0,758	0,688	0,772	0,939	0,737	0,596	0,746
KUJAWY-POMERANIA	0,944	0,717	0,568	0,716	0,847	0,580	0,543	0,609	0,898	0,705	0,568	0,711
POMERANIA	1,158	0,879	0,703	0,886	1,229	0,842	0,769	0,863	1,194	0,938	0,737	0,923
WARMIA-MASURIA	0,602	0,457	0,388	0,489	0,632	0,433	0,414	0,465	0,686	0,539	0,444	0,556

Source: own calculations

MEASURMENT OF SOCIO-ECONOMIC DEVELOPMENT – RESULTS

After arranging the regions using taxonomic measures z_i', it was examined by means of Spearman's rank correlation coefficient whether the arrangement depended on the method of standardization of diagnostic variables. It turned out

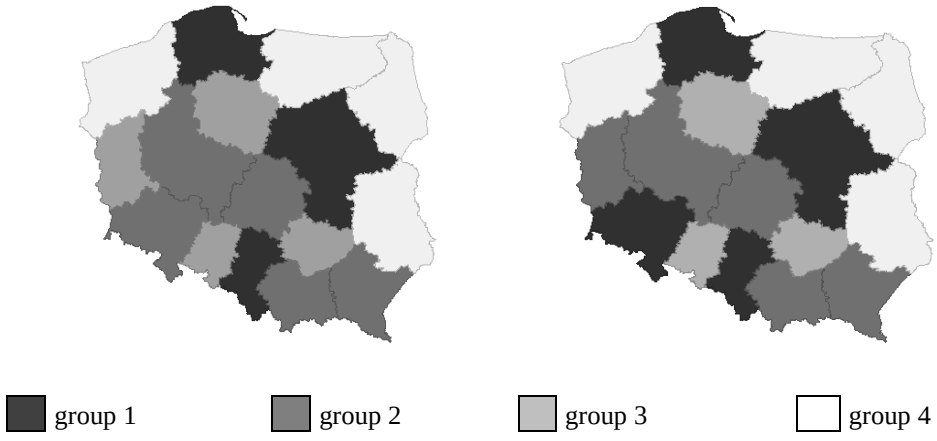
that for each analysed year the correlation coefficients were statistically significant and equal: 0,988; 0,997; 0,994 respectively, which suggests that the province arrangement according to the level of socio-economic development can be considered compatible in both adopted versions. Due to the fact, that the adopted version of standardization of diagnostic variables did not affect significantly the results of province ranking, the optimal variant (variant 2) was chosen and two methods - the standard deviation method and the division of the variation range into four pre-determined classes – were used to classify regions in 2010. The visualization of both classifications is shown in Figure 1. In the first method, the provinces were divided into four groups, including those regions for which synthetic indicator z_i obtained values in the following range (s_z – standard deviation):

Group 1: $\langle \bar{z} + s_z; \max_i z_i \rangle$
Group 2: $\langle \bar{z}; \bar{z} + s_z)$
Group 3: $\langle \bar{z} - s_z; \bar{z})$
Group 4: $\langle \min_i z_i; \bar{z} - s_z)$

According to this method the provinces were classified as follows:

- Group 1: Mazovia, Pomerania, Silesia,
- Group 2: Lower Silesia, Łódź, Małopolska, Podkarpacie, Wielkopolska,
- Group 3: Kujawy-Pomerania, Lubuskie, Opole, Świętokrzyskie,
- Group 4: Lublin, Podlasie, Warmia-Masuria, West Pomerania.

Figure 1. Classification of provinces in 2010 according to standard deviation method (the map on the left) and division of the variation range (the map on the right)



Source: own calculations

The classification results using the division of the variation range into the four classes were as follows:

- Group 1: Lower Silesia, Mazowia, Pomerania, Silesia,
- Group 2: Lubuskie, Łódź, Małopolska, Podkarpacie, Wielkopolska,
- Group 3: Kujawy-Pomerania, Opole, Świętokrzyskie,
- Group 4: Lublin, Podlasie, Warmia-Masuria, West Pomerania.

SUMMARY

When a set of diagnostic variables was determined, five variables from two groups were selected and the synthetic variables were used for the analysis of socio-economic development in provinces. Regardless of the standardization option, the synthetic variables allowed ranking the provinces according to the adopted criteria, from the best to the least developed ones. During the whole examined period the most developed regions were Mazovia Province and Silesia Province, and the least: Lublin Province and Warmia–Masuria Province.

Using the synthetic variables, regions with the same level of socio-economic development were grouped. The clustering methods led to very similar results. In the most developed regions group were: Mazovia Province, Pomerania Province and Silesia Province, and in the least developed regions group were: Lublin Province, Podlasie Province, Warmia-Masuria Province and West Pomerania Province. The difference in grouping can be seen when it comes to Lower Silesia Province, which, depending on the grouping belongs either to the first or the second group, and Lubuskie Province, which is either in the second or in the third group. There were no differences in the grouping of the other provinces.

The possibilities of using the analysed set of variables, which allowed assessing the level of socio-economic development in regions, are very broad. The conclusions of the study and the classification of regions can be used for further research. In the authors' opinion the selection of regions with a similar level of socio-economic development can be used to do research on entrepreneurship, its scale in relation to the generally understood conditions of life and economy. The described results can be the basis for analyzing the impact of different variables on economic conditions, conditions for conducting entrepreneurial activities and the intensity of the innovation process.

REFERENCES

Bizon W. (2011) Dobrobyt społeczno-ekonomiczny oraz gospodarka oparta na wiedzy w kontekście historycznych podziałów na Polskę A i B, E-mentor nr 4 (41).

Bratnicki M. (2011) Sprawdzanie teorii przedsiębiorczości, Przedsiębiorstwo przyszłości, Wyższej Szkoły Zarządzania i Prawa im. Heleny Chodkowskiej, nr 3 (8) lipiec, s. 9-18.

Dąbrowa M. (2011) Badanie poziomu życia – metodologia konstrukcji wybranych wskaźników, Zeszyty Naukowe MWSE w Tarnowie, Matematyka nr 1(17), s. 67-82.

Łapczyński M. (2005) Wpływ aktywności mieszkańców na poziom życia w gminach woj. małopolskiego, www.statsoft.pl/czytelnia/artykuly/Wplyw_aktywnosci.pdf

Malina A. (2004) Wielowymiarowa analiza przestrzennego zróżnicowania struktury gospodarki Polski wg województw, Wydawnictwo AE w Krakowie.

Strużycki M. red. (2004) Małe i średnie przedsiębiorstwa w gospodarce regionu, PWE, Warszawa.

Winiarczyk–Raźniak A., Raźniak P. (2011) Regional differences in the standard of living in Poland (based on selected indices), Procedia Social and Behavioral Sciences 19, s.31-36.

Zeliaś A. (red.) (2000) Taksonomiczna analiza przestrzennego zróżnicowania poziomu życia w Polsce w ujęciu dynamicznym, Wydawnictwo AE w Krakowie, Kraków.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 210 – 219

THE USE OF CORRESPONDENCE ANALYSIS IN THE EVALUATION OF THE ROLE OF FIBROUS AND MEDICINAL PLANTS IN PLANT PRODUCTION IN FARMS

Agnieszka Sompolska–Rzechuła
Department of Mathematics Applications in Economy
West Pomeranian University of Technology in Szczecin
e-mail: asompolska@zut.edu.pl
Grzegorz Spychalski
Department of Economic Sciences
Koszalin University of Technology
e-mail: grzegorz.spychalski@tu.koszalin.pl

Abstract: The paper presents the usage of multidimensional correspondence analysis to estimation the role of fibre and medicinal plants in farms plant production in the context of the determinants influencing the choice of these plants for crop rotation scheme. The source of data were the survey questionnaires collected among the farmers running agricultural activity in the period 2011-2012 in the Wielkopolska region. The correspondence analysis enabled to indicate relations a selected categorical variable and such categories as age, sex, education, number of person per household, farm area and farm income in 2010 and 2011.

Keywords: fibrous and medicinal plants, survey questionnaire, correspondence analysis

INTRODUCTION

The paper aims at the analysis of the role of fibrous and medicinal plants in the plant production at Polish farms in terms of the factors determining the choice of these plants for crop rotation schemes. The source of data for the analysis were survey questionnaires carried out at farms that actively run agricultural activity in Wielkopolska region in years 2011 and 2012.

In order to complete the research task multidimensional correspondence analysis was employed to assess relations between specific categories of variables.

The correspondence analysis allows for precise indication of simultaneous occurrence of two or more variables, which were measured in a nominal scale. Its advantage is the possibility of graphical presentation of concurrent occurrence of specific categorical variables. Despite its advantages, correspondence analysis is used relatively rarely in research and when used it is mostly applied in economic and social sciences.

METHODS

The correspondence analysis is a specialized tool for exploring data that presents the associations between the variables and objects, most often in a graphical form. It enables not only to analyze quantitative data but also the data measured in nominal and ordinal scales and does not have any requirements as for the number of the set of objects. The objective of using this method is to gain knowledge from data sets by analyzing correlation between specific variants of the observed variables.

The following procedure was employed for studying correlations between categorical variables regarding cultivation of fibrous and medicinal plants and the remaining categorical variables[1]:

- Determination of the Burt matrix – this method of recording data is most commonly used for correspondence analysis. As a result a symmetrical block matrix is obtained, where apart from the main diagonal contingency tables are prepared that represent two different variables and contain a number of objects with specific categories of these two variables. The diagonal matrices are placed on the main diagonal, where the non-zero values indicate the number of occurrences of a certain categorical variable,
- determination of the actual space of correlation of the *K* categorical variable, according to the following formula:

$$K = \sum_{q=1}^{Q} \left(J_q - 1 \right) \tag{1}$$

where: J_q - number of the categorical variable q $\left(q = 1,2,\ldots,Q\right)$, Q – number of variables,

- verification to what degree the values of the property space of lower dimension explains total inertia. The Greenacre's criterion was applied, which says that significant inertia are those main inertia of value higher than $1/Q$,
- modification of Eigen values according to the formula below:

[1] A detailed description of the method can be found in: [Stanimir 2005, Gatnar, Walesiak 2006, Machowska-Szewczyk, Sompolska-Rzechuła 2010, Sompolska-Rzechuła 2010].

$$\tilde{\lambda}_k = \left(\frac{Q}{Q-1} \right)^2 \cdot \left(\sqrt{\lambda_k} - \frac{1}{Q} \right)^2 \quad (2)$$

where: Q - number of variables, λ_k - k eigenvalue,

- application of the Ward's method for classification of categorical variables. To present graphically the correlation of variables in the dimension higher than 3 selected classification methods can be used. Categories of all analyzed variables must be defined as objects, where the variables are the values of projection coefficient for each category. Classification methods are also useful when the number of all the variants of variables is high and the distribution of the points in a figure does not allow for unambiguous definition of the classes. This study made use of one of the most common agglomerative classification of the results of correspondence analysis i.e. Ward's method [Ward 1963; Gordon 1999; Ostasiewicz 1998],
- graphical presentation of the categorical variable associations in two- or three-dimensional spaces.

CHARACTERISTICS OF THE RESEARCH MATERIAL

The role of fibrous and medicinal plants in plant production on farms, assessed with correspondence analysis, was determined on the basis of questionnaire answered by farmers active in agricultural activity in Wielkopolska region of Poland in years 2010 and 2011. Presently, agricultural producers choose the crop structure by analyzing natural and economic conditions. They apply the principle of income diversification.

In this respect farmers introduce industrial plants i.e. the so called specialty crops, which do not supply food products and have special role in several production-consumption [Wojdyła 2006].

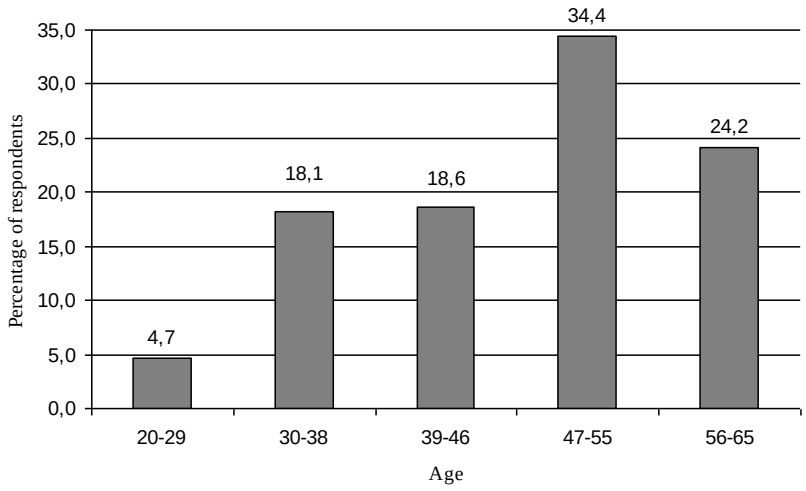
This group of crops includes fibrous and medicinal plants that are used for textile, protective, construction and medical applications. In Polish agriculture there is centuries long tradition of cultivation of flax, hemp and herbs, therefore the producers' skills are good and guarantee sufficient quality of the produced raw materials.

The survey questionnaire comprised two parts: 14 questions concerned the cultivation of fibrous and medicinal plants and 6 questions were demographic.

The total of 224 farmers responded to the questionnaire, provided their acceptance to take part in the survey.

Among the questioned farmers, the biggest group were people between 47-55 years old (34,4%), while the smallest group were the people between 20-29 years old (4,7%). The age structure of the respondents is presented in Figure 1.

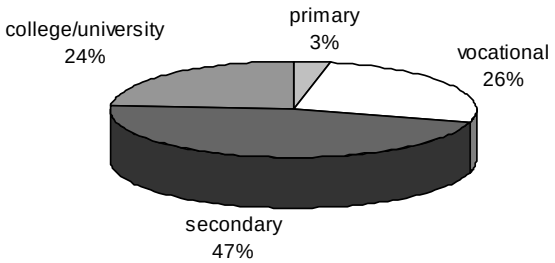
Figure 1. Age structure of the respondents



Source: own study

Among respondents, the largest group was individuals with high school education, while the smallest with primary school education (see Figure 2).

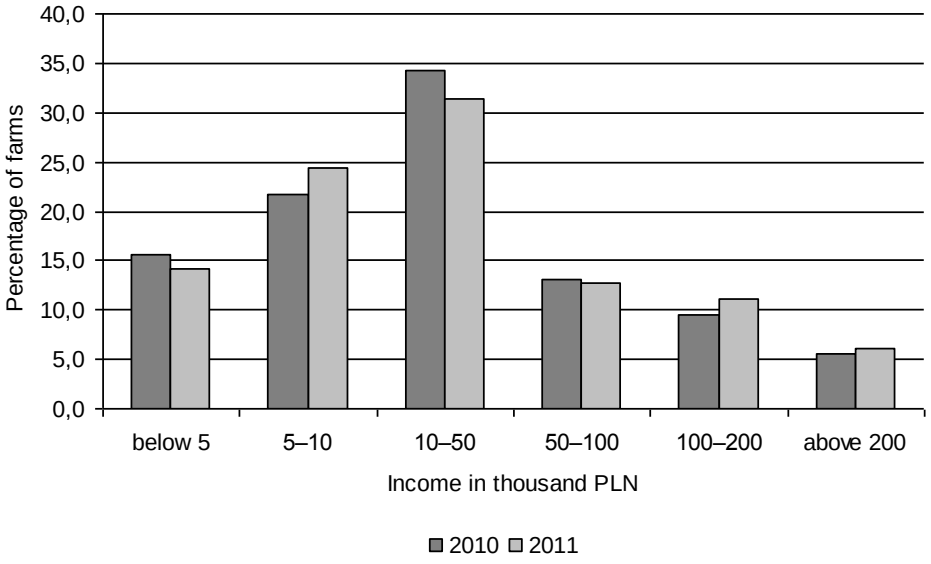
Figure 2. Education structure of the respondents



Source: own study

The farms studied in the questionnaire varied greatly in terms of their area with the smallest of 0,18 ha and the largest of 470 ha. The farms of 20 ha were the dominating group (15 farms). Similar situation was observed in case of the area of arable land, where the variation coefficient was more than 131%. The questioned farmers were also asked about the income levels in years 2010-2011. In both periods the most farms reached the income between 10-50 thousand PLN, in 2010 such income was reported for 34,3% of farms, while in 2011 – 31,5%. The income structure of farms in years 2010 and 2011 is presented in Figure 3.

Figure 3. The farm structure in terms of income levels in years 2010 and 2011



Source: own study

The technical part of the questionnaire concerning the role of cultivation of fibrous and medicinal plants in the plant production of farms included questions on cultivation of these plants, their position in the sowing structure, conditions for crop rotation schemes, knowledge on textile industry and herb processing industry in Poland, Polish research organizations active in this field and the expectations for the academic world.

As much as 18% of the questioned farmers declared that they grow fibrous (flax or hemp) or medicinal plants. At only 8% of the farms fibrous plants were cultivated, of which 9% of the farmers believed fibrous plants improve the soil structure and 5% that growing these plants is profitable. A similar situation was observed for medicinal plants. The highest number of farmers (42%) thought low popularity of fibrous plants is due to absence of that tradition for growing these plants. The same reason was given by 38% of questioned farmers regarding

absence of medicinal plants in the sowing structure of their farms. If they were to introduce a new plant, the highest group of farmers (51%) would opt for growing herbal plants, 33% - for flax, and 16% for fibrous hemp. As much as 56% of the respondents believe that cultivation of specialty crops can increase the farm income, 23% that it improves the soil quality, 16% respondents think it will facilitate sales of the yields. Half of the questioned farmers expect that academic world will provide advice while 43% that the science will provide tried and tested technology.

The survey questions were accompanied with the following sets of features and categories:

- cultivation of fibrous and/or medicinal plants: Y (yes), N (no),
- sex: F (female), M (male),
- age: A (20-38 years old), B (39-55 years old), C (more than 55 years old),
- education level: A (high school or college/university), B (vocational or primary school education),
- number of people in the household: A (1-3 individuals), B (3-6 individuals), C (more than 6 people),
- farm area: A (below 50 ha), B (50-100 ha), C (above 100 ha),
- area of arable land: A (below 50 ha), B (50-100 ha), C (above 100 ha),
- income per farm in 2010: A (below 50 thousand PLN), B (above 50 thousand PLN),
- income per farm in 2011: A (below 50 thousand PLN), B (above 50 thousand PLN).

Before applying multidimensional correspondence analysis we tested whether there is dependence between those features, what was confirmed in the results.

RESULTS

In case of studying the associations between categories of dependent variable i.e. cultivation of fibrous and/or medicinal plants and the other categories of variables associated with it a Burt matrix was obtained of the dimension 22×22. The dimension of actual answer correlation space was 13.

The next step involved verification how the eigenvalues of the space with lower dimension explains the total inertia. The results are presented in Table 1 that includes: eigenvalues λ_k, singular values γ_k, share of main inertia in total inertia (in percentage of λ_k / λ) and share of eigenvalues from *K* dimension in total inertia (accumulative percentage τ_K).

Table 1. Eigen and singular values and the degree of explanation of the total inertia

Number of K dimensions	Values		Percentage	
	Eigen λ_k	singular γ_k	inertia λ_k / λ	accumulated τ_K
1	0,3124	0,5589	21,6291	21,6291
2	0,2040	0,4516	14,1216	35,7507
3	0,1534	0,3917	10,6219	46,3726
4	0,1408	0,3753	9,7492	56,1218
5	0,1317	0,3629	9,1197	65,2415
6	0,1125	0,3354	7,7886	73,0302
7	0,1048	0,3237	7,2534	80,2836
8	0,0992	0,3149	6,8658	87,1494
9	0,0863	0,2938	5,9752	93,1246
10	0,0583	0,2414	4,0356	97,1602
11	0,0273	0,1652	1,8883	99,0485
12	0,0117	0,1084	0,8131	99,8615
13	0,0020	0,0447	0,1385	100,0000
	λ =1,4444			

Source: own study

According to Greenacre's criterion, the best dimension of projection of category variables is the one where eigenvalues fulfil the condition: $\lambda_k > 1/Q$. In the analyzed case, this value is 0,1111 for $Q = 9$. The data presented in Table 1 indicate that these are inertia for the R^6 dimension and total dimension in this case is 1,4444. Modification of eigenvalues was made according to the Greenacres's criterion. The values of modified eigenvalues and singular values and the degree of explanation of the total inertia are presented in Table 2.

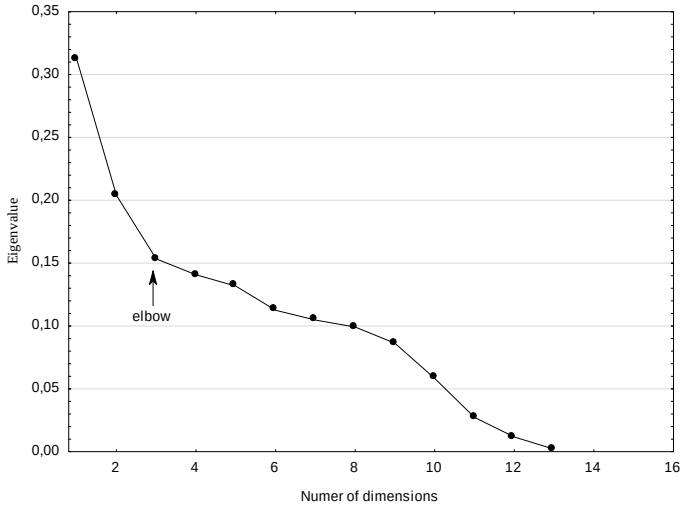
Table 2. Modified Eigen and singular values and the degree of explanation of the total inertia

Number of K dimensions	Values		Percentage	
	Eigen $\tilde{\lambda}_k$	singular $\tilde{\gamma}_k$	inertia $\tilde{\lambda}_k / \tilde{\lambda}$	accumulated $\tilde{\tau}_K$
1	0,2538	0,5038	34,6530	34,6530
2	0,1468	0,3831	20,0361	54,6892
3	0,0996	0,3157	13,6033	68,2925
4	0,0883	0,2972	12,0563	80,3487
5	0,0803	0,2833	10,9581	91,3068
6	0,0637	0,2523	8,6932	100,0000
	$\tilde{\lambda}$ =0,7325			

Source: own study

Additionally, a graph was prepared that shows eigenvalues (see Figure 4).

Figure 4. The graph presenting Eigen values

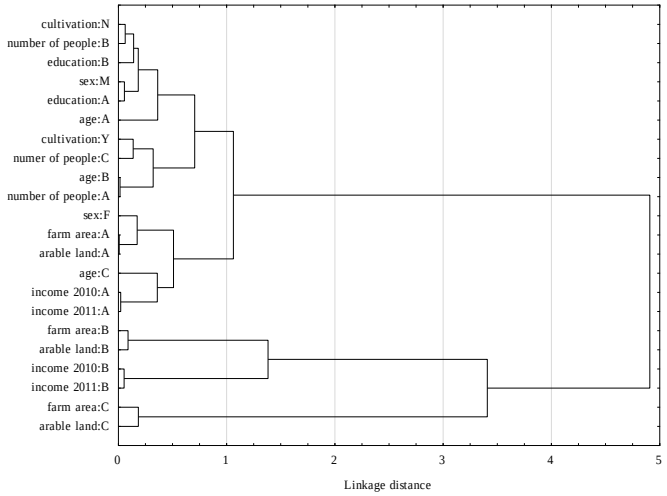


Source: own study

One of the methods of determining the number of eigenvalues, which indicate coordinates significant for projection with low dimension is the so called „elbow" criterion. A place is searched on the graph with all non-zero specific/proper values in descending order where a slight drop in these values is observed (a number of the eigenvalue is indicated, where the 'bend' is visible) [Stanimir 2005]. Figure 4 shows that 'the elbow' occurs for $k = 3$. Therefore, the correlation analysis between categorical variables will take place in three-dimensional space, which explains almost 68,3% of total inertia.

The correlations between categorical variables are presented in a dendrograph obtained with the Ward's method (Figure 5).

Figure 5. The dendrograph presenting the division of categorical variables according to the Ward's method



Source: own study

The use of correspondence analysis allowed for separating two groups of associations between categorical variables. Two groups were distinguished because of two categories of dependent variable i.e. cultivation of fibrous and/or medicinal plants. This variable is of dychotomic nature and takes two values Yes or No. The category 'non-cultivation of fibrous and/or medicinal plants' is mostly associated with the B category of the variable 'number of people per farm' i.e. between 3 and 6 people per farm. Moreover, absence of these crops is declared by the surveyed farmers with vocational and primary school education but also with high school and university education and men aged 20-38 years.

Cultivation of fibrous and/or medicinal plants is characteristic for the farms with more than 6 people in the household and also with 1-3 people. The age range of the farmers that grow fibrous and/or medicinal plants is 39-55 years. Weaker correlation is visible between the cultivation and the following categories of variables: income per farm both in 2010 and 2011, farm area and the area of arable land above 50 ha. These categories constitute a separate class not associated with the categorical variable of cultivation of fibrous and/or medicinal plants.

REFERENCES

Gatnar E., Walesiak M. (2006) Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, pp. 285-315.

Gordon A. D. (1999) Classification, Chapman & Hall/CRC, Boca Raton.

Machowska-Szewczyk M. Sompolska-Rzechuła A. (2010) Analiza korespondencji w badaniu postaw osób dokonujących zakupów przez Internet, Ekonometria 29, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 141, Wrocław, pp. 9-21.

Ostasiewicz W. (red.) (1998) Statystyczne metody analizy danych, Akademia Ekonomiczna we Wrocławiu, Wrocław, pp. 95-97.

Panek T. (2009) Statystyczne metody wielowymiarowej analizy porównawczej, Szkoła Główna Handlowa w Warszawie, pp. 22.

Sompolska–Rzechuła A. (2010) Zastosowanie metody analizy zgodności w badaniu jakości życia kobiet, Wiadomości Statystyczne, nr 1, pp. 53-64.

Stanimir A. (2005) Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław

Ward J. H. (1963) Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, 58, pp. 236-244.

Wojdyła T. (2006) Rośliny przemysłowe w przemyśle spożywczym oraz metody analiz stosowanych w ich przetwórstwie, Wydawnictwo Akademii Techniczno-Rolniczej, Bydgoszcz.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 220 – 228

WHO WANTS TO WORK LONGER?

Aleksander Strasburger, Olga Zajkowska
Department of Applied Mathematics
Warsaw University of Life Sciences – SGGW
e-mail: aleksander_strasburger@sggw.pl, olga_zajkowska@sggw.pl

Abstract: The observed collapse of demographic pyramid increases the tension on the social security systems, especially pensions. It implies a requirement to extend the retirement age. On the basis of Eurobarometer 65.1 we calculate difference between expected and preferred retirement age in Poland. Then we show the determinants of individual differences. Particular attention is paid to the problem of non-random missing observations.

Keywords: retirement decisions, ageing

INTRODUCTION

Within the last 25 years almost all of the countries in Central and Eastern Europe have experienced economic system transition. The centrally planned economies nearly collapsed and therefore hedging future pensioners was almost impossible. Additionally pension systems were based on assumption of stable demographical structure of societies. The so called Pay As You Go systems (defined benefit systems) use the property of similar inflows and outflows of the systems which allows for funding the pension system from general revenues [Góra, 2003]. Currently CEE countries face worsening demographic dependency ratio. The populations are aging — life expectancy grows faster than in the rest of Europe and additionally fertility ratios are significantly lower. Without sufficient reforms it would lead either to tax increase or lower replacement rate. Part of the solution is reduction of sector privileges, increase of labour market participation and discouragement of older people to exit the labour market. This leads to the questions which we try to answer in this article, that is what type of people are more willing to work longer? What are the features of people staying on the labour

market in old age? Additionally we try to investigate who has expectations about retirement age.

In the literature retirement decisions have been widely analyzed. Among determinants most often pointed out are financial factors- higher wage seems to be strong incentive to stay at labour market [Pellechio 1979, Hernoes 2000, Compton 2001, Antolín et al. 1998], good education, accumulated human capital and stable, unrisky employment (both in the sense of working conditions and contract type) are encouraging for later leave [Miniaci et al. 1998] or marital status [Bütler, 2004]- single women tend to be less willing to retire early. Quinn [1977] shows that eligibility for early retirement is also an important determinant.

METHODS AND DATA

We use microdata to estimate determinants expected and desired retirement age of individuals. Further we estimate the determinants of difference between those quantities. We stress the fact that not all individuals form expectations about their retirement. Not all of them have precise retirement age preferences. It implies existence of significant number of missing observations. We claim that process of preferences (and expectations) formation is crucial for final outcome. We claim that this might not be random and be dependent on income, type of employment or household composition of surveyed individuals. Also long horizon causing difficulties in precise predictions might be a possible issue. Therefore we also estimate determinants of expectations formation. We use Heckman's sample selection model [Cameron et al. 2007].

The model consists of two equations. First one is an assignment equation, which estimates probability for an individual of being assigned to the subsample of interest ($\Pr(y_1 = 1)$), that is the probability that variable of interest is observed. In our model we want to estimate expectations only on individuals who have formed ones. We assume that expectations are observable (that is y_1 =1) only if unobservable, latent willingness (y_1^{lat}) to form them has exceeded given threshold:

$$y_1 = \begin{cases} 1, \text{ if } y_1^{\text{lat}} > 0; \\ 0, \text{ if } y_1^{\text{lat}} \le 0. \end{cases}$$

Second equation describes only individuals with observable values of interest, so conditional on $y_1 = 1$.

$$y_2 = \begin{cases} y_2^{\text{lat}}, \text{ if } y_1^{\text{lat}} > 0; \\ 0, \quad \text{if } y_1^{\text{lat}} \le 0. \end{cases}$$

Data used in this study comes from survey Eurobarometr 65.1[1]. We decided to choose following countries: Bulgaria (with statutory retirement age for women and man 60 and 63 years respectively[2]), Czech Republic (59-63 and 63, DB[3]), Estonia (61.5 and 63, Point scheme), Lithuania (60 and 62.5), Latvia (62), Poland (60 and 65, NDC[4]), Romania (59 and 64), Hungary (62), Slovakia (62, Point scheme) and Slovenia (61 and 63). We have trimmed the sample to working individuals aged 18 and more. The sample consists of n = 4153 observations.

Definitions used in our study:

- defined expected retirement age – [binary] individual has expectations of own retirement age and has reported it, fraction in the subsample of working individuals with defined expectations is 78.3% (of n=4153), average age among individuals with undefined expectations is 36.54 years and among those with defined expectations- 40.95 years;
- defined expected and desired retirement age - [binary] individual has both expectations and preferences of own retirement age and has reported it, fraction in the subsample of working individuals is 73.46%;
- expected retirement age – expected retirement age reported by working individual, average for working men is 61.67 years (std dev.= 4.70) and for working woman average is 59.78 years (std dev.=3.86);
- desired retirement age – desired retirement age reported by working individual, average for working men is 57.17 years (std dev.= 6.35) and for working woman average is 55.04 years (std dev.=4.75)
- difference in expectations – difference between desired and expected retirement age, average for subsample with defined both expected and desired retirement age is equal to -4.57 years (std dev=5.01)

On the basis of simple descriptive statistics given above, it can be seen, that over 20% of working population in Central Europe is not thinking about their retirement.

