PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

Strona WWW „Przegląd Statystyczny":
http://www.przegladstatystyczny.pan.pl

SPIS TREŚCI

CONTENTS

STANISLAW MACIEJ KOT[1]

EQUIVALENCE SCALES
BASED ON STOCHASTIC INDIFFERENCE CRITERION: THE CASE OF POLAND[2]

1. INTRODUCTION

The aim of the paper is to offer economists the stochastic equivalence scale (*SES*) as a new tool for making comparisons with respect to inequality, welfare and poverty in a heterogeneous population of households. The definition of the *SES* is 'axiomatic' in the sense that it only postulates the properties of a function that can be recognised as an *SES*. Any function can be considered an *SES* if and only if it transforms the distribution of expenditures of given groups of households in such a way that the resulting distribution is *stochastically indifferent* to the distribution of the expenditures of a reference group of households. Here, the property of stochastic indifference is a criterion of the homogeneity of transformed expenditure distributions. This criterion is also used to develop a method to estimate the *SES*.

This paper was motivated by serious deficiencies in recent solutions to the problem of addressing inequality, welfare and poverty when households differ in attributes other than their expenditures; e.g., other relevant differences include households' sizes and demographic compositions and household members' disabilities. When heterogeneity among the households exists, a two-step procedure has traditionally been applied. In the first step, a reference household group and an equivalence scale are chosen. Then, the actual expenditures for individual groups of households are adjusted by the equivalence scale (Buhmann et al., 1988; Jones, O'Donnell, 1995). In the second stage, standard measures of inequality, welfare, and poverty are applied to the adjusted distribution. These stages are treated separately.

However, there are two serious reasons why the two-stage procedure is unsatisfactory. First, the homogenisation stage does not provide unambiguous results. Second, the two stages seem to be interdependent.

The ambiguity of the first stage is because there is no single 'correct' equivalence scale for adjusting expenditures or incomes (Coulter et al., 1992a). Jäntti,

Danziger (2000, p. 319) remark that 'there is no optimal method for deriving an equivalence scale'. In fact, many serious identification issues arise in the estimation of equivalence scales (see, in particular, Pollak, Wales, 1979, 1992; Blundell, Lewbel, 1991; Blackorby, Donaldson, 1993 and the surveys of Lewbel, 1997, and Slesnick, 1998). Indeed, without additional assumptions, there is no way of selecting an appropriate basis with which to choose the equivalence scale. The *independence of base* (IB) (or the *exactness of equivalence* scale) is one such assumption. Several papers have tested this assumption, but they ultimately rejected it (Blundell, Lewbel, 1991; Blundell et al., 1998; Dickens et al., 1993; Pashardes, 1995; Gozalo, 1997; Pedankur, 1999).

If equivalence scales are arbitrary or if the IB assumption is not fulfilled, then the equivalent expenditure distribution cannot be unambiguously assigned to the chosen reference household group. Thus, the resulting population of fictitious equivalent units cannot be treated as homogeneous with respect to the selected attribute of households.

The second objection concerns the interdependence of the two stages of the traditional method of studying households (Ebert, Moyes, 2003). There is evidence that the results of distributional comparisons are sensitive to the choice of equivalence scale (Coulter et al., 1992a, b).

All of these problems with the equivalence-scale approach have encouraged economists to search for alternative solutions. Atkinson, Bourguignon (1987) propose the sequential Lorenz dominance approach for comparing the living standards in populations that include diverse incomes and needs. However, Ebert, Moyes (2003) note that this approach has only had limited empirical success; studies seem to favour the conventional equivalence scale for taking a family's circumstances into account.

Donaldson, Pendakur (2003) propose a method that is based on the concept of the *equivalent expenditure function*. Following this approach, Ebert, Moyes (2003) suggest a normative method for adjusting household incomes, and this method accounts for the heterogeneity of income recipients' needs when measuring inequality and welfare.

The stochastic equivalence scale transfers the problem of homogenisation from an individual person's perspective to the distributional level, i.e., the object of adjustment is the probability distribution of expenditures rather than an individual household's expenditures. The *SES* makes an initially heterogeneous population of households homogeneous with respect to such distributional features as social welfare, inequality and poverty.

The rest of this paper is organised as follows: Section 2 contains the definition of stochastic equivalence scales and section 3 explains the relationship between *SES*s and welfare, inequality and poverty. Section 4 introduces pragmatic equivalence scales, which can be considered potential *SESs*. Section 5 gives a statistical test to verify whether a particular function can be recognised as an *SES* and a method for estimating *SESs*. Section 6 contains the empirical results of estimating non-parametric and parametric *SESs* for Poland in the years 2005–2010. Lastly, section 7 concludes.

2. THE CONCEPT STOCHASTIC EQUIVALENCE SCALES

The following 'paradigmatic' issue that arises during welfare comparisons addresses the problem with equivalence scales: it is difficult to determine how much money (x) an m-person household would need to be as well off as a single person who spends y each year. If the equivalence scale were of a relative type with a known deflator d, the answer to such a question would be $x = d \cdot y$. However, the arbitrariness of equivalence scales, which is discussed in the previous section, makes this equation unsolvable.

Donaldson, Pendakur (2003) propose the concepts of equivalent expenditure and an equivalence expenditure function for making welfare comparisons. Equivalent expenditure is the expenditure level which would make a single adult as well off as an m-member household; it may be written as a function of prices, expenditures and household characteristics. The corresponding equivalence scale is the actual expenditures divided by the equivalent expenditures. Similarly, the equivalent-expenditure function transforms the actual expenditures into equivalent expenditures. Donaldson, Pendakur (2003) maintain that: 'For welfare purposes, equivalent-expenditure functions permit the conversion of an economy with many household types into an economy of identical single individuals'. These authors show that, under certain conditions, equivalent-expenditure functions and their associated expenditure-dependent equivalence scales can be uniquely estimated from demand data. Following Donaldson and Pendakur's approach, Ebert, Moyes (2003) propose the concepts of equivalent income and the equivalent-income function.

The attractiveness of the equivalent-expenditure and equivalent-income concepts is due to the fact that their underlying assumptions are less demanding than the assumptions that are required by the equivalence-scale approach. However, some of these assumptions are not entirely convincing[3]. Although the new approach has resulted in many valuable theoretical achievements, some of the 'old' problems with equivalence scales still await solutions.

The concept of the stochastic equivalence scale (*SES*) offers an alternative potential solution to the problem of homogenising a heterogeneous population of households[4]. Let $h = [h_0, h_1, h_2, ..., h_m]$ be the vector of an $m + 1$-level household attribute other than expenditures where $m + 1 > 2$, e.g., a household's size or its demographic composition would be a suitable attribute. Suppose that the population of all households (which will henceforth be called the total population) is divided into $m + 1$ subpopulations according to a certain attribute. Let the h_0-attribute subpopulation be chosen as the reference household group and let the continuous random variable Y with the distribution

---

[3] For example, Ebert, Moyes (2003) notice that the *Between-Type Transfer Principle* rules out using *utilitarianism* to make a relevant social judgement in their model. See also Capéau, Ooghe (2007).

[4] Early version of the SES was presented in Kot (2012). However, the link between the SES and stochastic indifference was not mentioned in that paper.

function $G(y)$ (which we abbreviate as $Y \sim G(y)$[5]) represent the expenditure distribution of this reference subpopulation. Henceforth, Y will be called 'the reference distribution'. Let the set of m *continuous* random variables $X_i \sim F_i(x)$, $i = 1,...,m$, represent the expenditure distributions of the remaining m household subpopulations. From this point forwards, X_i will be called 'the evaluated distribution'.

Without loss of generality, the non-negative real-valued interval $[0,\infty)$ will be used as the domain of the considered random variables. However, the definition of the *SES* presented below would also be valid for all real numbers.

Suppose that $s(\cdot) = [s_1(\cdot),...,s_m(\cdot)]$ is a continuous and strictly monotonic real-valued vector function for which the inverse function $s^{-1}(\cdot) = [s_1^{-1}(\cdot),...,s_m^{-1}(\cdot)]$ exists, and suppose that this function is differentiable[6]. Let the random variable $Z_i = s_i(X_i) \sim H_i(z)$ be the transformation of the evaluated expenditure distribution X_i. Henceforth, the random variable $Z_i \sim H_i(z)$ will be called the 'transformed expenditure distribution'.

Denote by $Z \sim H(z)$ the following mixture of cumulative distribution functions $H_i(z)$:

$$H(z) = \sum_{i=1}^{m} \pi_i H_i(z), \text{ for all } z \geq 0, \tag{1}$$

where the weights π_i satisfy two conditions: $\forall i = 0,1,...,m$, $\pi_i > 0$, and $\sum_{i=1}^{m} \pi_i = 1$. A single π_i weight can be interpreted as the proportionate size of the ith subpopulation relative to the size of the total population[7].

Definition 2.1. With the above notations, the function $s(\cdot)$ will be called the *stochastic equivalence scale* (*SES*) if and only if the following equation holds:

$$\forall z \geq 0, \ H(z) = G(z). \tag{2}$$

When the function $s(\cdot)$ is an *SES*, the transformed expenditure distributions Z will be called 'the equivalent expenditure distributions'.

We call the defined equivalence scales 'stochastic' to underline the fact that they transform random variables. 'Classical' equivalence scales can be called 'individualistic' because they transform individual expenditures (or incomes) in pairs.

The above definition of an *SES* is axiomatic in the sense that it only postulates the criterion for a function to be recognised as an *SES*. This definition does not describe how an *SES* should be constructed or the conditions of its existence. In other words, *any* function $s(\cdot)$ that fulfils condition (2) has to be recognised as an *SES*.

---

[5]   We reserve capital letters for random variables and lowercase letters for the values that are taken by these variables.

[6]   These are standard conditions when transforming continuous random variables.

[7]   The number of persons or an equivalent unit is used when calculating π_i wages.

Naturally, the definition of an *SES* also applies when $m = 1$, i.e., when only one group of households is compared to the reference group. It is easy to see that condition (2) will automatically be fulfilled when the equation

$$H_i(z) = G(z) \tag{3}$$

holds for all $z \geq 0$ and for all $i = 1,...,m$. However, in general, equality (2) does not imply equality (3).

The relative *SES* can be defined as follows: let $d = [d_i]$, $i = 1,...,m$, be the vector of positive numbers called 'deflators' that transform the evaluated expenditure distributions $X_1,...,X_m$ thusly:

$$Z_i = X_i/d_i \sim H_i(z), \ i = 1,...,m. \tag{4}$$

Definition 2.2. Under the above notations, the vector d will be called the *relative SES* if and only if the deflators $d_1,...,d_m$ are such that equality (2) holds.

The concept of the *SES* has several advantages that make it an interesting alternative to the equivalence scales that have been developed so far. Some of these advantages are discussed below.

The validation of condition (2) can be tested using nonparametric statistical tests for equality between cumulative distribution functions.

As potential *SESs*, parametric and nonparametric functions of $s(\cdot)$ can easily be estimated on the basis of expenditure data[8]. Details of the statistical procedures that are used to test and estimate an *SES* are presented in section 5. It is worth adding that the usual procedure of extracting equivalence scales from the estimated model of a demand system is not necessary when estimating an *SES*.

Estimating an *SES* can potentially be useful when the relationship between household needs and socio-demographic attributes is ambiguous. It is generally accepted that the larger a household is, the greater its needs are. However, the classification of household types with respect to household needs may be problematic when more than a single attribute must be accounted for, e.g., a household including a single mother who has two children or a household with two parents and one disabled child.

The *SES* offers a solution to this problem. Assume that there are m groups of evaluated households that are selected with respect to certain criteria, but assume that these households are not necessarily selected with respect to their needs. Let the reference group also be selected. We can apply the nonparametric relative *SES* (4) with m deflators $d_1,...,d_m$. Let the estimates of these deflators be arranged in ascending order, i.e., $\bar{d}_{(1)},...,\bar{d}_{(m)}$. Then, these deflators will rank the evaluated household groups according to their needs.

---

[8] The parametric and nonparametric equivalence scales are defined in section 3.

3. THE RELATION BETWEEN AN *SES* AND WELFARE, INEQUALITY AND POVERTY

There are well known relations between stochastic dominance and economic inequality and between welfare and poverty (e.g., Davidson, 2008). In contrast, the concept of an *SES* satisfies the criterion of stochastic indifference as a symmetric factor of stochastic dominance. Hence, one can find the relation between an *SES* and the above-mentioned aspects of income or expenditure distributions.

Consider two non-negative[9] distributions of incomes or expenditures X_A and X_B and suppose that they are characterised by the cumulative distribution functions $F_A(x)$ and $F_B(x)$, respectively. Distribution X_B stochastically dominates distribution X_A in the first order if $F_A(x) \geq F_B(x)$ for any argument x (Davidson, 2008).

Higher orders of stochastic dominance can be defined in the following recursive way. Let $D^1(x) = F(x)$ and let $D^{s+1}(x) = \int_0^x D^s(t)dt$, for $s = 1,2,3,\ldots$ Distribution X_B dominates distribution X_A in order s if $D_A^s(x) \geq D_B^s(x)$ for all arguments x (Davidson, 2008).

It is easy to see that first-order stochastic dominance implies dominance in all higher orders. More generally, dominance in order s implies dominance in all orders higher than s (Davidson, 2008). However, it is not true that dominance in all orders higher than s implies dominance in order s.

Suppose that z is the poverty line in terms of incomes or expenditures. Then, $F(z)$ is the headcount ratio that measures the amount of poverty in a given distribution. The headcount ratio is higher in X_A than it is in X_B if and only if $F_A(x) > F_B(x)$ for all $x < z$.

We define the poverty gap for an individual with income x as $g(z,x) = max(z-x,0)$. Furthermore, we define the class of poverty indices over the poverty gaps as follows:

$$\Pi(z) = \int_0^z \pi(g(z,x))dF(x) \qquad (5)$$

Atkinson (1987).

It can be shown that, in the case of all indices (5) where π is differentiable and $\pi(0) = 0$, $\Pi_A(x) \geq \Pi_B(x)$ occurs if and only if X_B stochastically dominates X_A up to z in the first order for all $x \leq z$ (Atkinson, 1987; Foster, Shorrocks, 1988; McFadden, 1989). This class of indices and the corresponding headcount ratio will be denoted as P^1. Similarly, class P^2 is defined by convex increasing functions π where $\pi(0) = 0$. It can be shown that all of the indices in P^2 are greater for X_A than for X_B if and only if X_B stochastically dominates X_A up to the second order for all $x \leq z$. In general, we can define class P^s in any order s such that it contain indices (5) with the following properties: $\pi^{(s)}(x) \geq 0$ for $x > 0$, $\pi^{(s-1)}(0) \geq 0$, and $\pi^{(i)}(0) = 0$ for $i = 0,\ldots,s-2$. Then, $\Pi_A(x) \geq \Pi_B(x)$ for all $x \leq z$ and for all $\Pi \in P^s$ if and only if X_B stochastically dominates X_A up to z in order s (Davidson, Duclos, 2000).

---

[9]   We follow our restriction on the expenditure distributions' domain. In general, distribution functions are defined for all real arguments.

Stochastic dominance also allows welfare comparisons. Let U_1 denote the class of all von Neumann-Morgenstern-type utility functions u where $u' \geq 0$ (i.e., the function is increasing). Additionally, let U_2 denote the class of all utility functions in U_1 for which $u'' \leq 0$ (this condition implies strict concavity). Then, social welfare in distribution $X \sim F(x)$ can be defined as follows:

$$E[u(X)] = \int_0^\infty u(x)dF(x). \tag{6}$$

It can be seen that X_B implies that more social welfare will be awarded than does X_A if and only if X_A is stochastically dominated by X_B in the first order for all $u \in U_1$ and for all social welfare functions that have the form (6). When $u \in U_2$, all of the social welfare functions of this more restrictive class give an unambiguous ranking of two distributions if one dominates the other in the second order (Davidson, 2008).

A relation between stochastic dominance and inequalities can be obtained by means of Lorenz curves. Generalised Lorenz dominance is based on the generalised Lorenz curve (Shorrocks, 1983). Assuming $u \in U_2$, generalised Lorenz curves provide an unambiguous ranking of two distributions by means of the social welfare function (6). Generalised Lorenz dominance appears to be exactly the same as second-order stochastic dominance (Davidson, 2008).

All of these features of stochastic dominance also apply to stochastic indifference. We say that distribution X_A is stochastically indifferent to distribution X_B in the first order if and only if the following identity holds:

$$F_A(x) = F_B(x), \text{ for all } x \geq 0. \tag{7}$$

If we integrate both sides of identity (7) from 0 to x, we will obtain stochastic indifference in the second order, i.e.:

$$\int_0^x F_A(t)dt = \int_0^x F_B(t)dt, \text{ for all } x \geq 0. \tag{8}$$

Recursive integrations result in stochastic indifference in all higher orders. Thus, stochastic indifference in the first order implies that all higher orders are stochastically indifferent. Moreover, this implication also works in reverse. This fact can be shown by differentiating the integrals we have already obtained. Hence, stochastic indifference provides stronger results than stochastic dominance.

Condition (2) means that an *SES* guarantees the first-order stochastic indifference between the evaluated distribution and the reference distribution. This indifference implies that there is stochastic indifference at all higher orders, and at all lower orders.

All of the above considerations justify the following corollary:

Corollary 3.1.

Let X be the distribution of expenditures of the evaluated group of households, let Y be the distribution of expenditures of the reference group of households, and let $Z = s(X)$. If s is the corresponding *SES*, then the following equivalent conditions hold:
a) Z is stochastically indifferent to Y
b) Social welfare (6) in Z is exactly the same as in Y for all $u \in U_2$
c) Poverty in Z is exactly the same as poverty in Y for all poverty lines
d) Inequalities in Z are exactly the same as the corresponding inequalities in Y.

One may ask what type of homogeneity an *SES* provides. If an initial heterogeneous population of households consists of $m + 1$ subpopulations (including a reference subpopulation), then the adjustment of each m distinct expenditure distributions by the *SES* will result in new fictitious subpopulations that are homogeneous with respect to utilitarian social welfare, inequality and poverty.

Another advantage of the *SES* is related to Ebert, Moyes' (2003) opinion that the choice of equivalence scale and additional social judgements cannot be treated as two independent issues. Stochastic dominance appears as the base for the *SES* as well as for the measurement of inequality, social welfare, and poverty. Therefore, the *SES* links the problem of the homogenisation of a heterogeneous population of households with normative judgements.

4. PRAGMATIC EQUIVALENCE SCALES

Practitioners have developed various forms of equivalence scales, despite theoretical controversies (Coulter et al., 1992b). Indeed, from the theoretical point of view, these 'pragmatic' scales are arbitrary. However, even with this reservation, we can still treat them as potential *SES*s.

Several nonparametric and parametric pragmatic scales are in common use. Suppose that the total population of households is divided into $m + 1$ subpopulations and that one of them plays the role of a reference group. The nonparametric equivalence scale is the set $\{d_1,...,d_m\}$ of deflators (4), which are equal to the number of equivalent units that are attached to the ith group of households, $i = 1,...,m$. These deflators can be estimated by the method that is presented in section 5.

Jenkins, Cowell (1994) describe the parametric equivalence scale class as '…a set of scales sharing a common functional form and for which parametric variations change the scale rate relativities for households of a different type'. Following this definition, we will use the term *parametric SES* when the m deflators $d_1,...,d_m$ in equation (4) are a certain function of household attributes with several parameters. We denote this function as $d = d(\boldsymbol{h},\boldsymbol{\theta})$, where $\boldsymbol{h}$ is the vector of household attributes and $\boldsymbol{\theta}$ is the vector of the parameters. We denote the transformed distribution of expenditures as $Z = X / d(\boldsymbol{h},\boldsymbol{\theta})$, or equivalently, in the abbreviated form of $Z = X / d$, $Z \sim H(z)$,

where X is the distribution of the expenditures of the evaluated group of households. According to definitions 2.1 and 2.2, this transformation will be the relative parametric *SES* if and only if $H(z) = G(z)$ for all $z \geq 0$, where $Y \sim G(y)$ describes the expenditure distribution of the reference group.

Certain forms of the parametric deflator $d(\boldsymbol{h},\boldsymbol{\theta})$ are especially popular in practical applications. The *power deflator* has the following form:

$$d = h^{\theta_1}, \; 0 \leq \theta_1 \leq 1, \tag{9}$$

where h is the household size (Buhmann et al., 1988).

The parameter θ_1 is usually set arbitrarily. The per capita (with $\theta_1 = 1$) and square root (with $\theta_1 = 0.5$) equivalence scales appear to be the most popular equivalence scales (OECD, 2008).

Let a and k denote the numbers of adults and children, respectively. The *A-C* ('adults-children') deflator is defined as follows:

$$d = (a + \theta_1 k)^{\theta_2}, \; \theta_1, \; \theta_2 > 0, \tag{10}$$

where θ_1, θ_2 are parameters (Coulter et al., 1992a, and independently, Cutler, Katz, 1992).

According to Cutler, Katz (1992), θ_1 is a constant reflecting the resource cost of a child relative to an adult. Parameter θ_2 reflects the overall economies of scale with respect to households' sizes. Jenkins, Cowell (1994) refer to $(a + \theta_1 k)$ as the 'effective household size'.

The *OECD-type* deflator can be defined as follows:

$$d = 1 + \theta_1 \cdot (a - 1) + \theta_2 \cdot k, \; \; \theta_1, \; \theta_2 > 0, \tag{11}$$

where θ_1 and θ_2 are parameters.

The OECD scale assigns a value of 1 to the first adult and a value of θ_1 to every subsequent adult, whereas θ_2 is the weight that is attached to each child. The 'old OECD' scale (which is also referred to as the 'Oxford scale') assumes that $\theta_1 = 0.7$ and $\theta_2 = 0.5$ (OECD, 1982). This equivalence scale was in common use during the 1980s and early 1990s. In the late 1990s, the Statistical Office of the European Union (EUROSTAT) adopted the so-called OECD-modified (or augmented) equivalence scale, which assumes that $\theta_1 = 0.5$ and $\theta_2 = 0.3$. This scale was first proposed by Haagenars et al. (1994).

In addition, we can experiment with new versions of parametric scales. For instance, the logarithmic deflator is defined as follows:

$$d = 1 + \theta_1 \cdot \log h, \; \theta_1 > 0, \tag{12}$$

where h is the household size and θ_1 is the parameter to be estimated. The elasticity ε of this deflator with respect to the household size h is:

$$\varepsilon = \frac{1}{1/\theta_1 + \log h}. \tag{13}$$

We may note that this elasticity is a diminishing function of the household size.

We can treat the pragmatic equivalence scales as potential *SES*s. In our approach, these scales are estimated but not definitively designated. We will recognise them as *SES*s if condition (2) is fulfilled.

5. STATISTICAL ISSUES CONCERNING AN *SES*

There are two statistical issues related to *SESs*: testing whether a function $s(\cdot)$ can be recognised as an *SES* and estimating parametric and non-parametric *SESs*. These two problems require random samples of expenditures per household. Micro-data are the most suitable data for this purpose, though grouped data may also be used.[10] We assume that the general population consists of $m + 1$ disjointed household populations and that each household represents a different type of household. One of these populations, which usually comprises one-person households, is treated as the reference population. The remaining m populations are called the evaluated populations.