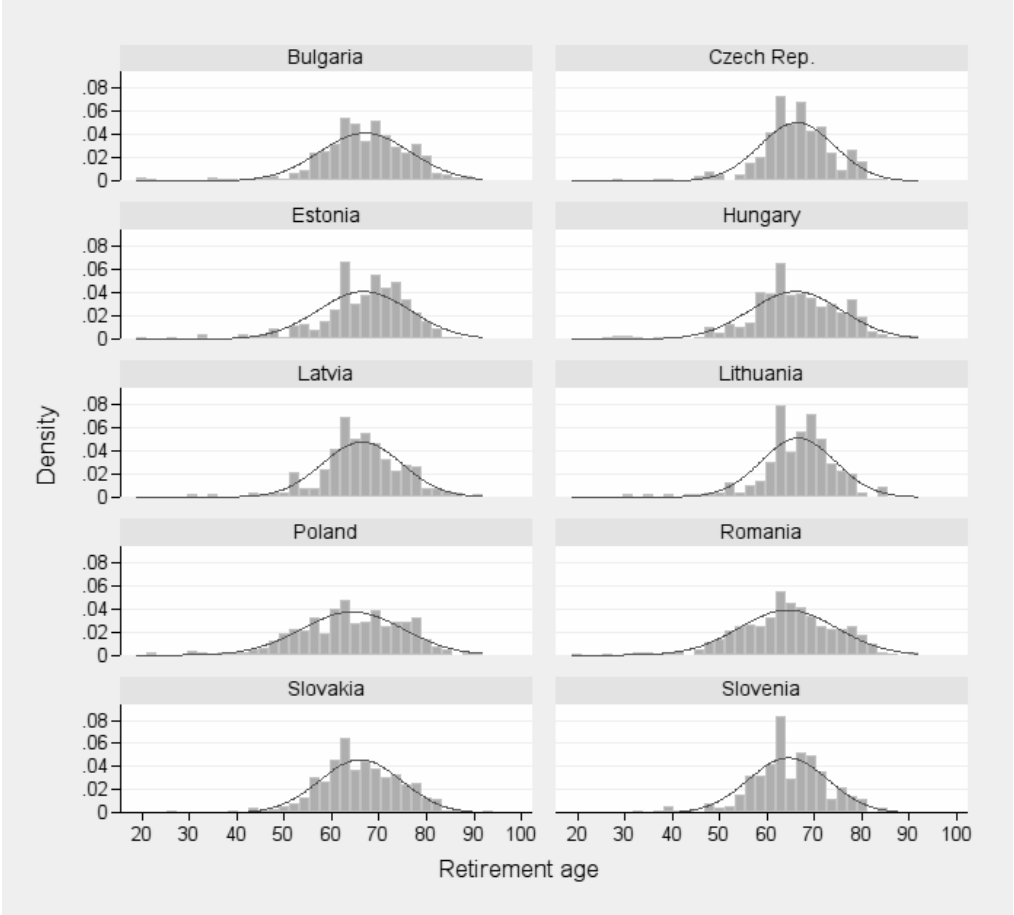
1 European Commission (2012): Eurobarometer 65.1 (2006). TNS OPINION \& SOCIAL, Brussels. GESIS Data Archive, Cologne. ZA4505 Data file Version 1.0.1, doi:10.4232/1.10973

2 Retirement age data comes from Eurostat

3 DB- Defined Benefit plan

4 NDC- Notional Defined Contribution plan

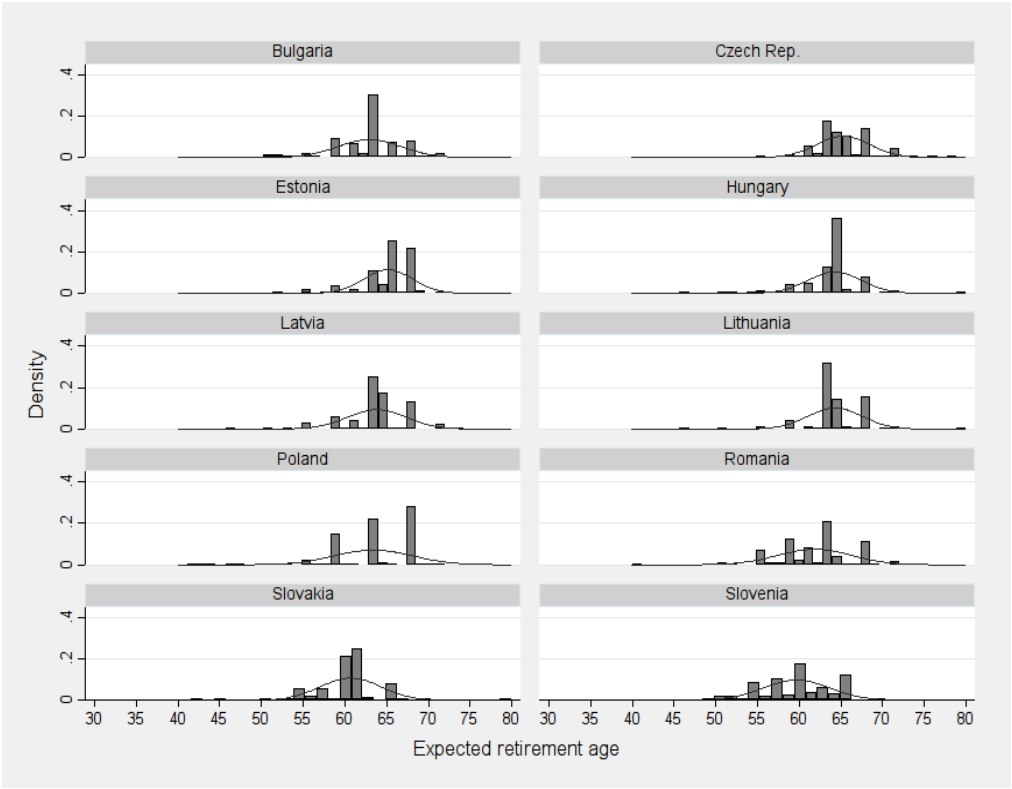
Figure 1. Distribution of retirement age by country (n = 4153)



Source: own calculations

In Figure 1 the fractions of retired individuals for each country are shown, together with the normal curve with parameters adjusted to the subsample observations visualized on the graph. It can clearly be seen that subpopulations of retired individuals by country differ between each other. Poland shows similar patterns to Romania , Bulgaria and partly to Hungary, where fraction of individuals younger than 60 is significant. More dense distributions are observed for Baltic countries, Czech Republic and Slovenia. One should be careful with interpretation of fraction of individuals older than 65, because the distribution does not distinguish effect of late exiting labour market and the life survival rate.

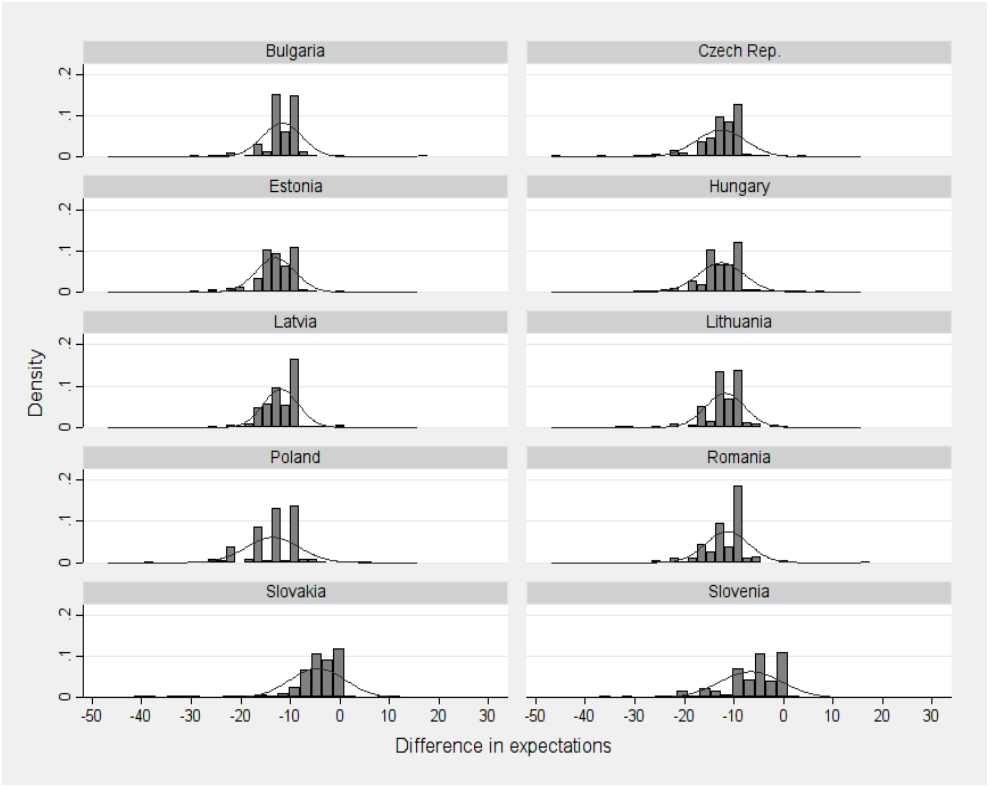
Figure 2. Expected retirement age by country (n = 3099)



Source: own calculations

In Figure 2 expectations concerning the retirement age are presented. It may be inferred that working individuals tend to form expectations on the basis of statutory retirement age. Most of them expects to leave the labour market as soon as they will become entitled to do so (each country has the largest peak at number equal to statutory retirement age or 2 peaks if statutory age differs conditional to gender). In Poland and Romania more peaks can be observed, which implies large fraction of working individuals entitled to earlier labour market exit and significant difference between statutory retirement age for men and women.

Figure 3. Difference between desired and expected retirement age by country



Source: own calculations

In all countries in the sample we can observe that desired retirement age is on average in the population lower than expected. It means individuals want to exit labour market earlier and are forced to stay by the pension system. What is worth mentioning are dissimilarities observed among Poland and Hungary. This is surprising, because this countries had very similar pension systems at the moment when data for this study was collected. It indicates that not only pension system incentives are important in lifecycle working patterns preferences, but some other factors might exert influence. Identification of these extra factors is the main aim of the present work.

Variables used in the study:

- gender – [binary] 1 if individual is a woman, fraction of females in the sample: 50,57%;
- age – individuals' age, ranges from 18 to 76 years, with mean equal to 40 years;
- age5 – age mod 5;

- years of edu – completed years of education, ranges from 8 to 50, with mean equal to 13.82 years;
- children – number of children under 14, ranges from 0 to 4, 27.96% of individuals have at least one child under 14 in the household;
- partner – [binary] 1 if individual declares having husband or partner in the household, fraction in the sample: 72.96%;
- income cat – income category [scale 1-10, low-high], mode=8;
- early retirement option – [binary] 1 if individual is entitled to earlier retirement, fraction in the sample: 16.37% (in female subsample: 15.95%);
- hours worked – average number of hours worked per week,
- public sector - [binary] 1 if individual is employed in public sector, fraction: 35.9% (in female subsample: 44.95%)
- fixed term contract - [binary] 1 if individual is employed on a fixed term contract, fraction in the sample: 16.2%.

RESULTS

We have estimated model explaining individuals' differences between desired and expected retirement age with respect to selection caused by possessing such preferences and expectations[5]. The results are shown in Table 1.

Table 1. Difference between desired and expected retirement age.

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
difference in expectations						
partner	0,385	0,225	1,71	0,09	-0,056	0,825
income category	0,074	0,034	2,16	0,03	0,007	0,142
years of education	0,110	0,030	3,61	0,00	0,050	0,169
fixed-term contract	-0,630	0,271	-2,33	0,02	-1,161	-0,099
children	-0,304	0,132	-2,30	0,02	-0,564	-0,045
age	0,119	0,010	11,36	0,00	0,099	0,140
early retirement option	-1,180	0,227	-5,19	0,00	-1,626	-0,734
const	-10,731	0,764	-14,05	0,00	-12,228	-9,234

5 Number of observations n = 4153, number of censored observations n' = 1102

Table 1. -continuation	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
defined expected and desired retirement age						
age	0,018	0,002	8,67	0,00	0,014	0,021
partner	0,103	0,049	2,12	0,03	0,008	0,199
income cat	0,023	0,007	3,18	0,00	0,009	0,038
public sector	0,220	0,046	4,81	0,00	0,130	0,309
early retirement option	0,198	0,047	4,21	0,00	0,106	0,291
hours worked	0,007	0,001	5,62	0,00	0,005	0,010
const	-0,843	0,116	-7,27	0,00	-1,070	-0,616
ath rho	0,085	0,080	1,07	0,29	-0,071	0,241
lnsigma	1,674	0,013	126,77	0,00	1,648	1,700
rho	0,085	0,079			-0,071	0,237
sigma	5,332	0,070			5,195	5,471
lambda	0,453	0,423			-0,376	1,282
LR test of indep. eqns. (rho = 0): chi2(1) = 0.99					Prob > chi2 = 0.3188	

Source: own calculations

As can be inferred from the first part of the model (seen in the lower part of Table 1), probability of having defined both expected and desired retirement by individual grow with age, which is rather intuitive, older individuals are simply closer to statutory retirement age and more willing to make retirement plans. Having partner also increases chances of making predictions which might be signal of responsibility for the spouse. Also work in a public sector increases the probability since career paths are more predictable than in private sector. Additionally there is strong positive effect on probability caused by early retirement option. That may imply that retirement privileges are important feature while choosing the profession. Further hours worked on average per week and individual's household income category increase the probability.

Second part of the model (seen in the upper part of Table 1) describes the difference between desired and expected retirement age. It means that negative value of dependent variable in the model occurs when individual wants to retire earlier than he/she expects to basing on statutory retirement age and possible privilege of earlier retirement option. The positive value of the dependent variable implies that individual is willing to work longer than she/he expects to. Further positive values of estimates are associated with variables being stimulants of longer labour market activity while negative estimates imply variables contributing to earlier retirement. So children (and in consequence grandchildren) decrease the engagement into labour market activity, probably by providing opportunity costs of work. Fixed term contracts and retirement privileges also have negative effect on working preferences —- leaving labour market with stable income reduces individual insecurity. Greater willingness to work rises with age, better education, higher income category of the household and a husband (or wife).

What we found interesting is that there are no statistically significant gender differences in the model. Part of the explanation is that in most of the countries in Central Europe, the statutory retirement age does not differ significantly between men and women. Case of Poland and Romania are an exception. Also having children seems not to make significant difference, which might be an important result but needs further investigation. Additionally we were not able to show a significant influence of job features (like higher accident risk) on the formed preferences.

CONCLUSIONS

We have investigated determinants of expected and desired retirement age of individuals in Central and Eastern Europe. We have shown that on average people are not willing to work longer than they are due according to the retirement age, they also tend to exercise privileges to earlier exit of labour market. Therefore increase of statutory retirement age might be politically costly reform. But on the other hand retirement preferences change with age. That might be caused by the ability of replacement rate and welfare reduction estimation. Also better educated individuals are more willing to work longer. Further research should stress on long term motivation and human capital formation.

REFERENCES

Antolín, P., Scarpetta S. (1998) Microeconometric analysis of the retirement decision, Germany, OECD Economic Department Working Papers, No. 204.

Bütler, M., Huguenin, O. i Teppa, F. (2005) What Triggers Early Retirement? Results from Swiss Pension Funds, DNB Working Papers, No. 41.

Cameron A. C., Trivedi P. K. (2007) Microeconometrics. Methods and applications, Cambridge University Press.

Góra M. (2003) System emerytalny, Wydawnictwo PWE, Warszawa.

Hernoes, E., Sollie M. i Strom S. (2000) Early retirement and economic incentives, ,The Scandinavian Journal of Economics, 102 (3), pp. 481-502.

Kotowska I. E., Przyszłość demograficzna a zasoby pracy [w:] Społeczno-ekonomiczne następstwa rozwoju procesów demograficznych do 2035 roku, Warszawa 2009.

Kotowska I. E. [red.] (2009) Strukturalne i kulturowe uwarunkowania aktywności zawodowej kobiet w Polsce, Wydawnictwo Naukowe SCHOLAR, Warszawa.

Miniaci, R., Stancanelli E. (1998) Microeconometric analysis of the retirement decision, United Kingdom, OECD Economic Department Working Paper, No. 206.

Pellechio, A. (1979) Social Security and retirement: evidence from the Canada time series, NBER Working Paper, No. 351.

Quinn, J. F. (1977) Microeconomic Determinants of Early Retirement: A Cross-Sectional View of White Married Men The Journal of Human Resources , Vol. 12, No. 3, pp. 329-346.

European Commission (2012) Eurobarometer 65.1 (2006) TNS OPINION & SOCIAL, Brussels. GESIS Data Archive, Cologne. ZA4505 Data file Version 1.0.1, doi: 10.4232/1.10973.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 229 – 239

FOREIGN EXCHANGE RATES IN CENTRAL EUROPEAN ECONOMIES: NONLINEARITIES IN ADJUSTMENT TO INTEREST RATE DIFFERENTIALS

Anna Sznajderska
National Bank of Poland
e-mail: Anna.Sznajderska@nbp.pl

Abstract: The aim of the paper is to examine the relation between foreign exchange rates and interest rate differentials in Poland, the Czech Republic, and Hungary. The exchange rate equations are inspired by the uncovered interest rate parity (i.e. the UIP condition). The results of empirical studies are usually contrary to the UIP condition. One of the explanations of this puzzle is the existence of certain nonlinearities. The nonlinearities appear because of transaction costs, central bank interventions, limits of speculations, hysteresis, or changes in risk perception. I estimate smooth transition autoregressive models. The threshold variable is an interest rate differential or a level of economic activity. I examine the exchange rates of USD and EUR and 1-, 3- and 6- months and 5- years interest rates. I also test various proxies for risk premium.

Keywords: foreign exchange rates, uncovered interest rate parity, STAR models

INTRODUCTION

The paper concerns the relations between exchange rates and interest rate differentials in Poland, the Czech Republic, and Hungary. The analyzed equations are based on the uncovered interest rate parity (i.e. UIP). According to the UIP condition expected gains from investing in two analogous assets in two different countries should be identical. Thus, the expected change of exchange rate in k-periods should be equal to the difference between domestic and foreign k-period interest rates. The UIP condition postulates that high interest rate currencies should depreciate in relation to low interest rate currencies.

But the results of the empirical studies are inconclusive or reject the UIP condition (see summary of the conducted research in Omer et al. 2012). It is so-called forward premium puzzle. Froot (1990) reports that the average β[1] coefficient for 75 published research equals -0.88. The strong negative correlation between exchange rate and interest rate differential (i.e. β= -1), means that after an increase of domestic interest rate by 1% the exchange rate appreciates by 1% within a year.

There are many explanations of this phenomenon, but their success in practical applications is very limited. Firstly, it is time-varying risk premium. For example, an increase in domestic interest rate could cause an increase in risk aversion to investments in domestic assets and, thus, could have no effect on the exchange rate. Secondly, the investors' expectations might not be rational, because of, for instance, certain expectational errors (learning or peso problems). Thirdly, part of investors slowly reacts for the changes in interest rates, because they have to reconsider their decisions or they cannot react faster. Chinn and Meredith (2005) argue that the negative relation between the exchange rate and interest rates characterizes the short-term data, whereas the positive relation with slope coefficient insignificantly different from unity characterize the long-term data. The authors show that such result is consistent with the standard structural model, in which long-term interest rates react differently on exchange rate shocks than short-term interest rates.

Moreover, the forward premium puzzle can be explained by certain nonlinearities, which I analyze in this paper. The nonlinearities can appear because of transaction costs, central bank interventions[2], limits of speculation, and changes in risk perception. Only when the expected gains from investing in domestic assets are high enough, they will attract speculative capital. The level of risk perception depends on the phase of the business cycle, for instance, during the recent financial crisis high level of risk aversion caused strong depreciation of currencies of Central European economies.

LITERATURE

We can distinguish two groups of studies on nonlinearities in the UIP condition. The first group uses simpler method that allows for discrete switching from one regime to another. Bansal (1997) and Bansal and Dahlquist (2000), using this method, concern a regression of an exchange rate change on a positive and a negative forward premium. Bansal (1997) carries a study for a group of advanced economies[3] and shows that β coefficient is negative for positive interest rate

[1] β denotes the coefficient of interest rate differential in the exchange rate equation: $\Delta s_{t+k} = \alpha + \beta(i_{t+k} - i^*_{t+k})$, according to UIP conditionα=0, β=1.

[2] in particular these unexpected [Moh et al. 2005]

[3] The author defines the exchange rate as a price of a unit of domestic currency in dollars.

differentials and positive for negative interest rate differentials[4]. In both cases he rejects the hypothesis that $\beta = 1$. Bansal and Dahlquist (2000) do not find similar relation for emerging economies and argue that the results depend on the risk premium and country specific attributes, such as per capita GNP, average inflation rates, inflation volatility, and sovereign ratings.

The second group of studies uses smooth transition models, which allow for a smooth transition from one regime to another. Investors in different periods make their decisions. Only when the potential profits are high enough, the investors will change their assets' portfolio. While when the potential profits are relatively low, less investors will be willing to trade. Also the investors might need time to observe the profitable trading possibility and assess information and transaction costs.

Sarno et al. (2005) study nonlinearities in the UIP condition concerning five major US dollar exchange rates in period from 1985 to 2002. The authors use exponential smooth transition function. The results indicate that β coefficient is negative when small deviations from UIP appear, and it is positive when large deviations appear. Precisely, they argue that when Sharpe ratios[5] are small than the deviation from market efficiency is statistically significant and persistent, however, too small economically to attract speculative capital. On the other hand, when Sharpe ratios are large, they attract speculative capital, and then the spot forward regression satisfies the UIP condition.

Similar study for ten currencies from 1978 to 2002 is conducted by Baillie and Kiliç (2006). The authors apply, in contrast to Sarno et al. (2005), a logistic smooth transition function. Their results also show that only relatively high level of interest rate differential generates the results consistent with the UIP condition.

METHODOLOGY

The no-arbitrage condition might be written as:

$$(1+i_{t+k}) = (1+i^{*}_{t+k})\frac{E(S_{t+k} \mid F)}{S_t}, \qquad (1)$$

where i_{t+k} and i^{*}_{t+k} are domestic and foreign nominal interest rates between t and t+k, S_t is the nominal exchange rate (the price of foreign currency in units of domestic currency), $E(S_{t+k} \mid F)$ is the expected exchange rate in t+k given the information set F at time t. Assuming that $\log(1+x) \approx x$, for x close to zero, and rational expectations, I logarithm the equation (1) and obtain:

[4] Similar results for an absolute value of interest rate differential are obtained by Bilson (1981).

[5] We can interpret Sharpe ratio as the expected excess return from the strategy per unit of risk.

$$i_{t+k} = i^*_{t+k} + \log(S_{t+k}) - \log(S_t), \tag{2}$$

I then denote the natural logarithms as the variables in lowercase letters and obtain the regression usually tested in the literature (so-called Fama regression):

$$\Delta s_{t+k} = \alpha + \beta(i_{t+k} - i^*_{t+k}) + \varepsilon_{t+k}. \tag{3}$$

In the paper I also estimate the following exchange rate equation, which has better statistical properties than equation (3), i.e.:

$$s_{t+k} = \varphi s_t + \alpha + \beta(i_{t+k} - i^*_{t+k}) + \varepsilon_{t+k}. \tag{4}$$

Next I test if adding a proxy for risk premium helps to obtain results that are consistent with UIP condition:

$$s_{t+k} = \varphi s_t + \alpha + \beta(i_{t+k} - i^*_{t+k}) + \gamma(risk\_premium) + \varepsilon_{t+k}. \tag{5}$$

Similarly to Sarno et al. (2005) and Baillie and Kiliç (2006) I estimate the smooth transition models:

$$s_{t+k} = \varphi s_t + \alpha + \beta_1(i_{t+k} - i^*_{t+k}) + \beta_2(i_{t+k} - i^*_{t+k})F(z_t, \gamma, c) + \varepsilon_{t+k}, \tag{6}$$

where in case of logistic transition function:

$$F(z_t, \gamma, c) = \frac{1}{1 + \exp(-\gamma(z_t - c)/\sigma(z_t))}, \qquad \gamma > 0, \tag{7}$$

or in case of exponential transition function:

$$F(z_t, \gamma, c) = 1 - \exp(-\gamma(z_t - c)^2), \qquad \gamma > 0, \tag{8}$$

z_t is a transition variable, namely it is a differential between domestic and foreign interest rates or a level of economic activity, γ is a slope parameter, c is a location parameter, $\sigma(z_t)$ is a standard deviation of z_t. All parameters are estimated using constrained maximum likelihood method.

For equation (4) I test linearity against smooth transition models, applying two tests [see van Dijk et al. 2000] LM1 and LM2. These tests are based on estimating auxiliary regressions, which have simpler form than logistic or exponential functions. In order to derive LM1 test one uses a first-order Taylor approximation, and to derive LM2 test a second-order Taylor approximation. When the linearity hypothesis is rejected, smooth transition models are estimated. I choose logistic or exponential function depending on the p-value of LM1 and LM2 tests. If LM1 test has smaller p-value I choose a logistic transition function, and if LM2 test has smaller p-value I choose an exponential transition function.

Additionally, I test if the equation (4) is nonlinear according to differential between interest rates using the following simple threshold regression:

$$s_{t+k} = \varphi s_t + \alpha + \beta_1(i_{t+k} - i^*_{t+k})I_{t+k}{}^{+} + \beta_2(i_{t+k} - i^*_{t+k})I_{t+k}{}^{-} + \varepsilon_{t+k}, \tag{9}$$

$$I_{t+k}^{+} = \begin{cases} 1 & when \quad (i_{t+k} - i_{t+k}^{*}) \geq \tau, \\ 0 & otherwise, \end{cases} \quad I_{t+k}^{-} = \begin{cases} 1 & when \quad (i_{t+k} - i_{t+k}^{*}) < \tau, \\ 0 & otherwise. \end{cases}$$

DATA

I use monthly data. The analysis is for Poland, the Czech Republic, and Hungary. There is dual currency system in these countries. The euro plays a central role in trade, and the dollar in financial transactions. The exchange rate is the price of USD or EUR in units of domestic currency (PLN, CZK, HUF).

When USD exchange rates are used, there are different starting dates for each country. Thus, the sample starts when the floating exchange regime was implemented, for Poland in April 2000 and for the Czech Republic in June 1997, or when the widening of the band to +/-15% occurred, for Hungary in May 2001.

When EUR exchange rates are used, the sample starts in January 2004, when all three countries joined the European Union. Marcinkowska et al. (2009) indicate this date as the date when the structural change appeared, and the dollar was replaced by the euro as the base currency for foreign currency transactions. The sample ends in December 2012.

I use 1-, 3-, 6- month money market rates, namely WIBOR, PRIBOR, BUBOR, and LIBOR and EURIBOR, as well as 5- year government bonds. The level of economic activity is measured as an output gap, that is the difference between logarithms of seasonally adjusted GDP and its trend. The quarterly data were disaggregated to monthly frequencies using the Fernandez method. I also use the other method of calculating the output gap using industrial production index. As a proxy of risk premium I apply: output gap, returns on stock market indices (WIG 20 index, PX index, BUX index), the difference between the returns on domestic and foreign (S&P 500, Euro Stoxx 50) stock market indices, the ratio of gross government debt to GDP, and the ratio of current account to GDP. The data were obtained from the webpages of the relevant central banks, OECD, and IMF.

RESULTS

Table 1 shows the results of estimating symmetric models (see equations (3), (4), and (5)). In case of the equation (3) β coefficients of interest rate differential are statistically insignificant for Poland and the Czech Republic, and they are negative and statistically significant for short-term interest rates for Hungary. This models seems to fit the data worst (c.f. high values of sums of the squared errors - SSR). In case of the equations (4) and (5) for Poland β coefficients are positive and statistically significant for dollar and 1-month interest rate for euro.

These positive β coefficients are consistent with the UIP condition. The coefficients on interest rate differentials for the Czech Republic and Hungary are

statistically insignificant or, in a few cases when euro exchange rate is considered, statistically significant and negative.

The equation 5 includes some proxies for risk premium. I have chosen the variables that are statistically significant and fit the data best. In case of Poland these are: the output gap calculated using industrial production index and returns on WIG20 index, in case of the Czech Republic these are: the ratio of current account balance to GDP and output gap, and in case of Hungary these are: returns on BUX index and the ratio of gross government debt to GDP. Nevertheless, the addition of these variables does not help to obtain the results consistent with the UIP condition. Still, the results of estimating the equation 5 give positive β coefficients only in case of Poland.

Table 1. Symmetric models

			Equation (3)			Equation (4)			Equation (5)		
			β	Wald test	SSR	β	Wald test	SSR	β	Wald test	SSR
Poland	USD	1M	0,64	0,61	0,195	2,97*	0,16	0,184	3,24*	0,09	0,179
		3M	0,55	0,53	0,195	2,70*	0,20	0,186	3,03*	0,11	0,180
		6M	0,49	0,49	0,195	2,57*	0,25	0,186	3,02*	0,13	0,181
		5Y	0,33	0,58	0,195	2,43	0,47	0,190	3,62*	0,19	0,184
	EUR	1M	-0,54	0,39	0,062	3,58*	0,22	0,058	3,46*	0,23	0,056
		3M	-1,01	0,27	0,062	3,12	0,31	0,058	3,05	0,32	0,056
		6M	-1,49	0,19	0,062	2,64	0,48	0,059	2,46	0,51	0,057
		5Y	-3,31	0,20	0,061	-1,52	0,45	0,059	-1,80	0,38	0,057
Czech Republic	USD	1M	0,68	0,65	0,191	0,95	0,94	0,190	0,91	0,92	0,188
		3M	0,54	0,54	0,191	0,80	0,80	0,190	0,75	0,78	0,189
		6M	0,44	0,47	0,191	0,70	0,72	0,190	0,64	0,70	0,189
		5Y	-0,07	0,72	0,162	-0,15	0,70	0,161	1,31	0,93	0,158
	EUR	1M	0,18	0,74	0,024	-0,66	0,52	0,023	-6,36*	0,02	0,021
		3M	0,20	0,75	0,024	-0,36	0,59	0,023	-5,27*	0,03	0,021
		6M	-0,02	0,68	0,024	-0,55	0,53	0,023	-4,98*	0,03	0,021
		5Y	-2,16	0,11	0,023	-0,33	0,61	0,023	-2,03	0,27	0,022
Hungary	USD	1M	-1,42	0,03	0,184	-0,38	0,30	0,176	-0,06	0,35	0,167
		3M	-1,78	0,01	0,183	-0,69	0,21	0,176	-0,41	0,23	0,167
		6M	-2,05*	0,01	0,182	-0,90	0,17	0,176	-0,71	0,17	0,167
		5Y	0,89	0,97	0,185	0,83	0,94	0,176	-0,49	0,47	0,167
	EUR	1M	-2,31*	0,01	0,050	-2,06*	0,01	0,049	-0,65	0,25	0,044
		3M	-2,61*	0,00	0,049	-2,35*	0,00	0,049	-0,75	0,23	0,044
		6M	-2,95*	0,00	0,049	-2,68*	0,00	0,048	-0,86	0,23	0,044
		5Y	-1,67	0,30	0,051	-1,19	0,44	0,050	0,44	0,83	0,044

Source: own calculations;

Wald test denotes p-value of Wald test for the null hypothesis $\beta = 1$;

* denotes statistical significance of the parameter

Now, I concentrate only on the equations 4 and 5. I carried out similar analysis for the equation 3, but in no case the obtain results showed positive relation between the change of exchange rate and interest rate differential.

I, next, estimate the models in the form of equation 9 to test if dividing the sample into two subsamples depending on the value of interest rate differential helps to obtain positive and statistically significant β coefficients. Various thresholds τ from the set of values of interest rate differential were tested. Table 2 presents the results for τ equal to 0.9 quantile.

The asymmetric effects are detected only in case of Poland. The Wald test rejects the null hypothesis of $\beta_1 = \beta_2$ for euro exchange rate in case of 1-, 3-, and 6-months interest rates, and for dollar exchange rate in case of 1-month interest rate. The interest rate differential seems to have larger impact on the exchange rate when it is relatively low ($\beta_2 > \beta_1$).

Table 2. Threshold model – Equation (9)

			β_1	β_2	Wald test	SSR
Poland	USD	1M	2,39*	4,73*	0,02	0,180
		3M	2,52*	3,16*	0,55	0,185
		6M	2,22*	3,44*	0,32	0,185
		5Y	2,27	3,02	0,64	0,190
	EUR	1M	1,91	7,52*	0,00	0,055
		3M	1,40	6,50*	0,01	0,056
		6M	1,40	5,67*	0,03	0,057
		5Y	-3,17	0,93	0,08	0,058
Czech Republic	USD	1M	0,72	2,09	0,53	0,189
		3M	0,59	1,70	0,61	0,190
		6M	0,46	1,59	0,60	0,190
		5Y	-2,75	0,90	0,57	0,160
	EUR	1M	-4,40	0,73	0,40	0,022
		3M	-3,61	0,64	0,45	0,023
		6M	-3,45	0,27	0,57	0,023
		5Y	0,25	-0,42	0,92	0,023
Hungary	USD	1M	-0,33	-0,57	0,88	0,176
		3M	-1,25	1,18	0,16	0,173
		6M	-1,14	-0,03	0,40	0,175
		5Y	0,93	-0,30	0,55	0,176
	EUR	1M	-1,93	-0,79	0,39	0,049
		3M	-2,13*	-0,65	0,23	0,048
		6M	-2,49*	-1,25	0,37	0,048
		5Y	-1,18	-1,15	0,98	0,050

Source: own calculations;

Wald test denotes p-value of Wald test for the null hypothesis $\beta_1 = \beta_2$;

* denotes statistical significance of the parameter

Then, linearity of the exchange rate equations against smooth transition models is tested. The threshold variables is an interest rate differential or a level of economic activity. Table 3 presents p-values of LM1 and LM2 tests. The tests indicate strong nonlinearity of euro exchange rate equations in Poland. Dollar exchange rate equations in the Czech Republic and Hungary seem to be nonlinear according to the level of economic activity. Also the euro exchange rate equation for 5-year government bonds in the Czech Republic appears to be nonlinear. Smooth transition models (i.e. STAR models) are estimated for the equations for which LM-type tests show nonlinearity. Table 3 presents the results of estimating the STAR models.

Table 3. LM-type tests for STAR nonlinearity

			USD			EUR		
Equation:			(4)	(5)	(4)	(4)	(5)	(4)
Threshold variable:			differential	differential	output gap	differential	differential	output gap
Poland	1M	LM1	0,23	0,54	0,46	0,07*	0,09*	0,00*
		LM2	0,10*	0,17	0,63	0,07*	0,07*	0,00*
	3M	LM1	0,31	0,72	0,44	0,03*	0,03*	0,00*
		LM2	0,20	0,38	0,71	0,05*	0,05*	0,00*
	6M	LM1	0,35	0,80	0,54	0,04*	0,05*	0,00*
		LM2	0,26	0,46	0,83	0,03*	0,03*	0,00*
	5Y	LM1	0,61	0,97	0,34	0,13	0,09*	0,09*
		LM2	0,50	0,38	0,25	0,31	0,25	0,14
Czech Republic	1M	LM1	0,53	0,52	0,11	0,94	0,20	0,13
		LM2	0,37	0,21	0,07*	0,81	0,28	0,13
	3M	LM1	0,63	0,69	0,12	0,91	0,39	0,17
		LM2	0,35	0,19	0,09*	0,78	0,40	0,21
	6M	LM1	0,87	0,98	0,22	0,46	0,78	0,28
		LM2	0,36	0,18	0,12	0,36	0,21	0,37
	5Y	LM1	0,64	0,40	0,04*	0,21	0,05*	0,85
		LM2	0,87	0,70	0,07*	0,34	0,12	0,02*
Hungary	1M	LM1	0,15	0,22	0,07*	0,17	0,23	0,30
		LM2	0,28	0,40	0,13	0,35	0,47	0,26
	3M	LM1	0,14	0,24	0,05*	0,25	0,25	0,33
		LM2	0,28	0,44	0,13	0,43	0,49	0,35
	6M	LM1	0,19	0,32	0,07*	0,47	0,34	0,48
		LM2	0,38	0,56	0,18	0,39	0,62	0,58
	5Y	LM1	0,86	0,96	0,27	0,59	0,83	0,97
		LM2	0,87	0,71	0,44	0,77	0,97	0,61

Source: own calculations;
differential means interest rate differential;
* denotes p-value less than 0.1

The estimation of logistic smooth transition models for Poland for euro exchange rate shows, similarly as the estimation of equation 9, that higher level of interest rate differential or higher level of economic growth generate the regimes where the difference between domestic and foreign interest rate has weaker impact on the exchange rate ($\beta_1+\beta_2 < \beta_1$). Interestingly, for the Czech koruna – dollar exchange rate in case of 5-years bonds the positive relation between the exchange rate and interest rate differential ($\beta_1+\beta_2>0$) is found for the time periods where the level of economic activity is relatively high. βcoefficients for Hungary are always negative and their absolute value is higher in the time periods where the interest rate differential is relatively high.

Similarly, the exponential smooth transition models for Poland show that higher absolute value of interest rate differential, generates the regimes where the differential does not affect the exchange rate (β_2 statistically insignificant). While for the Czech koruna – dollar exchange rate and short term interest rates the results show positive and statistically significant relation between the exchange rate and interest rate differential in the time periods in which the absolute value of the level of economic activity is relatively high.

Table 4. Smooth transition models– Equation (6)

			Equation	Threshold variable	β_1	β_2	c	SSR
Logistic smooth transition models								
Poland	EUR	1M	(4)	output gap	3,95*	-29,09*	0,025	0,051
Poland	EUR	3M	(4)	differential	10,17*	-6,68*	0,003	0,054
Poland	EUR	3M	(5)	differential	9,65*	-6,31*	0,003	0,052
Poland	EUR	3M	(4)	output gap	3,58*	-31,86*	0,025	0,052
Poland	EUR	6M	(4)	output gap	3,68*	-28,28*	0,025	0,052
Poland	EUR	5Y	(5)	differential	29,19	-34,00	0,001	0,052
Poland	EUR	5Y	(4)	output gap	0,43	-18,38*	0,025	0,053
Czech Republic	USD	5Y	(4)	output gap	-13,07*	19,73*	-0,005	0,154
Czech Republic	EUR	5Y	(5)	differential	1,89	-0,08*	-0,003	0,021
Hungary	USD	1M	(4)	output gap	-0,14	-11,98*	0,056	0,164
Hungary	USD	3M	(4)	output gap	-0,31	-11,82*	0,057	0,164
Hungary	USD	6M	(4)	output gap	-0,36	-11,62*	0,058	0,164

Table 4. -continuation

			Equation	Threshold variable	β_1	β_2	c	SSR
Exponential smooth transition model								
Poland	USD	1M	(4)	differential	5,15*	-30,96	0,005	0,182
Poland	EUR	1M	(4)	differential	7,64	-628,78	0,002	0,056
Poland	EUR	6M	(4)	differential	7,82*	-475,10	0,002	0,057
Poland	EUR	1M	(5)	differential	7,71	-637,23	0,002	0,054
Poland	EUR	6M	(5)	differential	7,45*	-431,95	0,002	0,055
Czech Republic	USD	1M	(4)	output gap	-1,05	5,41*	-0,014	0,187
Czech Republic	USD	3M	(4)	output gap	-1,33	5,21*	-0,014	0,187
Czech Republic	USD	6M	(4)	output gap	-1,42	5,42*	-0,015	0,188
Czech Republic	EUR	5Y	(4)	output gap	-7,09	8,11	-0,036	0,022

Source: own calculations;
differential means interest rate differential;
* denotes statistical significance of the parameter

CONCLUSIONS

The study concerns the relationship between an exchange rate and an interest rate differential in Poland, the Czech Republic, and Hungary. The main aim of the paper is to test whether allowing for certain nonlinear effects enables to obtain results consistent with the UIP condition, i.e. a positive relation between an exchange rate and an interest rate differential.

The results show that the Polish zloty – euro exchange rate equations and the Polish zloty – dollar exchange rate equation for 1-month interest rate are nonlinear. Precisely, in the periods where the level of interest rate differential or the level of economic growth is relatively high, the interest rate differential has weaker impact on the exchange rate. In case of the Czech koruna – dollar exchange rate allowing for nonlinear effects gives positive β coefficients in the periods where the level or the absolute level of economic activity is relatively high. In case of the Czech koruna – euro and the Hungarian forint – euro or –dollar exchange rates the results point to, difficult to explain, negative relation between the exchange rate and the interest rate differential. Additionally, including certain proxies for risk premium does not change the results.

REFERENCES

Baillie R.T., Kiliç R. (2006) Do asymmetric and nonlinear adjustments explain the forward premium anomaly?, Journal of International Money and Finance, 25, pp. 22-47.

Bansal R. (1997) An exploration of the forward premium puzzle in currency markets, The Review of Financial Studies, 10, pp. 369-403.

Bansal R., Dahlquist M (2000) The forward premium puzzle: different tales from developed and emerging economies, Journal of International Economics, 51, pp.115-144.

Chinn M. D., Meredith G. (2005) Testing uncovered interest parity at short and long horizons during the Post-Bretton Woods Era, NBER Working Papers 11077.

Froot K. A. (1990) Short rates and expected asset returns, NBER Working Paper, 3247.

Marcinkowska-Lewandowska W. (red. nauk.), Rubaszek M., Serwa D. (2009) Analiza kursu walutowego, Wydawnictwo C.H. Beck.

Moh Y. (2006) Continuous-time model of uncovered interest parity with regulated jump-diffusion interest differential, Applied Economics, 38(21), pp. 2523-2533.

Omer M., de Haan J., Scholtens B.(2012) Testing uncovered interest rate parity using LIBOR, CESifo Working Paper Series, 3839, CESifo Group Munich.

Sarno L., Valente G., Leon H. (2006) Nonlinearity in deviations from uncovered interest parity: an explanation of the forward bias puzzle, Review of Finance, 10, pp. 443-482.

van Dijk D., Teräsvirta T., Franses P.H. (2002) Smooth transition autoregressive models - a survey of recent developments, Econometric Reviews, 21, pp. 1–47.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 240 – 250

MULTIVARIATE DECOMPOSITIONS FOR VALUE AT RISK MODELLING

Ryszard Szupiluk, Piotr Wojewnik
Department of Business Informatics, Warsaw School of Economics
e-mail: rszupi@sgh.waw.pl, piotr.wojewnik@gmail.com
Tomasz Ząbkowski
Department of Informatics
Warsaw University of Life Sciences – SGGW
e-mail: tomasz_zabkowski@sggw.pl

Abstract: This paper presents the application of independent component analysis (ICA) for value at risk modelling (VaR). The probabilistic models fitted to hidden components from the time series help to identify the independent factors influencing the portfolio value. An important issue here is the choice of the ICA algorithm, especially taking into account the characteristics of the instruments with respect to higher-order statistics. The proposed ICA-VaR concept has been tested on transactional data of selected stocks listed on Warsaw Stock Exchange.

Keywords: multivariate decompositions, value at risk modelling, independent components analysis

VALUE AT RISK MODELING

One of the most popular concept of investment risk modelling is the concept of Value at Risk, which consists of estimating the risk for a specified time horizon at a given probability [Jorion 2001, Jajuga 2001, JP Morgan, 1995]. Although the concept itself is simple and intuitive, it is associated with a fundamental problem of estimating the probability that a given financial instrument (or portfolio) will reach specific values in the future. In this area, one of the most commonly used approaches are these based on simulations. They aim to find the possibly best mathematical model for the instrument based on historical data, and then performing predictive simulation of the model behaviour. This opens up a discussion whether the model is adequately fitted to empirical data. In practice,

the model fit is a compromise between the characteristics of the empirical data, a priori assumptions about the nature of the original phenomenon and properties of the mathematical apparatus [Bollerslev et al. 1992, Shiryaev 1999].

Since the Markowitz publications [Markowitz 1952] the phenomenon of uncertainty and risk was mainly interpreted in terms of financial instruments volatility and correlation. Additionally, the rational expectations hypothesis justified the random nature of the changes in financial instruments, making popular the models which were based on Gaussian distribution. Bearing in mind that the variance/covariance fully identifies the Gaussian distribution, volatility expressed by the variance gives complete statistical information about the phenomenon of uncertainty and risk. Taking into account also the relationship between Gaussian distribution and Central Limit Theorem a comprehensive conceptual system was established and it prevails in the description of uncertainty and risk over the last decades.

The experience of recent years and the current situation on the financial markets, however, indicate some limitations of this conceptual apparatus. It turned out that models based on Gaussianity and correlations cannot recognize some critical features of the financial markets such as rare events. Historical and actual volatility described by variance is often not applicable for the future situation reasoning, which, in fact, can change dramatically due to rare events such as market breakdown. As a result, unexpected change on financial instruments is not so much associated with the current volatility, but it is rather due to the occurrence of rare events such as panic in the stock market or bankruptcy of a large financial institution.

That types of phenomena and behaviour can be observed in the short-term scales. As a result, we need to look for mathematical apparatus that can better deal with the rare phenomenon. Although the number of works devoted to rare events is substantial, the problem is still an open research issue [Embrechts et al. 1997, Harvey 2013].

One reason for the difficulty of modelling specific market behaviour may be the fact that they do not occur as isolated events. In other words, a crisis or crash is frequently preceded by a long process that develops on a seemingly normally functioning market. Similarly, after the collapse, its effect lasts for a long time. In addition, it should be noted that market instruments are generally interrelated, although the nature, timing and scope of this relationship can be difficult to determine and predict. As a result, we can assume that the morphology of such time series is so complex and the only one signal analysis can be very confusing. This is natural motivation for time series decomposition into components associated with its particular characteristics for their individual properties modelling.

There are two main approaches for decompositions: (a) one-dimensional that is based on a decomposition of the time series (e.g. trend, cycles, noise) or (b) multi-dimensional, taking into account and exploring the relationships between few

different signals. In the following discussion we will focus on a multi-dimensional approach that recently led to a number of interesting decomposition methods with number of practical application. In particular, we consider the following method of VaR supported by ICA decomposition:

1. Collect the original time series into one multivariate variable;
2. Decompose the multivariate variable into hidden independent components (separation stage);
3. Choose components for further analysis (filtration stage);
4. Estimate the probability distributions for each component;
5. Perform simulation for each component;
6. Do re-mixing of the components using reverse system to decomposition (separation);
7. Calculate VaR for given portfolio.

This concept establishes the general research framework, in which components can be identified using different mathematical characteristics. In case of the components identified as noises the algorithm can also realize the filtration [Szupiluk 2004].

The following discussion will focus primarily on classic independent component analysis [Hyvärinen et al. 2001]. This method of decomposition has many practical applications including the blind signal separation problem [Cichocki and Amari, 2002]. However, in contrast to the relatively clear results obtained with a principal component analysis - PCA [Jolliffe 1986], ICA needs the deeper insight into the characteristics of the used algorithms. The motivation for ICA decomposition is due to the fact that blind separation is one of the most general methods of separation, exploring higher-order statistics. This is particularly important in case of rare events observed in financial time series, for which the kurtosis is one of the most important criterion of their assessment.

CHARACTERISTICS OF INDEPENDENT COMPONENT ANALYSIS

In the classical meaning independent component analysis is formulated as a method that allows separation of a multi-dimensional observation vector $\mathbf{x} = [x_1, x_2, ..., x_n]^{\mathrm{T}}$ into statistically independent components $\mathbf{y} = [y_1, y_2, ..., y_m]^{\mathrm{T}}$. It is assumed that the estimation of the independent components is performed using a linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is separating matrix.

Independent Component Analysis has been widely used for modelling economic and financial phenomena [Back and Weigend 1997]. Also, there are number of publications indicating the effectiveness of ICA for risk analysis [Chen et al. 2007, Wu et al. 2006]. Independent component analysis can be considered twofold. In first case, it can be assumed as a strict statistical decomposition

method, which allows to extract independent components from the multidimensional observations. In the second case, ICA can be treated as a method to solve blind separation problem.

ICA as a statistical method

The assessment of signals independence requires the knowledge of their probability distributions, which is a relatively complex task in case of financial time series. In addition, ICA models assume that independent components are mixed and hidden and their distributions are, by definition, unknown. As a result, the practical notion of statistical independence of components obtained in the ICA is not precise and the final effect is verifiable to a limited extent. Unlike the principal components analysis with linear algebra apparatus, separation of independent components requires adoption of certain criteria, concepts, principles and characteristics, which exploration (optimization) may result in certain numerical algorithms. The most popular approaches include the minimization of mutual information, entropy maximization, non-Gaussianity maximization (measured by negentropy or kurtosis) or non-linear decorrelation.

One of the standard algorithms for finding the matrix **W** are Natural Gradient [Amari et al. 1997]

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu(t)\left[\mathbf{I} - E\left\{\mathbf{f}(\mathbf{y})\mathbf{y}^{\mathrm{T}}\right\}\right]\mathbf{W}(t). \tag{1}$$

and FASTICA [Hyvärinen et al. 2001]

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \mu(t)\mathbf{D}\left[E\left\{\mathbf{f}(\mathbf{y}(t))\mathbf{y}^{T}(t)\right\} - \mathrm{diag}\left(E\{f(y_i)y_i\}\right)\right]\mathbf{W}(t) \tag{2}$$

where E is expectation operator, $\mu(t)$ is a learning rate, $\mathbf{f}(\mathbf{y}) = \left[f_1(y_1),..., f_n(y_n)\right]^{\mathrm{T}}$ and $\mathbf{D} = \mathrm{diag}\left(1/E\{f(y_i)y_i\} - E\{f'(y_i)\}\right)$ are the vector and the matrix with nonlinearities of the:

$$f_i(y_i) = -\frac{\partial \log(p_i(y_i))}{\partial y_i} \tag{3}$$

where $p_i(y_i)$ is *pdf* of signal y_i.

The key issue in these algorithms is the selection of non-linearity (3). One of the simplest but well working rules are these using higher-order statistics and based on the observation that the non-linearity (3): (i) takes the linear form for the Gaussian distribution; (ii) for the distributions with the slope higher than Gaussian it is growing faster than linear; (iii) for the distributions with the slope lower than Gaussian it is growing slower than linear. Taking the kurtosis $\kappa_4(y_i)$ as distribution slope parameter we obtain a rule of non-linearity selection [Cichocki et al. 2007]:

$$f_i(y_i) = \begin{cases} y_i^3 & \text{for} \quad \kappa_4(y_i) > 0, \\ \tanh(y_i) & \text{for} \quad \kappa_4(y_i) < 0, \\ y_i & \text{for} \quad \kappa_4(y_i) = 0. \end{cases} \tag{4}$$

The above rule can be effective for simple, unimodal and symmetric cases, taking into account a small number of signals with relatively different distributions. In more complex cases it requires more accurate non-linearity determination what makes it difficult since the distribution of independent component is a priori unknown. As a possible solution we can propose either, heuristic methods, parametric models or adaptive techniques. Nevertheless, one of the most versatile approaches is a method based on the Extendend Generalized Lambda Distribution (EGLD) in which it is possible to model a wide range of distributions with different kurtosis and skewness parameters [Karian et al. 1996, Karvanen et al. 2002].

It should be noted that at the theoretical level, with certain assumptions and simplifications, the criteria that establish the ICA algorithms lead to a situation in which the mutual independence is approximated or reduced to a fourth-order statistical relationships. In this case, for decorrelated and symmetric data ICA can be expressed as [Comon 1994]:

$$\max_{\mathbf{W}} \sum_{i=1}^{n} J(y_i) \approx \min_{\mathbf{W}} I(\mathbf{y}) \approx \max_{\mathbf{W}} \left(\sum_{i=1}^{n} \kappa_4^2(y_i) \right) \tag{5}$$

where $J(y_i)$ is negentropy of the signals y_i; $I(\mathbf{y})$ is join mutual entropy.

Therefore, it can be assumed that, although the ICA algorithms derived from different criteria have their particular numerical specificity, in practice the effect is linear and non-linear decorrelation, resulting in removal of the second and fourth order statistics. These observations led to the development of ICA methods based on fourth-order statistics, including also tensor approach, e.g. JADE algorithm [Cardoso 1999].

ICA as a blind signal separation method

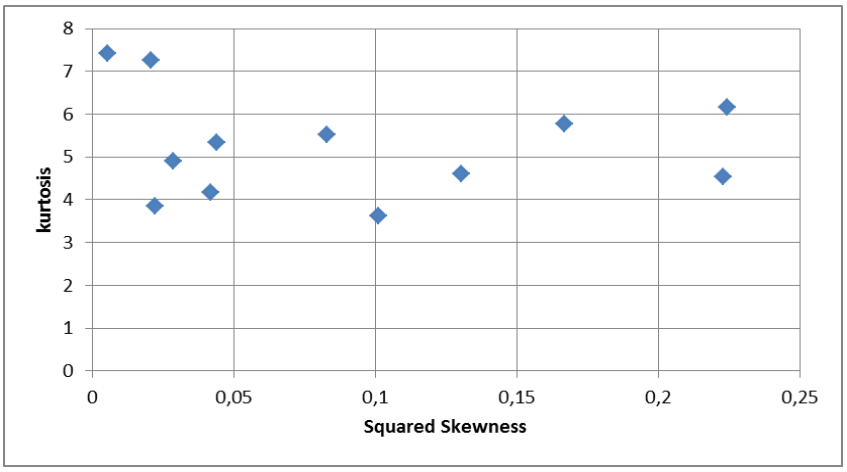
The problem of blind signal separation is defined as the extraction of unknown signals mixed in the unknown system, with limited a priori knowledge. If we additionally assume that the sought source signals are statistically independent and the mixing is a linear combination of the signals, then the problem of blind signal separation and independent component analysis are identical. However, to separate the real signal, then we need a criterion to measure the independence. This is conditioned by the approximation accuracy using fourth order statistics of the estimated signal probability distribution function. Unfortunately, the form of this distribution is not known a priori in separation problem. Therefore, there is also a question of accuracy assessment in the separation process. While in some practical applications e.g. physical separation

of speech signals, the evaluation can be quite simple; then in case of financial time series (signals), the situation is quite complex. First of all, a separation using different algorithms within ICA is possible. If their results comply then we can state that satisfactory solution was achieved, otherwise it will be difficult to assess which solution is better.

Consequently, an essential issue of ICA algorithm choosing is to fit properties and characteristics of the algorithm to a given problem. In case of financial time series it is a non-stationarity of time series with respect to the higher order statistics.

Figure 1 shows an example of kurtosis and squared skewness distribution calculated on logarithmic returns of the WIG20 index (for different periods, covering time span from 1994 till 2012). Kurtosis is calculated in a standardized form and for a Gaussian distribution it has zero value.

Figure 1. The points represent WIG20 index characteristics (kurtosis and squared skewness for different length of time windows (minimum 2 years) and covering time span from 1994 till 2012



Source: own calculation

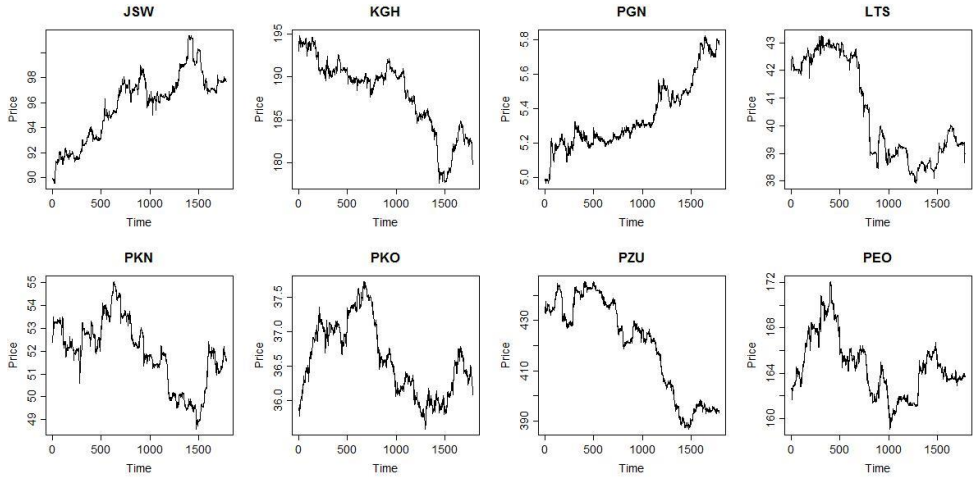
This example from Warsaw Stock Exchange shows that even for single time series the statistical characteristic is volatile and highly influenced by chosen time-window.

EXPERIMENT ON EMPIRICAL DATA

In this section an experiment on financial data was conducted. We considered eight stocks from Warsaw Stock Exchange covering the time span 19/12/2012 – 18/01/2013. Data consisted of 1783 observations of 5 minute data. Half of them (19/12/2012 – 07/01/2013) was used for decomposition matrix and

probability density functions (PDF's) estimation and the second half (08/01/2013–18/01/2013) was used for testing. A portfolio of stocks consisted of JSW, KGH, PGN, LTS, PKN, PKO, PZU, PEO. For their characteristics, please see Fig. 2.

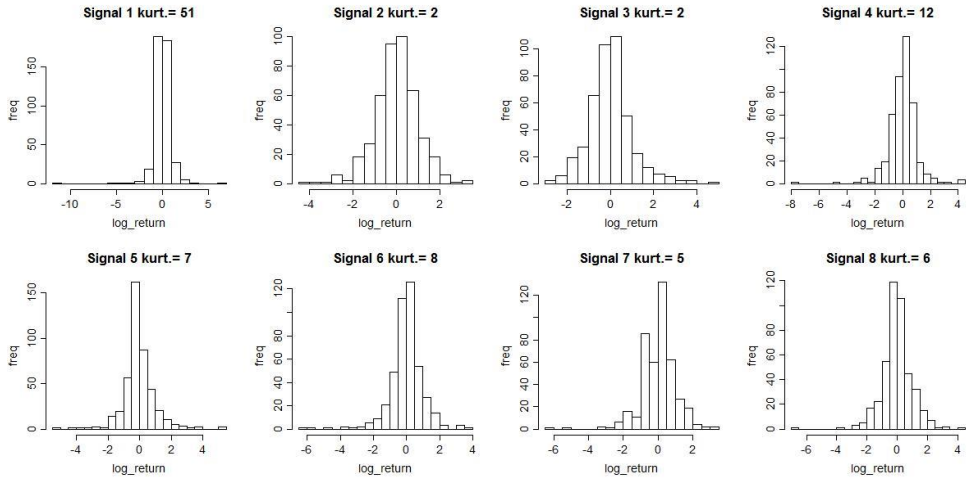
Figure 2. Stocks used in the experiment



Source: own preparation

Histograms of latent signals obtained as a result of JADE decomposition calculated on log returns for all stocks (time span for analysis 19/12/2012–07/01/2013) are presented on Fig. 3. We can observe that kurtosis of the signals ranges from 2 to 51 in this sample.

Figure 3. Histograms of latent components obtained after decomposition



Source: own preparation

The PDFs of latent signals were estimated using t-scale location distribution with shape parameter *df*, location parameter *m* and scale parameter *s*. Degrees of freedom ranges from 1 to 4. Please see Table 1 for details.

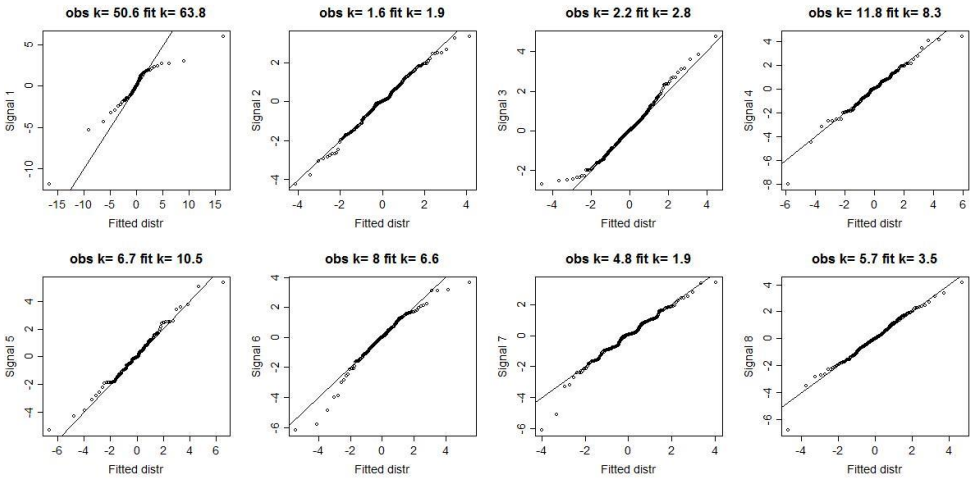
Table 1. PDF characteristics of latent signals

m	s	df
-0.0031	0.2110	1.1382
0.0171	0.7546	4.2239
-0.0445	0.7035	3.5227
0.0360	0.5241	2.3067
-0.0352	0.5056	2.1233
0.0429	0.5553	2.4985
0.0251	0.7277	4.1791
-0.0076	0.6633	3.2060

Source: own calculation

QQ plots of observed PDF's vs. fitted distribution are presented in Figure 4. Fitted kurtosis ranges from 2 to 60 and it is close to expectations. Therefore, we can conclude that used distributions seem to be an adequate choice.