In this section, we will use the notations and symbols defined in section 2. Whereas the random variable $Y \sim G(y)$ will represent the expenditure distribution in the general reference population, the expenditure distributions in the general evaluated populations will be represented by m random variables, $X_i \sim F_i(x)$, $i = 1,...,m$. If the vector function $s(\cdot) = [s_1(\cdot),...,s_m(\cdot)]$ is a potential *SES,* then the random variables $Z_i = s_i(X_i) \sim H_i(z)$ will refer to the transformed income distributions in the general evaluated populations. The random variable $Z \sim H(z)$ will describe a mixture (1) of distributions Z_i in the general population and the symbols $\widehat{G}(y)$, $\widehat{F}_i(x)$, $\widehat{H}_i(z)$ and $\widehat{H}(z)$ will denote the sample distribution functions.

The random samples are defined as follows: the sample of size l that is from the general reference population will be denoted as $(y_1,...,y_l)$. Whereas the random sample from the i-th evaluated general population will be denoted as $(x_1,...,x_{n_i})_i$, the transformed version of this sample will be denoted as $(z_1,...,z_{n_i})_i$, $i = 1,...,m$. The total sample size of all of the evaluated groups will be denoted as $n = n_1 +,...,+ n_m$ and weights $\pi_i = n_i / n$ are assigned to the ith group for each $i = 1,...,m$. Let $(z_1,...,z_n)$ denote the pooled sample of all of the transformed values with their corresponding weights $(\pi_1,...,\pi_n)$.

We calculate the empirical distribution function $\widehat{G}(\cdot)$ for the reference distribution Y using the sample $(y_1,...,y_l)$; similarly, we calculate the empirical distribution function $\widehat{H}(\cdot)$ using the pooled sample $(z_1,...,z_n)$ and the corresponding weights $(\pi_1,...,\pi_n)$.

---

[10] In this paper, we use micro-data when presenting statistical methods.

To verify that function $s(\cdot) = [s_1(\cdot),...,s_m(\cdot)]$ is an *SES*, i.e., to check whether identity (2) holds, we need to test the null statistical hypothesis

$$H_0\colon H(z) = G(z) \tag{14}$$

against the alternative hypothesis

$$H_a\colon H(z) \neq G(z) \tag{15}$$

for all $z \geq 0$.

We will verify one of these hypotheses using the Kolmogorov-Smirnov (*K-S*) test:

$$U = \max_z | \hat{H}(z) - \widehat{G}(z) | \sqrt{\frac{l \cdot n}{l + n}}, \text{ for all } z \geq 0 \tag{16}$$

Smirnov (1939). Under the null hypothesis, the *U* statistic (16) has an asymptotic Kolmogorov-λ distribution (Kolmogorov, 1933).

The *p-value* of the *K-S* test (16), i.e., $p = P(U \geq u_{calc})$, is a convenient tool for testing these hypotheses on the selected significance level α, where u_{calc} is the calculated value of the *U*-test in the sample. If $p \leq \alpha$, we reject the null hypothesis H_0 and accept H_a, and as a result, the function $s(\cdot)$ cannot be recognised as an *SES*. If $p > \alpha$, we accept the null hypothesis, and therefore, we recognise function $s(\cdot)$ as an *SES*.

The proposed method of estimating *SESs* uses the *U*-test as a loss function. Suppose that the function $s(\cdot)$ is a potential *SES*. This function may be non-parametric, e.g., it may take the form of a set of deflators $d = [d_1,...,d_m]$, or it may be parametric and depend on k parameters $\theta = [\theta_1,...,\theta_k]$. We will use the symbols $s(\cdot|d)$, $s(\cdot|\theta)$ or simply s when the context of the estimation is obvious. Let $z_1,...,z_n$ be the sequence of the evaluated expenditures that are adjusted by the function s, i.e., let $z_j = s(x_j|d)$ or $z_j = s(x_j|\theta)$, $j = 1,...,n$. Let $\widehat{G}(z)$ and $\hat{H}(z \mid s)$ denote the empirical distribution functions of the reference expenditures and the adjusted expenditures, respectively. We propose the estimator s^* of s that minimises the *K-S* statistic $U(s)$, i.e.:

$$U(s^*) = \min_s \max_z | \hat{H}(z \mid s) - \widehat{G}(z) | \sqrt{\frac{n \cdot l}{n + l}}, \text{ for all } z \geq 0. \tag{17}$$

When we compare the reference distribution Y with a particular evaluated distribution Z_i, we can substitute for the terms $\hat{H}(z \mid s)$ and n on the right-hand side of equation (17) with $\hat{H}_i(z \mid s)$ and n_i, respectively, for each $i = 1,...,m$.

Alternatively, we can estimate the *SES* by using *p-values* as a function of s, i.e., by using $p(s)$:

$$p(s) = P[U(s) \geq u_{calc}(s)]. \tag{18}$$

This *SES* estimator is s^* and it maximises the *p-value* (18):

$$p(s^*) = \max_{s} P[U(s) \geq u_{calc}(s)] \wedge p(s^*) > \alpha. \tag{19}$$

However, it should be noted that s^* estimators do not always exist. Although $p(s)$ usually reaches a maximum, the condition $p(s^*) > \alpha$ might be violated. In practice, the estimator s^* can be found using the grid-search method.

6. STOCHASTIC EQUIVALENCE SCALES FOR POLAND 2005–2010

In this section, we will use expenditure distributions to estimate the *SESs*. The monthly micro-data come from the Polish Household Budget Surveys for the years 2005–2010. The expenditures are expressed in constant 2010 prices. The data were collected annually by the Central Statistical Office of Poland. We assume a 5% significance level in all of the analysed cases.

Table 1 presents estimates of nonparametric relative *SESs* (4) when the household groups are distinguished according to the number of members (i.e., according to the households' sizes). The deflators d_h are estimated separately for each household group.

Table 1

The estimated deflators of the non-parametric *SESs* for various household sizes.
(95% confidence intervals in parentheses)

Household size	2005	2006	2007	2008	2009	2010
1	1	1	1	1	1	1
2	1.625 (1.601;1.656) $p = 0.704$	1.616 (1.605;1.628) $p = 0.180$	1.679 (1.670;1.700) $p = 0.383$	1.709 (1.692;1.733) $p = 0.557$	1.681 (1.662;1.721) $p = 0.611$	1.686 (1.662;1.728) $p = 0.618$
3	1.889 (1.879;1.900) $p = 0.324$	1.942 (*;*) $p = 0.027$	2.048 (2.024;2.081) $p = 0.603$	2.115 (2.095;2.158) $p = 0.356$	2.071 (2.055;2.121) $p = 0.263$	2.068 (2.062;2.093) $p = 0.090$
4	2.045 (2.003;2.070) $p = 0.303$	2.077 (2.069;2.088) $p = 0.118$	2.242 (2.211;2.295) $p = 0.648$	2.293 (2.285;2.309) $p = 0.111$	2.250 (2.232;2.283) $p = 0.202$	2.217 (2.202;2.276) $p = 0.228$
5	2.045 (2.003;2.007) $p = 0.303$	2.111 (2.106;2.115) $p = 0.061$	2.260 (2.260;2.270) $p = 0.065$	2.334 (2.324;2.347) $p = 0.096$	2.275 (*;*) $p = 0.007$	2.266 (*;*) $p = 0.006$
6 or more	2.179 (2.173;2.197) $p = 0.104$	2.303 (*;*) $p = 0.008$	2.474 (*;*) $p = 0.038$	2.485 (*;*) $p = 0.022$	2.466 (*;*) $p = 0.006$	2.543 (*;*) $p = 0.007$
p (K-S)	0.54071	0.07230	0.29703	0.06393	0.07053	0.03726
p (K-W)	0.82307	0.93336	0.89588	0.63157	0.83789	0.93808

Note: p(K-S): p-value in Kolmogorov-Smirnov test; *p(K-W)*: p-value in Kruskal-Wallis test.
Source: Polish Household Budget Survey, 2005–2010, own calculations.

An analysis of the results presented in table 1 shows that almost all of the estimated deflators can be recognised as *SESs*. The exceptions are two estimates for sizable household groups (specifically, households with five or more members). The second row from the bottom in table 1 shows that these exceptions influence the overall *K-S* test (15) only for the year 2010. However, the p-values of the overall Kruskal-Wallis test, which is presented in the last row of table 1, are greater than the significance level $\alpha = 0.05$. Thus, all of the estimated deflators can be recognised as *SESs*.

Three features of the estimated equivalence scales are remarkable. First, these nonparametric equivalence scales are very flat in comparison with the per capita scale. As a result, Polish households exhibited large economies of scale in the years 2005–2010. Second, the 95% confidence intervals are very narrow for the estimated deflators. Consequently, the proposed method to estimate the non-parametric scales is quite accurate. Third, equivalence scales vary over time.

Table 2 contains the estimates of the one-parameter pragmatic scales, i.e., the power scale (9) (Buhmann et al., 1988) and the logarithmic scale (12).

Table 2

The estimates of the one-parameter *SESs* (95% confidence intervals in parentheses)

Year	Power: $d = h^{\theta_1}$		Logarithmic: $d = 1+ \theta_1 log\ h$	
	θ_1	*p-value*	$\theta_1.$	*p-value*
2005	0.51872 (0.51267; 0.52559)	0.38150	0.75681 (0.73640; 0.77773)	0.61788
2006	0.53853 (0.53661;0.54090)	0.11294	0.79389 (0.78919; 0.80228)	0.11506
2007	0.58891 (0.58230; 0.59701)	0.47660	0.89671 (0.88455; 0.92414)	0.37624
2008	0.61021 (0.60595; 0.62193	0.24119	0.93199 (0.92955; 0.93490)	0.070294
2009	0.59684 (0.59176; 0.60856)	0.24987	0.90404 (0.90101; 0.91010)	0.071374
2010	0.59454 (0.59143; 0.60220)	0.14223	0.90424 (*; *)	0.03389

Note: h – household size.

Source: Polish Household Budget Survey, 2005–2010, own calculations.

Analysis of the results in table 2 shows that all of the power equivalence scales can be recognised as *SESs* because all of the corresponding *p-values* are greater than the significance level of 0.05. The estimates of θ_1 are less than one in every year under consideration. This is an indication of the economies of scale that were enjoyed by

Polish households in the years 2005–2010. However, the effect of economies of scale seems to diminish because the parameter θ_1 slowly increases in this period. It is also noteworthy that the value $\theta_1 = 0.5$ of power scale (9) is outside of the 95% confidence interval. Thus, a widely used 'square-root' pragmatic scale cannot be recognised as an *SES* of Polish households in this period.

The estimates of the logarithmic scale satisfy the *SES* condition because the *p-values* are greater than 0.05 for all of the years except 2010. This 'experimental' scale convinces us that the class of pragmatic equivalence scales could be enriched using new, more flexible forms.

Table 3 presents estimates of two two-parameter pragmatic scales: *A-C* (10) and OECD-Type (11). For comparison, the last two columns contain *p-values* of the *K-S* test (16) for two OECD scales, namely, 'old' and 'augmented', and their parameters θ_1 and θ_2 are not estimated, but they are arbitrarily designated.

Table 3

The estimates of the two-parameter *SESs*

Year	Adults-Children: $d = (a + \theta_1 k)^{\theta_2}$			OECD-Type: $d = 1+ \theta_1\,(a - 1) + \theta_2 k$			OECD 'old'	OECD augmented
	$\theta_1.$	θ_2	*p-value*	θ_1	θ_2	*p-value*	*p-value*	*p-value*
2005	0.48636	0.58616	0.76538	0.47261	0.14523	0.53367	0.00000	0.00000
2006	0.58051	0.59131	0.21795	0.49394	0.15152	0.17369	0.00000	0.00000
2007	0.61263	0.63990	0.58117	0.47035	0.32487	0.42113	0.00000	0.03268
2008	0.61080	0.6680	0.16365	0.53568	0.28392	0.32424	0.00000	0.00037
2009	0.61764	0.64322	0.12233	0.50327	0.29312	0.27139	0.00000	0.26723
2010	0.64658	0.63789	0.08588	0.48417	0.32362	0.21231	0.00000	0.14914

Note: a – the number of adults, k – the number of children under the age of 18, OECD 'old': $\theta_1 = 0.7$, $\theta_2 = 0.5$, OECD-augmented: $\theta_1 = 0.5$, $\theta_2 = 0.3$.

Source: Polish Household Budget Survey, 2005–2010, own calculations.

The estimates of the parameter θ_1 of the *A*-C equivalence scale show that the resource cost of a child in proportion to an adult increases from approximately 49% in 2005 to 65% in 2010. The elasticity of this scale with respect to the 'effective household size', i.e., $a + \theta_1 k$, increased in 2005–2008 and then decreased in the two subsequent years that were examined.

The parameters of the estimated OECD-type scales also vary over time. The weights that are assigned to the second and subsequent adults in a household fluctuate but remain close to 0.5. The weights assigned by this scale to each child under the age of 18 tend to increase from 0.15 to 0.32. In general, these estimates differ from the parameters of both the 'old' and the 'augmented' OECD scales. It is worth

noting that the 'old' OECD scale cannot be recognised as an *SES* for all of the years under consideration, whereas the 'augmented' OECD scale can be recognised as an *SES* only for the years 2009 and 2010.

7. CONCLUSIONS

Thus far, the individualistic paradigm of consumer behaviour theory has been unable to solve the problem of homogenising a population of households that differ in all respects other than their expenditures. The adjustment of the expenditures of individual households fails when it is based on a pairwise equalisation of household utilities. Moreover, the separation of the adjustment procedure from the normative evaluations of welfare, inequality and poverty makes the pairwise equalisation quite untrustworthy.

The stochastic equivalence scale addresses the problem of homogenisation at the 'distributional' level, where only the probability distributions of expenditures are objects of adjustment, and therefore, individual household expenditures are not adjusted. An *SES* makes an initially heterogeneous population of households homogeneous with respect to such distributional features as social welfare, inequality and poverty. We use the term 'distributional' because social welfare, inequality and poverty indices characterise the distribution of expenditures but not of individuals. For these reasons, our equivalence scales are called 'stochastic'.

The axiomatic formulation of the *SES* is quite general. It does not specify one definite form of the scale, but it does define the properties that should be satisfied by a certain function for it to be recognised as an *SES*. The validation of these properties can be verified using statistical tests. The possibility of estimating parametric and non-parametric *SES*s also opens new perspectives for experimenting with other forms of such scales.

It should be emphasised that the actual form of an SES function is not important; only the fact that a function is an *SES* function matters. Thus, we do not have to search for an optimal *SES*. For instance, regardless of whether the *SES* is a power scale, an *A-C* scale or an OECD-type scale, it will always provide the same equivalent distribution of expenditures. Obviously, each of these scales (as well as other scales) can shed additional light on an evaluated distribution. For instance, an *A-C* or OECD-type scale can be useful in evaluating the cost of children.

It is worth noting that stochastic equivalence scales open new perspectives for comparisons of welfare, inequality and poverty between various geographic regions and/or periods. For this purpose, we need to choose one common reference group of households and express all of the distributions of expenditures in terms of comparable equivalence units.

The application of *SESs* is easy in practice. Estimating parametric and nonparametric *SESs* requires typical statistical data and standard statistical packages.

However, we encountered an unexpected problem when we applied an *SES* to the disposable income distributions in Poland in the years 2005–2010. Although we obtained plausible estimates of parametric and nonparametric equivalence scales, *none* of these scales could be recognised as *SES*s. Further research will be required to explain this occurrence.

REFERENCES

Atkinson A. B., (1987), On the Measurement of Poverty, *Econometrica,* 55, 749–764.

Atkinson A. B., Bourguignon F., (1987), Income Distribution and Differences in Needs, in: Feiwel G. R., (ed.), *Arrow and the Foundations of the Theory of Economic Policy*, Macmillan, London.

Blackorby C., Donaldson D., (1993), Adult-Equivalence Scales and the Economic Implementation of Interpersonal Comparisons of Well-Being, *Choice and Welfare*, 10, 335–361.

Blundell R. W., Duncan A., Pendakur K., (1998), Semiparametric Estimation of Consumer Demand, *Journal of Applied Econometrics*, 13, 435–461.

Blundell R. W., Lewbel A., (1991), The Information Content of Equivalence Scales, *Journal of Econometrics*, 150, 49–68.

Buhmann B., Rainwater L., Schmaus G., Smeeding T., (1988), Equivalence Scales, Well-Being, Inequality, and Poverty: Sensitivity Estimates Across Ten Countries Using the Luxembourg Income Study (LIS) Database, *Review of Income and Wealth*, 34, 115–142.

Capéau B., Ooghe E., (2007), On Comparing Heterogeneous Populations: Is There Really a Conflict Between Welfarism and a Concern for Greater Equality in Living Standards? *Mathematical Social Sciences*, 53, 1–28.

Coulter F. A. E., Cowell F. A., Jenkins S. P., (1992a), Differences in Needs and Assessment of Income Distributions, *Bulletin of Economic Research*, 44, 77–124.

Coulter F. A. E., Cowell F. A., Jenkins S. P., (1992b), Equivalence Scale Relativities and the Extent of Inequality and Poverty, *Economic Journal*, 102, 1067–1082.

Cutler A. E., Katz L. F., (1992), Rising Inequality? Changes in the Distribution of Income and Consumption in the 1980s, *Economic Review*, *Papers and Proceedings* 82, 546–551.

Davidson R., (2008), Stochastic Dominance, in: Durlauf S. N., Lawrence E. B., (eds.), *The New Palgrave Dictionary of Economics*, Second Edition, Palgrave Macmillan.

Davidson R., Duclos J., (2000), Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality, *Econometrica*, 68, 1435–1464.

Dickens R., Fry V., Pashardes P., (1993), Nonlinearities, Aggregation and Equivalence Scales, *Economic Journal*, 103, 359–368.

Donaldson D., Pendakur K., (2003), Equivalent-Expenditure Functions and Expenditure-Dependent Equivalence Scales, *Journal of Public Economics*, 88, 175–208.

Ebert U., Moyes P., (2003), Equivalence Scales Reconsidered, *Econometrica*, 71, 319–343.

Foster J. E., Shorrocks A. F., (1988), Poverty Orderings, *Econometrica*, 56, 173–177.

Gozalo P., (1997), Nonparametric Bootstrap Analysis with Implementation to Demographic Effects in Demand Functions, *Journal of Econometrics*, 81, 357–393.

Haagenaars A, de Vos K., Zaidi A., (1994), *Poverty Statistics in the Late 1980s, Research Based on Microdata*, Office for Official Publications of the European Communities, Luxembourg.

Jones A., O'Donnell O., (1995), Equivalence Scales and the Costs of Disability, *Journal of Public Economics*, 56, 273–289.

Jäntti M., Danziger S., (2000), Poverty in Advanced Countries, in: Atkinson A. B., Bourguignon F., (eds.), *Handbook of Income Distribution*, North-Holland, Amsterdam.

Jenkins S. P., Cowell F. A., (1994), Parametric Equivalence Scales and Scale Relativities, *The Economic Journal*, 104, 891–900.

Kolmogorov A. N., (1933), Sulla Determinazione Empirica di Una Legge di Distribuzione, *Giornale delli Insitute Italiano degli Attuari*, 4, 83–91.

Kot S. M. (2012), *Towards the Stochastic Paradigm of Welfare Economics*, (in Polish: *Ku stochastycznemu paradygmatowi ekonomii dobrobytu*), Impuls, Cracow.

Kruskal W. H., Wallis W. A., (1952), Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association*, *47*, 583–621.

Lewbel A., (1997), Consumer Demand Systems and Household Equivalence Scales, in: Pesaran M. H., Schmidt P., (eds.), *Handbook of Applied Econometrics*, Volume II, *Microeconomics*, Blackwell Publishers Ltd, Oxford.

McFadden D., (1989), Testing for Stochastic Dominance, in: Fromby T. B., Seo T. K., (eds.), *Studies in the Economics of Uncertainty*, Springer-Verlag, New York.

OECD, (1982), *The List of Social Indicators*, Paris.

OECD, (2008), *Growing Unequal? Income Distribution and Poverty in OECD Countries*, Paris (www.oecd.org/els/social/inequality / www.oecd.org/els/social/inegalite).

Pashardes P., (1995), Equivalence Scales in a Rank-3 Demand System, *Journal of Public Economics*, 58, 143–158.

Pendakur K., (1999), Estimates and Tests of Base-Independent Equivalence Scales, *Journal of Econometrics*, 88, 1–40.

Pollak R. A., Wales T. J., (1979), Welfare Comparisons and Equivalence Scales, *American Economic Review*, 69, 216–221.

Pollak R. A., Wales T. J., (1992), *Demand System Specification and Estimation,* Oxford University Press, London.

Shorrocks A. F., (1983), Ranking Income Distributions, *Economica*, 50, 3–17.

Shorrocks A. F., (1998), Deprivation Profiles and Deprivation Indices, Jenkins S. P., Kapteyn A., van Praag B. M. S., (eds.), Ch. 11 in *The Distribution of Household Welfare and Household Production*, Cambridge University Press.

Slesnick D., (1998), Empirical Approaches to the Measurement of Welfare, *Journal of Economic Literature*, 36, 2108–2165.

Smirnov N. W., (1939), Sur les Ecarts de la Courbe de Distribution Empirique, *Comptes Rendus de l'Academie des Sciences Paris*, 6, 3–26.

SKALE EKWIWALENTNOŚCI BAZUJĄCE NA KRYTERIUM
STOCHASTYCZNEJ INDYFERENCJI: PRZYPADEK POLSKI

Streszczenie

Artykuł przedstawia koncepcję *stochastycznych skal ekwiwalentności* (*SES*), która bazuje na kryterium stochastycznej indyferencji. SES jest dowolną funkcją, która transformuje rozkład wydatków określonej grupy gospodarstw domowych w taki sposób, że wynikowy rozkład jest stochastycznie indyferentny wobec rozkładu wydatków grupy gospodarstw odniesienia. Kryterium stochastycznej indyferencji jest także wykorzystane dla opracowania metody estymacji SES. Oszacowano nieparametryczne i parametryczne SES na podstawie Polskich Budżetów Gospodarstw Domowych za lata 2005–2010.

Słowa kluczowe: skale ekwiwalentności, stochastyczna indyferencja, estymacja, rozkład wydatków

EQUIVALENCE SCALES BASED ON STOCHASTIC INDIFFERENCE CRITERION: THE CASE OF POLAND

A b s t r a c t

The paper presents the concept of the *stochastic equivalence scale* (*SES*), which is based on the stochastic indifference criterion. The *SES* is any function that transforms the expenditure distribution of a specific group of households in such a way that the resulting distribution is stochastically indifferent from the expenditure distribution of a reference group of households. The stochastic indifference criteria are also used in developing the method of the estimation of the *SES*. Non-parametric and parametric *SESs* are estimated using the Polish Household Budget Survey for the years 2005–2010.