Figure 4. QQ plots observed PDF's vs. fitted distribution



Source: own preparation

Finally, the VaR calculation results on test data (08/01/2013–18/01/2013) are presented in Table 2. It presents the percentage of cases exceeding specified VaR level using 100 000 simulations based on given decompositions and also in comparison to VaR level estimated on historical data (19/12/2012–07/01/2013).

In other words, we show VaR level not taking into account the value of a specific stock in this portfolio but total portfolio value.

Table 2. VaR results of portfolio consisted of eight stocks

VaR level	EGLD ICA VaR	Jade VaR	Historical VaR
0,01	0 %	1,2%	2,1%
0,025	1,4%	2,1%	4,2%
0,05	4,6%	6,0%	7,9%
0,1	9,7%	10,9%	12,7%

Source: own calculation

We showed that in an environment characterized by non-linearities and the occurrence of interactions between the stocks, the ICA methodology can advantageously reveal an underlying structure in financial time series for the purpose of risk measurement. The results obtained on the test dataset indicate the advantage of ICA approach over VaR calculation based on historical data. The best results were achieved for EGLD ICA method, which allows to simulate distributions in a wide range of kurtosis and skewness.

CONCLUSIONS

In this paper the application of independent component analysis in the multidimensional decomposition framework for VaR was proposed. In contrast to the relatively explicit algebraic decomposition methods, different ICA algorithms have their own characteristics, what can significantly influence the results. In particular, it refers to the case of ICA as a blind source separation method. Therefore, the choice of appropriate algorithm with respect to the third and fourth order statistics is the key issue here. These higher order statistics play an important role in ICA algorithms both, for the optimization task and for the numerical implementation of the algorithm. In the case of financial instruments we can observe both the instability of these statistics and the volumes which are significantly different from the Gaussian distribution. In such circumstances, it seems appropriate to select an adaptive algorithm for non-linearity selection taking into account a wide range of kurtosis and skewness. One of them is a system based on EGLD distribution. For the VaR modelling EGLD adoption led to better results than JADE, although the last one is recognized as one of the most popular and effective algorithms.

In this study, we focused mainly on distribution fitting and forecasting. Nevertheless, it should be noted that under the proposed approach a filtration and elimination of the specific components, can be also considered.

Acknowledgments: This work was funded by the National Science Center in Poland based on decision number DEC-2011/03/B/HS4/05092.

REFFERENCES

Amari S., Cichocki A., Yang H.H. (1996) A new learning algorithm for blind signal separation, Advances in Neural Information Processing Systems NIPS-1995, MIT Press, Cambridge MA, pp. 757–763.

Back A.D., Weigend A.S. (1997) A first application of independent component analysis to extracting structure from stock returns, Int. Journal of Neural Systems 8(4), pp. 473–484.

Bollerslev, T., Chou, R.Y., Kroner, K.F. (1992) ARCH Modelling in Finance: A Review of the Theory and Empirical Evidence, Journal of Econometrics 52, pp. 5–59.

Cardoso J.F. (1999) High-order contrasts for independent component analysis, Neural Computation 11, pp. 157–192.

Chen Y., Hardle W., Spokoiny V. (2007) Portfolio value at risk based on independent component analysis Journal of Computational and Applied Mathematics 205(1), pp. 594–607.

Cichocki A., Amari S. (2002) Adaptive Blind Signal and Image Processing, John Wiley, Chichester.

Cichocki A., Sabala I., Choi S., Orsier B., Szupiluk R. (1997) Self adaptive independent component analysis for sub-Gaussian and super-Gaussian mixtures with unknown number of sources and additive noise, Proc. of NOLTA-97, vol. 2, Hawaii USA, pp. 731–734.

Comon P. (1994) Independent component analysis, a new concept?, Signal Processing 36, pp. 287-314

Embrechts P., Klüppelberg C., Mikosch T. (1997) Modelling Extremal Events for Insurance and Finance. Berlin, Springer.

Harvey A.C. (2013) Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series, Cambridge University Press.

Hyvärinen A., Karhunen J., Oja E. (2001) Independent Component Analysis, John Wiley.

Jajuga K. (2001) Value at Risk, Rynek Terminowy13, pp. 18-20.

Jolliffe T. (1986) Principal Component Analysis, Springer-Verlag.

Jorion P. (2001) Value at Risk, McGraw-Hill.

JP Morgan (1995) Riskmetrics Technical Document, 3rd ed., New York

Karian Z.A., Dudewicz E.J., McDonald P. (1996) The extended generalized lambda distribution system for fitting distributions to data: history, completion of theory, tables, applications, the Final Word on Moment Fits, Communications in Statistics - Simulation and Computation 25, pp. 611–642.

Karvanen J., Eriksson J., Koivunen V. (2002) Adaptive Score Functions for Maximum Likelihood ICA. VLSI Signal Processing 32, pp. 83–92.

Kouontchou P., Maillet B. (2007) ICA-based High Frequency VaR for Risk Management ESANN'2007 proceedings - European Symposium on Artificial Neural Networks, pp. 385-390.

Markowitz H.M. (1952) Portfolio Selection, The Journal of Finance 7, pp. 77–91.

Shiryaev A.N. (1999) Essentials of stochastic finance: facts, models, theory, World Scientific, Singapore.

Szupiluk R., Wojewnik P., Zabkowski T. (2004) Model Improvement by the Statistical Decomposition, Lecture Notes in Computer Science 3070, pp. 1199–1204.

Wu E.H., Yu P.L., Li W.K. (2006) Value at risk estimation using independent component analysis-generalized autoregressive conditional heteroscedasticity (ICA-GARCH) models, International Journal of Neural Systems 16(5), pp. 371–382.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 251 – 260

EVALUATION OF THE EFFICIENCY OF FLEXICURITY IMPLEMENTATION IN OECD COUNTRIES

Andrzej Szuwarzyński
Department of Management, Gdansk University of Technology
e-mail: Andrzej.Szuwarzynski@zie.pg.gda.pl

Abstract: Flexicurity is a policy of flexible and secure labour market. It has been the subject of many analyses, however, a coherent evaluation methodology is difficult to specify. The purpose of this paper is to propose a Data Envelopment Analysis based model for the evaluation of the efficiency of flexicurity implementation in OECD countries. The results will be used to create the ranking of countries, to determine changes in time, and to identify the reasons for inefficiency. On top of that, it will be possible to formulate recommendations for decision makers.

Keywords: flexicurity, Data Envelopment Analysis, active labour market policy, lifelong learning, composite indicators

INTRODUCTION

Flexicurity is to become the target policy governing the labour market in the European Union (EU). The European Development Strategy – Europe 2020 – foresees that the implementation of flexicurity will result in lower unemployment and lower segmentation of the labour market. Also, flexicurity-related activities are undertaken by many non-EU countries. Although the issue of flexicurity implementation has been the subject of many studies, it has been emphasised that the distinct effects of the flexicurity policy are not easy to measure, which is seen as the key methodological issue with strong implications for forming the labour market policy [Wilthagen 2012].

The aim of this paper is attempt to create an efficiency evaluation model for activities relating to the implementation of flexicurity. The model is based on the Data Envelopment Analysis (DEA) thus making it possible to determine the efficiency of implementation in the compared countries. The OECD countries have been evaluated by taking into account parameters which characterise the concept

of flexicurity. The source data for the evaluation have been taken from the OECD labour market database for the years 2000-2010.

Flexicurity is defined as an integrated strategy aimed at simultaneous increase in the flexibility and the security of the labour market [Commission 2007]. This combination of a flexible labour market model on the one hand and a social security model on the other hand is based on consecutive transitions during person's professional life: from the completion of education to starting the professional career period, to subsequent changes of jobs, to periods of unemployment, and to retirement.

The historical roots of flexicurity lie in the Netherlands and Denmark. The main driving force behind making labour relations flexible were the economic needs of modern organisations which, in order to remain competitive, must be able to adapt to changes quickly and easily. In 2007, the European Commission defined common flexicurity implementation principles, which encompass four basic dimensions [Commission 2007]:

- Flexible and reliable contractual arrangements (FCA), achieved through modern labour law, collective agreements and work organisation.
- Comprehensive lifelong learning (LLL) strategies, ensuring that all employees are always able to adapt to the changes on the labour market.
- Active labour market policies (ALMP), providing assistance in dealing with the changes and making it possible to shorten the periods of unemployment.
- Modern social security systems (MSS), ensuring adequate income support (unemployment benefits, retirement pensions and healthcare services).

The proper functioning of the labour market has been the focus of attention of all decision makers. Forming policies requires measurement and evaluation. One example of flexicurity measurement is the study prepared under the European Commission project [Manca et al. 2010] which applied a simple methodology of measurement based on Composite Indicators (CIs) [Hoffman et al. 2008]. CIs are calculated according to the indicators measuring four above mentioned dimensions of flexicurity.

USING DATA ENVELOPMENT ANALYSIS TO CREATE COMPOSITE INDICATORS

CIs are regarded as a useful tool for the analysis of public policies. They integrate a large amount of information in a transparent and comprehensible format, which is easy to interpret by the general public [Shen et al. 2011]. CIs are crucial for taking operational decisions as well as for forming policies. Composite Indicators are created by mathematical aggregation of the set of individual indexes according to the CIs construction rules [Hoffman et al. 2008].

DEA is a nonparametric technique of mathematical programming which enables the measurement of relative efficiency of a homogeneous group of objects

called Decision Making Units (DMUs) [Charnes at al. 1978]. The efficiency measurement is based on the determination of relation between multiple inputs and multiple outputs of a given entity's functioning in the context of a given goal, with the use of linear programming techniques [Cooper et al. 2011].

The efficiency measurement consists in determining reference objects and comparing all other objects to them. Consequently, the relative efficiency of DMUs is measured by classifying them as fully efficient on the basis of available data [Cooper et al. 2011]. The first and the most commonly used DEA formulation is the CCR model [Charnes et al. 1978], where the efficiency measure of each DMU is obtained as the maximum of the quotient of weighted outputs to weighted inputs. The measure of efficiency θ_o for the DMU reference group $(j=1,...,n)$ is calculated for the outputs $(y_{rj},\ r=1,...,s)$ and inputs $(x_{ij},\ i=1,...,m)$, which may be expressed with the following formula:

$$\max_{v,u} \theta_o = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \tag{1}$$

subject to:

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_r x_{ij}} \leq 1 \quad j=1,...,n \tag{2}$$

$$u_r, v_i \geq 0 \quad r=1,...,s \quad i=1,...,m, \tag{3}$$

where: u_r, v_i are variable weights which are determined by solving the above problem on the basis of data from all DMUs.

DEA is also used for constructing CIs, because this method facilitates aggregating many indicators without referring to the *a priori* knowledge of their weights [Shen et al. 2011]. Each DMU receives its own set of best possible weights assessing the relative performance of a particular DMU. The DEA-based structure of Composite Indicators has been subject to many studies [e.g. Lovell et al. 1995, Cherchye et al. 2009, Despotis 2005].

The problem in using DEA to construct CIs for the evaluation of macroeconomic policies is the determination what should be the inputs. Here the concept of the helmsman (or central planning board) is used. It was first introduced by Koopmans, who examined the issues of attainment of efficiency under a regime of decentralised decisions [Koopmans 1951], where each country has control tools for conducting its own macroeconomic policy. The outcomes depend on one input only, i.e. the macroeconomic decision-making apparatus, which is called the helmsman [Lovell 1995]. Consequently, in the DEA model, the vector of inputs is limited to a dummy variable, whose value equals to one for every DMU [Lovell et al. 1995, Despotis 2005]. Such a model may be a tool for the aggregation of several indicators into the general CI without reference to inputs, assuming that all

countries have the same level of capacity to achieve full efficiency [Cherchye et al. 2009]. The evaluation of the performance of the examined unit differs from the evaluation of efficiency, because only the outputs are assessed and how they have been achieved is irrelevant. Consequently, the model is simplified and since the inputs are not converted into outputs, the process should be described as the measurement of effectiveness rather than the measurement of efficiency [Cooper at al 2009].

The use of the CCR model without additional weight restrictions enables each DMU to achieve the most beneficial possible result in the efficiency score, which is often related to zero values of weights, which are not acceptable in real-life applications [Roll, Golany 1993]. In practice, in many cases, classic DEA models evaluate inefficient units using the reference points on the frontier of the production possibility set (PPS), which are not Pareto-efficient. These models assign zero weights to optimal multipliers which means that not all sources of inefficiency are taken into account [Ramón et al. 2010].

The flexibility of weights is considered as one of DEA's main advantages, although full flexibility is also a disadvantage because important factors may be ignored in the analysis. This DEA deficiency may be remedied by the means of weight restrictions, which also improve the discrimination between the examined DMUs and consequently the number of efficient DMUs is reduced [Angulo-Meza, Lins 2002].

The process of imposing weight restrictions is highly case-dependent and there are no general rules in this area [Roll, Golany 1993]. However, the weight restrictions may be determined by referring to the opinion of experts [Cherchye et al. 2009] or on the basis of the value of variables of evaluated DMUs [Ramón et al. 2010, Roll, Golany 1993]. The following technique may be applied [Roll, Golany 1993]:

1. An unbounded CCR model is initiated and average weights u_r and v_i for outputs and inputs are obtained.
2. The magnitude of variability is established within the weights for the same factor as the relation of the highest value to the lowest d:1.
3. The basic CCR model is extended by adding a set of restrictions of the type:

$$\frac{2 \times u_r}{1+d} \leq u_{rj} \leq \frac{2 \times d \times u_r}{1+d} \tag{4}$$

4. The bounded model is initiated.

Average weights are calculated from the reduced vector of weights by ignoring extreme values [Roll, Golany 1993] or by using only the extreme efficient DMUs. [Angulo-Meza, Lins 2002]. Similarly, restrictions may be imposed on virtual outputs. This is the share of the entire virtual DMU_j output devoted to r output or in other words, the "importance" attached to output r for DMU_j may be restricted to the range between $[\varphi_r, \psi_r]$ in the following form [Allen et al. 1997]:

$$\phi_r \leq \frac{u_r y_{rj}}{\sum_{r=1}^{s} u_r y_{rj}} \leq \psi_r \tag{5}$$

STRUCTURE OF THE MODEL

The goal of the examination is to determine the degree of performance in implementing flexicurity in OECD countries for the years 2000, 2005 and 2010. Due to insufficient data, six out of 34 OECD countries were excluded: Chile, Iceland, Israel, Mexico, New Zealand and Turkey. An output oriented, weight bounded model was used. A single constant input with the value equal to one was adopted (helmsman). The basic output variable is the harmonised unemployment rate (UNEMPL). The following variables characterise four main flexicurity dimensions:

- LLL_GDP - percentage of GDP allocated for the training of employees,
- ALMP_GDP - percentage of GDP allocated for active labour market policies,
- MSS_GDP - percentage of GDP allocated for the unemployment benefits,
- EPL_TOT- complex Employment Protection Legislation (EPL) index, which includes various aspects of legal regulations protecting employees.

For UNEMPL and EPL_TOT outputs, inverse values were used in order to fulfil the DEA requirement for the direction of preference for output variables (the higher the value is the better).

The restrictions on virtual outputs were determined on the basis of the unbounded CCR model, calculating mean values of weights attached to outputs for extreme efficient DMUs, i.e. such which were fully efficient and had zero slacks. The value *d* from formula (4) was determined on the basis of the model output variability range analysis. Based thereon, the lower and upper bounds on virtual outputs were formulated. By adding the restrictions (5) for all outputs of the model, subsequent calculations were made with the output oriented Assurance Region Global model (ARG) with constant returns to scale, available in the Saitech's software: DEA Solver Learning version 3.0.

RESULTS AND INTERPRETATION

Calculations were made for 28 OECD countries, for which complete data were available. The scores for three examined years, according to the CCR model, are presented in Table 1, in the CCR column.

Full efficiency was achieved by 4 countries in 2000, by 7 countries in 2005 and by 10 countries in 2010. In the examined period only 3 countries maintained full efficiency: Denmark, the Netherlands and the United States; 4 countries recorded a drop in efficiency and all other countries recorded growth. The results of the CCR model cannot, however, constitute the basis for a reliable evaluation

of efficiency because of the occurrence of zero weights. This may be exemplified by fully CCR-efficient countries having only one non-zero weight in five outputs (e.g. Luxemburg in 2000 and Norway in 2010 have a non-zero weight only for the UNEMPL variable, which means that that variable dominated the results while all four remaining variables were not used in the evaluation).

Table 1. Efficiency scores according to CCR and ARG models

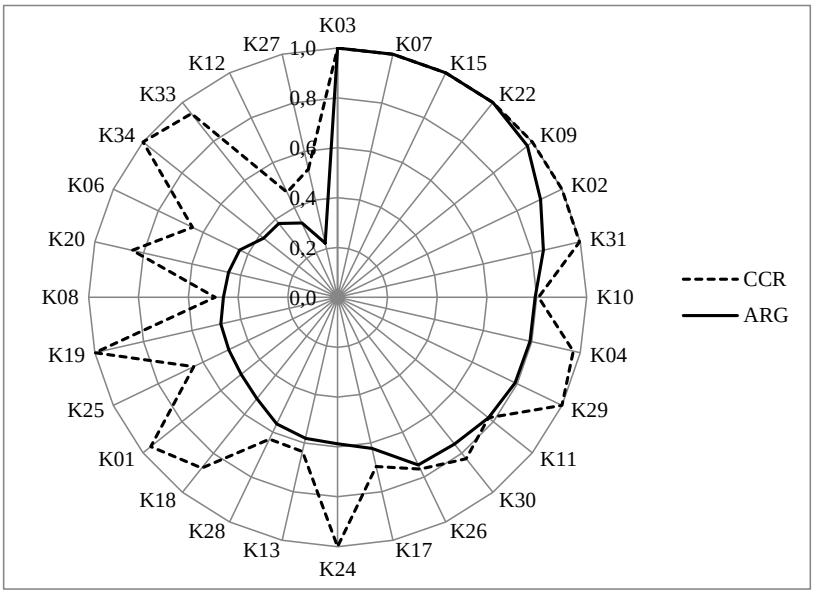
Item	DMU	Coutry	CCR			ARG		
			2000	2005	2010	2000	2005	2010
1	K01	Australia	0,7851	0,9096	0,9604	0,2326	0,3325	0,4954
2	K02	Austria	0,8415	0,8588	1,0000	0,6698	0,6557	0,9051
3	K03	Belgium	0,8901	1,0000	1,0000	0,8206	0,8470	1,0000
4	K04	Canada	0,9482	0,9377	0,9707	0,9126	0,8512	0,7929
5	K06	Czech Republic	0,4312	0,5679	0,6467	0,3522	0,2916	0,4373
6	K07	Denmark	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	K08	Estonia	0,3779	0,5351	0,4922	0,2601	0,1025	0,4589
8	K09	Finland	0,8739	0,8212	1,0000	0,8178	0,7517	0,9751
9	K10	France	0,6841	0,6795	0,8069	0,6689	0,6382	0,7929
10	K11	Germany	0,7983	0,8675	0,7749	0,7741	0,7316	0,7749
11	K12	Greece	0,3854	0,4524	0,4682	0,3509	0,0536	0,3307
12	K13	Hungary	0,5549	0,6240	0,6341	0,5154	0,5178	0,5801
13	K15	Ireland	0,8838	1,0000	1,0000	0,8441	0,8883	1,0000
14	K17	Italy	0,5014	0,5911	0,6956	0,4894	0,5604	0,6213
15	K18	Japan	0,7054	0,9188	0,8761	0,5630	0,5017	0,5195
16	K19	Korea	0,6842	1,0000	1,0000	0,1706	0,1916	0,4813
17	K20	Luxembourg	1,0000	0,8586	0,8463	0,2447	0,6262	0,4500
18	K22	Netherlands	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
19	K24	Norway	0,8842	0,9820	1,0000	0,5099	0,6260	0,5873
20	K25	Poland	0,5277	0,5454	0,6417	0,4842	0,5088	0,4857
21	K26	Portugal	0,7257	0,6175	0,7641	0,6461	0,5837	0,7467
22	K27	Slovak Republic	0,5783	0,5521	0,5277	0,3292	0,4653	0,2229
23	K28	Slovenia	0,4705	0,6323	0,6311	0,3738	0,4529	0,5639
24	K29	Spain	0,5867	0,6239	1,0000	0,5700	0,6023	0,7904
25	K30	Sweden	0,9414	0,9052	0,8276	0,8974	0,8396	0,7533
26	K31	Switzerland	0,7524	0,9630	0,9966	0,6525	0,8461	0,8482
27	K33	United Kingdom	0,8339	1,0000	0,9422	0,5632	0,4421	0,3794
28	K34	United States	1,0000	1,0000	1,0000	0,5064	0,2562	0,3799

Source: own elaboration

Average values of virtual outputs and the ranges of variability on the virtual outputs were calculated as was mentioned above. For instance, the value of parameter *d* for the UNEMPL variable for all examined years equals to 4. The analysis of sensitivity to weight restriction range was also conducted. The narrowing of the variability range resulted in obtaining no optimal solution (lack

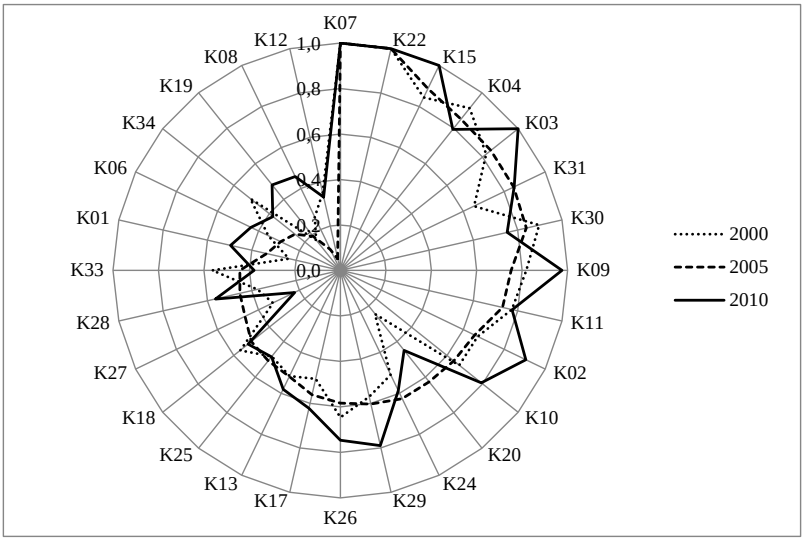
of DMUs with 100% efficiency). The radical widening of the range does not resulted in enlarging the number of fully efficient DMUs.

Figure 1. Comparison of efficiency scores for the year 2010



Source: own elaboration

Figure 2. Efficiency scores according to ARG model for the years 2000, 2005 and 2010



Source: own elaboration

For the so-determined restrictions, calculations were made with the Assurance Region Global model. The results are presented in Table 1, in the ARG column. The application of this model eliminated zero weights and non-zero slacks. This model has greater discrimination power so the number of full efficient DMUs dropped. In all three years, only Denmark and the Netherlands were efficient, just like in the CCR model. Figure 1 contains the comparison of results obtained in the CCR and ARG models for the year 2010.

Figure 2 shows the ARG results presented in the decreasing order according to values of 2005, which enables illustrating changes in subsequent years.

For decision makers, besides the ranking, it is important that the reasons for inefficiency be diagnosed and recommendations for situation improvement activities be formulated. Four countries were selected for the analysis: Greece, Ireland, Poland and Spain. They are characterised by high unemployment rates but very different efficiency scores – the effect of different labour market policies. The values observed for two benchmark countries (Denmark and the Netherlands) were presented, too. This is shown in Table 2 for the 2010 data.

Table 2. Example of efficiency scores and observed and projected output values

DMU	Country	Efficiency score	UNEMPL	Observed				Projected			
				MSS_GDP	ALMP_GDP	LLL_GDP	EPL_TOT	MSS_GDP	ALMP_GDP	LLL_GDP	EPL_TOT
K12	Greece	0,33	12,6	0,71	0,20	0,02	2,97	2,99	0,50	0,46	1,39
K25	Poland	0,49	9,7	0,34	0,65	0,04	2,30	2,30	0,98	0,44	1,60
K29	Spain	0,79	20,1	3,14	0,69	0,20	3,11	1,92	1,16	0,36	1,85
K15	Ireland	1,00	13,7	2,99	0,50	0,46	1,39	-	-	-	-
K07	Denmark	1,00	7,5	1,57	1,49	0,42	1,91	-	-	-	-
K22	Netherlands	1,00	4,5	1,75	1,09	0,13	2,23	-	-	-	-

Source: own elaboration

First four countries have the unemployment rates above the average (8.9% for all OECD countries in 2010) with very different efficiency scores (0.33-1.00). Analysing the observed values of the variables of the model (column „Observed" in Table 2), one may assess labour market policies conducted by these countries. Despite high unemployment rate (13.7%) Ireland has full efficiency - the effect of active policy, which is confirmed by high LLL_GDP and ALMP_GDP values (0.46% of GDP and 0.50% of GDP respectively). Very high value of MSS_GDP (2.99% GDP) – nearly threefold the OECD average, indicates very high social security of the unemployed. Of key importance is the low value of EPL_TOT, which reflects high flexibility of the labour market. Spain and Greece have the EPL_TOT value close to the maximum, which reflects low flexibility of their labour markets. Spain has a very high value of MSS_GDP (high expenditure on

social security) and ALMP_GDP value above the average, however, its LLL_GDP value is relatively low. This reflects a rather passive labour market policy. Greece has very low values of ALMP_GDP and LLL_GDP and a very low value of MSS_GDP, which may reflect the general weakness of the conducted labour market policy. In Poland, the situation is similar and the MSS_GDP value is close to the minimum (extremely low unemployment benefits). In Table 2, the "Projected" column shows how the output values should change for the inefficient countries in order to achieve full efficiency. All countries should make their legislation more flexible, as indicated by the changes required in EPL_TOT. In all cases, the expenditures on training should be increased (LLL_GDP): over 20 times in Greece and over 10 times in Poland. Also, the expenditures on active policy (ALMP_GDP) should be increased, but to a smaller extent. As far as MSS_GDP (social security) is concerned, a circa four-fold increase in Greece and an almost seven-fold increase in Poland are necessary. In Spain, however, the value should be reduced by about 40% (which indicates that the social security is too high).

SUMMARY

The obtained results allowed creating a ranking of evaluated countries. For the inefficient countries, the ways of determining the reasons for inefficiency as well as the way of forming recommendations for actions for these countries to achieve full efficiency were indicated. DEA method has proved well suited for this type of analysis and the results may be useful tool in the decision making process regarding the labour market policy.

REFERENCES

Allen R., Athanassopoulos A., Dyson R.G., Thanassoulis E. (1997) Weights restrictions and value judgements in Data Envelopment Analysis: Evolution, development and future directions, Annals of Operations Research, 73, pp. 13-34.

Angulo-Meza L., Lins M.P.E. (2002) Review of Methods for Increasing Discrimination in Data Envelopment Analysis, Annals of Operations Research, 116, pp. 225-242.

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units, European Journal of Operational Research, 2, pp. 429-444.

Cherchye L., Moesen W., Rogge N., Van Puyenbroeck T. (2009) Constructing a Knowledge Economy Composite Indicator with Imprecise Data, Hub Research Paper – Economics & Management, 2009/16, Katholieke Universiteit Leuven.

Commision of the European Communities (2007) COM(2007) 359 final, Towards Common Principles of Flexicurity: More and better jobs through flexibility and security, Brussels.

Cooper W.W., Ruiz J.L., Sirvent I. (2009) Selecting non-zero weights to evaluate effectiveness of basketball players with DEA, European Journal of Operational Research, 195, pp. 563-574.

Cooper W.W., Seiford L.M., Zhu J. (2011) Handbook on Data Envelopment Analysis, Springer, New York.

Despotis D.K. (2005) Measuring human development via Data Envelopment Analysis: the case of Asia and the Pacific, Omega-International Journal of Management Science, 33, pp. 385-390.

Hoffman A., Giovanni E., Nardo M., Saisana M., Saltelli A., Tarantola S. (2008) Handbook on Constructing Composite Indicators. Methodology and User Guide, OECD Publishing, Paris.

Koopmans T.C. (1951) Analysis of production as an efficient combination of activities, in: T.C. Koopmans (Ed.), Activity Analysis of Production and Allocation, Wiley, New York.

Lovell K.C.A. (1995) Measuring the macroeconomic performance of the Taiwanese economy, International Journal of Production Economics, 39, pp. 165-178.

Lovell K.C.A., Pastor J.T., Turner J.A. (1995) Measuring macroeconomic performance in the OECD: A comparison of European and non-European countries, European Journal of Operational Research, No 87, pp. 507-518.

Manca A.R., Governatori M., Mascherini M. (2010) Towards a set of Composite Indicators on Flexicurity: a Comprehensive Approach, European Commission, Joint Research Centre, European Communities.

Ramón N., Ruiz J.L., Sirvent I. (2010) A multiplier bound approach to assess relative efficiency in DEA without slacks, European Journal of Operational Research, 203, pp. 261-269.

Roll Y., Golany B. (1993) Alternate Methods of Treating Factor Weights in DEA, Omega-International Journal of Management Science, Vol. 21, No. 1, pp. 99-109.

Shen Y., Ruan D., Hermans E., Brijs T., Wets G., Vanhoof K. (2011) Modeling qualitative data in data envelopment analysis for composite indicators, International Journal of System Assurance Engineering and Management, 2 (1), pp. 21-30.

Wilthagen T. (2012) Flexicurity Practices in the EU – Which Way is Up?, Reflect Research Paper 12/002, Tilburg.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 261 – 271

THE ANALYSIS OF THE PHENOMENON OF SPATIAL AUTOCORRELATION OF INDICES OF AGRICULTURAL OUTPUT

Agnieszka Tłuczak
Department of Econometrics and Quantitative Methods
Opole University
e-mail: atluczak@uni.opole.pl

Abstract: The agricultural production depends on natural and economic conditions. Weak environmental conditions could be compensated by using the high technology, which requires capital. The agricultural production should evolve in a similar way in countries with similar natural conditions, i.e. spatial autocorrelation should take place. The aim of this article is to present the spatial autocorrelation of indices of agricultural output. The local and global I Moran's statistics were used and the changes in the dynamics of agricultural production in the EU in 2010-2011 were presented.