Keywords: equivalence scale, stochastic indifference, estimation, expenditure distribution

AGNIESZKA LIPIETA[1]

THE OPTIMAL PRODUCERS' ADJUSTMENT TRAJECTORY[2]

1. INTRODUCTION

Let us consider the private ownership economy (see e.g. Debreu, 1959; Mas-Colell et al., 1995). Let us also suppose that for some reasons (for instance new regulations, new technologies etc.) producers have to modify their technologies. It can result in a mild evolution within the production sector which does not, however, disturb the equilibrium. That evolution indicates at the time the current system of private ownership economies. Hence, this survey can be also viewed as an attempt to model changes of the producers' sphere of the private ownership economy (compare to Radner, 1972 or Magill, Quinzii, 2002).

This paper is a continuation of the research originated in Lipieta (2010) where a model of the private ownership economy with complementary commodities was presented. Later, changes of the production system of the private ownership economy were studied in a generalized form of the economy with complementary commodities, the so called economy with the reduced consumption sphere (see Lipieta, 2012).

The mapping describing changes introduced by producers will be called the producers' trajectory. Keeping equilibrium (where applicable) and minimization of the distance between the initial and final production plans are considered as the main criterion of the choice of the producers' trajectory. Hence, projections defined in commodity-price space $\mathbb{R}^{\ell}$ with maximum norm are used for modeling changes in the production sphere.

The paper consists of four parts. The second section presents the construction of the private ownership economy. The third part deals with the description of such a modification of the production sphere of the private ownership economy that does not disturb equilibrium. The fourth part presents the characterization of the best trajectory of changes of the economy under study with respect to the criterion of the distance minimization.

[1] Cracow University of Economics, 27 Rakowicka St., 31-510 Cracow, Poland, e-mail: alipieta@uek.krakow.pl.

2. THE MODEL

The private ownership economy defined in Debreu (1959) is studied in the form of a multi-range relational system which includes the combination of production and consumption systems (see Lipieta, 2010; 2013). The linear space $\mathbb{R}^\ell$ ($\ell \in \{1,2,\dots\}$) with the scalar product

$$(x \circ y) = (x_1, \dots, x_\ell) \circ (y_1, \dots, y_\ell) = \sum_{k=1}^{\ell} x_k \cdot y_k,$$

$x, y \in \mathbb{R}^\ell$, is the ℓ-dimensional commodity-price space. Suppose that two groups of agents viz. producers and consumers, operate in $\mathbb{R}^\ell$. Let $n \in \{1,2,\dots\}$ and
- $B = \{b_1, \dots, b_n\}$ be a finite set of producers,
- $y: B \ni b \to Y^b \subset \mathbb{R}^\ell$ be the correspondence of production sets, which to every producer b assigns a nonempty production set $y(b) = Y^b \subset \mathbb{R}^\ell$ of the producer's feasible production plans,
- $p \in \mathbb{R}^\ell$ be a price vector.

Definition 2.1. A two-range relational system $P_q = (B, \mathbb{R}^\ell; y, p)$, is called the quasi--production system.

Definition 2.2. If $P_q = (B, \mathbb{R}^\ell; y, p)$ is the quasi-production system, where

$$\forall\, b \in B \quad \eta^b(p) \stackrel{\text{def}}{=} \{y^{b*} \in y(b): p \circ y^{b*} = \max\{p \circ y^b: y^b \in y(b)\}\} \neq \emptyset,$$

then
- $\eta: B \ni b \to \eta^b(p) \subset \mathbb{R}^\ell$ is called the correspondence of supply at price system p,
- $\pi: B \ni b \to \pi(b) = p \circ y^{b*} \in \mathbb{R}$, where $y^{b*} \in \eta^b(p)$, is called the maximal profit function at price system p,
- the quasi-production system P_q is called the production system and denoted by

$$P_q = P = (B, \mathbb{R}^\ell; y, p, \eta, \pi).$$

The set $\eta^b(p)$ is called the set of optimal plans of producer b at given price vector p.

In the quasi-production system, the aim of producers is not specified in contrast to the production system, where the producers aim at profits maximization at given prices and technologies. In quasi-production system $P_q = (B, \mathbb{R}^\ell; y, p)$, the profit function of a producer b at price vector p, is of the form:

$$Y^b \ni y^b \to p \circ y^b \in \mathbb{R}.$$

Let $\hat{y}^b \in y(b)$ denote the plan realized by producer $b \in B$. If $\hat{y}^b$ is the optimal plan of producer b at given price vector p, then it will be noted by y^{b*} ($\hat{y}^b = y^{b*}$) and

$$p \circ \hat{y}^b = p \circ y^{b*} = \max\{p \circ y^b : y^b \in y(b)\}.$$

Now, the consumption sphere is defined. Let $m \in \{1, 2, \dots\}$ and
- $A = \{a_1, \dots, a_m\}$ be a finite set of consumers,
- $\Xi \subset \mathbb{R}^\ell \times \mathbb{R}^\ell$ be the family of all preference relations in $\mathbb{R}^\ell$,
- $\chi: A \ni a \to \chi(a) = X^a \subset \mathbb{R}^\ell$ be a correspondence of consumptions sets,
- $e: A \ni a \to e(a) \in \mathbb{R}^\ell$ be an initial endowment mapping,
- $\varepsilon \subset A \times (\mathbb{R}^\ell \times \mathbb{R}^\ell)$ be a correspondence, which assigns a preference relation $\preccurlyeq^a$ to every consumer $a \in A$ from set Ξ restricted to set $\chi(a) \times \chi(a)$,
- $p \in \mathbb{R}^\ell$ be a price vector.

Definition 2.3. The three-range relational system $C_q = (A, \mathbb{R}^\ell, \Xi; \chi, e, \varepsilon, p)$ is called the quasi-consumption system.

However, we assume that if consumer $a \in A$ has a possibility to maximize his preference relation on the budget set, then he uses that opportunity. It should be noted that the expenditures of every consumer $a \in A$ in quasi-consumption system C_q cannot be greater than the value

$$w(a) = p \circ e(a). \tag{1}$$

Vector (1) is called the wealth of consumer a.

Let $C_q = (A, \mathbb{R}^\ell, \Xi; \chi, e, \varepsilon, p)$ be the quasi-consumption system.

Definition 2.4. If at the given price vector $p \in \mathbb{R}^\ell$, for every $a \in A$

$$\beta(a) = \beta^a(p) = \{x \in \chi(a) : p \circ x \le w(a)\} \ne \emptyset. \tag{2}$$

$$\varphi(a) = \varphi^a(p) = \{x^{a*} \in \beta^a(p) : \forall x^a \in \beta^a(p)\ x^a \preccurlyeq^a x^{a*}, \preccurlyeq^a \in \Xi\} \ne \emptyset, \tag{3}$$

then
- $\beta: A \ni a \to \beta^a(p) \subset \mathbb{R}^\ell$ is the correspondence of budget sets at price system p, which to every consumer $a \in A$ assigns his set of budget constrains $\beta^a(p) \subset \chi(a)$ at price system p and initial endowment $e(a)$,
- $\varphi: A \ni a \to \varphi^a(p) \subset \mathbb{R}^\ell$ is the demand correspondence at price system p, which to every consumer $a \in A$ assigns the consumption plans maximizing his preference on the budget set $\beta^a(p)$,

– the quasi-consumption system C_q is called the consumption system and denoted by

$$C_q = C = (A, \mathbb{R}^\ell, \Xi; \chi, e, \varepsilon, p, \beta, \varphi).$$

Let $p \in \mathbb{R}^\ell$ be a price vector. The following definition may be assumed on the basis of the above:

Definition 2.5. The relational system $E_q = (\mathbb{R}^\ell, P_q, C_q, \theta, \omega)$, where
– $P_q = (B, \mathbb{R}^\ell; y, p)$ is the quasi-production system,
– the mapping $\theta: A \times B \to [0,1]$ satisfies,

$$\forall b \in B \ \sum_{a \in A} \theta(a, b) = 1, \tag{4}$$

– $C_q = (A, \mathbb{R}^\ell, \Xi; \chi, e, \varepsilon, p)$ is the quasi-consumption system in which

$$w(a) = p \circ e(a) + \sum_{b \in B} \theta(a, b) \cdot p \circ \hat{y}^b, \tag{5}$$

– $\sum_{a \in A} e(a) = \omega \in \mathbb{R}^\ell$ \hfill (6)

is called the private ownership economy.

If P_q is production system ($P_q = P$) and C_q is consumption system ($C_q = C$), then private ownership economy E_q will be called the Debreu economy and denoted by

$$E_p = E_q = (\mathbb{R}^\ell, P, C, \theta, \omega).$$

Number $\theta(a, b)$ indicates that part of the profit of producer b which is owned by consumer a. The private ownership economy E_q operates as follows. Let a price vector $p \in \mathbb{R}^\ell$ be given. Every producer b realizes a production plan $\hat{y}^b \in y(b)$. The profit of each producer b, by realization of the plan $\hat{y}^b$, is divided among all consumers according to function θ (see (4)). So, the expenditures of every consumer $a(a \in A)$ cannot be greater than value $w(a)$ (see (5)). If $\beta^a(p) \neq \emptyset$ (see (2)) and $\varphi^a(p) \neq \emptyset$ (see (3)), then consumer a chooses his consumption plan $\hat{x}^a = x^{a*} \in \varphi^a(p) \subset \chi(a)$ maximizing his preference on budget set $\beta^a(p)$. If $\beta^a(p) \neq \emptyset$ and $\varphi^a(p) = \emptyset$, then consumer a chooses his consumption plan $\hat{x}^a \in \beta^a(p)$, due to his own criterion. If $\beta^a(p) = \emptyset$, then we assume that $\hat{x}^a = 0 \in \mathbb{R}^\ell$. If

$$\sum_{a \in A} \hat{x}^a - \sum_{b \in B} \hat{y}^b = \omega, \tag{7}$$

then it is said that there is quasi-equilibrium in economy E_q and vector p is called the quasi-equilibrium price vector in that economy. Consequently, the sequence

$$((\hat{x}^a)_{a \in A}, (\hat{y}^b)_{b \in B}, p) \stackrel{\text{def}}{=} (\hat{x}^{a_1}, \dots, \hat{x}^{a_m}, \hat{y}^{b_1}, \dots, \hat{y}^{b_n}, p) \in (\mathbb{R}^\ell)^{m+n+1} \tag{8}$$

is called the state of quasi-equilibrium in economy E_q. If economy E_q is the Debreu economy ($E_q = E_p$, see def. 2.5), then the sequence

$$((x^{a*})_{a \in A}, (y^{b*})_{b \in B}, p) \overset{\text{def}}{=} (x^{a_1 *}, \dots, x^{a_m *}, \ y^{b_1 *}, \dots, y^{b_n *}, p) \in (\mathbb{R}^\ell)^{m+n+1}, \quad (9)$$

that satisfies

$$\sum_{a \in A} x^{a*} - \sum_{b \in B} y^{b*} = \omega, \tag{10}$$

is the state of equilibrium in Debreu economy E_p. If condition (10) is satisfied, then it is said that there is equilibrium in economy E_p and vector p is called the equilibrium price vector in that economy.

3. SYSTEM OF PRIVATE OWNERSHIP ECONOMIES

At first, we recall some properties of subspaces of $\mathbb{R}^\ell$ that will be in use later. Let $V \subset \mathbb{R}^\ell$ be a linear subspace of dimension $\ell - k$, $k \in \{1, \dots, \ell - 1\}$. Then there exist linearly independent vectors $g^1, \dots, g^k \in \mathbb{R}^\ell$ such that

$$V = \cap_{s=1}^{k} \ker \tilde{g}^s, \tag{11}$$

where, for $s \in \{1, 2, \dots, k\}$,

$$\tilde{g}^s : \mathbb{R}^\ell \ni (x_1, \dots, x_\ell) \to \sum_{l=1}^{\ell} g_l^s x_l \in \mathbb{R} \tag{12}$$

and

$$\ker \tilde{g}^s = \{x = (x_1, \dots, x_\ell) \in \mathbb{R}^\ell : \tilde{g}^s(x) = 0\}. \tag{13}$$

Now, we put the following definition:

Definition 3.**1** (see Lipieta, 2012). The private ownership economy $E_q = (P, C, \theta, \omega)$, in which condition

$$\forall a \in A \ X^a \subset V \tag{14}$$

is satisfied, will be called the private ownership economy with the reduced consumption system.
The private ownership economy $E_q = (P, C, \theta, \omega)$, in which condition

$$\forall b \in B \ Y^b \subset V \tag{15}$$

is satisfied, will be called the private ownership economy with the reduced production system.

If there is a proper subspace V of commodity-price space $\mathbb{R}^{\ell}$ ($\{0\} \neq V \subset \mathbb{R}^{\ell}$) such that conditions (14) and (15) are both satisfied in economy E_q, then this economy will be called the private ownership economy reduced to the subspace V.

The sets satisfying condition (14) or (15) are the linear sets (see for example Moore, 2007). Hence, the economy with the reduced consumption (production) system is also called the economy with linear consumption (production) sets.

Let us notice that assumption (14) has an economic interpretation. If the consumers are not interested in the consumption of a commodity $l_0 \in \{1, \dots, \ell\}$, then the coordinate l_0 is equal 0 in every plan $x^a \in X^a$, namely

$$\forall a \in A \ x_{l_0}^a = 0.$$

Hence,

$$\forall a \in A \ X^a \subset \ker \tilde{g}, \tag{16}$$

where $\tilde{g}$ is of the form (12), precisely

$$\tilde{g}: \mathbb{R}^{\ell} \ni (x_1, \dots, x_{\ell}) \to x_{l_0} \in \mathbb{R}. \tag{17}$$

Suppose that producers' output $l_0 \in \{1, \dots, \ell\}$ is not wanted by the consumers or it is a harmful commodity. The producers, for which l_0 is the output, have to modify their plans of action and stop producing that commodity. After modification, the condition (15) will be valid, namely

$$\forall b \in B \ Y^b \subset \ker \tilde{g}, \tag{18}$$

with $\tilde{g}$ is of the form (17).

If there exist two commodities $l_1, l_2 \in \{1, \dots, \ell\}$ such that

$$\exists c > 0 \ \forall a \in A \ \forall x^a = (x_1^a, \dots, x_{\ell}^a) \in X^a \ x_{l_1}^a = c \cdot x_{l_2}^a,$$

then commodities l_1 and l_2 are called complementary (see Lipieta, 2010). In that case, condition (16) is fulfilled with the functional of the form

$$\tilde{g}: \mathbb{R}^{\ell} \ni (x_1, \dots, x_{\ell}) \to x_{l_1} - c \cdot x_{l_2} \in \mathbb{R}. \tag{19}$$

Generally, if there exist numbers $c_1, c_2, \dots, c_{\ell} \in \mathbb{R}$ such that $\sum_{l=1}^{\ell}(c_l)^2 \neq 0$ and

$$\forall a \in A \ \forall x^a = (x_1^a, \dots, x_{\ell}^a) \in X^a \ \sum_{l=1}^{\ell} c_l x_l^a = 0, \tag{20}$$

we will say that the commodities for which $c_l \neq 0$ ($l \in \{1, \dots, \ell\}$) are dependent in the consumption sets. If condition (20) is satisfied, then (16) is fulfilled with functional $\tilde{g}$ of the form (12), namely

$$\tilde{g}: \mathbb{R}^\ell \ni (x_1, \dots, x_\ell) \to \textstyle\sum_{l=1}^{\ell} c_l x_l \in \mathbb{R}. \tag{21}$$

If $\tilde{g}$ is of the form (21), then set $\ker \tilde{g}$ is the linear subspace of $\mathbb{R}^\ell$ of dimension $\ell - 1$ (see (11)). Let us notice that if

$$\forall a \in A \quad x_{l_1}^a = x_{l_2}^a = 0,$$

then (14) is satisfied for subspace V defined, in the meaning of condition (11), by functionals $\tilde{g}^1$ and $\tilde{g}^2$ of the form (17) for l_0 equals respectively l_1 or l_2. Hence, in the sense of condition (20) the commodities l_1, l_2 are also the complementary ones.

In many real economies, the producers are obliged to reduce the amount of pollution emitted to the atmosphere. The amount of pollution increases with the quantities of goods. Hence, saying about the dependent commodities in production sets makes sense. As in the case of consumers, if there exist real numbers $c_1, c_2, \dots, c_\ell$ such that $\sum_{l=1}^{\ell}(c_l)^2 \neq 0$ and

$$\forall b \in B \quad \forall y^b = (y_1^b, \dots, y_\ell^b) \in Y^b \quad \textstyle\sum_{l=1}^{\ell} c_l y_l^b = 0, \tag{22}$$

we say that the commodities for which $c_l \neq 0$ ($l \in \{1, \dots, \ell\}$) are dependent in production sets.

Reducing the amount of an output l_0 in all production plans relies on decreasing in coordinate $y_{l_o}^b$ in every plan y^b of every producer b. If producers have to modify their technologies to get the desired dependency between quantities of some commodities, then they will change their plans of action to satisfy condition (15) with subspace V defined by using functionals of the form (21).

To sum up: legal requirements, new technologies, new fashions, inventions and many other reasons can contribute to the modification of production sets, to the sets satisfying (15), with subspace V of the form (11). Moreover, the rationality of producers' behavior implies that all profitable changes is production sphere are worth, in the opinion of producers, realizing.

Let us notice that the producers will not want to stop producing commodities used only in the producers' activities, namely the commodities that are outputs and inputs only for the producers. Although these commodities are not wanted by the consumers, they play an important role in the production sector.

In the next part of the paper, the procedure of such modification of the production sets will be presented that the modified producers' sets will satisfy condition (15) with a nontrivial subspace V of space $\mathbb{R}^\ell$. At first, some notations and definitions will

be introduced. Let $V \subset \mathbb{R}^{\ell}$, $V \neq \{0\}$ be a linear subspace. Then V^T means the linear subspace orthogonal to V, namely

$$V^T \overset{\text{def}}{=} \{x \in \mathbb{R}^{\ell} | \, \forall v \in V \colon x \circ v = 0\}.$$

Fix linearly independent functionals $\tilde{g}^1, \dots, \tilde{g}^k$ of the form (12) satisfying (11). In this situation, the system of equations

$$\tilde{g}^s(x) = g^s \circ x = \delta^{sr} \quad \text{for} \quad s, r \in \{1, \dots, k\}, \tag{23}$$

where $x \in \mathbb{R}^{\ell}$ and

$$\delta^{sr} = \begin{cases} 1 & \text{if} \quad s = r \\ 0 & \text{if} \quad s \neq r \end{cases}$$

is Kronecker delta, has a solution. We will denote a solution of (23) by $q^1, \dots, q^k \in \mathbb{R}^{\ell}$. Now, we define mapping $\tilde{Q} \colon \mathbb{R}^{\ell} \times [0,1] \to \mathbb{R}^{\ell}$ by the rules

$$\tilde{Q}(x, t) = x - t \cdot \sum_{s=1}^{k} \tilde{g}^s(x) \cdot q^s. \tag{24}$$

Notice that for every fixed $t \in [0,1]$ mapping $\tilde{Q}(\cdot, t) \colon \mathbb{R}^{\ell} \to \mathbb{R}^{\ell}$ is the linear and continuous operator. Moreover, $\tilde{Q}(\cdot, 1) \colon \mathbb{R}^{\ell} \to \mathbb{R}^{\ell}$, precisely

$$\tilde{Q}(x, 1) = x - \sum_{s=1}^{k} \tilde{g}^s(x) \cdot q^s \quad \text{for} \quad x \in \mathbb{R}^{\ell}$$

is the linear continuous projection from $\mathbb{R}^{\ell}$ into V, where

$$\forall x \in \mathbb{R}^{\ell} \, \tilde{Q}(x, 1) \in V \ \text{ and } \ \forall v \in V \, \tilde{Q}(v, 1) = v$$

(see for example Cheney, 1966). From now, the projection $\tilde{Q}(\cdot, 1)$ will be denoted by Q. Hence

$$Q(x) \overset{\text{def}}{=} \tilde{Q}(x, 1) = x - \sum_{s=1}^{k} \tilde{g}^s(x) \cdot q^s \quad \text{for} \quad x \in \mathbb{R}^{\ell}. \tag{25}$$

It should be noted that

$$\forall v \in V \ \forall t \in [0,1] \ \ \tilde{Q}(v, t) = v = Q(v). \tag{26}$$

We also say that vectors $q^1, \dots, q^k \in \mathbb{R}^{\ell}$ determine (or define) the mappings Q, $\tilde{Q}$ and $\tilde{Q}(\cdot, t)$ for every $t \in [0,1]$. The set of all projection from $\mathbb{R}^{\ell}$ into subspace V will be denoted by $\mathcal{P}(\mathbb{R}^{\ell}, V)$. Let us recall (see Cheney, 1966) that for every $Q \in \mathcal{P}(\mathbb{R}^{\ell}, V)$ there exist vectors $q^1, \dots, q^k \in \mathbb{R}^{\ell}$ satisfying (23) such that, the projection Q is of the form (25). The above defined objects lead us to the following:

Theorem 3.2. If $p \in \mathbb{R}^{\ell} \backslash V^T$ then there exists continuous and linear operator $\tilde{Q}: \mathbb{R}^{\ell} \times [0,1] \to \mathbb{R}^{\ell}$ of the form (24) satisfying

$$\forall x \in \mathbb{R}^{\ell} \; \forall t \in [0,1] \;\; p \circ x = p \circ \tilde{Q}(x,t). \tag{27}$$

Proof. Notice that if $p \notin V^T$ then vectors $p, g^1, ..., g^k$ are linearly independent and the system of equalities

$$\begin{cases} g^s \circ x = \delta^{sr} \\ p \circ x = 0 \end{cases} \quad s, r \in \{1, ..., k\}. \tag{28}$$

has a solution. The solution of (28) will be also denoted by $q^1, ..., q^k \in \mathbb{R}^{\ell}$. The operator $\tilde{Q}$ of the form (24), determined by vectors $q^1, ..., q^k$, satisfies the thesis of the theorem. □

The result of the theorem 3.2 implies the following:

Theorem 3.3. Let $P = (B, \mathbb{R}^{\ell}; y, p, \eta, \pi)$ be a production system. There is an operator $\tilde{Q}$ of the form (24) such that for every $b \in B$ and $y^{b*} \in \eta^b(p)$, vector $\tilde{Q}(y^{b*}, t)$ maximizes, for every $t \in [0,1]$, the profit of producer b at price p, on the modified production set

$$\tilde{Q}(Y^b, t) = \{\tilde{Q}(y^b, t) \in \mathbb{R}^{\ell} : y^b \in Y^b\}. \tag{29}$$

Proof. If $p \notin V^T$ then the thesis of the theorem is the immediate consequence of theorem 3.2. If $p \in V^T$, then by (26) we get that every operator $\tilde{Q} \in \mathcal{P}(\mathbb{R}^{\ell}, V)$ of the form (24) determined by vectors $q^1, ..., q^k \in \mathbb{R}^{\ell}$ calculated by (23) satisfies

$$\forall v \in V \; \forall t \in [0,1] \;\; p \circ v = p \circ \tilde{Q}(v,t) = 0 \tag{30}$$

which gives the result. □

The operator $\tilde{Q}$ by the thesis of theorem 3.3 is called the producers' adjustment trajectory.

Let $\tilde{Q}$ be an operator of form (24) and $P_q = (B, \mathbb{R}^{\ell}; y, p)$ be a quasi-production system. Replacing at every $t \in [0,1]$, producers' sets $Y^{b_1}, Y^{b_2}, ..., Y^{b_n}$ in quasi-production system P_q with the sets $\tilde{Q}(Y^{b_1}, t), \tilde{Q}(Y^{b_2}, t), ..., \tilde{Q}(Y^{b_n}, t)$, we receive also the quasi-production system. Such modified quasi-production system differs from the initial one (see def. 2.2) with correspondence of production sets. Additionally, if P_q ($P_q = P$) is the production system and $\tilde{Q}$ is the mapping by the thesis of theorem 3.3, then the modified production system is different from the initial one, also in the correspondence of supply at the given price system.

Fix $p \in \mathbb{R}^\ell$. Consider vectors $q^1, \ldots, q^k \in \mathbb{R}^\ell$ satisfying (28) if $p \in \mathbb{R}^\ell \backslash V^T$ and (23) if $p \in V^T$. Let $\tilde{Q}$ be an operator of form (24) determined by vectors $q^1, \ldots, q^k$ and $P_q = (B, \mathbb{R}^\ell; y, p)$ be a quasi-production system. For every $t \in [0,1]$, we have

Definition 3.**4**. The two-range relational system

$$P_q(q^1, \ldots, q^k; t) = (B, \mathbb{R}^\ell; y_t, p)$$

where

- $y_t: B \ni b \rightarrow \tilde{Q}(Y^b, t) \subset \mathbb{R}^\ell$ is the correspondence of production sets, which assigns the image of production set Y^b to every $b \in B$ producer by mapping $\tilde{Q}(\cdot, t)$, is called the modification of system P_q, at time t, determined by vectors $q^1, \ldots, q^k$.
  If quasi-production system P_q is the production system $P_q = P = (B, \mathbb{R}^\ell; y, p, \eta, \pi)$, then two-range relational system

$$P(q^1, \ldots, q^k; t) = (B, \mathbb{R}^\ell; y_t, p, \eta_t, \pi_t)$$

where additionally

- $\eta_t: B \ni b \rightarrow \eta_t^b(p) \subset \mathbb{R}^\ell$ is the correspondence of supply at the given price system p, which to every producer $b \in B$ assigns set $\eta_t^b(p)$ of production plans maximizing his profit, at the price system p, on the set $Q(Y^b, t)$,

$$\forall b \in B \;\; \eta_t^b(p) \stackrel{\text{def}}{=} \{Q(y^{b*}, t): p \circ y^{b*} = \max\{p \circ y^b: y^b \in Y^b\}\},$$

- $\pi_t: B \ni b \rightarrow \pi_t^b(p) \in \mathbb{R}$ is the maximal profit function at given price system p and

$$\pi_t^b(p) = p \circ Q(y^{b*}, t) \;\; \text{where} \; y^{b*} \in \eta^b(p) \; \text{for every} \; b \in B,$$

is called the modification of production system P, at time t, determined by vectors $q^1, \ldots, q^k$.

Definition 3.**5**. The relational system

$$E_q(q^1, \ldots, q^k; t) = (P_q(q^1, \ldots, q^k; t), C_q, \theta, \omega)$$

is called the modification of economy E_q at time t, determined by vectors $q^1, \ldots, q^k$. If E_q is the Debreu economy ($E_q = E_p = (\mathbb{R}^\ell, P, C, \theta, \omega)$), then the relational system

$$E_p(q^1, \ldots, q^k; t) = (P(q^1, \ldots, q^k; t), C, \theta, \omega)$$

is called the modification of economy E_q, at time t, determined by vectors $q^1, \ldots, q^k$.

It is apparent that if economy E_q satisfies the condition (14), then for every $t \in [0,1]$, economy $E_q(q^1, \dots, q^k; t)$ is the economy reduced to subspace V.

Let V be a linear subspace of $\mathbb{R}^\ell$ given, in the sense of condition (11), by functionals $\tilde{g}^1, \dots, \tilde{g}^k$ of the form (12). Let $p \in \mathbb{R}^\ell$ and $E_q = (\mathbb{R}^\ell, P_q, C_q, \theta, \omega)$ be a private ownership economy (see def. 2.5).

In this situation the following is true:

Theorem 3.6. Assume that $p \notin V^T$ and vectors $q^1, \dots, q^k \in \mathbb{R}^\ell$ satisfy (28). Let $\tilde{Q}$ be an operator of the form (24) determined by vectors $q^1, \dots, q^k$.

1. If the sequence $((\hat{x}^a)_{a \in A}, (\hat{y}^b)_{b \in B}, p)$ (see (8)) is the state of quasi-equilibrium in economy E_q and

$$\sum_{a \in A} \hat{x}^a - \omega \in V, \tag{31}$$

then the sequence

$$\left((\hat{x}^a)_{a \in A}, (\tilde{Q}(\hat{y}^b, t))_{b \in B}, p\right) = (\hat{x}^{a_1}, \dots, \hat{x}^{a_m}, \tilde{Q}(\hat{y}^{b_1}, t), \dots, \tilde{Q}(\hat{y}^{b_n}, t), p) \tag{32}$$

is the state of quasi-equilibrium in the private ownership economy $E_q(q^1, \dots, q^k; t)$.

2. If E_q is the Debreu economy, $E_q = E_p$, where the sequence $((x^{a*})_{a \in A}, (y^{b*})_{b \in B}, p)$ (see (9)) is the state of equilibrium in economy E_p and

$$\sum_{a \in A} x^{a*} - \omega \in V, \tag{33}$$

then the sequence

$$\left((x^{a*})_{a \in A}, \tilde{Q}(y^{b*}, t)_{b \in B}, p\right) = (x^{a_1*}, \dots, x^{a_m*}, \tilde{Q}(y^{b_1*}, t), \dots, \tilde{Q}(y^{b_n*}, t), p) \tag{34}$$

is the state of equilibrium in Debreu economy $E_p(q^1, \dots, q^k; t)$.

Proof.
1. Let $t \in [0,1]$ be given. By (27)

$$\forall b \in B \ \forall t \in [0,1] \ p \circ \hat{y}^b = p \circ \tilde{Q}(\hat{y}^b, t).$$

Hence, wealth $w(a)$ (see (5)) of every consumer $a \in A$ remains unchanged. Consequently condition

$$\forall \, a \in A \ \ p \circ \hat{x}^a \leq p \circ e(a) + \sum_{b \in B} \theta(a, b) \cdot (p \circ \tilde{Q}(\hat{y}^b, t))$$

is valid. By (7) and (31) we get that:

$$\sum_{b \in B} \hat{y}^b = \sum_{a \in A} \hat{x}^a - \omega \in V.$$

The linearity of the mapping $\tilde{Q}(\cdot, t)$ implies that

$$\tilde{Q}(\textstyle\sum_{b \in B} \hat{y}^b, t) = \textstyle\sum_{b \in B} \tilde{Q}(\hat{y}^b, t)$$

and

$$\textstyle\sum_{a \in A} \hat{x}^a - \textstyle\sum_{b \in B} \tilde{Q}(\hat{y}^b, t) = \omega. \tag{35}$$

From the above we infer that the first condition by the thesis of the theorem is fulfilled.

2. Let $t \in [0,1]$ be given. By theorem 3.3, vector $\tilde{Q}(y^{b*}, t)$ maximizes at price p the profit of every producer b on the production set $\tilde{Q}(Y^b, t)$. By (27), the wealth

$$w(a) = p \circ e(a) + \textstyle\sum_{b \in B} \theta(a, b) \cdot (p \circ y^{b*})$$

(see (5)) of every consumer remains unchanged. Consequently, the consumers' budgets sets are the same as in the initial economy. In this situation the inequality

$$\forall\, a \in A \quad p \circ x^{a*} \le p \circ e(a) + \textstyle\sum_{b \in B} \theta(a, b) \cdot (p \circ \tilde{Q}(y^{b*}, t)),$$

is satisfied. Hence vector x^{a*} also maximizes, at price p, the preference of every consumer a on the budget set $\beta^a(p)$. By (10) and (33) the following may be easily inferred:

$$\textstyle\sum_{b \in B} y^{b*} = \textstyle\sum_{a \in A} x^{a*} - \omega \in V.$$

By the linearity of the mapping $\tilde{Q}(\cdot, t)$ we get that

$$\textstyle\sum_{a \in A} x^{a*} - \textstyle\sum_{b \in B} \tilde{Q}(y^{b*}, t) = \omega \tag{36}$$

is valid, which gives the result.

$\square$

Now, we have:

Theorem 3.7. Let $\tilde{Q}$ be an operator of the form (24) determined by vectors $q^1, \ldots, q^k$ satisfying (23). Assume that condition (14) is satisfied in economy E_q, $p \in V^T$ and

$$\forall a \in A \ e(a) \in V. \tag{37}$$

1. If the sequence $((\hat{x}^a)_{a\in A}, (\hat{y}^b)_{b\in B}, p)$ (see (8)) is the state of quasi-equilibrium in economy E_q, then the sequence $\left((\hat{x}^a)_{a\in A}, (\tilde{Q}(\hat{y}^b, t))_{b\in B}, p\right)$ (see (32)) is the state of quasi-equilibrium in private ownership economy $E_q(q^1, \dots, q^k; t)$.

2. If E_q is the Debreu economy, $E_q = E_p$, where the sequence $((x^{a*})_{a\in A}, (y^{b*})_{b\in B}, p)$ (see (9)) is the state of equilibrium in economy Debreu E_p, then the sequence $\left((x^{a*})_{a\in A}, \tilde{Q}(y^{b*}, t)_{b\in B}, p\right)$ (see (34)) is the state of equilibrium in Debreu economy $E_p(q^1, \dots, q^k; t)$.

Proof. Fix $t \in [0,1]$. By assumptions (14), (37) and formulas (2)–(5), keeping in mind that $p \in V^T$, we get that

$$\forall\, a \in A \;\; \beta^a(p) = X^a \;\; \text{and} \;\; \forall\, a \in A \;\; e(a) \in X^a \subset V.$$

Moreover,

$$\forall t \in [0,1] \;\; \forall b \in B \;\; \forall y^b \in Y^j \;\; p \circ \tilde{Q}(y^b, t) = 0.$$

Hence, the condition

$$\forall\, a \in A \;\; 0 = p \circ \tilde{x}^a \leq p \circ e(a) + \sum_{b\in B} \theta(a,b) \cdot (p \circ \tilde{Q}(\hat{y}^b, t)),$$

or

$$\forall\, a \in A \;\; 0 = p \circ x^{a*} \leq p \circ e(a) + \sum_{b\in B} \theta(a,b) \cdot (p \circ \tilde{Q}(y^{b*}, t))$$

is satisfied, respectively. The conditions (35) or (36) can be proved in the same way as in the proof of theorem 3.6, which completes the proof.

$\square$

The immediate consequence of theorems 3.6 and 3.7 is the following:

Theorem 3.8. Let $p \in \mathbb{R}^\ell$ and $E_p = (\mathbb{R}^\ell, P, C, \theta, \omega)$ be the Debreu economy satisfying conditions (14) and (37) with a proper subspace $V \subset \mathbb{R}^\ell$ of the form (11). There is an operator $\tilde{Q}$ of the form (24) such that if the sequence $((x^{a*})_{a\in A}, (y^{b*})_{b\in B}, p)$ (see (9)) is the state of equilibrium in economy E_q, then the sequence $\left((x^{a*})_{a\in A}, \tilde{Q}(y^{b*}, t)_{b\in B}, p\right)$ (see (34)) is the state of equilibrium in Debreu economy $E_p(q^1, \dots, q^k; t)$.

Proof. The proof goes in the same way as the proofs of the second parts of the thesis of theorems 3.6 and 3.7.

$\square$

Notice that, the quasi-production system $P_q(q^1, \ldots, q^k; 1)$ – the component of the economy $E_q(q^1, \ldots, q^k; 1)$, is the image of the quasi-production system P_q – the component of economy E_q by projection $Q(\cdot) = \tilde{Q}(\cdot, 1)$ (see (25)) determined by vectors $q^1, \ldots, q^k$. In this sense we say that both: the quasi-production system $P_q(q^1, \ldots, q^k; 1)$ and the economy $E_q(q^1, \ldots, q^k; 1)$ are determined by vectors $q^1, \ldots, q^k$ (def. 3.5).

Remark 3.9. Consider a Debreu economy $E_p = \left(\mathbb{R}^\ell, P, C, \theta, \omega \right)$ satisfying conditions (14) and (37) with a subspace $V \subset \mathbb{R}^\ell$. Let vectors $q^1, \ldots, q^k \in \mathbb{R}^\ell$ determine mapping $\tilde{Q}$ by the thesis of theorem 3.8. Theorem 3.8 guarantees the existence of equilibrium in every economy $E_p(q^1, \ldots, q^k; t)$ for $t \in [0,1]$ if equilibrium exists in initial economy E_p. Observe that mapping $\breve{Q} \colon \mathbb{R}^\ell \times \mathbb{R}_+ \to \mathbb{R}^\ell$

$$\breve{Q}(x, t) = x - t \cdot \sum_{s=1}^{k} \tilde{g}^s(x) \cdot q^s$$

is the semi-dynamical system (see Sibirskij, Szube, 1987) in space $\mathbb{R}^\ell$. For every $t \geq 0$, the production system $P(q^1, \ldots, q^k; t)$ (see def. 3.4) is, besides price system p, the image of production system P from economy, by mapping $\breve{Q}(\cdot, t)$. Similarly, the relational system

$$E_p(q^1, \ldots, q^k; t) = (P(q^1, \ldots, q^k; t), C, \theta, \omega),$$

is the Debreu economy. If p is the equilibrium price vector in economy E_p then sequence

$$(x^{a_1*}, \ldots, x^{a_m*}, \breve{Q}(y^{b_1*}, t), \ldots, \breve{Q}(y^{b_n*}, t), p^*) \in (\mathbb{R}^\ell)^{m+n+1}$$

is the state of equilibrium in economy $E_p(q^1, \ldots, q^k; t)$. Hence, mapping $\breve{Q}$ lets us put the whole systems of economies $\{E_p(q^1, \ldots, q^k; t) \colon t \geq 0\}$ "in motion", where variable t means time. In the course of that motion, the production system P from economy E_p is changed in time, but the rest of relational systems in every economy $E_p(q^1, \ldots, q^k; t) \colon t \geq 0$ are not changed. At $t = 1$ economy $E_p(q^1, \ldots, q^k; 1)$ is, besides the equilibrium price system (which may be but not necessary), contained in the subspace V.

The recipe for producers' adjustment trajectory can be forced by the market or it can be set and driven by a person or an institution. If each producer modifies his activities according to the same trajectory of the form (24), then the transformation of the production sector with keeping equilibrium in the economy will be successful. If some of producers choose a different trajectory than the others do, then generally (despite particular cases) equilibrium will not exist at point $t = 1$. In summary, the potential producers' disagreement on the choice of the trajectory (24) or the exclusion of even one producer from the modification process may cause disequilibrium in the economy at point $t = 1$.

4. THE OPTIMAL PRODUCERS' ADJUSTMENT TRAJECTORY

Now, let us focus on the comparison of producers' adjustment trajectories as well as on the characterization of the trajectory under study, optimal under the criterion of distance minimizations.

Let E_q be a private ownership economy (see def. 2.5) and $p \in \mathbb{R}^\ell$ be the price system in economy E_q. Consider a proper linear subspace $V \subset \mathbb{R}^\ell$, $V \neq \{0\}$ defined as in (11). Let us notice that the system of equalities (23) has only one solution if and only, if $k = \ell$ (ℓ – number of commodities, k – number of functionals describing subspace V). Hence, for $V \neq \{0\}$ the number of functionals defining subspace V (see (11)) is less than the number of commodities ($k < l$). If $k \in \{1, \dots, \ell - 1\}$ then system of equalities (23) has infinitely many solutions. If $p \notin V^T$, then system of equalities (28) has only one solution if and only, if $k = \ell - 1$. If $k \in \{1, \dots, \ell - 2\}$ then the system of equalities (23) has infinitely many solutions. Hence, the producers, who want to change their production plans, have often infinitely many possibilities of choice of trajectories of the form (24).

Assuming that producers want to, or have to change their production activity, other problems (questions) arise. How to compare producers' adjustment trajectories? Which of them are the best or satisfactory enough for the producers? This certainly depends on the criterion of the choice. We assume that the producers, keeping in mind the necessity of modifying their plans to plans contained in subspace V, want also to minimize the costs. It results in changing the activities of producers as little as possible. It means that the difference between the respective coordinates of every production plan and its modification will be also properly small. Moreover, the producers' plans contained in subspace V should remain unchanged. Hence, we determine for every $x = (x_1, \dots, x_\ell) \in \mathbb{R}^\ell$, the norm

$$\|x\| = max\{|x_l| : l \in \{1, 2, \dots, \ell\}\}. \tag{38}$$

Assume that the given subspace $V \subset \mathbb{R}^\ell$ is defined (see (11)) by functionals $\tilde{g}^1, \dots, \tilde{g}^k$ of the form (12). Let $Q \in \mathcal{P}(\mathbb{R}^\ell, V)$, $Q(x) = x - \sum_{s=1}^k \tilde{g}^s(x) \cdot q^s$ be the projection (see (25)) determined by vectors $q^1, \dots, q^k$ satisfying (23). It is well known (see Cheney, 1966) that, for every $x \in \mathbb{R}^\ell$

$$dist(x, V) \leq \|(Id - Q)(x)\| \leq \|Id - Q\| dist(x, V), \tag{39}$$

where

$$\|Id - Q\| = \sup\{\|(Id - Q)(x)\| : x \in \mathbb{R}^\ell \wedge \|x\| \leq 1\}. \tag{40}$$

The number $\|Id - Q\|$ can be interpreted as the distance (by (39)) between the initial economy E_q and its final modification $E_q(q^1, \dots, q^k; 1)$. Hence the projection Q is

identified with the producers' adjustment trajectory $\tilde{Q}(x,t) = x - t \cdot \sum_{s=1}^{k} \tilde{g}^s(x) \cdot q^s$ (see (24)). By this reason, the projection Q is also called the producers' adjustment trajectory and the number $\|Id - Q\|$ is called the coefficient of the change of the economy E_q determined by trajectory Q. We also say that the projection Q realizes the number $\|Id - Q\|$.

It is apparent, by (39) and (40), that $\|Id - Q\| \geq 1$. Moreover, if the norm $\|Id - Q\|$ is not large, then the production plans and their modifications are close in terms of distance. Hence, the mapping $Q \in \mathcal{P}(\mathbb{R}^\ell, V)$ for which number $\|Id - Q\|$ is the smallest possible, is the producers' adjustment trajectory optimal (best) under the criterion of distance minimization. It will be called the optimal producers' adjustment trajectory.

Keeping in mind the motivations of producers and properties of projections presented above, we define the preference relation of producer $b \in B$ in the set $\mathcal{P}(\mathbb{R}^\ell, V)$. Let $b \in B$ and $V \subset \mathbb{R}^\ell$, $V \neq \{0\}$ be a subspace of the form (11), defined by functionals $\tilde{g}^1, \ldots, \tilde{g}^k$ given by (12). Let us notice, that without loss of generality we can assume that for every $s \in \{1, \ldots, k\}$, where $k \in \{1, \ldots, \ell - 1\}$, $\sum_{l=1}^{\ell} |g_l^s| = 1$. Define, for every functional $\tilde{g}^s$, $s \in \{1, \ldots, k\}$, the set

$$supp\,\tilde{g}^s = \{l \in \{1, \ldots, \ell\}: g_l^s \neq 0\}$$

The changes in producers' activities, which imply that condition (15) is satisfied, are called the adjustment of technologies to subspace V. It is said that a producer b is neutral to the adjustment of technologies to subspace V, if

$$l \in \bigcup_{s=1}^{k} supp\,\tilde{g}^s \implies \forall y^b \in Y^b \; y_l^b = 0. \tag{41}$$

So, the initial plans of producers, who are neutral to the adjustment of technologies to subspace V, are contained in V. The set of producers neutral to the adjustment of technologies to subspace V will be denoted by B_0.

To model the changes in the production system relaying on the adjustment of technologies to subspace V, it is worth assuming that there is at least one producer who is not neutral to the adjustment of technologies to V ($B \backslash B_0 \neq \emptyset$) as well as that every producer from set $B \backslash B_0$ will modify his production plans under the criterion of distance minimization. Let $Q_1, Q_2 \in \mathcal{P}(\mathbb{R}^\ell, V)$. Taking the above criterion into consideration, the preference relation of producer $b \in B \backslash B_0$ in set $\mathcal{P}(\mathbb{R}^\ell, V)$ is defined as follows

$$Q_2 \preccurlyeq^b Q_1 \iff \|Id - Q_1\| \leq \|Id - Q_2\|. \tag{42}$$

Hence, for every producer from set $B \backslash B_0$ is assigned the same preference relation of the form (42). By the fact that every producer $b \in B_0$ will not change his production set, all the producers' adjustment trajectories (projections from set $\mathcal{P}(\mathbb{R}^\ell, V)$) are indifferent for him. The above indifference will be traditionally marked, for $Q_1, Q_2 \in \mathcal{P}(\mathbb{R}^\ell, V)$, by

$$Q_2 \sim^b Q_1. \tag{43}$$

Condition (43) means that

$$Q_2 \leqslant^b Q_1 \quad \text{and} \quad Q_1 \leqslant^b Q_2$$

for $b \in B_0$. Let us recall that every producer $b \in B \backslash B_0$ wants to, or has to modify his production plans under the criterion of distance minimization. Additionally, according to the rationality assumption he will choose the optimal producers' adjustment trajectory provided such exists. Combining conditions (42) and (43), we define the preference relation $\leqslant$ in set $\mathcal{P}(\mathbb{R}^\ell, V)$ by the rule

$$Q_2 \leqslant Q_1 \Longleftrightarrow \forall b \in B \ \ Q_2 \leqslant^b Q_1. \tag{44}$$

The relation $\leqslant$ defined in (44) is the producers' preference relation in the set of defined producers' adjustment trajectories.

We will explain that the producers' preference relation defined in (44) will have a maximal element. Notice that, the dimension of commodity-price space $\mathbb{R}^\ell$ is finite, then the problem of the distance minimization in set $\mathcal{P}(\mathbb{R}^\ell, V)$ has a solution (see for example Cheney, 1966). Precisely, there is a projection $Q_0 \in \mathcal{P}(\mathbb{R}^\ell, V)$ such that

$$\|Id - Q_0\| = inf\{\|Id - Q\| : Q \in \mathcal{P}(\mathbb{R}^\ell, V)\} \tag{45}$$

(see Lewicki Odyniec, 1990). The projection $Q_0 \in \mathcal{P}(\mathbb{R}^\ell, V)$ satisfying (45), is called the cominimal projection (see Lipieta, 1999). It is obvious that the cominimal projection Q_0 minimizes the distance between the initial economy E_q and its final modifications $E_q(q^1, \dots, q^k; 1)$. That means that Q_0 is the maximal element of the producers' preference relation defined in (44) and it minimizes the coefficient of the change of the economy.