Keywords: spatial autocorrelation, indices of agricultural output, the European Union

INTRODUCTION

The agriculture in EU is diversified in terms of the agrarian structure. This is due to mostly to natural conditions and the level of advancement of structural transformation. There are general trends of structural change to reduce the number of farms and to stimulate the growth area of farms. The ability to effectively compete on the community market is a slow processes and require the mobilization mechanisms stimulating at EU level and at national level[1].

On the basis the agricultural census in the European Union we can conclude that the number of farms fell by almost 20% in the years 2003-2010, the area

[1] Babiak J., Zmiany w strukturze rolnictwa krajów Unii Europejskiej, Rocznik Integracji Europejskiej, 2010, nr 4, https://repozytorium.amu.edu.pl/jspui/bitstream/10593/1512/1/babiak.pdf

of land used for agricultural purposes decreased by almost 2% in this same period. The average area of farms has increased from 12 ha in 2003 to 14 ha in 2010. More than 80% of total number of farms is located in Romania, Italy, Poland, Spain, Greece, Hungary and France[2].

Meat and other animal products in the EU-27 represented 156.5 billion € in 2011, i.e. 41% of the total value of farm production and 11% more than in 2010. However, this increase must not be seen as a sign of recovery from the 2009 crisis, since feed costs increased dramatically in 2011, thus further hampering farmers' income. Animal feed is indeed the most important livestock production cost factor and represented in 2011 up to 83% of the farm gate value of poultry. The EU-27 farm animals are fed with 470 billion t of feedstuffs, thereof app. half are roughages produced on farm, 10% are grains produced on farm, 10% are purchased feed materials and 30% are industrial compound feed[3].

The production of meat in the EU-27 increased by 1.4% between 2010 and 2011, thus offsetting the dramatic contraction of production in 2009, due in particular to the drop in EU consumption for all categories of animal products except poultry. Pig production increased despite the high feed costs which continue to rise in 2012. The meat consumption in the EU-27 is stable around 90 kg/capita/year. Poultry meat is the second most consumed meat in the EU-27 with 23.3 kg/capita/year in 2011, far behind pig meat (41.2 kg/capita/year). The EU livestock sector contributes positively to the commercial balance, in particular pork and cheese, with self sufficiency ratio of resp. 109 and 106[4].

In 2011, the EU cereal harvest reached a usable production of 285.7 million tonnes, due to favourable yields, mainly in maize (+8.9%). Animal feed use slightly decreased to 167 million tonnes, resulting in an almost unchanged domestic use of 271.3 million tonnes. In 2011, the real value of EU crop production is estimated to have increased by 7.5% due to higher prices (5.7%) and volumes (1.7%). Prices rose for most crops markedly for cereals (18.3%), oilseeds (15.1%), forage plants (12.8%) and protein crops (11.6%) with the exception of fresh vegetables (-9.7%), olive oil (-1.4%) and flowers (-1.2%). Most products recorded higher volumes, in particular sugar beet (11%), wine, potatoes and fruits while lower volumes were recorded for protein crops (-16.3%)[5].

---

[2] Struktura rolnictwa w Unii Europejskiej, Bieżąca informacja o rolnictwie na świecie Nr 49/2011, http://www.minrol.gov.pl/pol/Informacje-branzowe/Opracowania-publikacje/Informacje-o-rolnictwie-na-swiecie/biezaca-informacja-o-rolnictwie-na-swiecie-nr-49-2011

[3] http://www.fefac.eu/file.pdf?FileID= 39499

[4] ibidem; Global livestock production system, Rome, 2011www.fao.org/docrep014/i2414e/ I 2414e.pdf

[5] Agriculture in the European Union Statistical and economic Information 2011, http://ec.europa.eu/agriculture/statistics/agricultural/2011/pdf/full-report_en.pdf; Global

METHODOLOGY

Since the 1950s, several spatial methods of analysis have been developed and modified to improve our ability to detect and characterize spatial patterns. These stem from several fields of study, having more or less different goals, mathematical approaches and underlying assumptions[6].

In its most general sense, spatial autocorrelation is concerned with the degree to which objects or activities at some place are similar to other objects or activities located nearby. Its existence is reflected in the proposition which Tobler (1970) has referred to as the "*first law of geography: everything is related to everything else, but near things are more related than distant things*". Spatial autocorrelation can be interpreted as a descriptive index, measuring aspects of the way things are distributed in space, but at the same time it can be seen as a causal process, measuring the degree of influence exerted by something over its neighbors'.

The aim of the analysis is to determine the spatial interrelationships and interactions between neighboring objects, in this case the EU countries. Observations made at different locations may not be independent. For example, measurements made at nearby locations may be closer in value than measurements made at locations farther apart. Spatial autocorrelation measures the correlation of a variable with itself through space, it can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another and negative - occurs when dissimilar values occur near one another[7].

The Moran's index and Geary's coefficient summarize the strength of associations between responses as a function of distance, and possibly direction. These indices are usually applied in ecology and geographical sciences. Fortin et al., for example, used these spatial autocorrelation coefficients to compare the capacity of different sampling designs and sample sizes to detect the spatial structure of a sugar-maple tree density data set gathered from a secondary growth forest. Moran's index is one of the oldest indicators of spatial autocorrelation. It is applied to zones or points which have continuous variables associated with their

crop production review, 2011 http://www.usda.gov/oce/weather/pubs/Annual/GlobalCropProduction Review2011.pdf

6 Anselin, L. (1995). Local indicators of spatial autocorrelation – LISA, Geographical Analysis 27, 93 – 115; Cressie, N.A.C. (1993). Statistics for Spatial Data ,Wiley, New York Perry, J.N. (1995). Spatial analysis by distance indices, Journal of Animal Ecology 64, 303 – 314

7 Gunaratna N., Liu Y., Park J., Spatial Autocorrelation, http://www.stat.purdue.edu/~bacraig/ SCS/Spatial%20Correlation%20new.doc; Wang J., Zhang Z., Su B., Zhang L., A case research on economic spatial distribution and differential of agriculture in China, An application to Hunan province based on the data of 1999, 2006 and 2010, Agricultural Sciences, Vol.3, No.8, 996-1006 (2012)

intensities. For any continuous variable, x_i, a mean can be calculated and the deviation of any observation from that mean can also be calculated. The statistic then compares the value of the variable at any one location with the value at all other locations. It is formally defined by

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{1}$$

where: $\bar{x}$ is the mean of the x variable, w_{ij} are the elements of the weight matrix[8], and S_0 is the sum of the elements of the weight matrix: $S_0 = \sum_i \sum_j w_{ij}$.

Moran's index varies between –1.0 and +1.0. When nearby points have similar values, the cross-product is high; and when nearby points have dissimilar values, the cross-product is low. In other words, an I value which is high indicates more spatial autocorrelation than an I which is low[9]. In the absence of autocorrelation and regardless of the specified weight matrix, the expectation of Moran's I statistic is $-1/(n-1)$, which tends to zero as the sample size increases. For a row-standardized spatial weight matrix, the normalizing factor S_0 equals n (since each row sums to 1), and the statistic simplifies to a ratio of a spatial cross product to a variance. A Moran's I coefficient larger than $-1/(n-1)$ indicates positive spatial autocorrelation, and a Moran's I less than $-1/(n-1)$ indicates negative spatial autocorrelation[10].

Geary's C statistic[11] is based on the deviations in responses of each observation with one another:

$$C = \frac{n-1}{2S_0} \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \tag{2}$$

---

[8] The weight matrix can be specified in many ways: (1) the weight for any two different locations is a constant, (2) all observations within a specified distance have a fixed weight, (3) K nearest neighbors have a fixed weight, and all others are zero, (4) weight is proportional to inverse distance, inverse distance squared, or inverse distance up to a specified distance.

[9] Silva E. Da, Silva A., De Paiva A., Nunes R, Diagnosis of lung nodule using Moran's index and Geary's coefficient in computerized tomography images, Pattern Analysis and Applications January 2008, Vol. 11, Issue 1, pp 89-99

[10] Gunaratna N., Liu. Y., Park J., op.cit.; Plant R. E., Spatial Data Analysis in Ecology and Agriculture Using R, CRC Press, 2012

[11] Geary R.C., The Contiguity Ratio and Statistical Mapping, The Incorporated Statistician, 1954, 5 (3), pp 115–114

The values of C typically vary between 0 and 2. The theoretical value of C is 1, that indicates that values of one zone are spatially unrelated to the values of any other zone. Values less than 1 (between 0 and 1) indicate positive spatial autocorrelation while values greater than 1 indicate negative spatial autocorrelation.

This coefficient does not provide the same information of spatial autocorrelation given by Moran's index, because it emphasizes the differences in values between pairs of observations comparisons rather than the covariation between the pairs. So the Moran's index gives a more global indicator whereas the Gearys coefficient is more sensitive to differences in small neighborhoods[12].

Moran's I is a more global measurement and sensitive to extreme values of, whereas Geary's C is more sensitive to differences in small neighborhoods. In general, Moran's I and Geary's C result in similar conclusions. However, Moran's I is preferred in most cases since Cliff and Ord (1975, 1981) have shown that Moran's I is consistently more powerful than Geary's C[13].

In addition to the global statistics the local statistics of spatial autocorrelation were calculated. It can be assumed that the interpretation of the local statistics are similar to the global statistics. If we get a negative value for the local Moran's statistics, we can conclude that the i-th country is surrounded by countries (neighbors) which are different from each other due to the test feature. In the case of the positive talk about similar countries (neighbors) in the i-th country setting. Local statistics are called LISA statistics. Local Moran statistic is given by formula:

$$I(w) = \frac{(x_i - \bar{x})\sum_{i=j}^{n} w_{ij}(x_i - \bar{x})}{\sum_{i=j}^{n}(x_i - \bar{x})^2} \tag{3}$$

RESULTS AND CONCLUSIONS

The study included 27 member states of the European Union, statistical data were taken from Eurostat databases and the World Bank. The following variables were taken under consideration:

x_1 – indices of agricultural crop output at producer prices – 2010;
x_2 - indices of agricultural crop output at producer prices – 2011;
x_3 - indices of agricultural animal output at producer prices – 2010;
x_4 – indices of agricultural animal output at producer prices – 2011.

[12] Silva E. Da, Silva A., De Paiva A., Nunes R, op.cit.
[13] Gunaratna N., Liu Y., Park J., Spatial Autocorrelation, http://www.stat.purdue.edu/~bacraig/ SCS/Spatial%20Correlation%20new.doc

Table 1 shows the statistical characteristics of the variables.

Table 1 Statistical characteristics of the variables

variables	x_1	x_2	x_3	x_4
mean	6731	7334	5076	5574
standard deviation	9652,6	10169,3	6406,6	7070,5
coefficient of variation	143,4%	138,7%	126,2%	126,9%
min	45	50	68	69
max	37668	38839	22452	24720

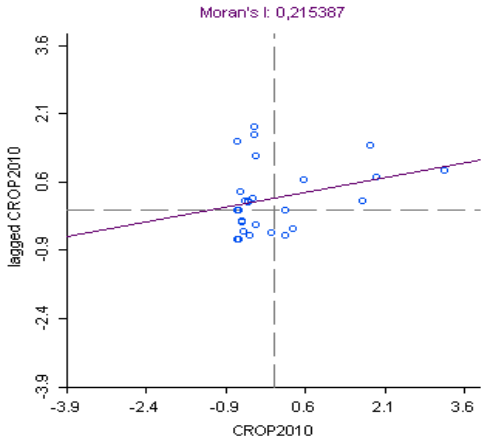
Source: own calculation based on EUROSTAT data

Analyzing the results contained in Table 1, it is clear that variables taken under consideration diversify the area in terms of growth of agricultural production (value of the coefficients of variation exceeds the value of 100%). The data shows an increasing trend of average growth of indices of agriculture animal and crop output at producer prices in the EU member states.

The study of spatial autocorrelation of indices of agricultural animal and crop output at producer prices have been carried out under the assumption of contact matrix W. The calculated value of the global I Moran's statistics indicates that in the adopted study period a moderate spatial autocorrelation can be observed.

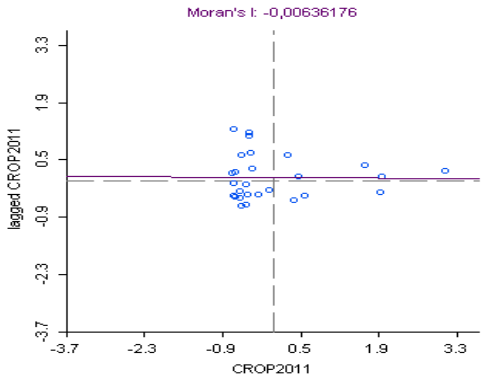
It is either positive, that is, there is a tendency to focus on individuals with similar levels of indices of agriculture animal and crop output at producer prices. All obtained values of I Moran's statistics are statistically significant (p-value <0.05) (Fig. 1-4).

Figure 1. Moran's I scatterplot for the variable x_1



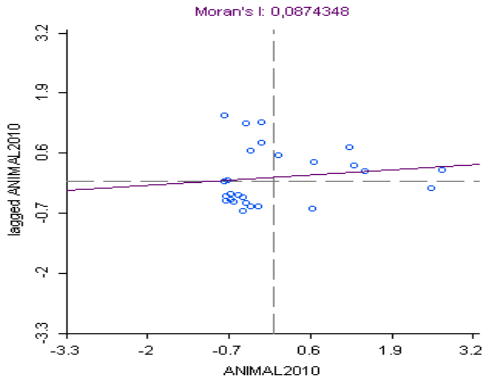
Source: calculations in the GeoDa based on EUROSTAT data

Figure 2. Moran's I scatterplot for the variable x_2



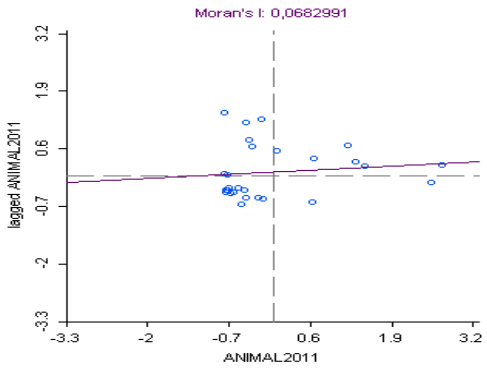
Source: calculations in the GeoDa based on EUROSTAT data

Figure 3. Moran's I scatterplot for the variable x_3



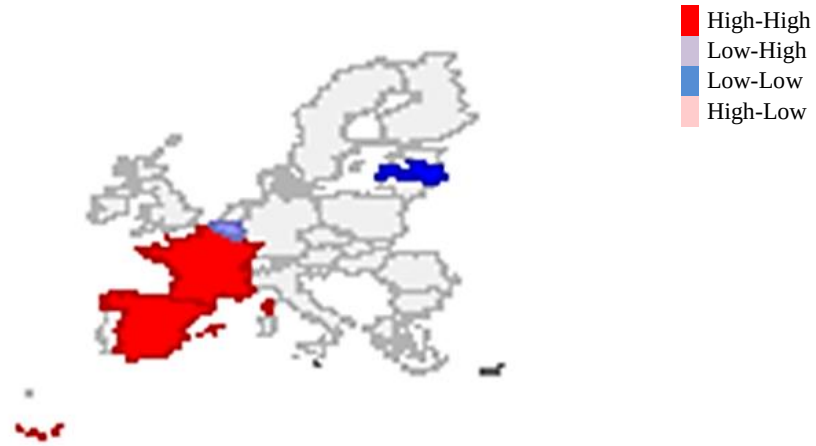
Source: calculations in the GeoDa based on EUROSTAT data

Figure 4. Moran's I scatterplot for the variable x_4



Source: calculations in the GeoDa based on EUROSTAT data

Figure 5. Map of affiliations of objects to quarters of Moran scatterplot (variable x_1)

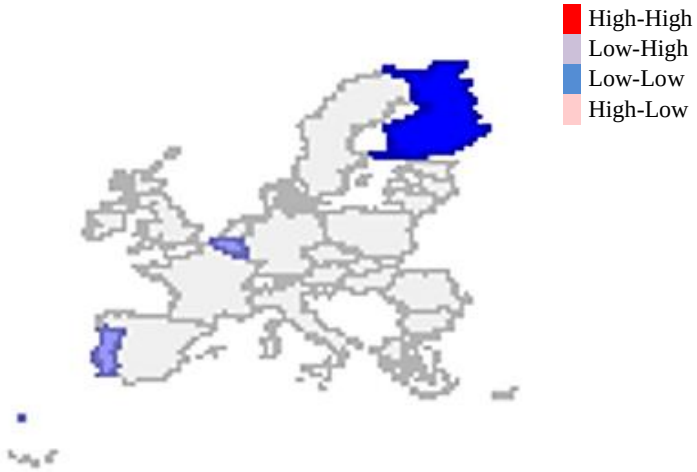


Source: calculations in the GeoDa based on EUROSTAT data

Figure 6. Map of affiliations of objects to quarters of Moran scatterplot (variable x_2)

Source: calculations in the GeoDa based on EUROSTAT data

Figure 7. Map of affiliations of objects to quarters of Moran scatterplot (variable x_3)



Source: calculations in the GeoDa based on EUROSTAT data

Figure 8. Map of affiliations of objects to quarters of Moran scatterplot (variable x_4)

Source: calculations in the GeoDa based on EUROSTAT data

The four quadrants in the Moran scatter plot provide a classification of four types of spatial autocorrelation. Areas that are significant are labelled with these categories in the "High-High/Low-Low" dataset produced in the Moran analysis, and are colored in the Moran scatter plot and Local Moran maps as well.

The map contains information on only those locations that have a significant Local Moran statistic. While every region in the dataset will be represented in the Moran scatterplot, only those with Local Moran statistic p-values <0.05 are significant. Any island locations are considered missing values because they have no adjacent neighbors.

Figures 5-8 shows that the space can be divided into clusters with similar values of local I Moran's statistics. Clustering of countries with similar I Moran's statistics indicates the existence of spatial autocorrelation. The direction of the relationship was changing in the analyzed period, which leads to the conclusion about the need for in-depth research and an explanation of the reasons for this phenomenon.

SUMMARY

The I Moran's and Gettis statistics indicate the type and strength of spatial dependency, which allows to identify the structures and changes. On the basis of the positive I Moran's statistics (statistically significant) it can be concluded the positive spatial autocorrelation of indices of agricultural crop and animal output at producer prices in 2010 and 2011.

The neighboring countries in the European Union were similar in terms of crop and animal agricultural output at producer prices. At the same time it should be noted that there the need for further studies decomposition crop and animal agricultural output at producer prices in the European Union.

REFERENCES

Agriculture in the European Union Statistical and Economic Information 2011, http://ec.europa.eu/agriculture/statistics/agricultural/2011/pdf/full-report_en.pdf

Arfa N., Rodriguez C., Daniel K., Shonkwiler J.S., Spatial Structure of Agricultural Production in France: the Role of the CAP, http://www.oecd.org/agriculture/44832556.pdf

Anselin, L. (1995) Local indicators of spatial autocorrelation – LISA, Geographical Analysis 27, pp. 93 – 115;

Babiak J. (2010) Zmiany w strukturze rolnictwa krajów Unii Europejskiej, Rocznik Integracji Europejskiej, nr 4 https://repozytorium.amu.edu.pl/jspui/bitstream/10593/1512 /1/babiak.pdf

Cressie, N.A.C. (1993). Statistics for Spatial Data ,Wiley, New York.

Geary, R. C., (1954) The Contiguity Ratio and Statistical Mapping, The Incorporated Statistician, 5 (3): 115–14.

Global crop production review, 2011 http://www.usda.gov/oce/weather/pubs/Annual/GlobalCropProduction Review2011.pdf

Global livestock production system (2011) Rome, www.fao.org/docrep014/i2414e/I 2414e.pdf

Gunaratna N., Liu Y., Park J., Spatial Autocorrelation, http://www.stat.purdue.edu/~bacraig/ SCS/Spatial%20Correlation%20new.doc

http://www.fefac.eu/file.pdf?FileID= 39499

Perry, J.N. (1995) Spatial analysis by distance indices, Journal of Animal Ecology 64, pp. 303 – 314.

Plant R. E. (2012) Spatial Data Analysis in Ecology and Agriculture Using R, CRC Press.

Silva E. Da, Silva A., De Paiva A., Nunes R. (2008) Diagnosis of lung nodule using Moran's index and Geary's coefficient in computerized tomography images, Pattern Analysis and Applications January 2008, Vol. 11, Issue 1, pp 89-99.

Struktura rolnictwa w Unii Europejskiej, Bieżąca informacja o rolnictwie na świecie Nr 49/2011, http://www.minrol.gov.pl/pol/Informacje-branzowe/Opracowania-publikacje/Informacje-o-rolnictwie-na-swiecie/biezaca-informacja-o-rolnictwie-na-swiecie-Nr-49-2011

Wang J., Zhang Z., Su B., Zhang L. (2012) A case research on economic spatial distribution and differential of agriculture in China, An application to Hunan province based on the data of 1999, 2006 and 2010, Agricultural Sciences, Vol.3, No.8, pp. 996-1006.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 272 – 278

GENDER PAY GAP IN POLAND – BLINDER-OAXACA DECOMPOSITION

Olga Zajkowska
Department of Applied Mathematics
Warsaw University of Life Sciences – SGGW
e-mail: o.zajkowska@gmail.com

Abstract: Providing equal opportunities to both men and women has become an important policy issue to most developed countries' governments. One of it's main dimensions is gender equality on the labor market. That includes: market participation, employment and wages. In a workplace equal treatment can be observed through wages.
The aim of this paper is to investigate determinants of wage differentials between women and men of equal productivity on Polish labor market. Using Blinder-Oaxaca decomposition allows to: measure wage differences, return on individual characteristics and gender pay discrimination.

Keywords: Blinder-Oaxaca Decomposition, gender pay gap, Polish labor market, gender wage differential, outcome differential

INTRODUCTION

In the last 20 years structure of Polish labor market changed substantially. Labor supply started to be evaluated by market and given a price (wage). Heterogeneous individuals became rewarded with wage for their productivity dependent on education level and other individual characteristics. Soon it was observed by researchers, that like in most labor markets over the world, female earnings on average differ from male earnings [Polachek 2009]. Broad survey study of reasons discussed in the literature on gender pay gap is described in detail by Weichselbaumer and Winter–Ebner [2005]. Assuming that wage structure reflects productivity and value of market factors, the question which follows is whether this differences can be explained with objective factors or are an effect of uneven treatment with respect to gender.

The aim of this study is to verify the hypothesis that there exists a gender pay gap between full time employed men and women and that these differences are not entirely driven by productivity characteristics.

DATA

Data used in this study is based on International Social Survey Programme 2002: Family and Changing Gender Roles III, which is a representative study covering several 'household-related' topics. For the purpose of this study, sample is limited to full-time employed individuals from Poland aged 18-60. It consists of 219 males and 217 females (n=436). Individuals working part-time were excluded from the sample to avoid cases where hourly wage is affected by smaller total of hours worked, that is individuals with additional constraints in the labor supply function or strong preferences of leisure over consumption. Variable describing wage is constructed on the basis of question: "Taking into consideration the last 12 months, please tell me what was your average monthly income/ earnings from your job or business after taxes in PLN?"
Potential experience is defined as difference between age and years of schooling. Higher or much higher income are dummy variables based on self-reported comparison of own and spouse's earnings.

Table 1. Descriptive statistics of males in the sample (n=219)

Variable	Number of responses	Mean	Standard Deviation	Median
wage (net) [PLN]	219	1388,24	89,3429	1100
master degree [binary]	219	0,1142	0,3187	0
secondary education [binary]	219	0,3425	0,4756	0
years of education [years]	219	11,4429	2,5932	10
potential experience [years]	219	20,6849	9,5344	21
trade union [binary]	219	0,1370	0,3446	0
household production [hours]	167	13,9102	16,7511	10
age [years]	219	39,8995	9,0877	41
much higher income than spouse [binary]	219	0,3927	0,4895	0
higher income than spouse [binary]	219	0,5616	0,4973	1
more household chores than spouse [binary]	219	0,1324	0,3397	0
private employer [binary]	219	0,4292	0,4961	0

Source: own calculations

Table 2. Descriptive statistics of females in the sample (n=217)

Variable	Number of responses	Mean	Standard Deviation	Median
wage (net) [PLN]	217	1127,48	57,6742	1000
master degree [binary]	217	0,2350	0,4250	0,0000
secondary education [binary]	217	0,4793	0,5007	0,0000
years of education [years]	217	12,6590	2,9113	12,0000
potential experience [years]	217	20,8940	10,2231	22,0000
trade union [binary]	217	0,2120	0,4097	0,0000
household production [hours]	149	18,2013	13,4275	15,0000
age [years]	217	41,0691	9,5801	41,0000
much higher income than spouse [binary]	217	0,1475	0,3554	0,0000
higher income than spouse [binary]	217	0,2350	0,4250	0,0000
more household chores than spouse [binary]	217	0,4147	0,4938	0,0000
private employer [binary]	217	0,2811	0,4506	0,0000

Source: own calculations

As can be seen from the tables above full-time working males earn on average substantially more than full-time working females (1388,24 and 1127,48 PLN of net wage). Although women are ones better educated (larger fraction in subpopulation of both master degree and secondary education attained, on the other hand average number of years of completed education differ less than 1,5 year). Also more females are members of trade unions. Both of these factors should work in favour of female work valuation on the labor market. On the other hand, self-reported involvement in household production measured in hours spent on domestic chores is substantially higher for females (on average 18,2 hours per week) than males (13,91 hours per week). Also men are more often employed in private than public sector. What is more only 1 in 4 women reports earnings higher than earnings of their partners. This is consistent with intuition of higher involvement in intra household activities is correlated with relative smaller wage of a spouse. But this is also contradiction with human capital wage explanation. Therefore in the further part of this research differences in wages are decomposed.

METHODOLOGY

In labor economics fundamental for analysis of differences in wages is Mincer wage equation is defined as follows:

$$Y = \alpha_0 + \rho_s s + \beta_0 x + \beta_1 x^2 + \varepsilon \quad (1)$$

where wage logarithm (Y) is a stochastic function of schooling s (measured with completed years of education), experience x and experience squared x^2. It uses formulated in the 70. theory claiming that human capital influencing productivity is reflected in wages [Mincer 1974]. If required data on labour market experience is unavailable in the dataset, potential experience (years of education and age of school enrollment are subtracted from current age of individual) is estimated and used as proxy. If differences in rewarding human capital factors of two groups or any other kind of discrimination is expected, two Mincer equation are estimated. To analyze the underlying reasons of expected differences method proposed by Oaxaca (Oaxaca 1973, 1977) is applied. Than expected values (means) are compared:

$$R = E(Y_{male}) - E(Y_{female}) \tag{2}$$

If the difference is statistically significant, the decomposition is performed. The difference is assumed to consist of 3 factors defined as follows:

$$R = E + C + I \tag{3}$$

Endowments factor (E), that is what would be the mean increase (or decrease) if the discriminated group (females) had the same characteristic as favoured group (males):

$$E = [E(Y_{male}) - E(Y_{female})]'\beta_{female} \tag{4}$$

Coefficient factor (C) which quantifies the change in discriminated groups' (females') wages if favoured group (male) coefficients were applied to them (that is if the labor market rewarded them as men are rewarded):

$$C = E(X_{female})'(\beta_{male} - \beta_{female}) \tag{5}$$

Interaction factor (I) measures simultaneous effect of previous two factors:

$$I = [E(X_{male}) - E(X_{female})]'(\beta_{male} - \beta_{female}) \tag{6}$$

Equivalently β* -parameter of Mincer function over population can be introduced. It describes 'true value' of productivity factors. It is based on assumption that discrimination can be both negative (against one group) or positive (in favour of other group). After some simple algebra it can be shown decomposition into explained and unexplained part of the wage difference:

$$R = P + U \tag{7}$$

$$P = [E(X_{male}) - E(X_{female})]'\beta^* \tag{8}$$

$$U = E(X_{male})'(\beta_{male} - \beta^*) + E(X_{female})'(\beta^* - \beta_{female}) \tag{9}$$

where the first part of the formula (P) describes productivity differences between two analyzed groups, second part (U) describes: average „discrimination in favour" of group in relatively better situation ($E(X_{male})'(\beta_{male} - \beta^*)$), that is males in this study and average „discrimination against" group in relatively worse situation ($E(X_{female})'(\beta^* - \beta_{female})$) [Oaxaca and Ransom 1994].

RESULTS

Results of group-specific regressions are shown below (Table 3 for males and Table 4 for females). Initially model for group 1 (males), n=219:

Quality of the model can be described with following characteristics:

Table 3. Estimates of wage equation for males (n=219, R-squared = 0.1966)

ln total wage	Coefficient	Standard Error	t	P>|z|	95% Conf. Interval	
years of education	0,1292	0,0184	7,0200	0,00	0,0929	0,1655
potential experience	0,0432	0,0202	2,1500	0,033	0,0035	0,0830
potential experience squared	-0,0010	0,0005	-2,0900	0,038	-0,0020	-0,0001
constant	5,1403	0,3296	15,5900	0,00	4,4905	5,7900

Source: own calculations

Further model for group 2 (females) was estimated, n=217:

Table 3. Estimates of wage equation for females (n=217, R-squared = 0.3118)

ln total wage	Coefficient	Standard Error	t	P>|z|	95% Conf. Interval	
years of education	0,1125	0,0123	9,13	0,000	0,0882	0,1368
potential experience	0,0301	0,0124	2,43	0,016	0,0057	0,0546
potential experience squared	-0,0006	0,0003	-1,19	0,048	-0,0012	-4.37e-06
constant	5,1364	0,2128	24,13	0,000	4,7169	5,5560

Source: own calculations

As can be seen from group-specific regressions, men are rewarded higher by the labor market than women. They have higher return of each additional year of education- 0,13 versus 0,11 coefficient respectively. Probable reason might be that female choose less remunerate educational paths (so called valuative discrimination). This is also consistent with crowding hypothesis. Notice: the coefficients describe differences in logarithms of net wages and don't have exact economic interpretation, they just stress differences. Also potential experience on the labor market is paid better for men than for women (0,043 and 0,30). This last

phenomenon might be explained by existence maternity leave which is impossible to capture within most datasets- women are not asked how much time they spent out of the labor force due to children (but existence of children in the household turned out to be statistically insignificant variable). Long term unemployment is an omitted factor in this analysis.

Finally results of Blinder-Oaxaca decomposition is presented (Table 5).

Table 5. Blinder-Oaxaca decomposition estimates (n=436)

ln total wage	Coef.	Std. Err	z	P>|z|	95% Conf. Interval	
Overall:						
Group 1	6,9875	0,0486	143,71	0,000	6,8922	7,0828
Group 2	6,8560	0,0387	177,24	0,000	6,7839	6,9356
Difference	0,1278	0,0621	2,06	0,040	0,0060	0,2496
Endowments	-0,1295	0,0326	-3,98	0,000	-0,1934	-0,0657
Coefficients	0,2715	0,0588	4,62	0,000	0,1562	0,3866
Explained	-0,1344	0,0339	-3,97	0,000	-0,2008	-0,0680
Unexplained	0,2622	0,0551	4,75	0,000	0,1541	0,3703

Source: own calculations

Blinder-Oaxaca decomposition shows that in analyzed sample the difference in wages between groups is statistically significant. That implies uneven situation of males and females on the labor market. The endowment factor implies that if women had the same productivity characteristics (in Mincer sense), their wage would be lower- the wage logarithm would be on average 0,13 lower. Further, if the labor market rewarded women like it rewards men, their wages would be on average substantially higher- the increase in logarithm of wage would be 0,27. But the most warning estimates are those describing explained and unexplained part of the wage gap- the unexplained coefficient is in absolute values twice as high as coefficient of explained part of the equation.

CONCLUSIONS

Blinder-Oaxaca decomposition performed on full-time working polish individuals aged 18-60 shows that there is statistically significant difference in wages per month of men and women. Productivity factors are rewarded differently- men have higher return to both schooling and potential labor market experience. But even if standard Mincer factors influencing wages are controlled, there still exists unexplained gender pay gap which can be explained as discrimination, cultural stereotypes, different family roles or preferences. The decomposition of wage differences into explained show that gender pay gap is important but

unsolved problem. Therefore more detailed researcher is needed in order to obtain the reasons why does the market undervalue female labor in comparison to men and to address the policies adequately to reasons of inequalities, not just the consequences.

REFERENCES

Jann B. (2008) OAXACA: Stata module to compute the Blinder-Oaxaca decomposition, Statistical Software Components S456936, Boston College Department of Economics, revised 25 Aug 2011.

Jann B. (2008) A Stata implementation of the Blinder-Oaxaca decomposition, ETH Zurich Sociology Working Papers 5, ETH Zurich, Chair of Sociology, revised 14 May 2008.

Blinder A. S. (1973) Wage Discrimination: Reduced Form and Structural Estimates, Journal of Human Resources 8(4): 436-55.

Elder T. E., Goddeeris J. H., Haider S. J. (2009) Unexplained Gaps and Oaxaca-Blinder Decompositions IZA Discussion Paper No. 4159

International Social Survey Programme 2002: Family and Changing Gender Roles III (ISSP 2002), B. Cichomski, Affiliation: Institut for Social Studies, Warsaw, Poland

Hassenboehmer S., Sinning M. (2010) Distributional Changes in the Gender Wage Gap, IZA Discussion Paper No. 5303

Mincer J. (1974) Schooling, Experience, and Earnings, New York, NBER Press.

Oaxaca R. (1973) Male-female wage differences in urban labor markets, International Economic Review, Vol. 14, No. 3

Oaxaca R. (1977) The Persistence of Male-Female Earnings Differentials, in: Juster T. (ed.) Distribution of Economic Well-Being, NBER Press

Oaxaca R. L. & Ransom, M. R. (1994) On discrimination and the decomposition of wage differentials, Journal of Econometrics, Elsevier, vol. 61(1), pp. 5-21, March

Polachek S. W. and Xiang J. (2009) The Gender Pay Gap across Countries: A Human Capital Approach, SOEP papers on Multidisciplinary Panel Data Research 227, DIW Berlin, The German Socio-Economic Panel (SOEP)

Weichselbaumer, D. and Winter-Ebmer R. (2005) A Meta-Analysis of the International Gender Wage Gap, Journal of Economic Surveys, Wiley Blackwell, vol. 19(3), pp. 479-511, 07.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 279 – 288

INVESTORS' PREFERENCES AND PAYOFFS FROM STRUCTURED PRODUCTS

Wojciech Zatoń
Centre for Central Banking and Financial Intermediation
University of Lodz
e-mail: wojciech.zaton@uni.lodz.pl

Abstract: The attractiveness of structured products is mainly due to the fact that the diversity of their structures corresponds to the specific preferences of different groups of investors. To determine the appropriate investment product for an investor it is necessary to identify investor preferences, attitude toward risk. The paper contains an analysis of investors preferences in relation to the payoffs from the typical structured products. Investors' preferences are based both on the rational choice theory and the prospect theory.

Keywords: structured financial products, utility function, prospect theory

INTRODUCTION

Structured financial products have gained great popularity since the last decade of the 20th century. At the peak of the last global bull market of 2007, the estimated value of the investment in structured products exceeded 1000 billion EUR. The total volume of sales of these products in Poland in 2011 amounted to over 10 billion PLN. The attractiveness of structured products is mainly due to the fact that the diversity of their structures corresponds to the specific preferences of different investors. To determine the appropriate investment product for an investor it is necessary to identify the payoff from the product and the investor preferences, attitude toward risk and loss. The aim of this paper is to show that different investors' preferences may determine the choices of particular structured products and also indicate whether to invest in a given structured product or in an asset underlying this product.

After outlining theoretical background referring to the structured products, both the rational choice theory and the prospect theory, example structured

products are presented, expected utilities and certainty equivalents of their payoffs are then computed and finally concluding remarks are presented.

THEORETICAL BACKGROUND

Structured products are synthetic investment instruments[1]. They are tailor-made investments created to meet specific needs that cannot be met from the standardized financial instruments available in the markets generally. They are composed of several elements or component parts, each providing a specific exposure or protection for the investor. In general a structured product is constructed by combining a bond (typically a zero-coupon bond) and financial options on the underlying asset (single or many). There may be a broad scope of these underlying assets, the same as list of assets that options may be written for - particular stock, stock indexes, commodities, exchange rates etc. In most cases, these underlying assets and the derivatives markets which offer exposure to these assets are not easily accessible to individual investors.

The combination of a bond and a financial option results in the capital protection secured by the bond and possible return outcomes provided by the option. The investment is structured in the sense that the investor knows the package of underlying assets included in the product and the method of calculating gains and risk from investment in the specific product. This enables investors to establish a payoff from a structured product which shows a risk profile of such product based on different forms of capital protection: full, partial or conditional. The typical product has a payoff based on the performance of an underlying asset. The investor may benefit from its good performance and simultaneously may receive a minimum guaranteed level of the capital protection in the case of poor underlying returns. The cost of the guarantee is generally covered by modifying the payoff usually capping it at some level or reducing participation in underlying returns.

Structured products fall into three broad categories[2]:

- Capital Protection
- Yield Enhancement
- Participation

It is generally considered that capital protection products carry the least risk of the three and participation products carry the most risk.

Capital protected products offer full capital protection of the initial investment and others offer partial or conditional capital protection. Conditional capital protection is commonly linked to the performance of the underlying asset.

[1] For a broad review of structured products see Blumke A. (2011) Jak inwestować w produkty strukturyzowane, Wolters Kluwer Polska, Warszawa

[2] The terms used may depend on the issuer.

If specific conditions are met, for example if the price of the underlying asset falls below an agreed threshold during the investment period, the capital protection disappears and the investor may incur a loss at maturity.

Yield enhancement products offer no protection for the initial investment. The purpose is to generate a fixed return that is higher than a bond. Yield enhancement products may be desirable to investors when the market for the underlying is rather stable. The yield potential can be above this market, however the capital may be at risk

Participation products are very closely tied to the underlying assets. They offer leveraged upside potential or downside protection with no or only partial and conditional capital protection. Usually no coupon is paid on these instruments and a return at maturity is calculated by multiplying the performance of the underlying asset by a fixed percentage, called the participation rate.

Structured products can have features such as a barrier or multiple barriers. A barrier causes a certain feature of an instrument to come into effect once a predetermined condition is met. This is usually based on the price movement of the underlying asset.

Structured products have a strong consumer appeal since they combine upside potential with downside protection in a single instrument. As there is a large variety of structured products offered by financial institutions, virtually every investor may find the one which is suitable to her or his needs and preferences.

Utility functions enable to measure investor's preferences for wealth and the amount of risk they are willing to undertake in the hope of attaining greater wealth. In general a utility function widely used in economics as representative for a rational investor is a twice-differentiable function of wealth $U(W)$ which has the properties of non-satiation $(U'(W) > 0)$ and risk aversion $(U''(W) < 0)$. The principle of expected utility maximization states that a rational investor, when faced with a choice among a set of competing feasible investment alternatives, acts to select an investment which maximizes his expected utility of wealth.

In the paper four basic and popular utility functions are considered and a function which reflects behavioral finance theory which has an S-shape around a point of reference (the S-shaped utility function).

The important property of a utility function is an assumption about how the investor's preferences change with a change in wealth. The Arrow-Pratt measures of absolute (ARA) and relative (RRA) risk aversion may be applied to examine this behaviour.

$$ARA(W) = \frac{-U'(W)}{U(W)} \tag{1}$$

$$RRA(W) = W \cdot \frac{-U'(W)}{U(W)} = W \cdot ARA(W) \tag{2}$$

The list of considered utility functions with their properties is shown in Table 1.

Table 1. Functional forms of utility functions and their properties

Name	Form	Type of ARA	Type of RRA
Logarithmic	$\ln(W)$	Decreasing (DARA)	Constant (CRRA)
Power	W^{α}	Decreasing (DARA)	Constant (CRRA)
Exponential	$-e^{-\alpha \cdot W}$	Constant (CARA)	Decreasing (DRRA)
Quadratic	$W-\alpha \cdot W^2$	Increasing (IARA)	Increasing (IRRA)

Source: own elaboration

The applications of utility functions with CRRA are frequent supported by the classical studies by Friend and Blume (1975) and Pindyck (1988). Some studies also recognize the potential of DRRA utility function [Ogaki and Zhang 2001].

A number of studies indicate that behavioral factors such as loss aversion affect the investment decision of the individual investor. Prospect theory (PT) presented by Kahneman and Tversky (1979) and further extended into cumulative prospect theory (CPT) in 1992 [Tversky, Kahneman 1992] examines how people make decisions involving risk. Many later studies (see for example [Fischer 2007], [Doebeli and Vanini 2010], [Hens and Bachmann 2010]) show that the behavioral approach is useful in analyzing investor preferences with respect to investment in structured products.

The prospect theory assumes that the investor does not focus on absolute levels of final wealth W but rather view the gains and losses measured as from a certain reference point ΔW. The utility of gains and losses are evaluated separately. In the region of gains, the utility function (originally in PT called a value function) is concave whereas in the region of losses it is convex[3]. In other words, investors are risk averse over gains and risk seeking over losses. Moreover, an investor's value function is steeper for losses than for gains pointing that keeping the magnitude equal, losses hurt more than gains please.

The functional form of value function in CPT proposed by Tversky and Kahneman (1992) and used later in this paper is expressed as[4]:

$$v(\Delta W)=\begin{cases}(\Delta W)^{\alpha} & \text{if } \Delta W \geq 0\\ -\lambda \cdot (-\Delta W)^{\beta} & \text{if } \Delta W < 0\end{cases} \tag{3}$$

[3] This kind of function is also called the S-shaped utility function

[4] For the review of other functional forms for the value function as well as for decision weights described in the next paragraph proposed by other researchers see Glimcher P., Camerer C., Fehr E., Poldrack R. (2009) Neuroeconomics: Decision Making and the Brain, Elsevier Academic Press, London, 145-170

The parameters obtained by Tversky and Kahneman (1992) in their study were the following: $\alpha = 0{,}88,\ \beta = 0{,}88, \lambda = 2{,}25$ and λ is a coefficient of loss aversion.

In CPT, to obtain the value of a prospect (reported as CPT utility in the next section of the paper), the value of an outcome $v(\Delta W)$ is weighted not by its probability (as is the case in the utility theory) but by the decision weight instead. These weights reflect the fact that moderate to high probabilities are underweighted (which intensifies risk aversion for gains and risk-seeking for losses in this range) and low probabilities are overweighted (which reverses the attitude towards risk in this range of probabilities and leads to risk seeking for gains and risk aversion for losses).

The functional forms of decision weights with their originally estimated parameters proposed by Tversky and Kahneman (1992) and employed in the paper are as follows:

$$w^{+}(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{\frac{1}{\gamma}}} \qquad w^{-}(p) = \frac{p^{\delta}}{(p^{\delta} + (1-p)^{\delta})^{\frac{1}{\delta}}} \qquad (4)$$

$$\gamma = 0{,}61,\ \delta = 0{,}69$$

DATA ANALYSIS AND RESULTS

Two structured products issued on Polish financial market and based on the Warsaw Stock Exchange Index WIG20 were analyzed. One of them is an example of a capital protected product with a barrier while the second is representative for the class of participation products.

Zyskuj z WIG20 - Strategia 100 (Profit with WIG20 – strategy 100) product

This product is a kind of a capital guaranteed product with a knock-out barrier. This structured product is an example of a Shark note. It is built using a zero-coupon bond plus an up-and-out call option. Such option has a barrier embedded, which, if breached, causes the option to "die" All participation accumulated once the barrier is breached is lost. Zyskuj z WIG20 product, like most Shark notes includes a rebate (6,5% in this case), which is a return for the investor if the barrier has been breached. The short description of this product is presented in Table 2 below[5].

[5] Full termsheet for this product is available at http://www.analizy.pl/fundusze/produkty-strukturyzowane/produkt/PSALR009/Zyskuj-z-WIG20--strategia-100.html . Accessed: 20/06/2013

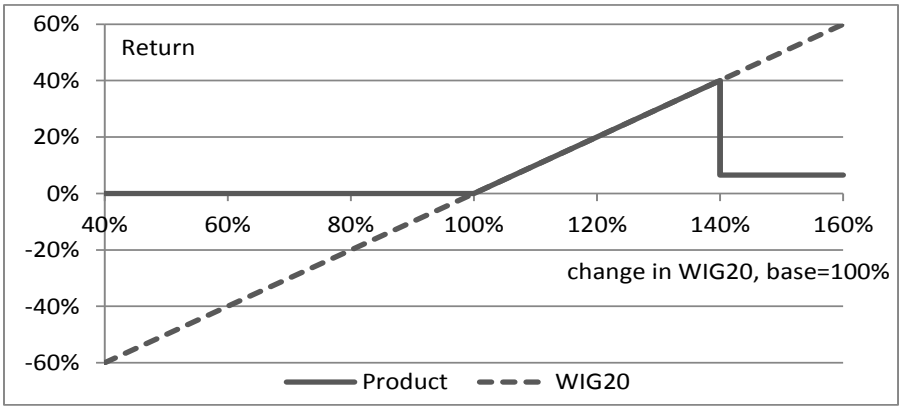
Table 2. Basic parameters for Zyskuj z WIG20 – Strategia 100 product

Issuer	Alior Bank SA
Currency	PLN
Unit certificate issue price	100 PLN
Subscription period	26/07/2010-20/08/2010
Initial fixing date	27/08/2010
Final fixing date	27/08/2012
Redemption date	03/09/2012
Underlying asset	WIG20 index
Two-year investment offering 100% capital protection on the Redemption Date. The investor participates in the 100% growth of WIG20 index (underlying asset) until the WIG20 reaches 140% (barrier) at any time of the investment compared to its value at the initial fixing date. If the barrier is reached the investor receives a coupon of 6,5%..	

Source: product termsheet

The payoff from Zyskuj z WIG20 – Strategia 100 product is illustrated in Figure 1.

Figure 1. Payoff from Zyskuj z WIG20 – Strategia 100 product



Source: own elaboration based on product termsheet

WIG20 Twin Win product

This product is an example from the class of participation products and attractive looking investment profiting both the rise of WIG20 index (unlimited) and fall (capped). The short description of this product is presented in Table 3[6].

[6] Full termsheet for this product is available at http://www.gpw.pl/info_produkty_strukturyzowane?isin=AT0000A10550&ph_tresc_glowna_start=show. Accessed: 20/06/2013

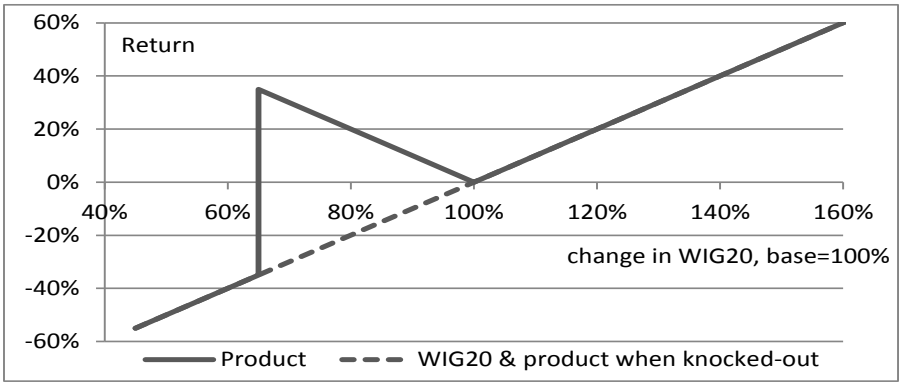
Table 3. Basic parameters for WIG20 Twin Win product

Issuer	Raiffeisen Centrobank AG
Currency	PLN
Unit certificate issue price	1000 PLN
Subscription period	15/04/2013-24/04/2013
Initial fixing date	26/04/2013
Final fixing date	27/10/2015
Redemption date	30/10/2015
Underlying asset	WIG20 index
Two and a half-year investment offering 100% participation in the increase of WIG20 index. The decrease of WIG20 is turned to profit 1:1 until the WIG20 falls 35% (barrier) at any time of the investment compared to its value at the initial fixing date. If the barrier is reached the product return tracks exactly movement in the WIG20 index.	

Source: product termsheet

The payoff from WIG20 Twin Win product is illustrated in Figure 2.

Figure 2. Payoff from WIG20 Twin Win product



Source: own elaboration based on product termsheet

The structure of such product is constructed by means of a long zero strike call option, and long two down-and-out put options, where the strike is set at-the-money. If the barrier is knocked-out the Twin-Win transforms itself in a certificate tracking the WIG20 index.

The simulations of returns for two described products were carried out. Note, that Zyskuj z WIG20 product has already expired[7] whereas WIG20 Twin Win is in course of investment period. The starting period of simulations was 02/01/2000 for both products. For Zyskuj z WIG20 there were 2170 simulations including 500 overlapping observations each (two-year investment period, the last simulation ended just before final subscription day 20/08/2010). For WIG20 Twin Win there

[7] Final value of one certificate paid at redemption date was 100 PLN (return 0%)

were 2715 simulations, each lasting for two and a half-year investment period, the last one ended before final subscription day 24/04/2013. Details about simulation returns are presented in Tables 4 and 5.

Table 4. Distribution of returns for WIG20 index and Zyskuj z WIG20 product

Range of return	Average return in the range for:		Probability
	WIG20	Product	
$R < 0\%$	-31,1%	0	0,438
$R \in \langle 0\%; 40\%)$	20,5%	20,5%	0,149
$R \geq 40\%$	65,6%	6,5%	0,413

Source: own calculations

Table 5. Distribution of returns for WIG20 index and WIG20 Twin Win product

Range of return	Average return in the range for:		Probability for:	
	WIG20	Product	WIG20	Product
$R \leq -35\%$	-42,8%	-41,9%	0,175	0,172
$R \in (-35\%; 0\%\rangle$	-17,9%	-25,5%	0,232	0,132
$R \geq 0\%$	57,1%	49,9%	0,593	0,696

Source: own calculations

Based on the distributions of returns, expected utilities and certainty equivalents were computed for the investments in WIG20 and both products[8]. The results are shown in Tables 6 and 7. For exponential and quadratic utility functions results for two different levels of risk aversion are reported whereas for CPT utility two different coefficients of loss aversion were examined.

Generally the results show that for the given sample, investments in both products as well as directly in WIG20 (asset underlying) could be considered attractive except for few cases with high level of risk or loss aversion. There is also not much difference in results regarding different functional forms of utility function and absolute and relative risk aversion. However there is a noticeable difference comparing investments in WIG20 and in particular product. Volatility of returns from Zyskaj z WIG20 is so low that most investors prefer direct investment in much riskyWIG20 index. The capital protection seems insufficient reward for potential profit lost. On the contrary WIG20 Twin Win is preferred in all cases to the direct investment in WIG20, but according to strong loss aversion this preference is weakest for investors exhibiting behavioral biases (for higher loss aversion both investments in WIG20 and WIG20 Twin are for them unattractive).

[8] For the purpose of calculations, final levels of wealth reflecting distributions of returns were scaled for quadratic and exponential utility functions. For CPT utility gains and loss were considered taking 100 (initial investment) as a reference point. Certainty equivalents in all cases are comparable to the initial investment of 100.

Table 6. Expected utilities and certainty equivalents for WIG20 index and Zyskuj z WIG20 product

Functional form	Expected utility for:		Certainty equivalent for:	
	WIG20	Product	WIG20	Product
Logarithmic	4,678	4,658	107,57	105,46
Power $\alpha = 0{,}5$	10,586	10,274	112,06	105,56
Exponential $\alpha = 0{,}5$	-0,572	-0,590	111,58	105,55
Exponential $\alpha = 2$	-0,139	-0,122	98,67	105,12
Quadratic $\alpha = 0{,}5$	0,109	0,100	115,06	105,38
Quadratic $\alpha = 2$	0,085	0,083	108,88	105,19
CPT $\lambda = 1{,}25$	5,355	4,336	106,73	105,30
CPT $\lambda = 2{,}25$	-3,196	4,336	98,51	105,30

Source: own calculations

Table 7. Expected utilities and certainty equivalents for WIG20 index and WIG20 Twin Win product

Functional form	Expected utility for:		Certainty equivalent for:	
	WIG20	Product	WIG20	Product
Logarithmic	4,729	4,755	113,24	116,12
Power $\alpha = 0{,}5$	10,858	10,972	117,19	120,38
Exponential $\alpha = 0{,}5$	-0,556	-0,548	117,48	120,08
Exponential $\alpha = 2$	-0,126	-0,118	103,37	106,60
Quadratic $\alpha = 0{,}5$	0,114	0,116	120,51	122,48
Quadratic $\alpha = 2$	0,089	0,090	114,69	117,57
CPT $\lambda = 1{,}25$	5,912	6,660	107,53	108,63
CPT $\lambda = 2{,}25$	-2,571	-1,276	98,84	99,48

Source: own calculations

CONCLUDING REMARKS

Structured financial products have gained more and more popularity in recent years, however due to their variety and complexity there is much to be done for the analysis of their expected utility for investors with different preferences. The results presented in this paper are only little contribution in this field. Many popular structured products use behavioral factors, like loss aversion. Hens and Rieger (2008) show that the currently most popular products cannot be explained even within the framework of prospect theory, but only when taking into account probability mis-estimation. This suggests the possible extensions of the studies.

REFERENCES

Blumke A. (2011) Jak inwestować w produkty strukturyzowane, Wolters Kluwer Polska, Warszawa.

Doebeli B., Vanini P. (2010) Stated and Revealed Investment Decisions Concerning Structured Products. Journal of Banking and Finance 34 (6), 1400-1411.

Fischer R. (2007) Do Investors in Structured Products Act Rationally, European Business School Working Paper. SSRN eLibrary.

Friend I., Blume M. E. (1975) The Demand for Risky Assets. American Economic Review 65 (5), 900-922.

Glimcher P., Camerer C., Fehr E., Poldrack R. (2009) Neuroeconomics: Decision Making and the Brain, Elsevier Academic Press, London.

Hens T., Bachmann K. (2010) Psychologia rynku dla doradców finansowych, CeDeWu Wydawnictwa Fachowe, Warszawa

Hens T., Rieger M. (2008) The dark side of the moon: structured products from the customer's perspective, National Centre of Competence in Research Financial Valuation and Risk Management, Working Paper No. 459, http://www.econbiz.de/archiv1/2008/47197_structured_products_perspective.pdf, accessed: 20/05/2013

Kahneman D., Tversky A. (1979) Prospect Theory: An Analysis of Decision Under Risk, Econometrica 47, 1979, 263-291.

Ogaki M., Zhang Q. (2001) Decreasing Relative Risk Aversion and Tests of Risk Sharing, Econometrica 69 (2), 515-526.

Pindyck R. S. (1988) Risk Aversion and the Determinants of Stock Market Behavior, Review of Economic Studies 70 (2), 183-190.

Tversky A., Kahneman D. (1992) Advances in Prospect Theory: Cumulative Representation of Uncertainty, Journal of Risk and Uncertainty 5, 297-323.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 289 – 297

FORECASTING OF INDIVIDUAL ELECTRICITY USAGE USING SMART METER DATA

Tomasz Ząbkowski, Krzysztof Gajowniczek
Department of Informatics
Warsaw University of Life Sciences – SGGW
e-mail: tomasz_zabkowski@sggw.pl, krzysztof_gajowniczek@sggw.pl

Abstract: Forecasting electricity usage is an important task to provide intelligence to the smart gird. The customers will benefit from metering solutions through greater understanding of their own energy consumption and future projections, allowing them to better manage costs of their usage. In this proof of concept paper, we show the approach for short term electricity load forecasting for 24 hours ahead, calculated on the individual household level. In this context authors will develop an approach to the analysis and prediction using Multivariate Adaptive Regression Splines (MARSplines).

Keywords: smart metering systems, short term energy forecasting, multivariate adaptive regression splines

INTRODUCTION

Smart metering is a quite new topic that has grown in importance all over the world and it appears to be a remedy for rising prices of electricity. One of the most important challenge of smart metering is to encourage users to use less electricity through being better informed about their consumption patterns.

Forecasting electricity usage is an important issue to provide intelligence to the smart gird. Accurate forecasting will enable a utility provider to plan the resources and also to take control actions to balance the electricity supply and demand. The customers will benefit from metering solutions through greater understanding of their own energy consumption and future projections, allowing them to better manage costs of their usage.

In this proof of concept paper, our contribution is the approach for short term electricity load forecasting for 24 hours ahead, not on the aggregate but on the

individual household level. The individual customer load profile is influenced by a number of factors, such as devices' operational characteristics, users' behaviours, economic factors, time of the day, day of the week, holidays, weather conditions, geographic patterns and random effects. In this context authors develop an approach to the analysis and prediction of smart metering data using such modelling techniques as Multivariate Adaptive Regression Splines (MARSplines) [Friedman 1991], to capture the factors responsible for accurate short term forecasting in smart metering applications.

Over the last decades different methods have been applied to forecasting the electric load demand. Some of the most popular include time series analyses with autoregressive integrated moving average (ARIMA) [Brockwell and Davis 2002], fuzzy logic [Song et al. 2005], artificial neural network (ANN) [Beccali et al. 2004], [Castillo et al. 2001], [Hippert et al. 2001] and support vector machines (SVM) [Lv et al. 2006]. Majority of them is devoted to analysis of larger loads such as region or the country grid, and therefore, forecasting is achieved with relatively high accuracy [Alfares and Nazeeruddin 2002], [Khotanzad et al. 2002], [Weron 2006].

Leveraging smart metering to support energy efficiency on the individual user level brings research challenges in monitoring usage and providing accurate load forecasting. However, it should be noted that forecasting loads of individual smart meter is not common practice since the volatility of the system is high thus resulting in high error rates [Javed et al. 2012].

CHARACTERISTICS OF DATA

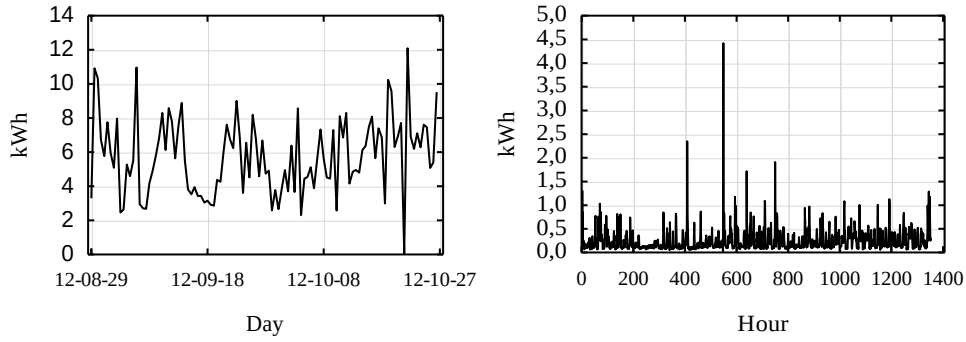
Electricity measurements data were prepared using Mieo HA104 meter installed in one of the households in Warsaw, Poland, for the purpose of SMEPI project[1]. The household consisted of two adult people and a child. The household lived in a flat which was equipped in various home appliances including washing machine, refrigerator, dishwasher, iron, electric oven, two TV sets, audio set, pot, coffee maker, desk lamps, computer, and a couple of light bulbs. The data were gathered during 60 days, starting from 29 August until 27 October 2012.

Original source data contained the electricity usage readings of the Mieosmart meter at every second, every minute and every hour. From these readings, we extracted the hour loads (in kilowatt hour - kWh) for the purpose of short-term load forecasting.

Data characteristics showing the daily and hourly data readings for the analyzed period are illustrated in Figure 1.

[1] SMEPI – Smart Metering Poland, a Hi-Tech project to develop smart metering solutions partially financed by National Centre for Research and Development (NCBiR) and led by Vedia S.A in cooperation with GridPocket and Faculty of Applied Informatics and Mathematics at SGGW.

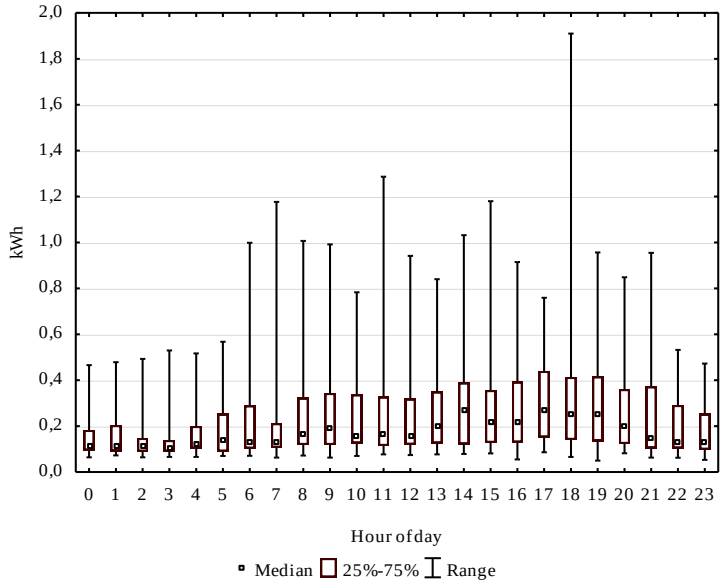
Figure 1. Daily and hourly load in kWh



Source: own preparation

Taking into account that forecasting loads of individual smart meter may be associated with high volatility [Javed et al. 2012] we prepared the box and whisker plot for each of 24 hours using load data over all 60 day, please see Figure 2. The whiskers show the minimum and maximum value in a given hour and box encloses 50% of the total data (top edge represents 75th quartile and bottom edge 25th quartile and line in the middle is the median). The results show that the volatility is rather high (especially during day hours) what can have impact on forecast accuracy.