Unfortunately, the formula for cominimal projections as well as the number (45) are not known besides some cases. However, if properties (18) or (22) are fulfilled, then the problem of indicating such cominimal projection Q_0, for which $\|Id - Q\| = 1$, becomes quite simple. Namely, the following is true:

Theorem 4.1 (theorem 3.1 in Lipieta, 1999). Let $V \subset \mathbb{R}^\ell$, $V \neq \{0\}$ be a linear subspace of the form (11) defined by functionals $\tilde{g}^1, \dots, \tilde{g}^k$ of the form (12) satisfying, for every $s \in \{1, \dots, k\}$, condition $\sum_{l=1}^\ell |g_l^s| = 1$. Then
- $\|Id - Q\| = 1$ if and only if $\bigcap_{s=1}^k supp\tilde{g}^s = \emptyset$.
- if $\|Id - Q\| = 1$, then the cominimal projection Q_0 is determined by vectors $q^1, \dots, q^k \in \mathbb{R}^\ell$ such that if $g_l^i \neq 0$ for some $i \in \{1, \dots, k\}$ and $l \in \{1, 2, \dots, \ell\}$, then for $s \in \{1, \dots, k\}$

$$q_l^s = \begin{cases} 1 & \text{for} \ \ s = i \land \ g_l^i > 0, \\ -1 & \text{for} \ \ s = i \land \ g_l^i < 0, \\ 0 & \text{for} \ \ s \neq i. \end{cases} \qquad (46)$$

The projection Q_0 determined by vectors $q^1, ..., q^k \in \mathbb{R}^\ell$ of the form (4.9), under the assumptions that $\bigcap_{s=1}^k supp \tilde{g}^s = \emptyset$ as well as, for every $s \in \{1, ..., k\}$, $\sum_{l=1}^\ell |g_l^s| = 1$, is the closest to identity mapping. Then, changes in the production sector induced by projection Q_0 are the least, among changes determined by other projections.

Consider a subspace $V \subset \mathbb{R}^\ell$ defined as in (3.11) with functionals $\tilde{g}^1, ..., \tilde{g}^k$ of the form (12). Let $\bigcap_{s=1}^k supp \tilde{g}^s = \emptyset$ and for every $s \in \{1, ..., k\}$, $\sum_{l=1}^\ell |g_l^s| = 1$. Assume that the initial Debreu economy $E_p = (\mathbb{R}^\ell, P, C, \theta, \omega)$ satisfies assumption α_1 or α_2, where
- α_1: $p \notin V^T$, $q^1, ..., q^k$ are the solution of (28) as well as (33) is fulfilled,
- α_2: $p \in V^T$, $q^1, ..., q^k$ satisfy (3.13) as well as conditions (14) and (37) are fulfilled.

Then projection Q determined by vectors $q^1, ..., q^k$ "moves" the economy E_p from its initial equilibrium state into another equilibrium state. On the basis of the above, if vectors $q^1, ..., q^k$ satisfy additionally condition (46), then projection Q determined by vectors $q^1, ..., q^k$, projection is the optimal producers' adjustment trajectory, $Q = Q_o$ which gives the least coefficient of the change of the considered economy. Similarly, mapping $\tilde{Q}_o$ of the form (24), determined by vectors of the form (46), is also the optimal producers' adjustment trajectory. Moreover, for every $t \in [0,1]$, the production system $P(q^1, ..., q^k; t)$ (see def. 3.4) as well as the whole economy $E_p(q^1, ..., q^k; t)$ is the image, by the mapping $\tilde{Q}_o$, of the initial production system P or the private ownership economy E_p respectively.

If vectors $q^1, ..., q^k$ of the form (46) are not orthogonal to the price vector $p \notin V^T$ (it means that condition (28) and consequently assumption α_1 are not satisfied), then there is no equilibrium in economy $E_q(q^1, ..., q^k; 1)$, despite some particular cases. Moreover, the economy $E_q(q^1, ..., q^k; 1)$ (see def. 2.5 and 3.5) does not have to be the Debreu economy.

Let us emphasize that if a producer, who is not neutral to the adjustment of technologies to subspace V, does not change his production plans according to mapping $\tilde{Q}$ of the form (24), then equilibrium could not exist at least at one $t \in (0,1)$. If every producer from set $B \backslash B_0$, under the assumption α_1 or α_2, follows the same trajectory $\tilde{Q}$ as well as the producers from set B_0 do not change their production plans, then there will be equilibrium in the economy with modified production system if it existed in the initial economy (see also theorem 3.8).

Let us notice that either new firms or commodities do not appear and are not eliminated from the producers' activities in the considered modifications of the economy under study. Moreover, the technologies are mildly modified as well as the prices are not changed. These result in the same profits. Hence, the production sector of the discussed economy evolves in the framework of the Schumpeterian circular flow (see Schumpeter, 1912; Lipieta, 2013).

5. CONCLUSIONS

The results of this research lead to simplifying the geometric structure of the initial economy. It is caused by the appearance of the linear dependency between quantities of some commodities in all producers' plans in the modified form of the economy or by the elimination of some harmful commodities from the production processes. Consequently, the correspondences and functions – the components of the final production system – depend, in fact, on fewer variables than in the beginning.

On the basis of the above, the prerequisites for the appearance of the optimal producers' adjustment trajectory under the criterion of distance minimization, with the smallest possibly the coefficient of the change of the considered economy, were presented. The definition of the mentioned trajectory also has been formulated.

If the changes of producers' activities are caused by other reasons than those considered in the paper, or the criterion for comparing producers' adjustment trajectories is different, then the recipe for producers' adjustment trajectories as well as the optimal producers' adjustment trajectories might be modified. The studies on designing the producers' adjustment trajectories, under other criterions, still remains within our research plans.

REFERENCES

Aliprantis C. D., (1996), *Problems in Equilibrium Theory*, Springer – Verlag Berlin, Heidelberg, German.
Cheney E. W., (1966), *Introduction to Approximation Theory*, Mc Grow Hill, New York.
Debreu G., (1959), *Theory of Value*, New York, Viley.
Lewicki G., Odyniec W., (1990), *Minimal Projections in Banach Spaces*, Lecture Notes in Mathematics, 1449, Springer-Verlag, Berlin/New York.
Lipieta A., (1999), Cominimal Projections in . *Journal of Approximation Theory*, 98, 86–100.
Lipieta A., (2010), The Debreu Private Ownership Economy with Complementary Commodities and Prices, *Economic Modelling*, 27, 22–27.
Lipieta A., (2012), The Economy with Production and Consumption Systems Changing in Time, *Przegląd Statystyczny*, 59 (3), 233–245.
Lipieta A., (2013), Mechanisms of Schumpeterian Evolution, in: Malawski A., (ed.), *Innovative Economy as the Object Investigation in Theoretical Economics*, Cracow University of Economics Press, 94–119.
Magill M., Quinzii M., (2002), *Theory of Incomplete Markets*, MIT Press, Cambridge.
Mas-Colell A., Whinston M. D., Green J. R., (1959), *Microeconomic Theory*, Oxford University Press, New York.
Moore J., (2007), *General Equilibrium and Welfare Economics*, Springer Berlin-Heidelberg-New Jork.
Radner R., (1972), Existence of Equilibrium of Plans, Prices and Price Expectations in a Sequence of Markets, *Econometrica*, 40 (2), 289–303.
Sibirskij K. S., Szube A. S., (1987), *Semidynamical Systems* (Topological Theory). Russian, Sztinica, Kiszyniów, Belknap Press of Harvard University Press.
Schumpeter J. A., (1912), *Die Theorie der Wirtschaftlichen Entwicklung*, Leipzig, Duncker & Humblot; English translations, The Theory of Economic Development, Cambridge, MA, Harvard University Press 1934 and A Galaxy Book, New York, Oxford University Press 1961.
Varian H. R., (1999), *Intermediate Microeconomics, A Modern Approach.* W. W. Norton & Company, New York, London.

OPTYMALNA TRAJEKTORIA DOSTOSOWAWCZA PRODUCENTÓW

S t r e s z c z e n i e

W artykule została zdefiniowana grupa ścieżek dostosowawczych (trajektorii), które opisują niezbędne zmiany w sferze produkcji, spowodowane koniecznością lub chęcią producentów dostosowania swojej działalności na rynkach do danych wymogów. Działalność producentów jest modelowana w ekonomii Debreu, a wymogi są zadane analitycznie przez funkcjonały liniowe.

Rozważane trajektorie ilustrują zmiany w systemie produkcji, które nie zaburzają równowagi w ekonomii w okresie transformacji, chociaż początkowy stan równowagi może ulec zmianie. W zbiorze omawianych trajektorii została zdefiniowana relacja preferencji, której element maksymalny tzw. optymalna trajektoria dostosowawcza producentów, przy pewnych założeniach, wyznacza kierunek najbardziej korzystnych dla producentów zmian, z punktu widzenia minimalizacji strat.

Słowa kluczowe: ekonomia z własnością prywatną, równowaga, zbiory liniowe, projekcje

THE OPTIMAL PRODUCERS' ADJUSTMENT TRAJECTORY

A b s t r a c t

The trajectories illustrating the necessary changes in the production sphere, which are caused by the necessity or the wish of producers, who adjust their activities to the given requirements, are analyzed in the paper. The producers' activities are modeled in the Debreu economy, while the requirements are given analytically, by using the linear functionals.

If the producers change their plans of action due to the considered trajectories, equilibrium in the economy will be kept, although the initial state of equilibrium can be replaced by the other one.

In the set of trajectories under study, the preference relation is defined. Under some assumptions, the maximal element of the above relation, so called the best producers' adjustment trajectory, indicates the best path of changes in producers' activities, under the criterion of losses minimization.

Keywords: private ownership economy, equilibrium, linear sets, projections

MARCIN PEŁKA[1], ANDRZEJ DUDEK[2]

THE COMPARISON OF FUZZY CLUSTERING METHODS FOR SYMBOLIC INTERVAL-VALUED DATA

1. INTRODUCTION

In general terms, clustering methods seek to organize certain sets of objects (items) into clusters in the way allowing objects from the same cluster be more similar to each other than to objects from other clusters. Usually such similarity is measured by some distance measure (e.g. Euclidean, Manhattan, etc.). Successful application of these methods has been confirmed in many different areas such as taxonomy, image processing, data mining, etc. In general, clustering techniques can be divided into two groups of methods – hierarchical (agglomerative or divisive) and partitioning (see, e.g. Gordon, 1999; Jain et al., 1999).

In cluster analysis, objects (patterns) are usually described by single-valued variables. This allows representing each object as a vector of qualitative or quantitative measurements, where each column represents a variable.

However, this kind of data representation is too restrictive to cover more complex data. If the uncertainty and/or variability of the data are to be taken into account, variables must assume sets of categories or intervals, including frequencies or weights in some cases.

The discussed data are primarily studied using *Symbolic Data Analysis* (SDA). The main aim of Symbolic Data Analysis is to provide suitable methods for managing aggregated or complex data, described by multi-valued variables, where the cells of a data table contain sets of categories, intervals, or weight (probability) distributions (see e.g. Billard, Diday, 2006; Bock et al., 2000).

Conventional hard clustering methods restrict each object of the data set to exactly one cluster. Fuzzy clustering generates a fuzzy partition using the idea of partial membership, expressed by the degree of membership of each object in a given cluster.

In terms of the real-valued data, Dunn (1973) presented one of the first fuzzy clustering methods applying an adequacy criterion based on the Euclidean distance. Bezdek (1981) generalized this method even further. Diday, Govaert (1977) offered

―――――――
[1] Wrocław University of Economics, Department of Econometrics and Computer Science, 3 Nowowiejska St., 58-500 Jelenia Góra, Poland, corresponding author – e-mail: marcin.pelka@ue.wroc.pl.
[2] Wrocław University of Economics, Department of Econometrics and Computer Science, 3 Nowowiejska St., 58-500 Jelenia Góra, Poland.

one of the first approaches to use adaptive distances in the partitioning of quantitative data. Gustafson, Kessel (1979) introduced the first adaptive fuzzy clustering, based on a quadric distance defined by fuzzy covariance matrix.

More recently, De Carvalho et al. (2006) introduced fuzzy c-means clustering algorithms based on adaptive quadratic distances. These distances can be defined by full as well as diagonal fuzzy covariance matrices (estimated globally), or by diagonal fuzzy covariance matrices (estimated locally for each cluster).

Finite-sample properties of spectral clustering have been theoretically studied by many scientists (see Ng et al., 2002; Shi, Malik, 2002; Meila, Shi, 2001; Chung, 1997; von Luxburg et al., 2005; von Luxburg, 2006; Kannan et al., 2000; Guattery, Miller, 1998; de Sa, 2005). Spectral clustering has the advantage of performing well in the presence of the non-Gaussian clusters. This method is also easy to implement. Furthermore, it is also not a disadvantage for the local minima presence (von Luxburg et al., 2005). Additionally, the convergence of the normalized spectral clustering is less difficult to handle than the unnormalized one (von Luxburg et al., 2005). The results obtained by spectral clustering frequently outperform the traditional approaches (see e.g. von Luxburg, 2006). It is due to the fact that spectral clustering makes no assumptions regarding the form of clusters − it can solve very general clustering problems (von Luxburg, 2006, p. 22).

Spectral clustering, however, has certain disadvantages. It can be quite unstable under different choices of parameters for the neighborhood graphs. Many different kernels can be used, each of them leading to different results (Gaussian kernel is used at most cases) − Karatzoglou (2006) presents the applications of different kernels in spectral clustering.

Another important task is to choose a good σ value for the kernel − in the paper published by Karatzoglou (2006) quite an efficient way for estimating the appropriate σ parameter has been proposed. σ is a scaling parameter which should minimize the sum of inter-cluster distances for a given number of clusters. Usually a heuristic algorithm is used to find the best σ value.

Cominetti et al. (2010) proposed a fuzzy spectral clustering algorithm for complex data, referred to as DifFUZZY − which combines the ideas of fuzzy clustering and spectral clustering. It is applicable to a larger class of clustering problems. DifFUZZY is better than traditional fuzzy clustering algorithms in handling "curved" and elongated data sets or those which contain different dispersion (see Cominetti et al., 2010, p. 1). Moreover, DifFUZZY does not require any prior information on the number of clusters. The algorithm of DifFUZZY may be divided into three main steps: 1) the construction of σ-neighborhood graph using the Euclidean norm, to be followed by applying this graph in determining the number of clusters. 2) computation of auxiliary matrices $\mathbf{W}$, $\mathbf{D}$, $\mathbf{P}$, the definition of which can be intuitively understood in terms of diffusion processes on graphs. Matrix $\mathbf{W}$ uses the idea of Gaussian kernel. Matrix $\mathbf{D}$ is defined as a diagonal matrix with diagonal elements equal to $\sum_{j=1}^{N} w_{ij}$ (where w_{ij} are

the elements of matrix **W**). Matrix **P** is calculated from the identity matrix, **W** and **D** matrices, as well as γ_2 which is an internal parameter of DifFUZZY. The default value of this parameter is 0.1. 3) calculation of membership values of even soft points.

Cominetti et al. (2010) showed that the fuzzy spectral algorithm DifFUZZY performs well in a number of data sets (both artificial and real) with sizes ranging from tens to hundreds of data points presenting dimensions as high as hundreds.

Interval-valued variables are needed, e.g. when an object represents a group of individuals and the variables used to describe it need to take the value which expresses the variability inherent in the description of a group. Such data arise in practical situations, e.g. recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. Another source of interval-valued data is the aggregation of huge databases into a reduced number of groups, the properties of which are described by interval-valued variables.

Construction of clustering methods for interval-valued data must take into account that for this kind of data operations of adding, subtracting, multiplying, squaring, calculation of means or calculation of variance are not defined. In literature of the subject some proposals of construction of similar operations and measures can be found can be found (see Billard, Diday, 2006, p. 69–142) and some author even use achievements of interval algebra (Moore, 1966). The most common approach, although, is to base the clustering algorithm on distance measure dedicated for symbolic Boolean object (see table 1) instead of direct calculation on interval variables.

Symbolic Data Analysis provides a number of fuzzy clustering algorithms for interval-valued data. El-Sonbaty, Ismail (1998) introduced a fuzzy *c*-means algorithm to cluster data on the basis of different types of symbolic variables. Yang et al. (2004) proposed fuzzy clustering algorithms for mixed features of symbolic and fuzzy data. De Carvalho (2007) introduced a fuzzy *c*-means and adaptive fuzzy *c*-means methods for interval-valued data, based on the general form of the Euclidean distance. De Carvalho, Tenório (2010) introduced fuzzy *k*-means clustering algorithms for interval-valued data based on adaptive quadric distances.

However, none of the fuzzy clustering methods for interval-valued data presented so far uses the spectral clustering approach. The spectral clustering algorithm proposed by Ng et al. (2002), based on spectral decomposition of the distance matrix, does not, in fact, represent a new clustering method, but rather a new way of preparing data for the well-known *k*-means method. We propose to adapt this popular way of preparing the inputted data to manage interval-valued symbolic data and then to apply the well-known fuzzy *c*-means clustering algorithm.

The recommended algorithm gives a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on a suitable Euclidean distance.

Simulation studies, with artificial and real data sets, confirm the usefulness of the suggested method when dealing data with different cluster structures, noisy variables and/or outliers.

The presented paper is organized as follows. Section 2 discusses three fuzzy clustering methods for symbolic interval-valued data – the fuzzy c-means clustering, the adaptive fuzzy c-means clustering, the fuzzy k-means clustering and compares them with the proposed spectral fuzzy c-means algorithm for clustering symbolic data.

To show the usefulness and stability of the proposed method, section 3 presents the results of evaluation studies with different synthetic interval-valued data sets, as well as the application with real interval-valued data sets. Section 4 offers the concluding remarks.

2. FUZZY CLUSTERING METHODS FOR SYMBOLIC INTERVAL-VALUED DATA

Let $\Omega = \{e_1,\dots,e_n\}$ be the set of n objects (patterns), where each object is indexed by k and described by p interval-valued variables $\{y_1,\dots,y_p\}$ where each variable is indexed by j. An *interval-valued variable X* (see e.g. Billard, Diday, 2006; Bock et al., 2000) is a correspondence defined from Ω in $\Re$ such that for each $k \in \Omega, X(k) = [a,b] \in \Im$, where $\Im\{[a,b]: a,b \in \Re, a \le b\}$ is the set of closed intervals defined from $\Re$. Each object k is represented as a vector of intervals $\mathbf{x}_k = (x_{k1},\dots,x_{kp})$, where $x_{kj} = [a_{kj}, b_{kj}] \in \Im$. In this paper, an interval data table $\{x_{kj}\}_{n \times p}$ is made up of n rows standing for n objects and p columns representing p symbolic variables. Each cell of this data table, called also a symbolic data table or a symbolic data matrix, contains an interval $x_{kj} = [a_{kj}, b_{kj}] \in \Im$. Such a symbolic data matrix serves as input data for the computation of a distance matrix through a distance measure suitable for interval-valued data.

As it has been mentioned in section 1 there are three main fuzzy clustering methods for symbolic interval-valued data (see de Carvalho, 2007; de Carvalho, Tenório, 2010):
1. Fuzzy c-means clustering.
2. Adaptive fuzzy c-means clustering.
3. Fuzzy k-means clustering.

Fuzzy c-means clustering for symbolic interval-valued data (IFCM) aims at furnishing the fuzzy partition of a data set and a corresponding set of prototypes, so that criterion W^1 measuring the fitting between clusters and their representatives (prototypes) is locally minimized, which is defined as follows:

$$W^1 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \left[(a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right], \tag{1}$$

where: a_{kj} and b_{kj} represent lower and upper bounds of an interval for an object, whereas α_{ij} and β_{ij} stand lower and upper bounds of an interval for cluster prototype.

Fuzzy c-means clustering for the interval-valued data is carried out in the following steps (see de Carvalho, 2007, p. 426):
1. Initialization. Fix number of clusters c, $2 \le c < n$, fix fuzzification parameter m, $1 < m < \infty$, fix iteration limit T, fix $\varepsilon > 0$. Initialize $u_{ik} (k = 1,\dots,n)$ and $(i = 1,\dots,c)$ of pattern k belonging to cluster P_i so that $u_{ik} \ge 0$ and $\sum_{i=1}^{c} u_{ik} = 1$.

2. $t = 1$.
3. Representation step. Membership degree u_{ik} of pattern k belonging to cluster P_i is fixed. Compute the class prototypes:

$$\alpha_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m \cdot a_{kj}}{\sum_{k=1}^{n} (u_{ik})^m}, \quad \beta_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m \cdot b_{kj}}{\sum_{k=1}^{n} (u_{ik})^m}. \tag{2}$$

4. Allocation step. The prototypes $\mathbf{g}_i$ of class P_i are fixed. Update the fuzzy membership degree as follows:

$$u_{ik} = \left[\sum_{h=1}^{c} \left(\frac{\sum_{j=1}^{p} \left[(a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right]}{\sum_{j=1}^{p} \left[(a_{kj} - \alpha_{hj})^2 + (b_{kj} - \beta_{hj})^2 \right]} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{3}$$

5. Stopping criterion. If $\left| W_{t+1}^1 - W_t^1 \right| \le \varepsilon$ or $t > T$ then stop, else $t = t + 1$ and go to step 3 (representation step).

    Adaptive fuzzy *c*-means clustering for the interval-valued data (IFCMADS) has the same purpose as fuzzy *c*-means clustering for symbolic interval-valued data, but criterion W^2 is defined as follows:

$$W^2 = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m \sum_{j=1}^{p} \lambda_{ij} \left[(a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right], \tag{4}$$

where: a_{kj} and b_{kj} represent lower and upper bounds of an interval for an object, whereas α_{ij} and β_{ij} stand for lower and upper bounds of an interval for cluster prototype. λ_{ij} are the weights defined by equation 5.

The adaptive fuzzy *c*-means clustering algorithm is carried out based on the following steps (see de Carvalho, 2007, p. 427):
1. Initialization. Fix c, $2 \le c < n$, fix m, $1 < m < \infty$, fix an iteration limit T, fix $\varepsilon > 0$. Initialize u_{ik} in the same manner as in the IFCM clustering algorithm.
2. $t = 1$.
3. Representation step:
    Stage 1: Membership degree u_{ik} is fixed. Compute cluster prototypes in the same way as in IFCM clustering algorithm.
    Stage 2: Membership degree u_{ik} is fixed and class prototypes are fixed. Compute the vector of weights λ_i as follows:

$$\lambda_{ij} = \frac{\left\{ \prod_{h=1}^{p} \left[\sum_{k=1}^{n} (u_{ik})^m \left((a_{kh} - \alpha_{ih})^2 + (b_{kh} - \beta_{ih})^2 \right) \right] \right\}^{\frac{1}{p}}}{\sum_{k=1}^{n} (u_{ik})^m \left((a_{kj} - \alpha_{ij})^2 + (b_{kj} - \beta_{ij})^2 \right)}. \tag{5}$$

4. Allocation step. Class prototypes and the vector of weights λ_i are fixed. Update the fuzzy membership degree u_{ik} of pattern k belonging to cluster P_i as follows:

$$u_{ik} = \left[\sum_{h=1}^{c} \left(\frac{\sum_{j=1}^{p} \lambda_{ij} \left[\left(a_{kj} - \alpha_{ij} \right)^2 + \left(b_{kj} - \beta_{ij} \right)^2 \right]}{\sum_{j=1}^{p} \lambda_{ij} \left[\left(a_{hj} - \alpha_{hj} \right)^2 + \left(b_{kj} - \beta_{hj} \right)^2 \right]} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{6}$$

5. Stopping criterion. If $\left| W_{t+1}^2 - W_t^2 \right| \le \varepsilon$ or $t > T$ then stop, else $t = t + 1$ and go to step 3 (representation step).