Figure 2. Box and whisker plot for electricity consumption over each of 24 hours



Source: own preparation

In this research, we focused on forecasting the electricity usage of a particular household for 24 hours ahead. In order to forecast the load we constructed a feature vector with variables as presented in Table 1.

Table 1. Variables used in forecasting

Variable no.	Description	Formula
1 - 24	Load of previous 24 hours	$W_{hi}, W_{h-1} \ldots W_{h-24}$
25- 28	Average load of previous 3, 6, 12, 24 hours	$\frac{1}{i}\sum(W_{hi}),\ i = 3,6,12,24$
29 - 32	Maximum load of previous 3, 6, 12, 24 hours	$\max\{W_{hi}\},\ i = 3,6,12,24$
33 - 36	Minimum load of previous 3, 6, 12, 24 hours	$\min\{W_{hi}\},\ i = 3,6,12,24$
37 - 40	Range of load of previous 3, 6, 12, 24 hours	$\max\{W_{hi}\} - \min\{W_{hi}\},\ i = 3,6,12,24$
41	Day of the week	D_w
42	Day part (morning, noon, afternoon, evening, night)	D_p
43	Temperature observed in each hour	T_{hi}

Source: own calculations

These 43 attributes were empirically derived. The individual, the average, the minimum, the maximum and the range loads information were obtained from the hourly load time series. The temperature information inside the flat, for each hour, was collected with Mieo smart meter.

FORECASTING METHOD

In the experiment we used Multivariate Adaptive Regression Splines (MARSplines) which is nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. MARSplines constructs the relation from a set of coefficients and so-called basis functions that are entirely determined from the regression data set. In a sense, the method is based on the divide and conquer strategy, which partitions the input space into regions, each with its own regression equation. In general, nonparametric models are adaptive and very flexible what can ultimately result in over fitting. Although such models can achieve very low error on training data, they have the tendency to perform very bad with new observations. To overcome this problem, MARSplines uses a pruning technique

(similar to pruning in classification trees) to limit the complexity of the model by reducing the number of its basis functions. As basis functions MARSplines uses two-sided truncated functions (hinge function) for linear or nonlinear expansion, which approximates the relationships between the response and predictor variables. A hinge function takes the form:

$$(x-t)_+ = \begin{cases} x-t, & x > t \\ 0, & x \leq t \end{cases} \tag{1}$$

where parameter t is a constant, called the knot, of the basis functions which defining the pieces of the segmented linear regression. It should be stressed, that only positive results of the respective equations are considered, otherwise the respective functions evaluate to zero.

MARSplines can be proposed even in situations where the relationship between the predictors and the dependent variables is non-monotone and difficult to approximate with parametric models, therefore seem much more capable of solving forecasting problem. The general MARSplines model equation is given as

$$P = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \tag{2}$$

where the summation is over the M nonconstant terms in the model. To summarize, P is predicted as a function of the predictor variables x. This function consists of an intercept parameter β_0 and the weighted byβ_m sum of one or more basis functions $h_m(x)$.

Implementing MARSplines involves a two-step procedure that is applied successively until a desired model is found. In the first step, the model is build, i.e. increase its complexity by adding basis functions until a preset maximum level of complexity has been reached. After implementing the forward stepwise selection of basis functions, a backward procedure is applied in which the model is pruned by removing those basis functions that are associated with the smallest increase in the goodness of fit.

The so-called Generalized Cross Validation error is a measure of the goodness of fit that takes into account not only the residual error but also the model complexity as well, which is given by:

$$GCV = \frac{\sum_{i=1}^{n}(W_{hi} - P_{hi})^2}{\left(1 - \frac{C}{n}\right)^2} \tag{3}$$

with

$$C = 1 + cd$$

where W_{hi} is the observed load in hour i and P_{hi} is the forecasted load in hour i, n is the number of observations in the original data set, d is the effective degrees of freedom which is equal to the number of independent basis functions and finally,c is the penalty for adding a basis function.

The MARS algorithm can be described and implemented as follows:

1. Start with the simplest model involving only the constant basis function (an intercept parameter β_0).
2. Search the space of basis functions, for each variable and for all possible knots, and add those which maximize a certain measure of goodness of fit (minimize prediction error).
3. Step 2 is recursively applied until a model of pre-determined maximum complexity is derived.
4. Finally, in the last stage, a pruning procedure is applied where those basis functions are removed that contribute least to the overall (least squares) goodness of fit.

EVALUATION MEASURES

To assess the model performance for forecasting, we used two measures: precision and accuracy [Javed 2012].

Precision shows how close the model is able to forecast to the actual load. To measure precision we used mean squared error (MSE) given by:

$$MSE = \frac{\sum_{i=1}^{n}(W_{hi} - P_{hi})^2}{n} \tag{4}$$

where W_{hi} is the observed load in hour i and P_{hi} is the forecasted load in hour i.

In case of accuracy, this measure shows how many correct forecasts the model makes. For this purpose we need to define a correct forecast as the value within a percentage range of the actual load. However, for very low loads, a percentage range may become insignificant. For instance, having a load of 0.1 kWh, a 10% correctness range would be 0.09–0.11 kWh and a forecast of 0.2 kWh will be considered as wrong, but in practice such forecast would be acceptable. To overcome this false loss of accuracy we set two scales to measure accuracy. We set a 10% range of error for accuracy, but if the load is smaller than 1 kWh then we consider range of ±0.10 kWh as range of acceptable forecast. Therefore, accuracy for hour i is given as:

$$\begin{aligned} AC = \sum 1\{W_{hi} > 1 \& |W_{hi} - P_{hi}| < P_{hi} \times 0.10\} \\ + \sum 1\{W_{hi} < 1 \& |W_{hi} - P_{hi}| < 0.10\}. \end{aligned} \tag{5}$$

ELECTRICITY USAGE FORECASTING

Before estimating and assessing the MARSplines model, we have randomly selected data into two samples. The first set (training) was used to estimate the model, while the second set (test) was used to validate the model. The training and the testing sample included 80% and 20% of the observations, respectively.

As loss function we chose the least squares estimator. In the most general terms, least squares estimation is aimed at minimizing the sum of squared deviations of the observed values for the dependent variables from those forecasted by the model. Technically, the least squares estimator is obtained by minimizing SOS (sum of squares) function:

$$SOS = \sum_{i=1}^{n} (W_{hi} - P_{hi})^2 \tag{6}$$

where W_{hi} is the observed load in hour i and P_{hi} is the forecasted load in hour i.

The calculations were prepared in Statistica ver. 10. Due the limitations of the theory and software in the experiment we build 24 models, each for single hour of a day. As a maximum number of basis functions we selected 70 functions, which mean that very complex model will be built in the third step of the algorithm. The forward step usually builds an over fitted model, therefore to build a model with better generalization ability, the backward step prunes the model taking into account c parameter (in our case c=2) which is the penalty for adding a one more basis function. A further constraint which we faced on the forward step is specification of a maximum allowable degree of interaction. Typically, only one or two degrees of interaction are allowed, but higher degrees can be used when the problem require it. Therefore, we choose value 20 as a maximum degree of interaction between independent variables.

The final results obtained by MARSplines and aggregated over all hours are shown in Table 2.

Table 2. Model results aggregated over all hours

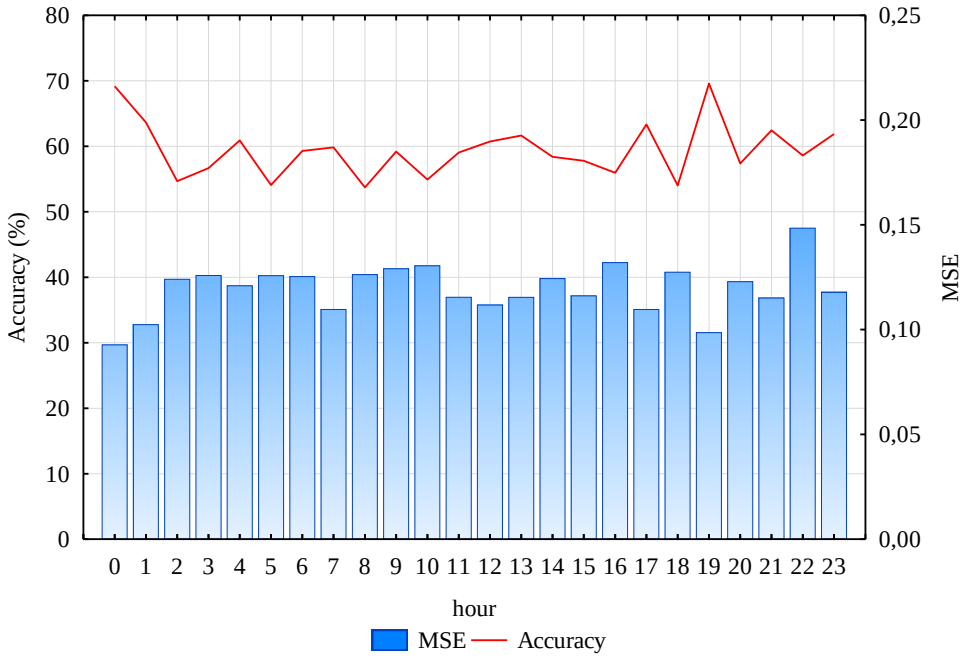
Measure	Training dataset	Test dataset
Accuracy (%)	62%	59%
MSE	0.10	0.11

Source: own calculations

For training sample, the accuracy which measures of how many correct forecasts the model makes is 62% and the precision of how close the model is able to forecast to the actual load (MSE) is 0.10. The results associated with the test set are close to these obtained on training set. For this sample MARSplines obtained 59% of accuracy and 0.11 for MSE.

Detailed results per single hour using proposed measures for test sample are shown in Figure 3.

Figure 3. Results in terms of accuracy and MSE for each single hour calculated on the test dataset



Source: own preparation

From the Figure 3 we can observe that almost all hours can be forecasted with relatively high accuracy (more than 50%) which is rather stable over all hours and it does not drops down unexpectedly.

The results presented above are promising but it should be underlined that forecasting on individual household level is perceived as a difficult task since the hourly and daily behaviour may change drastically due to different circumstances, e.g. using particular home appliances, weather conditions, holidays of household members. In larger populations such as local grid or region, smaller loads tend to neutralize to produce a stable time series but for an individual home load, the time series volatility is quite high, thus accurate forecasting becomes challenging task.

CONCLUSIONS AND FUTURE WORKS

In this paper, we presented an approach to forecast electricity load on individual household level what can potentially provide greater intelligence in smart metering systems. The result of MARS model used for 24 hours ahead short term load forecast shows that it has good performance and reasonable prediction accuracy can be achieved.

Accurate forecasting brings value added both, to a utility provider and individual customers. The first one can plan the resources and also to take control actions to balance the electricity supply and demand. The customers can benefit from metering solutions through greater understanding of their own energy consumption and future projections, allowing them to better manage costs of their usage.

As future work we see the need to undertake home appliance recognition problem under the nonintrusive appliance load monitoring (NIALM) concept. It may generate additional value to smart meters since the electricity usage of a household changes over time based on the operation of various appliances used by the family. Therefore, appliance detection might be additional input variable used for more accurate electricity usage forecasting.

REFERENCES

Alfares H.K., Nazeeruddin M. (2002) Electric load forecasting: literature survey and classification of method, International Journal of Systems Science, vol. 33(1), 3–34.

Beccali M., Cellura M., Brano V.L., Marvuglia A. (2004) Forecasting daily urban electric load profiles using artificial neural networks, Energy Conversion and Management, vol. 45, 2879–2900.

Brockwell P.J., Davis R.A. (2002) Introduction to Time Series and Forecasting, Springer.

Castillo E., Guijarro B., Alonso M. (2001) Electricity Load Forecast using Functional Networks, Report for EUNITE 2001 Competition, available at http://neuron.tuke.sk/competition/ on 2013-07-10.

Friedman J.H. (1991) Multivariate Adaptive Regression Splines, The Annals of Statistics, vol. 19, 1–141.

Hippert H.S, Pedreira C.E., Souza R.C. (2001) Neural networks for short term load forecasting: a review and evaluation, IEEE Transactions on Power Systems, vol. 16, 44–55.

Javed F., Arshad N., Wallin F., Vassileva I., Dahlquist E. (2012) Forecasting for demand response in smart grids: an analysis on use of anthropologic and structural data and short term multiple loads forecasting, Applied Energy, vol. 69, 150–160.

Khotanzad A., Zhou E., Elragal H. (2002) A neuro-fuzzy approach to short-term load forecasting in a price-sensitive environment, IEEE Transactions on Power Systems, vol. 17, 1273–1282.

Song K.B., Baek Y.S., Hong D.H., Jang G. (2005) Short-term load forecasting for the holidays using fuzzy linear regression method, IEEE Transactions on Power Systems, vol. 20, 96–101.

Weron R. (2006) Modeling and forecasting electricity loads and prices: A statistical approach, Wiley, Chichester.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 298 – 308

APPLICATION OF MULTIVARIATE DISCRIMINANT ANALYSIS FOR ASSESSMENT OF CONDITION OF CONSTRUCTION COMPANIES

Monika Zielińska–Sitkiewicz
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
e-mail: monika_zielinska_sitkiewicz@sggw.pl

Abstract: The construction is important in a market economy. From the development of the construction industry depends on large extent how the economy will function. Hence, the need for continuous monitoring of both – the market and the use of methods- that will objectively evaluate the quality of the construction companies. The paper contains consideration about usage discriminant analysis in financial audit of construction companies. 30 companies from construction sector, which are listed on the Warsaw Stock Exchange, were selected for study. The analysis encompassed financial data from balance sheets and from profit and loss account in the period from January 1, 2005 to December 31, 2012.

Keywords: Polish real estate market, construction company, financial ratios, discriminant analysis

INTRODUCTION

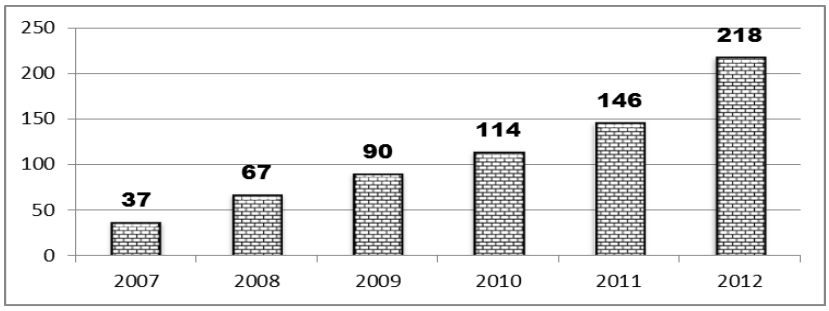
Important role of the construction industry in the economy results from the social function fulfilled by the industry that is expressed in the construction of housing and commercial premises, municipal infrastructure, including roads and investments aimed at environmental protection. Unfortunately, such function is not able to ensure the stability of the construction market, so it characterizes with big fluctuations. In the present times of globalisation the course of economic construction cycle in Poland points out directly to direct relationship with fluctuations in the Polish and global economy.

Current statistical data point out clearly to a slowdown in the construction market. Construction engineering recorded a fall of 9% in 2012 as compared to

2011, and the production in the road transportation infrastructure recorded a decrease amounting to 15.4%. Housing and commercial construction industries were the only areas that ended with a growth in 2012 of 10.8% and 10.6% respectively. Deteriorating results of the construction industry are the outcome of decreasing value of works in all the three types of construction companies. The lowest activity was recorded by the group of companies that build buildings, since the growth of the value of their works amounted to 0.1% only. Slightly higher growth was recorded by companies performing engineering works (by 2.1%) and specialist works (by 2.6%). For the first time after many years the construction and assembly production executed in Poland in 2012 was lower than in the previous year. The decrease amounted to 1.1%.

The economic situation in the construction industry is assessed more and more pessimistically, and growing problems with payment of liabilities led to bankruptcy of many construction companies.

Figure 1. Bankruptcies of construction companies in 2007-2012



Source: ASM – Centrum Badań i Analiz Rynku Sp. z o. o.

The most recent data point out to bankruptcy of 273 construction companies in 2012. It translates into a growth by 87.0% as compared to the previous year. What is more, it means seven times more bankruptcies than in 2007 (compare Figure 1). It should be emphasized that regardless of the type of operations of a company, payment backlogs were the main reason of bankruptcies.

Taking into account the fact that we observe further deterioration of the macroeconomic situation, it seems necessary to start analysing the condition of the construction industry for companies quoted at the Warsaw Stock Exchange, at least from the perspective of investors. So it is worth examining what picture of the construction industry may be seen in result of analysis of financial data from companies in that industry.

The aim of the article is to present some Polish models based on the discriminant analysis and to attempt to use them for general evaluation of the condition of 30 selected construction companies. Enterprises selected for the examination are quoted on the main market of the Warsaw Stock Exchange and their profit and loss account is made by type. The analysis referred to the following

models: Hołda's (1996), Mączyńska's (1994), Sojak and Stawicki's (1998), Gajdek and Stos's (2003) and Mączyńska and Zawadzki's (2006). Presented methods were prepared for the Polish market and they are highly estimated in the literature in respect to their forecasting value.

USE OF DISCRIMINANT ANALYSIS FOR ASSESSMENT OF THE CONDITION OF CONSTRUCTION COMPANIES

Discriminant models are used for early identification of the symptoms of deteriorating financial condition of an enterprise. A set of financial ratios used for a given model should decide about the condition and development opportunities of a given company. The discriminant analysis allows identification of ratios that well or badly reflect financial capabilities of a company.

The main forecasting tool is the discriminant function that has the following general form [Prusak 2005]:

$$Z = a_0 + a_1 X_1 + a_2 X_2 + ... + a_n X_n \tag{1}$$

where:

Z – value of the discriminant function,
X_i, for i = 1, 2, …, n – explanatory variables (financial indices),
a_i, for i = 1, 2, ..., n – coefficients of the discriminant function,
a_0 – constant.

The result is interpreted on the basis of comparison of calculated Z value of the discriminant function with a boundary value set by an author of a given model. Entities are classified as members of separable groups on the basis of Z value.

One of the first models estimated for the Polish conditions was the model prepared by A. Hołda. The period of research covered years 1993–1996, and the research covered 40 enterprises threatened with bankruptcy and 40 enterprises that had already gone bankrupt. The entities belonged to group 45-74 of the European Classification of Activities (e.g. construction industry, hotels and restaurants, financial intermediation and others). They were selected by industry and underwent ratio analysis. In the final form of the model the number of ratios was reduced to 5:

X_1 = current assets / short-term liabilities
X_2 = (borrowed capital / total liabilities) * 100
X_3 = (net profit / average value of total assets) * 100
X_4 = (average value of current liabilities / costs of production of sold products, goods and materials) * 360
X_5 = total revenues / average value of total assets

Economic entities in case of which the value of discriminant function in form of:

$$Z = 0{,}605 + 0{,}681 \cdot X_1 - 0{,}0196 \cdot X_2 + 0{,}00969 \cdot X_3 + \\ + 0{,}000672 \cdot X_4 + 0{,}157 \cdot X_5 \tag{2}$$

is higher than 0.1. are defined as very unlikely to go bankrupt, while if the value is lower than -0.3, they are very likely to go bankrupt. The range of uncertainty that is the range of Z function values that are very likely to be classified wrongly was set at the level from -0.3 to 0.1 [Hołda 2001].
The results of the classification for analysed construction enterprises are presented in Table 1.

Table 1. Results of Hołda's model (1996) for construction companies

Name of the company	2012-12-31	2011-12-31	2010-12-31	2009-12-31	2008-12-31	2007-12-31
TOTAL	**-0,296**	**0,299**	**0,517**	**0,615**	**0,575**	**0,611**
BUDIMEX	0,009	0,031	0,074	0,151	0,311	0,268
POLIMEXMS	-0,264	0,338	0,533	0,528	0,473	0,717
PBG W UPADŁOŚCI	-1,083	0,440	0,788	0,846	0,699	0,543
MOSTALWAR	0,031	0,221	0,599	0,653	0,505	0,620
TRAKCJA	0,287	0,247	1,175	1,760	0,888	0,452
HBPOLSKA W UPADŁOŚCI	-3,445	0,309	0,365	0,332	0,132	
ERBUD	0,623	0,645	0,867	0,816	0,786	1,156
ELBUDOWA	0,963	1,011	1,102	1,474	1,322	0,577
MOSTALZAB	0,746	0,530	0,469	0,536	0,367	1,094
POLAQUA	-0,096	0,474	0,143	0,685	0,746	
DSS W UPADŁOŚCI	-1,654	-1,765	0,058	-0,093		
ABMSOLID W UPADŁOŚCI	-2,175	-0,304	0,251	0,441	0,735	1,002
INSTALKRK	1,640	1,301	1,613	1,595	1,243	0,954
AWBUD	0,150	0,404	0,503			
PROJPRZEM	2,056	1,854	1,702	2,352	2,012	2,710
BIPROMET	1,019	0,550	0,866	0,795		
INTAKUS W UPADŁOŚCI	-0,719	-0,093	0,886	0,961		
UNIBEP SA	0,491	0,648	0,861	0,759	0,881	
ZUE SA	0,617	0,786	0,627			
ULMA CONSTRUCCION POLSKA SA	1,108	0,816	0,490	0,385	0,566	1,032
ENERGOMONTAŻ-POŁUDNIE SA	-6,136	0,025	0,616	0,680	0,594	1,032
P.A. NOVA SA	0,564	1,073	1,441	1,463	1,483	4,040
INTERBUD-LUBLIN SA	0,457	0,913	0,791			
PROCHEM SA	1,265	1,005	1,241	1,080	0,692	0,747
MOSTOSTAL-EXPORT SA	1,237	0,428	1,516	1,213	1,315	0,932
ELEKTROTIM SA	1,434	1,758	2,322			
CNT	1,026	1,150	1,577	1,340	1,325	2,604
ENERGOAPARATURA SA	0,999	1,058	1,396	0,826	0,554	0,571
BUDOPOL-WROCŁAW SA	-0,251	1,272	0,993	2,182	2,056	
RESBUD SA	1,497	0,225	0,987	0,999	1,179	
UNCERTAINTY **-0,3 <= Z <= 0,1**		BANKRUPT RISK **Z< -0,3**			GOOD **Z > 0,1**	

Source: own calculations

The following ratios were decisive in Hołda's function for evaluation of construction companies: (X_1) (basic liquidity ratio) and (X_2) (debt ratio), while asset profitability ratio (X_3) was equal zero. It should be pointed out that in case of bankrupt companies, problems occurred next year were detected in 2011 already, except companies Polimex and PBG. However, it should be taken into account that as the dominating shareholder, in its consolidated financial statement PBG includes disastrous results of HBPOLSKA and ENERGOMONTAŻ-POŁUDNIE.

The examination on the discriminant model adjusted to the Polish conditions was conducted by E. Mączyńska, as well. She adapted O. Jacobs's function used for the assessment of credit rating of entities. The form of the function was worked out relatively long time ago, but its forecasting values are high. Comments adopted

in the model took account of the meaning of individual ratios for the general financial conditions of a company [Mączyńska 1994].

Financial ratios used for the purpose of model construction have the following form:

X_1 = (gross result + depreciation) /total liabilities
X_2 = total assets / total liabilities
X_3 = gross result / total assets
X_4 = gross result /revenues from sales
X_5 = inventory/ revenues from sales
X_6 = revenues from sales / total assets

Results of the classification for analysed developers are presented in Table 2.

Table 2. Results of Mączyńska's model (1994) for construction companies

Name of the company	2012-12-31	2011-12-31	2010-12-31	2009-12-31	2008-12-31	2007-12-31
TOTAL	**-4,850**	**0,393**	**0,791**	**1,236**	**1,329**	**1,294**
BUDIMEX	1,105	1,308	1,484	1,238	0,984	0,294
POLIMEXMS	-5,408	0,595	0,932	1,208	1,017	1,138
PBG W UPADŁOŚCI	-22,266	0,828	1,194	1,518	1,690	1,521
MOSTALWAR	-0,753	-0,824	1,165	2,054	1,736	1,268
TRAKCJA	0,206	0,884	1,348	3,031	1,801	1,417
HBPOLSKA W UPADŁOŚCI	-119,685	0,503	0,425	1,413	1,102	
ERBUD	0,713	-0,086	0,783	1,738	0,603	1,535
ELBUDOWA	1,474	1,569	1,981	2,999	3,100	2,296
MOSTALZAB	0,111	1,274	0,923	1,713	1,846	3,165
POLAQUA	-3,841	1,121	-5,053	-0,985	0,963	
DSS W UPADŁOŚCI	-5,722	-12,764	-0,969	-1,460		
ABMSOLID W UPADŁOŚCI	-14,143	-1,964	0,566	0,970	1,358	1,390
INSTALKRK	1,792	1,823	1,973	2,610	2,610	2,451
AWBUD	-1,592	0,270	-0,230			
PROJPRZEM	1,998	0,571	-1,744	0,700	2,066	2,845
BIPROMET	1,388	1,125	0,617	0,537		
INTAKUS W UPADŁOŚCI	-4,959	-4,926	0,521	0,015		
UNIBEP SA	0,650	1,145	1,487	1,682	2,341	
ZUE SA	0,439	1,367	1,159			
ULMA CONSTRUCCION POLSKA SA	2,514	3,551	1,334	0,173	2,058	4,184
ENERGOMONTAŻ-POŁUDNIE SA	-44,048	-1,038	0,038	-0,308	1,373	1,730
P.A. NOVA SA	1,384	1,967	1,960	2,951	2,433	3,550
INTERBUD-LUBLIN SA	-0,213	1,181	1,678			
PROCHEM SA	-0,062	0,996	0,956	0,877	1,624	1,870
MOSTOSTAL-EXPORT SA	-5,409	-12,040	-2,259	-1,807	4,879	0,532
ELEKTROTIM SA	1,420	2,030	1,370			
CNT	1,395	1,085	1,281	-2,221	-2,221	1,359
ENERGOAPARATURA SA	1,594	1,305	2,598	2,173	1,485	0,131
BUDOPOL-WROCŁAW SA	-7,855	1,762	1,491	1,358	1,030	
RESBUD SA	-12,303	-4,469	0,131	-1,510	3,388	
VERY GOOD S>2	GOOD 1<S<=2		WEAK 0<S<=1		WRONG S<0	

Source: own calculations

Interpretation of a discriminant function:

$$Z = 1{,}5 \cdot X_1 + 0{,}08 \cdot X_2 + 10 \cdot X_3 + 5 \cdot X_4 - 0{,}3 \cdot X_5 + 0{,}1 \cdot X_6 \qquad (3)$$

should be based on the following principles: Z <= 0 is an enterprise threatened with bankruptcy within 1 year, if 0 < Z < 1 an enterprise is weak but not threatened with bankruptcy, is 1 <= Z <= 2 an enterprise is good and if Z >= 2 an enterprise is very good.

The ratios that were most decisive in an assessment of a given company in Mączyńska's model were turnover profitability (X_4) and sales margin (X_3). The

assessment of enterprises shows their large differentiation. All companies classified in the worst category of economic condition had very bad financial results in relation to their assets. On the basis of comparisons of the results of this model for a few former years it should be stated that it is possible to point out quite precisely to companies that will have serious problems sooner or later.

Contrary to the other models, Sojak and Stawicki's model consists of three classification functions for: good enterprises, average enterprises and enterprises threatened with bankruptcy. An analysis conducted by the researchers covered 58 enterprises and they computed 20 financial ratios on the basis of information from 1998. Then by means of data clustering the authors selected 11 ratios out of those 20 that are best for group discrimination. Then 7 best ratios were selected out of those 11 ratios [Prusak 2004]:

X_1 = (net financial result/average value of current assets) · 100;
X_2 = (current assets - inventory - accruals) /short-term liabilities;
X_3 = average working capital / average value of assets;
X_4 = (net financial result / average value of equity) · 100;
X_5 = (net financial result / average value of fixed assets) · 100;
X_6 = (net financial result + interests on borrowed capital– income tax) / average value of assets;
X_7 = current assets / short-term liabilities;

And the three following classification functions were constructed on the basis of them:

$$Enterprise_{wrong} = -0{,}1144 \cdot X_1 + 0{,}5178 \cdot X_2 - 20{,}4475 \cdot X_3 + \\ - 0{,}0661 \cdot X_4 + 0{,}0663 \cdot X_5 - 50{,}461 \cdot X_6 + 1{,}8358 \cdot X_7 - 11{,}6499 \quad (4)$$

$$Enterprise_{average} = -0{,}0586 \cdot X_1 - 3{,}3608 \cdot X_2 + 10{,}7088 \cdot X_3 + \\ + 0{,}01455 \cdot X_4 - 0{,}066 \cdot X_5 + 4{,}5837 \cdot X_6 + 0{,}24329 \cdot X_7 - 2{,}3393 \quad (5)$$

$$Enterprise_{good} = -0{,}0153 \cdot X_1 + 2{,}0482 \cdot X_2 + 9{,}637 \cdot X_3 + \\ + 0{,}1714 \cdot X_4 - 0{,}0091 \cdot X_5 - 15{,}78 \cdot X_6 - 0{,}0018 \cdot X_7 - 5{,}992 \quad (6)$$

Allocation to a respective group of enterprises depends on the highest positive ratio. Results of the classification for analysed construction companies are presented in Table 3. The group of ratios selected for the above functions focuses on profitability, and in result it refers directly or indirectly to inventories that are small in construction companies. Thus it may be observed that these models assessed more highly companies that invested their financial surpluses in the current activities. But they did not show any threats in case of two companies

undergoing bankruptcy proceedings in 2012, that is ABMSOLID and HBPOLSKA.