Fuzzy *k*-means clustering algorithms for the interval-valued data are based on adaptive quadric distances. Fuzzy *k*-means clustering algorithms optimize an adequacy criterion J measuring the fit between clusters and their prototypes, which is defined as (de Carvalho, Tenório, 2010, p. 2980):

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} \left(u_{ik} \right)^m d_{\mathbf{M}_k (\mathbf{x}_i, \mathbf{y}_k)}^2, \tag{7}$$

where $d_{\mathbf{M}_k (\mathbf{x}_i, \mathbf{y}_k)}^2 = \left(\mathbf{x}_{iL} - \mathbf{y}_{kL} \right)^T \mathbf{M}_k \left(\mathbf{x}_{iL} - \mathbf{y}_{kL} \right) + \left(\mathbf{x}_{iU} - \mathbf{y}_{kU} \right)^T \mathbf{M}_k \left(\mathbf{x}_{iU} - \mathbf{y}_{kU} \right)$ is a suitable adaptive quadric distance between vectors of intervals parameterized by a positive definite symmetric matrix $\mathbf{M}_k$, $\mathbf{x}_{iL}, \mathbf{y}_{kL} \left(\mathbf{x}_{iU}, \mathbf{y}_{kU} \right)$ represent lower (L) and upper (U) bounds of intervals.

There are three types of adaptive quadric distances possible to apply:
– Single adaptive quadric distances defined by a full pooled fuzzy covariance matrix $\mathbf{M}_k = \mathbf{M}$.
– Single adaptive quadric distances defined by a diagonal pooled fuzzy covariance matrix $\mathbf{M}_k = \mathbf{M} = Diag \left(\lambda^1, \ldots, \lambda^p \right)$.
– Cluster adaptive quadric distances defined by a full fuzzy covariance matrix.

The fuzzy *k*-means clustering algorithm, regardless of the distance type, is executed in four main steps (de Carvalho, Tenório, 2010, p. 2981–2982; 2983–2984):
1. Initialization. Fix c (number of clusters), $2 \le c < n$, fix m, $1 < m < \infty$, fix an iteration limit T, fix $\varepsilon > 0$. Initialize u_{ik} in the same manner as in IFCM or IFCMADS algorithms.
2. $t = t + 1$.
3. Definition of the best prototypes, u_{ik} and the corresponding vector of matrices $\mathbf{\Theta} = \{ \mathbf{M}_1, \ldots, \mathbf{M}_k \}$ are fixed. Compute the vector of prototypes in the same way as in IFCM algorithm.
4. Definition of the best distances. Both u_{ik} and the corresponding vector of prototypes are fixed. Compute the vector of positive definite symmetric matrices $\mathbf{\Theta}$.
5. Definition of the best partition. The vector of prototypes and the corresponding matrices of weights $\mathbf{\Theta}$ are fixed. Determine the fuzzy partition represented by u_{ik}:

$$u_{ik} = \left[\sum_{h=1}^{K} \left(\frac{d_{\mathbf{M}_k(\mathbf{x}_i, \mathbf{y}_k)}^2}{d_{\mathbf{M}_k(\mathbf{x}_i, \mathbf{y}_k)}^2} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{8}$$

6. Stopping criterion. If $\left| J_{t+1} - J_t \right| \le \varepsilon$ or $t > T$ then stop, else $t = t + 1$ and go to step 3.
   Spectral fuzzy c-means algorithm covers two steps:
a) spectral decomposition algorithm adapted to deal with an interval-valued data, where $\mathbf{E}'$ or $\mathbf{E}$ matrix is obtained,
b) fuzzy c-means clustering built upon the $\mathbf{E}'$ or $\mathbf{E}$ matrix.
   Spectral decomposition algorithm, adapted to deal with an interval-valued data set, takes the following steps (Ng et al., 2002):
1. Let $\mathbf{X}$ be a symbolic data table with n rows and p columns and let c be the number of clusters.
2. Let $\mathbf{S} = [s_{kl}]$ be a similarity matrix between the objects belonging to $\mathbf{X}$. The similarity matrix can be computed using the below equation:

$$s_{kl} = \frac{d_{kl}}{e^{\sigma^2}}, \tag{9}$$

where d_{kl} is a suitable dissimilarity measure computed on the pair of vectors of intervals and σ is a scaling parameter that should minimize the sum of inter-cluster distances for a given number of clusters. Usually a heuristic algorithm is used to find the best σ value.
3. From the similarity matrix $\mathbf{S} = [s_{kl}]$ compute the matrix of weights $\mathbf{W} = [w_{kl}]$ as follows:

$$w_{kl} = \begin{cases} \sum_{l=1}^{n} d_{kl}^2 & k \ne l, \\ 0 & k = l. \end{cases} \tag{10}$$

4. Then compute the Laplacian $\mathbf{L}$ matrix according to:

$$\mathbf{L} = W^{-\frac{1}{2}} \times \mathbf{S} \times \mathbf{W}^{-\frac{1}{2}}. \tag{11}$$

In the graph theory, $\mathbf{L}$ is treated as the algebraical representation of the graph created from the objects of $\mathbf{X}$.
5. Extract the first c eigenvectors of the Laplacian matrix to create the matrix $\mathbf{E} = [e_{kl}]$. Each eigenvector of $\mathbf{L}$ is a column of $\mathbf{E}$ (thus, matrix $\mathbf{E}$ is $n \times c$ dimensional). Alternatively, instead of matrix $\mathbf{E}$, the normalized matrix $\mathbf{E}' = [e'_{kl}]$ can be considered, which is computed as follows:

$$e'_{kl} = \frac{E_{kl}}{\sqrt{\sum_{i=1}^{n} E_{il}^2}}. \tag{12}$$

6. Finally, a standard clustering algorithm is applied on matrix $\mathbf{E}'$ or $\mathbf{E}$, if the normalization step is omitted to obtain a suitable clustering structure. In the presented paper, the well-known fuzzy c-means represents the considered standard clustering algorithm.

The variants of spectral clustering may differ depending on the applied kernel estimator type. Usually the Gaussian estimator, based on the squared Euclidean distance, is used (for classical data with a ratio and interval variables). For the purposes of interval-valued symbolic data it is suggested to apply the Gaussian estimator based on squared dissimilarity functions that that are suitable for interval-valued data (see Billard, Diday, 2006; Bock et al., 2000; Zelnik-Manor, Perona, 2004). Table 1 presents all suitable distance measures for symbolic interval-valued data available in R software. All of them (except C_1, which is the distance measure for hierarchical or logical dependent symbolic variables) will be used in evaluation studies (see section 3).

Parameter σ represents the key element in spectral clustering. There are many heuristic approaches allowing the selection of the best σ value (see e.g. Fisher, Poland, 2004; Poland, Zeugmann, 2006; Zelnik-Manor, Perona, 2004). Parameter σ can be selected using some descriptive statistics computed from the distance matrix. A better way for selecting it was proposed by Karatzoglou (2006) following which a σ that minimizes the total within the sum of squares of distances, computed between the objects for given u clusters, is searched for.

Fuzzy c-means clustering algorithm (FCM), proposed by Dunn (1973) and improved by Bezdek (1981), is a very well-known algorithm commonly applied to pattern recognition tasks. FCM clustering algorithm is based on the minimization of the following objective function:

$$J_m = \sum_{k=1}^{n} \sum_{i=1}^{C} u_{ik}^m \left\| \mathbf{x}_k - \mathbf{c}_i \right\|^2. \tag{13}$$

FCM algorithm has the following steps:
1. Initialize the membership degrees u_{ik} and form the fuzzy partition matrix $\mathbf{U}^{(0)} = [u_{ik}]$.
2. At the r-th step – compute the center vectors $\mathbf{c}_i^{(r)}$, with $\mathbf{U}^{(r-1)}$ kept fixed, as follows:

$$\mathbf{c}_i^{(r)} = \frac{\sum_{k=1}^{N} \left(u_{ik}^{(r-1)} \right)^m \cdot \mathbf{e}_k}{\sum_{k=1}^{N} \left(u_{ik}^{(r-1)} \right)^m}. \tag{14}$$

3. Compute $\mathbf{U}^{(r)}$, with $\mathbf{c}_i^{(r)}$ kept fixed, as follows:

$$u_{ik}^{(r-1)} = \frac{1}{\sum_{j=1}^{C} \left(\dfrac{\left\| \mathbf{e}_k - \mathbf{c}_k \right\|}{\left\| \mathbf{e}_i - \mathbf{c}_j \right\|} \right)^{\frac{2}{m-1}}}, \tag{15}$$

where: $\mathbf{e}_k$ and $\mathbf{e}_i$ are k-th and i-th elements of $\mathbf{E}'$ matrix.
4. If $\left\| \mathbf{U}^{(r)} - \mathbf{U}^{(r-1)} \right\| < \varepsilon$ then stop, else return to step 2.

Table 1.

Distance measures for Boolean symbolic objects available in R software

DistType & distance name	Elements of distance measure	Distance measure
U_2 Ichino-Yaguchi	$\phi(v_{ij}, v_{kj}) = \left\| v_{ij} \oplus v_{kj} \right\| - \left\| v_{ij} \otimes v_{kj} \right\| + \gamma(2 \cdot \left\| v_{ij} \oplus v_{kj} \right\| - \left\| v_{ij} \right\| - \left\| v_{kj} \right\|)$	$\sqrt[q]{\sum_{j=1}^{m} \phi(v_{ij}, v_{kj})^q}$
U_3 normalized Ichino-Yaguchi	$\psi(v_{ij}, v_{kj}) = \dfrac{\phi(v_{ij}, v_{kj})}{\left\| V_j \right\|}$	$\sqrt[q]{\sum_{j=1}^{m} \psi(v_{ij}, v_{kj})^q}$
U_4 weighted and normalized Ichino-Yaguchi	$\phi(v_{ij}, v_{kj})$ same as in U_2	$\sqrt[q]{\sum_{j=1}^{m} w_j \psi(v_{ij}, v_{kj})^q}$
SO_2 de Carvalho	$\psi(v_{ij}, v_{kj}) = \dfrac{\phi(v_{ij}, v_{kj})}{\mu(v_{ij} \oplus v_{kj})}$ $\phi(v_{ij}, v_{kj})$ same as in U_2	$\sqrt[q]{\sum_{j=1}^{m} \dfrac{1}{m} \left[\psi(v_{ij}, v_{kj}) \right]^q}$
SO_1 de Carvalho	$\alpha = \mu(v_{ij} \cap v_{kj}),$ $\beta = \mu\left[v_{ij} \cap c(v_{kj}) \right],$ $\chi = \mu\left[c(v_{ij}) \cap v_{kj} \right],$ $\delta = \mu\left[c(v_{ij}) \cap c(v_{kj}) \right]$ $d_1 = \dfrac{\alpha}{\alpha + \beta + \chi},$	$\sqrt[q]{\sum_{j=1}^{m} \left[w_j d_f(v_{ij}, v_{kj}) \right]^q}$
C_1 de Carvalho for hierarchical or logical dependent variables	$d_2 = \dfrac{2\alpha}{2\alpha + \beta + \chi},$ $d_3 = \dfrac{\alpha}{\alpha + 2(\beta + \chi)},$ $d_4 = \dfrac{1}{2}\left[\dfrac{\alpha}{\alpha + \beta} + \dfrac{\alpha}{\alpha + \chi} \right],$ $d_5 = \dfrac{\alpha}{\sqrt{(\alpha + \beta) + (\alpha + \chi)}}$ $d_f(v_{ij}, v_{kj}) = 1 - D_f$ $f = 1, \dots, 5$	$\sqrt[q]{\dfrac{\sum_{j=1}^{m} \left[w_j d_f(v_{ij}, v_{kj}) \right]^q}{\sum_{j=1}^{m} \delta(V_j)}}$

Table 1. (cont.)

DistType & distance name	Elements of distance measure	Distance measure
SO_3 de Carvalho	–	$[\pi(A_i \oplus A_k) - \pi(A_i \otimes A_k) + \gamma(2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k))]$
SO_4 normalized de Carvalho	–	$[\pi(A_i \oplus A_k) - \pi(A_i \otimes A_k) + \gamma(2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k))] / \pi(A^E)$
SO_5 normalized de Carvalho	–	$[\pi(A_i \oplus A_k) - \pi(A_i \otimes A_k) + \gamma(2\pi(A_i \oplus A_k) - \pi(A_i) - \pi(A_k))] / \pi(A_i \oplus A_k)$
H Hausdorff	–	$\left[\sum_{j=1}^{m} \left(\max \left\{ \left| \bar{v}_{ij} - \bar{v}_{kj} \right|, \left| \underline{v}_{ij} - \underline{v}_{kj} \right| \right\} \right)^2 \right]^{\frac{1}{2}}$
L_1 L_2	a) interval-valued variables $L_1(v_{ij}, v_{kj}) = \left| \bar{v}_{ij} - \bar{v}_{kj} \right| + \left| \underline{v}_{ij} - \underline{v}_{kj} \right|$ $L_2(v_{ij}, v_{kj}) = \left| \bar{v}_{ij} - \bar{v}_{kj} \right|^2 + \left| \underline{v}_{ij} - \underline{v}_{kj} \right|^2$ b) multinominal variables: $L_1(v_{ij}, v_{kj}) = \sum_{y_f} \left| q_i v_j(y_f) - q_k v_j(y_f) \right|$ $L_2(v_{ij}, v_{kj}) = \sum_{y_f} \left| q_i v_j(y_f) - q_k v_j(y_f) \right|^2$	$\sqrt[q]{\sum_{j=1}^{m} \left(L_q(v_{ij}, v_{kj}) \right)^q}$ $q = 1$ for L_1 $q = 2$ for L_2

Where $v_{ij}v_{kj}$ – realizations of symbolic variables (interval-valued or multinominal), $A_i = (v_{i1}, v_{i2}, \ldots, v_{im})$ and $A_k = (v_{k1}, v_{k2}, \ldots, v_{km})$ – i-th and k-th symbolic object described by m symbolic variables, γ – parameter from the range of [0;1], usually $\gamma = \frac{1}{2}$, $q = \{1, 2 \ldots\}$ (usually $q = 2$), $|\ |$ – for interval-valued data it is the length of the interval, for other variables it is the number of elements, w_j – weight for j-th variable, μ – interval length for interval-valued variables, $c(v_{ij})$ – complement of the symbolic variable V_j, $\alpha, \beta, \chi, \delta$ – agreement and disagreement measures for symbolic variables, $\pi(A_i)$ – description potential of i-th symbolic object, A^E – maximum symbolic object according to the descriptive potential, $\delta(V_i)$ – indicator function. It equals 1 when the variable is defined according to logical or hierarchical dependencies with other variables. It equals 0 in other cases. For L_1 and L_2 distance measures in the case of multinominal variables: q.

Source: Gatnar, Walesiak (2011, p. 20–23).

3. EVALUATION EXPERIMENTS

For the purposes of simulation study, four different data sets were prepared with the application of `cluster.Gen` and `genRandomClust` functions of `clusterSim` (Walesiak, Dudek, 2014) and `clusterGeneration` (Qiu, Joe, 2006) packages of `R` software. Models contain the known structure of clusters. Simulation models, generated following the application of `cluster.Gen` function differ in the number of true variables, the density of cluster shapes, the number of true clusters, the number of noisy variables.

In case of symbolic data the data sets can have different shapes – more or less spherical, rounded and non-classical – like smiley, worms, or cuboids that are well-known from `mlbench` package of R software. The general shape of symbolic interval-valued data mimics the desired shape (e.g. spheres, worms, cuboids, smiley, etc.).

In order to obtain the symbolic interval-valued variables, the data were generated twice for each model into sets A and B, while the minimal (maximal) value of $\left\{ x_{ij}^{A}, x_{ij}^{B} \right\}$ is treated as the beginning (the end) of an interval. The noisy variables are simulated independently, based on the uniformly distributed random variables. The variances of noisy variables, in the generated data sets, are required to be similar to non-noisy variables (see Milligan, Cooper, 1988; Qiu, Joe, 2006, p. 322).

The models generated by `genRandomClust` function represent data sets with the specified degree of separation (see Qiu, Joe, 2006; Qiu, Joe, 2006a). They differ in the number of true variables, the density and the shapes of clusters, the number of true clusters, the number of noisy variables. In order to build interval data – the obtained data is treated as the center of rectangle. The width and the height of the rectangle are drawn randomly within the of [1, 8].

Real data sets were also used to check the proposed method – well-known Ichiono's oils (Ichino, 1998), cars (de Carvalho et al., 2006) and the European Union countries (Dudek, 2013) data sets.

3.1. ARTIFICIAL DATA SETS

Four different artificial models are used:
1. **Model I**. The model is generated with the application of `clusterSim` package. It contains five clusters in 2 dimensions which are not well separated. The observations are independently drawn from a bivariate normal distribution with means (5,5), (–3,3), (3,–3), (0,0), (–5,–5) and the identity covariance matrix $\sum \left(\sigma_{jj} = 1, \sigma_{jl} = -0.9 \right)$.
2. **Model II**. The model is generated using `clusterSim` package. It contains five clusters in 3 dimensions which are not well separated. The observations are independently drawn from a multivariate normal distribution with means (5,5,5), (–3,3,–3), (3,–3,3), (0,0,0), (–5,–5,–5) and a covariance matrix $\sum$, where $\sigma_{jj} = 1 \left(1 \leq j \leq 3 \right)$ and $\sigma_{jl} = 0.9 \left(1 \leq j \neq l \leq 3 \right)$.

3. **Model III**. The model is generated with the application of `clusterGeneration` package. It also contains five clusters in five dimensions which are not well separated. The desired value of the separation index between (see Qiu, Joe, 2006, Qiu, Joe, 2006a) a cluster and its nearest neighboring cluster was equal to 0.03 and method to generate the covariance matrices for clusters was set to "`onion`".

4. **Model IV**. Model generated with application of `clusterGeneration` package. The model contains six clusters in four dimensions representing overlapping clusters. The desired value of separation index (see Qiu, Joe 2006; Qiu, Joe, 2006a) between a cluster and its nearest neighboring cluster was equal to 0.013 and the method responsible for generating the covariance matrices for clusters was set to "`c-vine`".

For each model 20 simulation runs, with different distance types, were performed. The mean (MR) and the standard deviation (SD) of the fuzzy variant of Rand Index, proposed by Hüllermeier, Rifqi (see Hüllermeier, Rifqi, 2009, p. 1296–1297), were calculated for these trials. The fuzzy variant of Rand index is calculated as follows (see Hüllermeier, Rifqi, 2009, p. 1296–1297):

$$Rand = 1 - dist(P,Q), \tag{16}$$

where: $dist(P,Q)$ – the distance on two fuzzy partitions P and Q defined on the normalized sum of degrees of discordance, calculated as follows:

$$d(P,Q) = \frac{\sum_{(x,x')\in C}\left|E_P(x,x') - E_Q(x,x')\right|}{n(n-1)/2}, \tag{17}$$

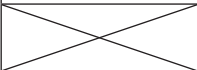$$E_P = 1 - \left\|P(x) - P(x')\right\|, \tag{18}$$
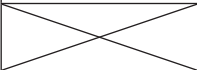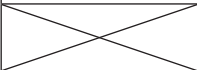
where: $\|\bullet\|$ is the proper distance measure on [0,1], $P_i(x)$ is the membership degree of x in i-th cluster.

Obviously, there are many other validity indices to be used in terms of the results of fuzzy clustering method – see for example Wang, Zhang (2007).

The results of these simulations for different distance measures, applicable for symbolic interval-valued data, are presented in the table 2. Then 20 simulations were also performed for each model with outliers and noisy variables. The mean (MR) and the standard deviation (SD) of the fuzzy version of Rand Index were also computed for these trials − the results are presented in table 3.

Table 2.

The results of simulations for models without noisy variables or outliers for spectral fuzzy *c*-means, fuzzy *c*-means, adaptive fuzzy *c*-means and fuzzy *k*-means clustering

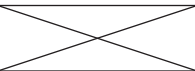|	Model I	Model II	Model III	Model IV
Distance measure	Spectral fuzzy *c*-means			
U_2	MR = 1 SD = 5.078e-13	MR = 1 SD = 4.023e-12	MR = 0.9999 SD = 2.150e-07	MR = 1 SD = 4.054e-06
U_3 & U_4 with equal weights	MR = 0.9997 SD = 6.092e-10	MR = 0.9999 SD = 7.324e-11	MR = 0.9999 SD = 9.221e-11	MR = 0.9999 SD = 4.994e-08
SO_2	MR = 0.9999 SD = 5.003e-11	MR = 0.9999 SD = 5.000e-10	MR = 0.9999 SD = 3.098e-10	MR = 0.9997 SD = 2.208e-11
SO_1 with d_1	MR = 1 SD = 6.005e-10	MR = 0.9996 SD = 4.004e-09	MR = 0.9994 SD = 8.003e-07	MR = 0.9998 SD = 7.011e-07
SO_1 with d_2	MR = 0.9994 SD = 4.213e-10	MR = 0.9921 SD = 6.437e-09	MR = 0.9970 SD = 8.478e-07	MR = 0.9977 SD = 1.987e-07
SO_1 with d_3	MR = 0.9999 SD = 6.827e-09	MR = 0.9967 SD = 8.748e-08	MR = 0.9998 SD = 4.874e-07	MR = 0.990 SD = 8.885e-06
SO_1 with d_4	MR = 0.9998 SD = 4.667e-06	MR = 0.9987 SD = 4.009e-04	MR = 0.9990 SD = 5.330e-05	MR = 0.9986 SD = 5.087e-05
SO_1 with d_5	MR = 0.9900 SD = 4.123e-05	MR = 0.9954 SD = 1.100e-05	MR = 0.9950 SD = 2.083e-05	MR = 0.9903 SD = 7.078e-04
SO_3	MR = 0.9999 SD = 1.913e-06	MR = 0.9999 SD = 4.878375e-08	MR = 0.9999 SD = 7.115587e-08	MR = 1 SD = 5.976251e-08
SO_4	MR = 0.9999 SD = 2.398e-06	MR = 0.9999 SD = 3.866e-07	MR = 0.9999 SD = 6.989e-06	MR = 0.9999 SD = 4.948e-05
SO_5	MR = 0.9999 SD = 5.108e-06	MR = 0.9999 SD = 3.006e-06	MR = 0.9999 SD = 6.072e-05	MR = 0.9999 SD = 6.089e-05
H	MR = 1 SD = 3.113e-08	MR = 1 SD = 3.410e-08	MR = 1 SD = 4.058e-08	MR = 0.9999 SD = 7.485e-08
L_1 and L_2	NA NA	NA NA	NA NA	NA NA
	Fuzzy *c*-means			
	MR =1 SD = 8.769e-13	MR = 1 SD = 3.004e-11	MR = 0.9999 SD = 0.0470	MR = 0.8916 SD = 0.0428
	Adaptive fuzzy *c*-means			
	MR = 1 SD = 1.152e-11	MR = 1 SD = 5.326e-11	MR = 0.9999 SD = 0.0344	MR = 0.8747 SD = 0.0426
	Fuzzy *k*-means			
	MR = 0.9999 SD = 7.944e-07	MR = 0.9999 SD = 4.595e-07	MR = 0.9999 SD = 0.0004	MR = 0.9994 SD = 0.0005

Where MR – mean fuzzy Rand index, SD – standard deviation of fuzzy Rand index, NA – value could not be calculated.

Source: authors' compilation.

Table 3.

The results of simulations for models with noisy variables and/or outliers for spectral fuzzy *c*-means, fuzzy *c*-means, adaptive fuzzy *c*-means, fuzzy *k*-means clustering
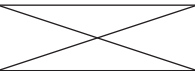
	Model I +1 noisy variable & 25% outliers	Model II +45% outliers	Model III +2 noisy variables	Model IV +1 noisy variable & 25% outliers
Distance measure	Spectral fuzzy *c*-means			
U_2	MR = 1 SD = 8.637e-12	MR = 1 SD = 6.463e-13	MR = 0.9953 SD = 0.0003	MR = 0.9837 SD = 0.0016
U_3 & U_4 with equal weights	MR = 0.9986 SD = 4.493e-10	MR = 0.9989 SD = 5.424e-09	MR = 0.9945 SD = 1.091e-10	MR = 0.9960 SD = 6.900e-07
SO_2	MR = 0.9999 SD = 3.472-07	MR = 0.9999 SD = 4.982e-11	MR = 0.9999 SD = 1.387e-10	MR = 0.9999 SD = 3.478e-06
SO_1 with d_1	MR = 1 SD = 3.389e-11	MR = 0.9987 SD = 1.137e-10	MR = 0.9967 SD = 3.873e-10	MR = 0.9940 SD = 5.839e-07
SO_1 with d_2	MR = 0.9973 SD = 4.325e-10	MR = 0.9910 SD = 6.434e-09	MR = 0.9949 SD = 8.532e-05	MR = 0.9956 SD = 1.133e-07
SO_1 with d_3	MR = 0.9999 SD = 3.764e-08	MR = 0.9967 SD = 7.837e-07	MR = 0.9998 SD = 1.387e-06	MR = 0.9900 SD = 3.424e-05
SO_1 with d_4	MR = 0.9987 SD = 2.228e-05	MR = 0.9968 SD = 5.576e-05	MR = 0.9976 SD = 3.347e-06	MR = 0.9950 SD = 4.437e-05
SO_1 with d_5	MR = 0.9967 SD = 3.887e-06	MR = 0.9917 SD = 3.001e-06	MR = 0.9933 SD = 1.378e-05	MR = 0.9900 SD = 2.873e-05
SO_3	MR = 0.9999 SD = 8.551e-07	MR = 0.9999 SD = 4.197e-07	MR = 0.9999 SD = 3.313e-07	MR = 0.9999 SD = 3.422491e-07
SO_4	MR = 0.9997 SD = 4.766e-05	MR = 0.9998 SD = 6.616e-06	MR = 0.9998 SD = 5.428e-05	SD = 0.9999 SD = 7.774e-05
SO_5	MR = 0.9987 SD = 2.135e-05	MR = 0.9988 SD = 8.663e-05	MR = 0.9998 SD = 4.849e-05	MR = 0.9998 SD = 8.117e-05
H	MR = 0.9999 SD = 8.530e-08	MR = 0.9999 SD = 0.021e-07	MR = 0.999 SD = 1.076e-07	MR = 0.9999 SD = 8.592e-07
L_1 and L_2	NA NA	NA NA	NA NA	NA NA
	Fuzzy *c*-means			
	MR = 0.7551 SD = 0.2048	MR = 0.7916 SD = 0.0637	MR = 0.8521 SD = 0.0042	MR = 0.6732 SD = 0.3093
	Adaptive fuzzy *c*-means			
	MR = 1 SD = 1.136e-03	MR = 0.8934 SD = 4.273e-04	MR = 0.9107 SD = 2.235e-06	MR = 0.9999 SD = 0.0034
	Fuzzy *k*-means			
	MR = 0.9983 SD = 0.0004	MR = 0.9978 SD = 7.500e-05	MR = 0.9994 SD = 2.342e-04	MR = 0.9972 SD = 5.345e-04

Where: all elements are the same as in table 2.

Source: authors' compilation.

A car symbolic interval data set consists of 33 objects (car models) described by 8 interval-valued variables, 2 categorical multi-nominal variables and one nominal variable (de Carvalho et al., 2006). In this application, only 8 interval-valued variables – *Price*, *Engine Capacity*, *Top Speed*, *Acceleration*, *Step*, *Length*, *Width* and *Height* – were considered for clustering purposes. This data set was clustered 20 times into 4 clusters using all of the applicable distances. The best results were obtained for unnormalized Ichino and Yaguchi, Hausdorff and normalized Ichino and Yaguchi distance measures. The mean Rand index for fuzzy data for the normalized Ichino and Yaguchi distance equals 0.9999983, the standard deviation is 1.384948e-06. For the Hausdorff distance the mean Rand index equals 1, whereas its standard deviation is 0, the same result is achieved in case of Ichino and Yaguchi distance.

The Ichino's oils data set consists of 8 oils and fats described by 8 interval-valued variables (Ichino, 1988) – *Specific gravity*, *Freezing point*, *Iodine value* and *Saponification value*. This data set was 20 times clustered into 2 clusters by applying all distances. The best results were obtained for Hausdorff, Ichino and Yaguchi and De Carvalho distances. In case of Hausdorff and the De Carvalho distances the mean Rand index equals 1 and its standard deviation is 0 or nearly 0. In case of Ichnino and Yaguchi distance the mean Rand index equals 0.9898985 and its standard deviation is 0.03109184.

The European Union (EU) – data set consists of 27 European Union countries described by 8 interval-valued variables representing innovation indicators within the EU countries (Dudek 2013) – *R & D expenditures*, *Enterprises with innovation activity*, *Expenditures on education*, *Internet access*, *Patents per million of citizens*, *e-Administration accessibility indicator*, *Broadband Internet*, *High-technology trade (exports)*. In case of De Carvalho, Ichino and Yaguchi and the normalized Ichino and Yaguchi distance measures the mean Rand index equals 1 (or nearly 1) and its standard deviation is 3.88068e-07.

4. FINAL REMARKS

The main contribution of this paper is the introduction of spectral fuzzy *c*-means algorithm (SCFM) for the symbolic interval-valued data. Due to the fact that the discussed algorithm is based on spectral decomposition of the distance matrix, it can be easily applied to any other symbolic data types and selecting the suitable distance remains the only requirement. SCFM starts from the symbolic data matrix followed by a distance matrix calculation. Spectral decomposition of the distance table is performed, and then the well-known fuzzy *c*-means algorithm is applied.

The spectral fuzzy *c*-means clustering requires the selection of σ parameter for the kernel, which can turn out difficult. However, the solution proposed by Karatzoglou can be used along with the selection of distance measure for symbolic data. Experiments

show that Hausdorff (H) distance reaches the best results (in terms of Rand index mean) when dealing with data sets without noisy variables and outliers.

When clustering symbolic data with (or without) noisy variables and/or outliers SCFM with the application of Hausdorff (`H`) and De Carvalho (`SO_3`) distances generally reach better results than SCFM with the application of normalized Ichino and Yaguchi (`U_3`, `U_4`) distances. Unnormalized Ichino and Yaguchi (`U_2`) distance sometimes reaches similar results as Hausdorff and De Carvalho distance measures. Slightly worse results are reached for the data sets with noisy variables than for the data sets with outliers. The above result was expected due to the fact that this method is based on distance measurements. It can be omitted by using some sort of variable selection algorithm, e.g. HINoV for symbolic data, which is available in `clusterSim` package (see Walesiak, Dudek, 2014; Walesiak, Dudek, 2008), or Ichino and Yaguchi feature selection for symbolic data (see Dudek et al., 2014) available in `symbolicDA` package of R software.

The spectral fuzzy c-means clustering, due to spectral decomposition of data matrix, can deal quite easily with data sets that have some "non-classical" shapes, known from `mlbench` package of R software, like cubes, worms, etc. Furthermore, if the number of clusters is exceeding the actual number of clusters in the data set, the membership degree for those exceeding clusters is getting lower and lower by each iteration.

Experiments with artificial data sets, with different cluster structures or the set degree of cluster separation, without noisy variables and/or outliers, show that this method reaches quite stable results – in terms of fuzzy Rand index. The same results appear while dealing with real data sets and artificial data sets with noisy variables and/or outliers.

One problem (limitation) appears while attempting to apply `L_1`-type distance measure to symbolic interval-data with low degrees of cluster separation. SCFM could not calculate eigenvectors due to the fact that, in this case, objects are too close to each other. The only solution in such a situation, i.e. while trying to apply this method and this distance measure, is to add some slight noise. It will not change the cluster structure, however, calculating eigenvalues will be possible.

When compared to other fuzzy clustering methods for symbolic data the proposed methods usually offer quite good results (in terms of Rand index mean and its standard deviation).

REFERENCES

Bezdek J. C., (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

Billard L., Diday E., (2006), *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

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

Chung F., (1997), *Spectral graph theory*, Washington, Conference Board of the Mathematical Sciences.

Cominetti O., Matzavinos A., Samarasinghe S., Kulasiri D., Maini P. K., Erban R., (2010)*, DifFUZZY: A Fuzzy Spectral Clustering Algorithm For Complex Data Sets, International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, 1 (4), 402–417.

De Carvalho F. A. T., Souza R. M. C. R., Chavent M., Lechevallier Y., (2006), Adaptive Hausdorff Distances And Dynamic Clustering Of Symbolic Data, *Pattern Recognition Letters*, 27 (3), 167–179.

De Carvalho F. A. T., Tenório C. P., Cavalcanti Junior N. L., (2006), Partitional Fuzzy Clustering Methods Based On Adaptive Quadratic Distances, *Fuzzy Sets and Systems*, 157, 2833–2857.

De Carvalho F. A. T., (2007), Fuzzy C-means Clustering Methods For Symbolic Interval Data, *Pattern Recognition Letters*, 28 (4), 423–437.

De Carvalho F. A. T., Tenório C. P., (2010), Fuzzy K-means Clustering Algorithms For Interval-valued Data Based On Adaptive Quadric Distances, *Fuzzy Sets and Systems*, 161 (23), 2978–2999.

de Sa V. R., (2005), *Spectral Clustering With Two Views*, ICML Workshop on Learning with Multiple Views.

Diday E., Govaert G., (1977), Classification Automatique Avec Distances Adaptatives, *R.A.I.R.O. Informatique Computer Science*, 11 (4), 329–349.

Dunn J. C., (1973), A Fuzzy Relative ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics*, 3, 32–57.

Dudek A., (2013), *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wrocław University of Economics Publishing House, Wrocław.

Dudek A., Pełka M., Wilk J., (2014), The `symbolicDA` package, `http://www.R-project.org`.

El-Sonbaty Y., Ismail M.A., (1998), Fuzzy Clustering For Symbolic Data, *IEEE Transactions on Fuzzy Systems*, 6, 195–204.

Fischer I., Poland J., (2004), *New methods for spectral clustering*, Technical Report No. IDSIA-12-04, Dalle Molle Institute for Artificial Intelligence, Manno-Lugano, Switzerland.

Gatnar E., Walesiak M., (eds.), (2011), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Gordon A. D., (1999), *Classification*, Chapman and Hall/CRC, Boca Raton.

Guattery S., Miller G.L., (1998), On the Quality of Spectral Separators, *SIAM Journal on Matrix Analysis and Applications*, 19 (3), 701–719.

Gustafson D. E., Kessel W. C., (1979), *Fuzzy Clustering with Fuzzy Covariance Matrix*, Proceedings of IEEE Conference on Decision and Control, San Diego, CA, 761–766.

Hüllermeier E., Rifqi M., (2009), *A Fuzzy Variant of the Rand Index for Comparing Clustering Structures*, Proceedings of IFSA/EUSFLAT Conference '2009, 1294–1298.

Ichino M., (1988), *General Metrics for Mixed Features – The Cartesian Space Theory for Pattern Recognition*, Proceedings of the 1988 IEEE International Conference on Systems, Man and Cybernetics, 1, 494–497, International Academic Publishers Beijing.

Jain A. K., Murty M. N., Flynn P. J., (1999), Data Clustering: A Review, *ACM Computational Surveys*, 31 (3), 264–323.

Kannan R., Vempala S., Vetta A., (2000), *On Clusterings – Good, Bad and Spectral*, Technical Report, Computer Science Department, Yale University.

Karatzoglou A., (2006), *Kernel Methods. Software, Algorithms and Applications*, Doctoral thesis, Vienna University of Technology.

Malerba D., Esposito F., Gioviale V., Tamma V., (2001), *Comparing Dissimilarity Measures for Symbolic Data Analysis*, Pre-Proceedings of ETK-NTTS 2001, Hersonissos, 473-48.

Meila M., Shi J., (2001), *A Random Walks View of Spectral Segmentation*, 8-th International Workshop on Artificial Intelligence and Statistics (AISTATS).

Milligan G. W., Cooper M. C., (1988), A Study of Standardization of Variables in Cluster Analysis, *Journal of Classification*, 5 (2), 181–204.

Moore R..E., (1966), *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

Ng A., Jordan M., Weiss Y., (2002), On Spectral Clustering: Analysis and Algorithm, in: Dietterich T., Becker S., Ghahramani Z., (eds.), *Advances in Neural Information Processing Systems*, 14, MIT Press, 849–856.

Nieddu L., Rizzi A., (2005), Metrics in Symbolic Data Analysis, in: Vichi M., Monari P., Signani S., Montanari A., (eds.), *New Development in Classification and Data Analysis*, Springer-Verlag, Berlin-Heidelberg, 71–78.

Poland J., Zeugmann T., (2006), Clustering the Google Distance with Eigenvectors and Semidefinite Programming, in: Jantke K. P., Kreuzberger G., (eds.), *Diskussionsbeiträge*, *Institut für Medien und Kommunikationswisschaft*, Technische Universität Ilmenau, 21, 61–69, July 2006.

Shi J., Malik J., (2000), Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (8), 888–905.

Qiu W., Joe H., (2006), Generation of Random Clusters With Specified Degree of Separation, *Journal of Classification*, 23 (2), 315–334.

Qiu W., Joe H., (2006a), Separation Index and Partial Membership for Clustering, *Computational Statistics and Data Analysis*, 50, 585–603.

Qiu, W., Joe, H. (2010), The `clusterGeneration` package, `http://www.R-project.org`.

von Luxburg U., Bousquet O., Belkin M., (2005), *Limits of Spectral Clustering*, in: Saul L., Weiss Y., Bottou L., (eds.), *Advances in Neural Information Processing Systems (NIPS)* 17, Cambridge, MA: MIT Press, 857–864.

von Luxburg U., (2006), *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.

Walesiak M., Dudek A., (2008), Identification of Noisy Variables for Nonmetric and Symbolic Data in Cluster Analysis, in: Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R., (eds.), *Data Analysis, Machine Learning and Applications*, Springer-Verlag, Berlin-Heidelberg, 85–92.

Walesiak M., Dudek A., (2014), The `clusterSim` package, `http://www.R-project.org`.

Wang W., Zhang Y., (2007), On Fuzzy Validity Indices, *Fuzzy Sets and Systems*, 158, 2095–2117.

Zelnik-Manor L., Perona P., (2004), *Self-tuning Spectral Clustering*, Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04), `http://books.nips.cc/nips17.html`.

Yang M.-S., Hwang P.-Y., Chen D.-H., (2004*),* Fuzzy Clustering Algorithms for Mixed Feature Types, *Fuzzy Sets Systems*, 141, 301–317.

Yaguchi H., Ichino M., (1994), Feature Selection for Symbolic Data Classification, in: Diday E. Lechevallier Y., Schader M., Bertrand P., Burtschy B., (eds.), *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin-Heidelberg, 387–394.

PORÓWNANIE METOD KLASYFIKACJI ROZMYTEJ
DLA DANYCH SYMBOLICZNYCH INTERWAŁOWYCH

S t r e s z c z e n i e

Dane symboliczne interwałowe mogą znaleźć zastosowanie w wielu sytuacjach – np. w przypadku notowań giełdowych, zmianach kursów walut, itp. Celem artykułu jest porównanie trzech metod klasyfikacji rozmytej dla danych symbolicznych interwałowych – tj. rozmytej klasyfikacji *c*-średnich, adaptacyjnej rozmytej klasyfikacji *c*-średnich oraz rozmytej klasyfikacji *k*-średnich z rozmytą klasyfikacją spektralną. Rozmyta klasyfikacja spektralna stanowi połączenie podejścia spektralnego oraz klasyfikacji rozmytej *c*-średnich, dzięki czemu możliwe jest otrzymanie lepszych rezultatów (w sensie indeksu Randa dla klasyfikacji rozmytych). Przeprowadzone badania symulacyjne wskazują, że rozmyta klasyfikacja spektralna dla danych symbolicznych pozwala na uzyskanie lepszych wyników niż inne rozmyte metody

klasyfikacji dla tego typu danych jeżeli weźmiemy pod uwagę zbiory danych o różnej strukturze klas, która dodatkowo jest zniekształcana przez obserwacje odstające lub zmienne zakłócające.

Słowa kluczowe: klasyfikacja spektralna, klasyfikacja rozmyta, dane symboliczne interwałowe, analiza danych symbolicznych

THE COMPARISON OF FUZZY CLUSTERING METHODS FOR SYMBOLIC INTERVAL-VALUED DATA

Abstract

Interval-valued data can find their practical applications in such situations as recording monthly interval temperatures at meteorological stations, daily interval stock prices, etc. The primary objective of the presented paper is to compare three different methods of fuzzy clustering for interval-valued symbolic data, i.e.: fuzzy c-means clustering, adaptive fuzzy c-means clustering and fuzzy k-means clustering with fuzzy spectral clustering. Fuzzy spectral clustering combines both spectral and fuzzy approaches in order to obtain better results (in terms of Rand index for fuzzy clustering). The conducted simulation studies with artificial and real data sets confirm both higher usefulness and more stable results of fuzzy spectral clustering method, as compared to other existing fuzzy clustering methods for symbolic interval-valued data, when dealing with data featuring different cluster structures, noisy variables and/or outliers.

Keywords: spectral clustering, fuzzy clustering, fuzzy partition, interval-valued data, symbolic data analysis

MACIEJ RYCZKOWSKI[1]

EFFECTS OF BEING IN AN OCCUPATION – IS ISCO 1 DIGIT CLASSIFICATION ENOUGH TO MODEL WAGES IN POLAND?

1. INTRODUCTION

Some analyses of wages are carried out taking into consideration only non-occupational variables that, no doubt, affect wages, however, they lack considerations about the key factor, which is the employee's occupation. This is problematic since occupation should influence the level of wages the most as it reflects the level of responsibility within the company, is often interconnected with appropriate education, qualifications, skills, personality traits and more generally with all the other required at a given post human capital factors. Even the level of education seems secondary in comparison to occupation as education is just a mean by which an individual might increase his or hers chances of getting a better paid occupation. It can be assumed that occupation is a derivative of education, sex (although it should not be), job experience, age (although it should not be), knowledge and skills, and other personal characteristics. Therefore, omitting occupations or their structure (for aggregated data) while determining factors affecting wages might produce less accurate outcomes.

Nevertheless, there are only few empirical studies in Poland concerning the influence of employees' occupations on wages. In turn, those studies which consider the influence of occupations on wages, take into account the impact of occupations aggregated only into 9 big, basic groups (ISCO 1-digit level). Yet, there are no studies explaining the influence of (ISCO) occupations measured at the 2-digit level. The paper fills in this gap. According to the results of the econometric model, the standard deviation of the estimated ISCO 2 digits coefficients is 0.24, which accounts for as much as more than 91% of the arithmetical mean of those coefficients. The huge diversification of the estimated impact of particular occupations on wages is even more visible when considering each of the ISCO 1 digit groups separately. The standard deviation of the estimated ISCO 2 digit occupation coefficients in the group 6 of ISCO 1 digit (i.e. Skilled agricultural, forestry and fishery workers) is more than five times higher than the arithmetical mean of the estimated occupation coefficients

---

[1] Nicolaus Copernicus University, Faculty of Economic Sciences and Management, The Department of Economics, 13a Gagarina St., 87-100 Toruń, Poland, e-mail: m_ryczkowski@umk.pl; Statistical Office in Bydgoszcz, 86-066 Bydgoszcz, Poland.

within this group –which results in coefficient of variation of more than 600%. In the remaining big ISCO 1 digit groups the situation is remarkably better, but still far from satisfactory. The quotient of the analogous standard deviations to arithmetical means in the remaining ISCO 1 digit groups (i.e. coefficient of variation) varies from 129% (group 5, Service and sales Workers) to 12.6% in group 2. Generally only groups: 2 (Professionals), 3 (Technicians and associated Professional), 4 (Clerical support Workers) and 8 (Plant and machine operators, and assemblers) (i.e. four out of nine ISCO 1 digit groups) experience relatively low variability (measured by the quotient of the analogous standard deviations to arithmetical means), with the coefficient of variation lower than 30%. The obtained results indicate that models of wages basing on ISCO 1 digit might be potentially biased. In the chapter "empirical results" we additionally perform an exercise and estimate the model with only ISCO 1 digit occupations. The results indicate that taking into account occupations at the ISCO 2 digit level should allow to better capture the role of occupations, however this still might be not enough to properly and in detail describe their role (see, Pryor, 2013 for the intra-occupational wage dispersion for the USA at ISCO 4 digit level).

The literature on wage determinants in Poland is very limited. Especially scarce are the papers that explain the level of wages by incorporating ISCO 2 digit variables into the econometric models. The paper fills in this gap and contributes to explaining the still growing diversification of wages even within the same occupation. We believe that our outcomes might be of practical use. Detailed influence of occupations on wages might indicate proper direction of future career of young people and of those who wish to retrain. Moreover, it might be also an information for the government which policies to implement in order to attract young people into the most profitable and wanted in the labour market occupations.

2. LITERATURE REVIEW ON WAGE DETERMINANTS IN POLAND

Previous studies indicate the importance of a firm size, industry, wage-tenure relation, education, gender, age, occupation, race, union membership on the wage differences (see, for example Mortensen, 2003; Mouw, Kalleberg, 2010; Mysíková, 2012). Nevertheless, the impact of particular variables on wage determination in Poland has not gained considerable attention. One of the few papers exploring the impact of occupations on wages in Poland is written in polish '*Czynniki determinujące poziom wynagrodzenia*' by Śliwicki (2012), in which the author constructed a multidimensional logit model on the basis of a significant number of descriptive variables (NUTS 2 regions, nine aggregated groups of occupations, NACE 1 economic activities, levels of education, length of job experience, sex, type of employment contract, size of the establishment, and national and/or sectoral levels of wage bargaining). The study, conducted for the year 2012, proved that belonging to one of nine groups of occupations was statistically significant and influenced the level of wages. The author concluded that also sex (men had higher probability of entering the upper decile group

than women), education (the higher the level of education the greater the chance to enter the upper decile group), type of job contract, and voivodship had statistically significant impact on the level of wages (Śliwicki, 2012).