Table 3. Results of Sojak and Stawicki's models (1998) for construction companies

Name of the company	2012-12-31	2011-12-31	2010-12-31	2009-12-31	2008-12-31	2007-12-31
TOTAL	WRONG	Max<0	Max<0	Max<0	Max<0	GOOD
BUDIMEX	GOOD	GOOD	GOOD	GOOD	Max<0	Max<0
POLIMEXMS	WRONG	Max<0	Max<0	Max<0	Max<0	Max<0
PBG W UPADŁOŚCI	WRONG	Max<0	GOOD	GOOD	GOOD	GOOD
MOSTALWAR	Max<0	Max<0	Max<0	GOOD	GOOD	GOOD
TRAKCJA	Max<0	Max<0	GOOD	GOOD	GOOD	GOOD
HBPOLSKA W UPADŁOŚCI	GOOD	Max<0	Max<0	GOOD	GOOD	
ERBUD	GOOD	Max<0	GOOD	GOOD	GOOD	GOOD
ELBUDOWA	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
MOSTALZAB	Max<0	Max<0	Max<0	Max<0	Max<0	GOOD
POLAQUA	WRONG	Max<0	WRONG	Max<0	Max<0	
DSS W UPADŁOŚCI	WRONG	GOOD	Max<0	Max<0		
ABMSOLID W UPADŁOŚCI	GOOD	Max<0	Max<0	Max<0	GOOD	Max<0
INSTALKRK	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD
AWBUD	Max<0	Max<0	Max<0			
PROJPRZEM	GOOD	Max<0	Max<0	GOOD	GOOD	GOOD
BIPROMET	Max<0	Max<0	Max<0	Max<0		
INTAKUS W UPADŁOŚCI	WRONG	WRONG	Max<0	Max<0		
UNIBEP SA	Max<0	GOOD	GOOD	Max<0	GOOD	
ZUE SA	Max<0	Max<0	Max<0			
ULMA CONSTRUCCION POLSKA SA	Max<0	Max<0	Max<0	Max<0	Max<0	Max<0
ENERGOMONTAŻ-POŁUDNIE SA	GOOD	Max<0	Max<0	Max<0	Max<0	GOOD
P.A. NOVA SA	Max<0	Max<0	GOOD	Max<0	GOOD	GOOD
INTERBUD-LUBLIN SA	Max<0	Max<0	GOOD			
PROCHEM SA	Max<0	Max<0	Max<0	Max<0	Max<0	GOOD
MOSTOSTAL-EXPORT SA	Max<0	WRONG	Max<0	Max<0	GOOD	Max<0
ELEKTROTIM SA	GOOD	GOOD	GOOD			
CNT	GOOD	GOOD	GOOD	Max<0	Max<0	GOOD
ENERGOAPARATURA SA	GOOD	GOOD	GOOD	GOOD	Max<0	Max<0
BUDOPOL-WROCŁAW SA	WRONG	GOOD	GOOD	GOOD	GOOD	
RESBUD SA	WRONG	WRONG	Max<0	Max<0	GOOD	

Source: own calculations

Another model, that is Gajdek and Stos's model dated 2003, was constructed for the purpose of assessment of companies quoted on the Warsaw Stock Exchange. The system was worked out on the basis of a balanced sample consisting of 34 items: 17 bankrupt companies to which 17 "healthy" entities with similar business profile were assigned [Kisielińska, Waszkowski 2010].
Estimated linear discriminant model had the following form:

$$Z = -0,3342 - 0,000500 \cdot X_1 + 2,055200 \cdot X_2 + 1,726000 \cdot X_3 + \\ + 0,115500 \cdot X_4 \tag{7}$$

Four financial ratios were used in this model:
X_1 = average value of short-term liabilities / costs of sold production * 360 days;
X_2 = net profit / average asset value during a year;
X_3 = gross profit / net revenues from sales;
X_4 = total assets/ total liabilities;

Results of the classification for analysed developers are presented in Table 4.

The boundary value for the model is zero. If $Z < 0$, an enterprise is classified in a group of companies threatened with bankruptcy, if $Z > 0$ – in a group with good financial standing. The uncertainty range for the model is ⟨–0.49; 0.49⟩.
The decisive ratio in the assessment of the condition of construction companies was ratio (X_4), that is inverse debt rate but only in case of years when a company recorded small profit or loss. When companies recorded relatively high profits or losses, ratios (X_2) – asset profitability and (X_3) – return on sales were dominant. The model pointed out to the threat in respect to all bankrupt companies and by means of 2011 results it confirmed the crisis in the construction industry in 2012.

Table 4. Results of Gajdek and Stos's model (2003) for construction companies

Name of the company	2012-12-31	2011-12-31	2010-12-31	2009-12-31	2008-12-31	2007-12-31
TOTAL	**- 1,244**	**-0,142**	**-0,049**	**0,043**	**0,069**	**0,076**
BUDIMEX	-0,049	0,029	0,082	0,040	-0,028	-0,164
POLIMEXMS	-1,369	-0,075	-0,023	0,030	-0,003	0,031
PBG W UPADŁOŚCI	-5,237	0,002	0,081	0,166	0,186	0,171
MOSTALWAR	-0,394	-0,408	-0,008	0,143	0,094	0,034
TRAKCJA	-0,175	-0,016	0,157	0,402	0,196	0,068
HBPOLSKA W UPADŁOŚCI	-35,794	-0,131	-0,140	0,081	0,030	
ERBUD	-0,091	-0,251	-0,072	0,097	-0,097	0,173
ELBUDOWA	0,095	0,113	0,220	0,379	0,400	0,191
MOSTALZAB	-0,159	0,017	-0,037	0,099	0,168	0,398
POLAQUA	-1,114	0,004	-1,171	-0,364	0,012	
DSS W UPADŁOŚCI	-1,869	-3,084	-0,484	-0,607		
ABMSOLID W UPADŁOŚCI	-3,284	-0,648	-0,113	-0,048	0,065	0,134
INSTALKRK	0,208	0,206	0,251	0,329	0,320	0,340
AWBUD	-0,559	-0,152	-0,258			
PROJPRZEM	0,363	0,094	-0,325	0,178	0,327	0,632
BIPROMET	0,142	0,053	0,032	-0,007		
INTAKUS W UPADŁOŚCI	-1,620	-1,518	-0,074	-0,175		
UNIBEP SA	-0,076	-0,016	0,070	0,101	0,259	
ZUE SA	-0,110	0,065	0,070			
ULMA CONSTRUCCION POLSKA SA	0,287	0,511	0,056	-0,202	0,251	0,767
ENERGOMONTAŻ-POŁUDNIE SA	-5,794	-0,409	-0,185	-0,288	0,118	0,164
P.A. NOVA SA	0,196	0,345	0,382	0,576	0,463	0,931
INTERBUD-LUBLIN SA	-0,215	0,087	0,246			
PROCHEM SA	-0,066	0,073	0,131	0,043	0,103	0,167
MOSTOSTAL-EXPORT SA	-1,475	-2,605	-0,376	-0,496	0,866	0,074
ELEKTROTIM SA	0,116	0,257	0,213			
CNT	0,030	0,027	0,097	-0,545	-0,551	0,273
ENERGOAPARATURA SA	0,101	0,063	0,240	0,161	0,043	-0,143
BUDOPOL-WROCŁAW SA	-1,871	0,206	0,175	0,193	0,117	
RESBUD SA	-2,825	-1,106	-0,220	-0,450	0,413	
BANKRUPT RISK Z<=0			GOOD Z > 0			

Source: own calculations

The last presented Polish model is the model worked out by E. Mączyńska and M. Zawadzki in 2006. The authors conducted a research on a balanced sample of 80 companies quoted on the WSE, using financial statements for 1997–2001 and financial ratios computed on the basis of such statements. The research included 45 ratios characterising profitability, liquidity, debt level, operating efficiency and dynamics of company's growth. Four ratios were selected for the purpose of presented model:
X_1 = (gross result + depreciation) /total liabilities
X_2 = total assets / total liabilities
X_3 = gross result /total assets

X_4 = gross result /revenues from sales

Results of the classifications for analysed construction companies were presented in Table 5.

Table 5. Results of Mączyńska and Zawadzki's model (2006) for construction companies

Nazwa spółki	2012-12-31	2011-12-31	2010-12-31	2009-12-31	2008-12-31	2007-12-31
RAZEM BUDOWNICTWO	**-1,573**	**4,180**	**5,065**	**5,680**	**5,648**	**5,652**
BUDIMEX	3,746	4,205	4,435	4,270	4,386	3,676
POLIMEXMS	-1,908	4,672	5,273	5,458	5,094	5,757
PBG W UPADŁOŚCI	-9,976	5,089	5,735	6,268	6,233	5,620
MOSTALWAR	2,114	2,295	5,428	6,620	5,699	5,392
TRAKCJA	4,516	4,523	8,132	12,948	7,001	5,314
HBPOLSKA W UPADŁOŚCI	-15,665	4,082	4,308	4,954	4,088	3,373
ERBUD	4,648	3,640	5,544	6,625	4,845	7,014
ELBUDOWA	7,603	7,821	9,369	12,352	11,229	6,526
MOSTALZAB	5,694	5,785	5,067	6,530	5,991	10,642
POLAQUA	-0,572	5,685	-2,038	3,862	6,271	8,433
DSS W UPADŁOŚCI	-1,538	-7,060	3,423	2,743	3,739	
ABMSOLID W UPADŁOŚCI	-6,833	0,555	3,832	4,720	5,379	7,833
INSTALKRK	11,582	9,597	11,129	11,202	9,719	8,656
AWBUD	1,596	4,238	4,178	5,730		
PROJPRZEM	17,318	12,012	7,662	14,656	14,937	21,941
BIPROMET	9,161	6,747	9,388	8,577	6,524	
INTAKUS W UPADŁOŚCI	0,705	0,338	4,614	4,337	5,126	
UNIBEP SA	4,519	5,096	6,065	6,614	7,421	4,376
ZUE SA	5,165	6,601	5,779	5,353		
ULMA CONSTRUCCION POLSKA SA	7,385	7,860	4,547	3,606	6,002	10,395
ENERGOMONTAŻ-POŁUDNIE SA	-20,046	2,478	4,588	4,189	5,723	7,509
P.A. NOVA SA	7,136	10,884	10,145	19,106	15,249	28,407
INTERBUD-LUBLIN SA	3,946	5,938	5,862	6,141		
PROCHEM SA	9,588	8,702	10,547	9,209	7,036	7,629
MOSTOSTAL-EXPORT SA	5,227	-6,439	11,508	7,100	15,700	7,596
ELEKTROTIM SA	9,490	11,903	13,395	20,744		
CNT	7,378	6,984	8,798	2,851	2,851	14,682
ENERGOAPARATURA SA	7,041	7,029	10,245	6,950	5,361	3,877
BUDOPOL-WROCŁAW SA	-3,620	8,985	9,905	12,143	10,520	9,155
RESBUD SA	-5,748	-1,565	5,515	3,650	11,469	4,562
BANKRUPT RISK Z<0			GOOD Z>=0			

Source: own calculations

The discriminant function has the following form:

$$Z = 9{,}498X_1 + 3{,}566X_2 + 2{,}903X_3 + 0{,}452X_4 - 1{,}498 \quad (8)$$

The boundary value in this model is zero, and if Z < 0, a company is threatened. The decisive ratio in the assessment of the condition of construction companies was an inverse debt ratio (X_2). In case of companies that experience significant problems with debt repayment also ratio (X_1) –coverage of liabilities with financial surplus – had large impact on the final result of given company assessment. The model pointed out to certain bankrupt companies as threatened companies with certain delay since in 2012 only. Results for 2011 did not point out to any threat. The only exception here is DSS company that was undergoing bankruptcy proceedings in 2011 already.

SUMMARY

After an examination of financial data of 30 construction companies for 2007-2012 it should be stated that 2012 was the most difficult year for this

industry. The majority of companies recorded losses. Preparations for Euro 2012 were the cause of problems and bankruptcies of many enterprises. It should be noted that companies experiencing serious problems recorded also a significant increase of the ratio of purchased external services to revenues from sales, that is from the level of ca. 50% in years 2007-2011 up to 75% in 2012. Thus, problems of companies could also result from a resignation from own construction works in order to earn margins on purchased services.

The analysis of financial threats of examined companies points out to the fact that application of individual discriminant models does not ensure clear assessment of their economic condition. It happens that a model uses ratios that in combination with the others generate a negative or positive impact on assessment of a company. It may completely change perception of the financial condition of the same enterprise. Every analysed discriminant function is based on a different set of ratios and it analyses – in a better or worse way – the state of finances of construction companies. Three models: Hołda's, Sojak – Stawicki's and Mączyńska – Zawadzki's pointed out to companies threatened with bankruptcy too late since only in the year when such bankruptcy was announced. Indeed, it is a specific feature of construction companies that they "settle construction contracts" in compliance with IAS 11, while the majority of companies recognize revenues and costs on the basis of IAS 18. The specific feature of IAS 11 is the fact that non-invoiced revenues resulting from actual progress of construction works in relation to a contract value are recorded as revenues and costs. It results in the fact that revenues and costs are recognized in a given reporting period although they would be recognized in the following periods if the conditions were different. Thus, bad results of numerous companies in 2012 refer – at least partially – to construction contracts that will be finally completed in the upcoming years.

When assessing a given discriminant model one should focus on the history of financial results of a given company in the previous years. Only such analysis may point out to long-term factors determining company's operations that may result in financial problems in the future.

REFERENCES

Jajuga K., Walesiak M. (1998) Klasyfikacja i analiza danych. Teoria i zastosowania, AE, Wrocław.

Kisielińska J., Waszkowski A. (2010) Polskie modele do prognozowania bankructwa przedsiębiorstw i ich weryfikacja, EiOGŻ nr 82, Wydawnictwo SGGW, Warszawa.

Lichota W. (2009) Metody wczesnego ostrzegania o zmianach sytuacji finansowej przedsiębiorstw, Wiadomości Statystyczne nr 10, Warszawa.

Gajdka J., Stos D.(1996) Wykorzystanie analizy dyskryminacyjnej w ocenie kondycji finansowej przedsiębiorstw, praca zbiorowa pod red. R. Borowieckiego: Restrukturyzacja w procesie przekształceń i rozwoju przedsiębiorstw, AE, Kraków.

Hamrol M., Chodakowski J. (2008) Prognozowanie zagrożenia finansowego przedsiębiorstwa. Wartość predykcyjna polskich modeli analizy dyskryminacyjnej, Badania Operacyjne i Decyzje nr 3, Wrocław.

Hołda A. (2001) Prognozowanie bankructwa jednostki w warunkach gospodarki polskiej z wykorzystaniem funkcji dyskryminacyjnej Z_H, Rachunkowość nr 5.

Mączyńska E. (1994) Ocena kondycji przedsiębiorstwa (uproszczone metody), Życie gospodarcze nr 38.

Mączyńska E. Zawadzki M. (2006) Dyskryminacyjne modele predykcji bankructwa przedsiębiorstw, Ekonomista nr 2, Warszawa.

Prusak B. (2005) Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstwa, Difin, Warszawa.

Prusak B. (2004) Jak rozpoznać potencjalnego bankruta?, Wydawnictwo Politechniki Gdańskiej, Prace Naukowe Katedry Ekonomii i Zarządzania Przedsiębiorstwem, Tom 3, Gdańsk.

Sojak S. Stawicki J. (2001) Wykorzystanie metod taksonomicznych do oceny kondycji ekonomicznej przedsiębiorstw, Zeszyty Teoretyczne Rachunkowości, Tom 3, Warszawa.

Sukiennik M. (2007) Zastosowanie analizy dyskryminacyjnej do oceny stanu finansowego przedsiębiorstw, Referaty II KKMM, Kraków

The consolidated financial statements construction companies for the years 2007-2012.

QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 309 – 315

CONFIDENCE INTERVALS FOR FRACTION IN FINITE POPULATIONS: MINIMAL SAMPLE SIZE

Wojciech Zieliński
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
Department of the Prevention of Environmental Hazards and Allergology
Medical University of Warsaw
e-mail: wojciech_zielinski@sggw.pl

Abstract: Consider a finite population of N units. Let $\theta \in (0,1)$ denotes the fraction of units with a given property. The problem is in interval estimation of θ on the basis of a sample drawn due to the simple random sampling without replacement. It is of interest to obtain confidence intervals of a prescribed length. In the paper the minimal sample size which guarantees the length to not exceed the given value is calculated.

Keywords: confidence interval, sample size, fraction, finite population

Consider a population $\{u_1, \dots, u_N\}$ containing the finite number N units. Let M denotes an unknown number of objects in population which has an interesting property. We are interested in an interval estimation of M, or equivalently, the fraction $\theta = M/N$. The sample of size n is drawn due to the simple random sampling without replacement ($lpbz$ to be short). Let ξ_{bz} be a random variable describing a number of objects with the property in the sample. On the basis of ξ_{bz} we want to construct a confidence interval for θ at the confidence level δ. The main problem is to find minimal sample size n such that the expected length of the confidence interval is smaller than the given number $\varepsilon > 0$.
The random variable ξ_{bz} has the hypergeometric distribution [Johnson and Kotz 1969, Zieliński 2010]

$$P_{\{\theta,N,n\}}\{\xi_{bz} = x\} = \frac{\binom{\theta N}{x}\binom{(1-\theta)N}{n-x}}{\binom{N}{n}},$$

for integer x from the interval$\langle max\{0, n-(1-\theta)N\}, min\{n, \theta\, N\}\rangle$. Let $f_{\{\theta,N,n\}}(\cdot)$ be the probability distribution function, i.e.

$$f_{\{\theta,N,n\}}(x)=\begin{cases} P_{\{\theta,N,n\}}\{\xi_{bz}=x\}, & \text{for integer } x\in\langle \max\{0,n-(1-\theta)N\},\min\{n,\theta\}\rangle \\ 0, & \text{elsewhere,}\end{cases}$$

and let

$$F_{\{\theta,N,n\}}(x)=\sum\nolimits_{\{t\le x\}} f_{\{\theta,N,n\}}(t)$$

be the cumulative distribution function of ξ_{bz}.
The CDF of ξ_{bz} may be written as

$$1-\frac{\binom{n}{x+1}\binom{N-n}{\theta\, N-x-1}}{\binom{N}{\theta\, N}}\cdot \quad {}_3F_2[\{1, x+1-\theta\, N, x+1-n\},\{x+2,(1-\theta)N+x+2-n\};1],$$

where

$${}_3F_2[\{1, x+1-\theta\, N, x+1-n\},\{x+2,(1-\theta)N+x+2-n\};1]=\sum_{k=0}^{\infty}\frac{(a_1)_k(a_2)_k(a_3)_k}{(b_1)_k(b_2)_k}\frac{t^k}{k!}$$

and $(a)_k=a(a+1)\cdots(a+k-1)$.

A construction of the confidence interval at a confidence level δ for θ is based on the cumulative distribution function of ξ_{bz}. If $\xi_{bz}=x$ is observed then the ends $\theta_L=\theta_L(x,N,n,\delta_1)$ and $\theta_U=\theta_U(x,N,n,\delta_2)$ of the confidence interval are the solutions of the two following equations

$$F_{\{\theta_L,N,n\}}(x)=\delta_1, \qquad F_{\{\theta_U,N,n\}}(x)=\delta_2.$$

The numbers δ_1 and δ_2 are such that $\delta_2-\delta_1=\delta$. In what follows we take $\delta_1=\frac{1-\delta}{2}$ and $\delta_2=\frac{1+\delta}{2}$. Analytic solution is unavailable. However, for given x, n and N, the confidence interval may be found numerically. In the Table 1 there are given exemplary confidence intervals for $N=1000$ units, sample size $n=20$, confidence level $\delta=0.95$ and $\delta_1=0.025$.

Table 1. Confidence intervals for θ for $N=1000$, $n=20$, $\delta=0.95$

x	θ_L	θ_U	x	θ_L	θ_U	x	θ_L	θ_U
0	0.000	0.167	7	0.155	0.591	14	0.459	0.880
1	0.001	0.247	8	0.192	0.638	15	0.511	0.913
2	0.012	0.316	9	0.232	0.683	16	0.565	0.942
3	0.032	0.377	10	0.273	0.727	17	0.623	0.968
4	0.058	0.435	11	0.317	0.768	18	0.685	0.988
5	0.087	0.489	12	0.362	0.808	19	0.753	0.999
6	0.120	0.541	13	0.409	0.845	20	0.833	1.000

Source: own study

The real confidence level equals

$$conf_{N,n,\delta}(\theta) = \sum_{x=x_d}^{x_g} f_{\{\theta,N,n\}}(x)\,,$$

where

$$x_d = F^{-1}_{\{\theta,N,n\}}\left(1-\frac{\delta}{2}\right) \text{ and } x_g = F^{-1}_{\{\theta,N,n\}}\left(1+\frac{\delta}{2}\right).$$

Since the population is finite, the number of admissible values of θ is also finite. For example, for $x = 1$ admissible values of θ are $0.001, 0.002, \ldots, 0.247$. It means, that the number of units with the investigated property is one of $1, 2, \ldots$ or 247. Consider the length of the confidence interval

$$d(\xi_{bz}, N, n, \delta) = \theta_U(\xi_{bz}, N, n, \frac{1+\delta}{2}) - \theta_L(\xi_{bz}, N, n, \frac{1-\delta}{2}).$$

It is easy to note, that the length depends among others on the population size N and the sample size n.

Let $\varepsilon > 0$ be a given number. We are going to find minimal sample size n (for a given population size N) such that the length of the confidence interval is smaller than ε. More precisely, we are going to find minimal sample size n such that the expected length of confidence interval covering θ is smaller than ε, i.e.

$$L(\theta, N, n) = E_{\theta,N,n}\left(d(\xi_{bz}, N, n, \delta)1_{\left(\theta_L\left(\xi_{bz},N,n,\frac{1-\delta}{2}\right),\theta_U\left(\xi_{bz},N,n,\frac{1+\delta}{2}\right)\right)}(\theta)\right) \leq \varepsilon,$$

where

$$1_A(z) = \begin{cases} 1, & z \in A \\ 0, & z \notin A \end{cases}$$

Note that the expected length may be written as

$$L(\theta, N, n) = \sum_{x=x_d}^{x_g} d(x, N, n, \delta) f_{\theta,N,n}(x).$$

For given θ, N and ε the inequality

$$L(\theta, N, n) \leq \varepsilon$$

may be solved numerically. Exemplary solutions for $\theta = 0.05$ and $\varepsilon = 0.02$ are given in the Table 2.

Table 2. Minimal sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$

N	n_{min}	conf. level	Length	N	n_{min}	conf. level	length
500	410	0.970508	0.019380	5500	1352	0.956080	0.019994
1000	661	0.954932	0.019901	6000	1369	0.952244	0.019989
1500	836	0.957584	0.019865	6500	1401	0.955592	0.019990
2000	966	0.960184	0.019949	7000	1416	0.952569	0.019998
2500	1061	0.958530	0.019992	7500	1443	0.956406	0.019999
3000	1136	0.953237	0.019902	8000	1461	0.953642	0.019963
3500	1188	0.951010	0.019947	8500	1465	0.952202	0.019993
4000	1238	0.950565	0.019962	9000	1489	0.956288	0.019999
4500	1286	0.951242	0.019900	9500	1497	0.955144	0.019982
5000	1316	0.953628	0.019997	10000	1511	0.953024	0.019932

Source: own study

Note that the expected length is smaller than given ε. To obtain the expected length equal exactly ε a randomization is needed. This randomization is between sample sizes. Note that

$$L(\theta, N, n) \leq \varepsilon \leq L(\theta, N, n-1).$$

Let us choose in random the sample size:

$$sample\ size = \begin{cases} n-1, & with\ probability\ \gamma, \\ n, & with\ probability\ 1-\gamma, \end{cases}$$

where $\gamma = \frac{\varepsilon - L(\theta,N,n)}{L(\theta,N,n-1) - L(\theta,N,n)}$.

The expected length of the confidence interval after such randomization equals

$$\gamma\, L(\theta, N, n-1) + (1-\gamma) L(\theta, N, n) = \varepsilon.$$

For example, if $N = 1000$, $\varepsilon = 0.02$, $\delta = 0.05$ and $\theta = 0.05$ then

$$L(0.05,1000,661) = 0.019901\ and\ \ L(0.05,1000,660) = 0.020133.$$

Choosing $\gamma = 0.427158$ and applying the rule

$$sample\ size = \begin{cases} 660, & with\ probability\ 0.427158, \\ 661, & with\ probability\ 0.572842, \end{cases}$$

we obtain the confidence interval with the mean length equal to the prescribed accuracy.

In the Table 3 are given randomized minimum sample sizes. In the last column of the Table the probability γ is given.

Table 3. Randomized minimal sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$.

N	n_{min}	conf. level	length	$n_{min} - 1$	conf. level	length	γ
500	410	0.970508	0.019380	409	0.986565	0.020130	0.826673
1000	661	0.954932	0.019901	660	0.967684	0.020133	0.427158
1500	836	0.957584	0.019865	835	0.968619	0.020139	0.492457
2000	966	0.960184	0.019949	965	0.960078	0.020030	0.630721
2500	1061	0.958530	0.019992	1060	0.96769	0.020188	0.038677
3000	1136	0.953237	0.019902	1135	0.96203	0.020066	0.599829
3500	1188	0.951010	0.019947	1187	0.959462	0.020158	0.250663
4000	1238	0.950565	0.019962	1237	0.958750	0.020112	0.253571
4500	1286	0.951242	0.019900	1285	0.959274	0.020063	0.612891
5000	1316	0.953628	0.019997	1315	0.953633	0.020025	0.123186
5500	1352	0.956080	0.019994	1351	0.956157	0.020012	0.337816
6000	1369	0.952244	0.019989	1368	0.952291	0.020020	0.360976
6500	1401	0.955592	0.019990	1400	0.955660	0.020008	0.551318
7000	1416	0.952569	0.019998	1415	0.952588	0.020015	0.124424
7500	1443	0.956406	0.019999	1442	0.956399	0.020024	0.047004
8000	1461	0.953642	0.019963	1460	0.960810	0.020094	0.281286
8500	1465	0.952202	0.019993	1464	0.952217	0.020001	0.921897
9000	1489	0.956288	0.019999	1488	0.956253	0.020003	0.232190
9500	1497	0.955144	0.019982	1496	0.955181	0.020014	0.554152
10000	1511	0.953024	0.019932	1510	0.960018	0.020093	0.420791

Source: own study

Normal approximation. The hypergeometric distribution is analytically untractable, so in many applications the normal approximation of the distribution is used. The hypergeometric distribution, by Central Limit Theorem, may be approximated by a normal distribution with mean and variance equal to mean and variance of ξ_{bz}, i.e by the distribution $N(n\theta, \frac{N-n}{N-1}\, n\theta(1-\theta))$. To do so the rule of thumb $N\theta \geq 4$ is sometimes used [Bracha 1996], [Karliński 2003], [Zieliński 2010].

In practice two approaches are met. In the first one the confidence interval is obtained as a solution of

$$\left| \frac{\xi_{bz} - n\theta}{\sqrt{\frac{N-n}{N-1}\xi_{bz}(n-\xi_{bz})}} \sqrt{n} \right| \leq z_{\frac{1+\delta}{2}},$$

where z_q denotes the q-th quantile of the distribution $N(0,1)$. The confidence interval takes on the form

$$\theta_L^N = \left(\frac{\xi_{bz}}{n}\right) - e, \theta_U^N = \left(\frac{\xi_{bz}}{n}\right) + e$$

and

$$e^2 = \frac{z_{\frac{1+\delta}{2}}^2}{n^3} \xi_{bz}(n - \xi_{bz}) \left(\frac{N-n}{N-1}\right).$$

In the second approach the confidence interval is obtained as a solution of

$$\left| \frac{\xi_{bz} - n\theta}{\sqrt{\frac{N-n}{N-1}\theta \quad (n-\theta)}} \right| \leq z_{\frac{1+\delta}{2}},$$

The confidence interval ($z = z_{\frac{1+\delta}{2}}\sqrt{\frac{n(N-n)}{N-1}}$) has the form

$$\theta_L^N = \frac{z^2 + 2n\xi_{bz} - z\sqrt{z^2 + 4(n-\xi_{bz})\xi_{bz}}}{2(n^2+z^2)}$$

$$\theta_L^N = \frac{z^2 + 2n\xi_{bz} + z\sqrt{z^2 + 4(n-\xi_{bz})\xi_{bz}}}{2(n^2+z^2)}$$

Table 4. First normal approximation of sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$.

N	n_{min}	conf. level	length	N	n_{min}	conf. level	length
500	385	0.950162	0.019979	5500	1267	0.944224	0.019930
1000	624	0.949319	0.019965	6000	1294	0.948367	0.020000
1500	786	0.942629	0.019891	6500	1317	0.944391	0.019899
2000	906	0.949799	0.019976	7000	1320	0.941959	0.019994
2500	989	0.939873	0.019903	7500	1349	0.946901	0.019989
3000	1064	0.946046	0.019966	8000	1368	0.943899	0.019895
3500	1116	0.944653	0.019993	8500	1370	0.942252	0.019959
4000	1166	0.944888	0.019952	9000	1394	0.947670	0.019984
4500	1204	0.946749	0.019986	9500	1400	0.946198	0.019989
5000	1243	0.949173	0.019993	10000	1419	0.943788	0.019869

Source: own study

Table 5. Second normal approximation of sample sizes for $\theta = 0.05$, $\varepsilon = 0.02$, $\delta = 0.95$.

N	n_{min}	conf. level	length	N	n_{min}	conf. level	length
500	384	0.948488	0.019988	5500	1261	0.939422	0.019866
1000	619	0.936714	0.019932	6000	1286	0.942464	0.019930
1500	790	0.935830	0.019619	6500	1310	0.939736	0.019843
2000	901	0.946060	0.019987	7000	1329	0.947937	0.019996
2500	989	0.934991	0.019777	7500	1336	0.941714	0.019980
3000	1063	0.939874	0.019824	8000	1360	0.939704	0.019852
3500	1113	0.939006	0.019889	8500	1361	0.938450	0.019935
4000	1162	0.939416	0.019860	9000	1385	0.942431	0.019925
4500	1195	0.943374	0.019998	9500	1386	0.941560	0.019991
5000	1236	0.942720	0.019908	10000	1409	0.939789	0.019844

Source: own study

Comparing minimal sample sizes obtained in the exact solution (Table 2) to normal approximations (Tables 4 and 5) it is seen that application of approximate

confidence intervals needs smaller sample sizes. But the confidence level of approximate confidence intervals does not keep the nominal confidence level, so the risk of wrong conclusions is too high (greater than nominal).

In some applications the Binomial approximation to the hypergeometric distribution is applied. Comparison of all approximations may be found in [Zieliński 2011].

REFERENCES

Bracha Cz. (1996) Teoretyczne podstawy metody reprezentacyjnej, PWN, Warszawa.

Johnson N. L., Kotz S. (1969) Discrete distributions: distributions in statistics, Houghton Mifflin Company, Boston.

Karliński W. (2003) Nowe techniki w kontroli wykonania budżetów państwa, Kontrola Państwowa 5/2003, 101-124.

Zieliński W. (2010) Estymacja wskaźnika struktury, Wydawnictwo SGGW, Warszawa.

Zieliński W. (2011) Comparison of confidence intervals for fraction in finite populations, Metody Ilościowe w Badaniach Ekonomicznych XII, 177-182.