Another research on wage determinants in Poland (together with the Czech Republic, Latvia, and Lithuania), this time for the year 2002, was carried out by Magda et al. (2011). The authors analyzed a similar to Śliwicki's (2012) set of data, however, they extended the occupation variable to the ISCO 2-digit and industry to NACE 2 levels, but decreased the number of regions to the NUTS 1 level. The authors found substantial differences in earnings across sectors in Poland and the other countries under consideration, even when taking into account a wide range of employee, job, and employer characteristics. The size of the company affected wages in Poland positively from almost 17% (50–249 employees) to almost 27% (1000+ employees). Despite the fact that the authors treated occupations as regressors, the question about the influence of occupations (at the ISCO 2-digit level) on wages remains unanswered as occupational variables were incorporated for the purpose of detailed calculation of economic activities' impact on wages (Magda et al., 2011).

A recent paper on wages in Poland was authored by Cieślik, Rokicki (2013a). Although the authors did not refer to the influence of occupations on the level of wages, using the standard two stage least squares estimation method they ascertained that, in compliance with the New Economic Geography (NEG) approach, economic potential accounts for about one third of the explained wage variation. This leaves room for the impact of individual characteristics, they concluded (Cieślik, Rokicki, 2013a). One of such characteristics is, no doubt, occupation.

Also Brülhart, Koenig (2006) received results in line with NEG. Using regional data for the years 1996–2000, the authors concluded that in the Czech Republic, Hungary, Poland, Slovakia, and Slovenia, market access variables explained up to 43% of the variance in regional fixed effects. This suggests that market access variables are significant explanatory factors of the spatial patterns of wages with high influence of big and capital cities (Brülhart, Koenig, 2006). Notwithstanding, Brülhart, Koenig (2006) did not analyze the possible influence of occupations on the level of wages.

Adamchik, King (2007) found that in 2001 in Poland full-time workers realized on average 86% of their potential earnings. However, their attempt to identify determinants of wages in Poland yielded mixed results for the prepared choice of explanatory variables (Adamchik, King, 2007). Unfortunately neither did their research include occupations, nor posts into the set of explanatory variables.

Another paper that may be added to the list of papers dealing with determinants of wages in Poland was authored by Basu et al. (2005). The authors analyzed wages in Poland (as well as in the Czech Republic, Hungary, and Slovakia) during transition from planned to market economy. Although they obtained interesting results and found little evidence of labor hoarding, they used aggregated data with no division to occupations. As a result, there is no possibility to draw conclusions from this paper on the significance of this crucial wage determinant (Basu et al., 2005).

3. DATA AND METHODOLOGY

The data come from the Structure of Earnings Survey (2012 edition) which is carried out in Poland every two years. The data consist of full and part–time employed persons who worked the whole month in October 2012. Data come from sample survey that covers companies with the number of the employed with 10 persons and more. The exact sampling selection scheme, generalization method and accuracy of the estimates of the selected parameters may be found in (CSO, 2014). Nevertheless some caveats of the research presented in this paper must be noted. First, as from the research are excluded companies with less than 9 employees, the conclusions given in the paper might differ for smaller companies. Secondly, as we measure only those employed, the paper is leaving aside those who are unemployed or who were made redundant. Therefore, any conclusions that, for instance, better education allows to earn more – concerns those who are employed and the impact of education on the *probability* of being employed in companies with more than 9 employees is not considered. Finally, into the research are not included people working on civil contracts and those in unregistered economy – therefore the conclusions are valid only for those in registered economy with full or part-time employment contracts.

The model includes 84 descriptive variables, with gross theoretical monthly wage in October 2012 being the dependent variable. The theoretical monthly wage is designed to allow for better comparisons of wages between particular employees. It is so because the total amount paid to the employee for its job is calculated as a sum of different bonuses, paid over-hours, extra money for a shift work etc. Theoretical wage takes all of them into account. Additionally the variable represents an employee's wage recalculated as if he or she worked the whole month of October (adjustment for breaks due to for example sick leaves, which would otherwise decrease the wage received in the analyzed month) and had full-time employment contract. In case an employee worked less hours in a given month due to for example an illness (and received therefore lower wage), wage is recalculated as if the employee had worked whole month. Moreover, some companies might work slightly different number of days in a month or year and the wage is also appropriately adjusted. For details of how theoretical monthly wage is calculated see appendix 1 and methodological notes in (CSO, 2014). The caveat of using the theoretical monthly wage is however loosing the possibility to verify whether part-time workers receive some wage penalty in comparison to full-time employees. Indeed, empirical results indicate that the penalty exists. For example, Magda et al. (2011) confirmed it for the analyzed Central and Eastern European Countries[2]. In our research it is however not considered.

The descriptive variables include variables that describe personal characteristics and variables representing terms of job agreement. The first set includes age, job experience in the entity, level of education, and sex, while the second set comprises occu-

---

[2] However, they had no data for Poland.

pation (2-digit ISCO) and type of job agreement (unlimited-term employment contract, limited-term employment contract, task oriented contract or probation contract). As economic activity of the employer is another important factor affecting wages, variables determining economic activities in accordance with the Statistical Classification of Economic Activities NACE Rev. 2, were also incorporated into the model.

Differences in wages between voivodships might be a result of various regional potentials that can be split into social potential (demographic trends, social capital, living conditions), economic potential (gross domestic product per capita, exports, foreign direct investments, innovations, productivity), institutional potential (economic freedom, transaction costs, trust in central and local government actions, share of private ownership), and environmental potential (existence of local labor market areas, tourism, geopolitical indicators, public transportation). (Strategy of socio-economic development for western Poland 2020, 2008). The influence of these potentials on the level of wages was measured by incorporating 16 regional dummies at the NUTS 2 level into the model.

The theoretical model we use takes the form of an extended Mincer (Mincer, 1974) wage regression. Mincer assumed an equalizing differences model of equilibrium wage function:

$$\log y = a_0 + a_1(t - S) + a_2(t - S)^2 + rS + other\ terms, \tag{1}$$

which is known as 'human capital earnings function' (HCEF). $t - S$ stands for years since completion of schooling as t is the age when people finish their earnings with schooling S. The HCEF is often used as a basis for modeling effects of schooling--related factors. However in the more recent literature it is usually proved that cubic or quartic values of $t - S$ years of post-schooling experience give better results. Original Mincer-type regression understates early career earnings growth and overstates mid-career earnings growth (Murphy, Welch, 1990).

To avoid this problem our model takes into consideration years of post-schooling experience as well as its squared (the quadratic experience terms describe the concavity of the earnings profile; see Willis, 1986) and cubic values and takes the form of the following equation[3]:

---

[3] It must be noted that age is often linked with job experience, which is natural for older workers to had worked more than younger workers. Nevertheless when an individual has breaks in his/her job career than both variables (i.e. age and job experience) might differ. Therefore, we consider in the equation both age and job experience. We do not implement squared and cubed age (as in case of experience) as we would otherwise measure the potential penalty for older (more experienced) workers twice, i.e. people with considerable job experience are also those of appropriately older age. A partial confirmation of this may be that a correlation coefficient between squared age and squared job experience was statistically significant and amounted up to 84%.

$$\ln w_i = \alpha + \beta AGE_i + \sum_{n=1}^{3} \chi_n EXPERIENCE_i^n + \delta EDUCATION_i + \phi SEX_i +$$
$$+ \varphi POST_i + \gamma CONTRACTTYPE_i + \eta NACE_i + \lambda NUTS2_i + \varepsilon_i, \tag{2}$$

where w_i represents the gross theoretical monthly wage of the individual i. AGE, EXPERIENCE, EDUCATION, and SEX represent personal characteristics of the employee. Both age and job experience are measured in years. Eight dummies are included for the level of education and one dummy variable for sex. Thirty four dummies stand for post held within the company, three dummies for contract type, seventeen dummies for type of activity of the company (NACE), and fourteen dummies for the NUTS 2 regions. α, β, χ, δ, ϕ, φ, γ, η, λ are parameters to be estimated and ε_i is the error term. The equation for Poland is estimated by OLS (White, 1980) with heteroscedasticity-consistent standard errors.

4. EMPIRICAL RESULTS

The model was estimated on the basis of 725215 records. Almost all of the considered coefficients are statistically significant at the 1% level (only coefficients for 'task oriented employment contract' and 'cleaners and helpers' are statistically significant at the 10% level, while only Wielkopolskie voivodship turned out to be statistically insignificant), see table 1.

Table 1.

Earnings equations (with ISCO 2 digit occupations) for the year 2012 (model I)

Exogenous variables	Coefficient	Std. error
Intercept	7.127	0.007
Individual characteristics		
Age	0.003	0.000
Job experience	0.024	0.000
Squared job experience	-0.001	0.000
Cubed job experience	0.000	0.000
Higher education (tertiary studies) with a degree of at least doctor	0.464	0.004
Master's degree, physician's degree or any other degree of equal status	0.352	0.003
Higher education (tertiary studies) with engineer's degree, bachelor, economist with diploma or any other degree of equal status	0.220	0.003
Post-secondary	0.115	0.002

Exogenous variables	Coefficient	Std. error
Vocational secondary	0.083	0.001
General secondary	0.096	0.002
Basic vocational	0.008	0.002
Lower secondary	0.071	0.010
Primary and incomplete primary	Ref	Ref
Male	0.155	0.001
Female	Ref	Ref
Kind of job agreement		
Unlimited-term employment contract	0.086	0.005
Limited-term employment contract	-0.034	0.005
Task oriented employment contract	-0.017	0.009
Probation contract	Ref	Ref
Economic activity of the employer		
NACE Rev.2 A, Agriculture, forestry and fishing	0.302	0.006
NACE Rev.2 B, mining and quarrying	0.695	0.004
NACE Rev.2 C, Manufacturing	0.193	0.003
NACE Rev.2 D, Electricity, gas, steam and air conditioning supply	0.403	0.004
NACE Rev.2 E, Water supply; sewerage, waste management and remediation activities	0.193	0.004
NACE Rev.2 F, Construction	0.134	0.004
NACE Rev.2 G, Wholesale and retail trade; repair of motor vehicles and motorcycles	0.103	0.003
NACE Rev.2 H, Transportation and storage	0.162	0.003
NACE Rev.2 I, Accommodation and food service activities	0.037	0.005
NACE Rev.2 J, Information and communication	0.297	0.004
NACE Rev.2 K, Financial and insurance activities	0.300	0.004

Table 1. (cont.)

Exogenous variables	Coefficient	Std. error
NACE Rev.2 L, Real estate activities	0.118	0.004
NACE Rev.2 M, Professional, scientific and technical activities	0.202	0.004
NACE Rev.2 N, Administrative and support service activities	Ref	Ref
NACE Rev.2 O, Public administration and defence; compulsory social security	0.137	0.003
NACE Rev.2 P, Education	0.009	0.003
NACE Rev.2 Q, Human health and social work activities	0.017	0.003
NACE Rev.2 R, Arts, entertainment and recreation	-0.029	0.005
Nace rev.2 S, Other service activities	0.142	0.008
Post held within the company by the employee		
Managers		
Chief executives, senior officials and legislators	1.085	0.005
Administrative and commercial managers	0.761	0.004
Production and specialised services managers	0.670	0.004
Hospitality, retail and other services managers	0.494	0.006
Professionals		
Science and engineering professionals	0.381	0.004
Health professionals	0.447	0.004
Teaching Professional	0.437	0.004
Business and administration professionals	0.424	0.004
Information and Communications technology professionals	0.537	0.005
Legal professionals	0.511	0.005
Technicians and associated Professional		
Science and engineering associated professionals	0.353	0.004
Health associate professionals	0.213	0.005
Business and administration associated professionals	0.289	0.004

Exogenous variables	Coefficient	Std. error
Legal, social, cultural and related associate professionals	0.221	0.006
Information and communications professionals	0.261	0.008
Clerical support Workers		
General and keyboard clerks	0.192	0.004
Customer services clerks	0.138	0.005
Numerical and material recording clerks	0.173	0.004
Other clerical support workers	0.159	0.005
Service and sales Workers		
Personal services workers	0.094	0.005
Sales workers	0.068	0.004
Personal care workers	0.043	0.007
Protective services workers	-0.034	0.005
*Skilled agricultural, forestry and fishery worker*s		
Market-oriented skilled agricultural workers	0.094	0.012
Market-oriented skilled forestry, fishery and hunting workers	-0.159	0.022
Craft and related trades workers		
Building and related trades workers, excluding electricians	0.142	0.005
Metal, machinery and related trades workers	0.257	0.004
Handicraft and printing workers	0.129	0.006
Electrical and electronic trades workers	0.273	0.005
Food processing, wood working, garment and other craft and related trades workers	0.029	0.004
Plant and machine operators, and assemblers		
Stationary plant and machine operators	0.246	0.004
Assemblers	0.198	0.005
Drivers and mobile plant operators	0.179	0.004
Elementary occupations		
Cleaners and helpers	0.013	0.004
Labourers in mining, construction, manufacturing and transport	0.126	0.004
Food preparation assistants	0.038	0.007
Refuse workers and other elementary workers	Ref	Ref

Table 1. (cont.)

Exogenous variables	Coefficient	Std. error
NUTS 2 regions (voivodships)		
Dolnośląskie	-0.036	0.002
Kujawsko-Pomorskie	-0.037	0.002
Lubelskie	-0.099	0.002
Lubuskie	-0.033	0.003
Łódzkie	-0.036	0.002
Małopolskie	-0.014	0.002
Mazowieckie	0.133	0.002
Opolskie	-0.030	0.003
Podkarpackie	-0.127	0.002
Podlaskie	-0.097	0.003
Pomorskie	0.051	0.002
Śląskie	0.017	0.002
Świętokrzyskie	-0.106	0.003
Warmińsko-mazurskie	-0.047	0.003
Wielkopolskie	s.i.	
Zachodniopomorskie	Ref	Ref

$R^2 = 0.53$; adjusted$R^2 = 0.53$; $\bar{y} = 8.15$; st.dev.(y) = 0.530; Ftest(p – value) = 0; V = 6.5%,
Chi-squared(2) = 71067.026(p – value = 0), standard error = 0.36.
Ref stands for reference variable, s.i. stands for statistically insignificant.

Source: own calculations.

According to empirical results, wages increase by 0.3% with every year of life. In line with this result, a minor support may be found for the positive impact of age on wages. In previous empirical studies the influence of age on wages was often positive (see, for example Van Ours, Stoeldraijer, 2010), which by comparing with assumed decreasing productivity of older workers led some to conclusions that older workers are often overpaid (Hellerstein, Neumark, 2004). The issue of joint relations between productivity and age is however behind the scope of this paper.

The impact of job experience on wages is 8 times higher than the impact of age since wages increase by 2.4% per every additional year of job experience. This result confirms that when an employee is more experienced, he or she is more valuable for the employer. It also means that for companies less important is the age of the employee and what matters is the job experience. Hence an increase in age and job experience transfers into increase in wages. The obtained results confirm the concavity

of the observed earnings in line with expectations (see, Mincer 1974), the estimated coefficients of job experience and squared job experience are respectively positive and negative, although with very small value of the second one. Likewise, Magda et al. (2011) found that each year of additional prior potential job experience increased the wage by 3%, which is a little higher value than ours. Nevertheless in their model each additional year of seniority to the company increased the wage by 1.2%. Having considered the lower impact of seniority to the company on wages would probably lower the estimated by us total job experience coefficient to values somehow closer to estimated by Magda et al. (2011). The favorable effect of experience and age found by us and other authors does not, however, eliminate labour-market disadvantages that are also age-related. One must, for example, take into account that people over a particular age may be more prone to redundancy. Having between 50 and 60 years for women and 50 to 65 years for men reduces the probability of being employed by almost 11% (Ryczkowski, 2012). The higher redundancy risk for older people is not measured by this model, as the model takes into account only those employed. Therefore, according to empirical results, an additional 10 years of age and job experience transform into 3% and 24% increase in wages respectively. The estimated coefficients of squared and cubed job experiences are slightly negative and statistically significant. This may mean that older people with substantial job experience are punished on the labour market due to their advanced age, but the estimated penalty is very little. The estimated penalty might be so small due to the mechanism of rewarding faithful employees in line with the efficiency wage theory (Stiglitz, 1974) when the rewarding exceeds any possible negative consequences of age-related productivity downfalls.

Empirical results confirm that sex affects wages at a benefit of men in Poland (Adamchik, Bedi, 2001; Rokicka, Ruzik, 2010; O'Darchai, 2011; Śliwicki, Ryczkowski, 2014). Men receive 15% higher wages than women. On the other hand, Magda et al. (2011) obtained that being a woman decreased wages by 15%. In turn, empirical research by Śliwicki and Ryczkowski (2014) proves that women get wages lower than men's wages by 10.13% to 14.6% due to discrimination depending on the selected methodology. In comparison, the estimated GPG for Poland of Mysíková (2012) was smaller and amounted only to 8.6% on the data set from the year 2008. Both results are however not fully comparable and it cannot be concluded that GPG increased recently. Mysíková (2012) used for the analysis data coming from Labour Force Survey which in terms of wages are less accurate – many data concerning wages miss or are given only in brackets and may be a subject to a survey –bias, however data concerning wages which come from the Structure of Earning Survey (like in the paper of Śliwicki, Ryczkowski, 2014) although are more reliable (as they are gathered from the companies with a resulting from the law obligation to provide them to public statistical services), concern only companies with more than 9 employees. In sum, the result that men receive 15% higher wages than women in Poland is broadly in line with the empirical papers measuring the gender pay gap and may suggest that the pay gap may be higher in companies with more than 9 employees. Nevertheless, it must be

noted that the overall level of *real* discrimination might be lower than the obtained in literature values as a result of sociological, psychological, and social factors that were not taken into account for the quoted decomposition in the analyzed papers.

According to the model, people with master's degree or higher have a considerably greater chance to earn more. The degree of doctor (or higher) increases the wage by 46.4% and master's degree boosts wages by 35.2%. Empirical results confirm that the level of education is a key factor affecting wages and our results are close to the ones of other authors (see, for example: Śliwicki, Zwara, 2012[4]). Likewise, Magda et al. (2011) indicated that the higher the level of education, the greater was its positive influence on wages, with the highest value of 49% for workers with university and non-university higher education. Marcinkowska et al. (2008) ascertained that the difference of wages in Poland between people with higher education and those with primary education increased considerably after the transformation and in the year 1996 amounted already up to 41%, while in 1987 it was only 23%. Our results show that the more educated the potential employee the higher his or hers chance to earn more. Empirical results contest the view of the depreciation of the master degree on the labour market in Poland. By the phrase 'depreciation of a master degree' we mean often raised opinion of a decreasing importance of this title. Such an unfavorable perception resulted from the growing share of population with master degree, especially among young people above 25 and below 30 years old. The share of people with master degree in the population of people above 25 years old is ca. 24% in Poland, while in the bracket of people below 30 years old, the share rises up to over 42%[5]. One must, however, take into account that estimations were carried out on a sample comprising only those employed. Thus, no conclusions can be drawn on the influence of the level of education on the chances for being employed. However, in enterprises where the number of workers is above 10, employees with master's degree are in a remarkably better situation as regards to wages than people with a lower level of education. The reason behind this might be identified by Marcinkowska et al. (2008) the so called skilled biased technical change which increases the demand for qualified labour or the cause may reflect more rapid human capital accumulation and higher returns to job search (Bagger et al., 2014).

Unlimited-term employment contract translates to an 8.6% wage increase. Limited-term employment contract and task oriented employment contract, by contrast, have penalties of –3.4% and –1.7% respectively. Employees falling under one of the two latter options earn less as they are often replacement workers or may have an employ-

---

[4]   Authors find that higher education and having a doctor title altogether increase on average the mean wage in kujawsko-pomorskie voivodship by almost 1404 zlote in comparison to other levels of completed education. It is about 50% of the average wage at that time, which indicate that our estimations undervalue the masters degree, nevertheless Śliwicki and Zwara (2012) used more aggregated measure of education and included into masters degree also people with doctoral and PhD title plus their sample was limited only to one voivodship.

[5]   Data for the fourth quarter of 2014. Own calculations basing on Labour Force Survey database.

ment contract for an *implicite* probationary period. After the trial period, when employees prove their worth for the company, they will get a better paid unlimited-term employment contract. Moreover, the form of limited term employment and probationary employment concerns mostly graduates which additionally decreases the level of wages. Also Magda et al. (2011) found that fixed term employment contracts contribute to a wage penalty. The penalty according to the research for selected eastern and western European countries varied between 5.8% for Norway up to 27.4% for Italy. For Poland the penalty for limited term employment contract amounted high 15%. The big gap between our results and those of Magda et al. (2011) may result from different methodologies among which could be included the fact that they verified only two kinds of contracts (unlimited and limited-term employment contract), while we included additionally task oriented employment contracts and probation contracts, also other variables in both models differed. The impact of particular employment contracts on wages in Poland remains then inconclusive.
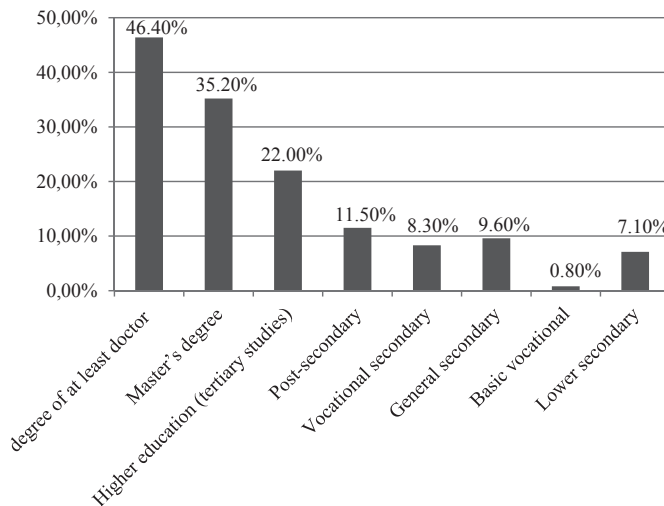

Figure 1. The influence of the level of education on wages in the year 2012

Source: own calculations.

According to the empirical results, we show that in Poland exist strong sectoral wage premia. The reason behind this might be the country-specific degree of corporatism (Magda et al., 2011). In our research mining and quarrying was the economic activity of the enterprise that most prominently boosted wages – by 69.5%. Among the subsequent most favorable economic activities were: 'electricity, gas, steam, and air conditioning supply', 'agriculture, forestry, and fishing', 'financial and insurance activities', 'information and communication'. In comparison, the best paid economic activity – 'public administration and defence together with compulsory social security' – augmented wages by 13.7%. It might seem surprising that 'human health

and social work activities' and 'education' had the weakest influence on wages. One must, however, take into account that more detailed research concerning employers' economic activities would be necessary as each of them consists of specific subsections. Moreover, workers falling under these economic activities represent different occupations and perform different tasks. Education, for instance, comprises teachers as well as cleaners and helpers – one of the worst paid occupational groups. In education the majority of teachers are women, which additionally decreases wages as empirical results indicate that women earn less.
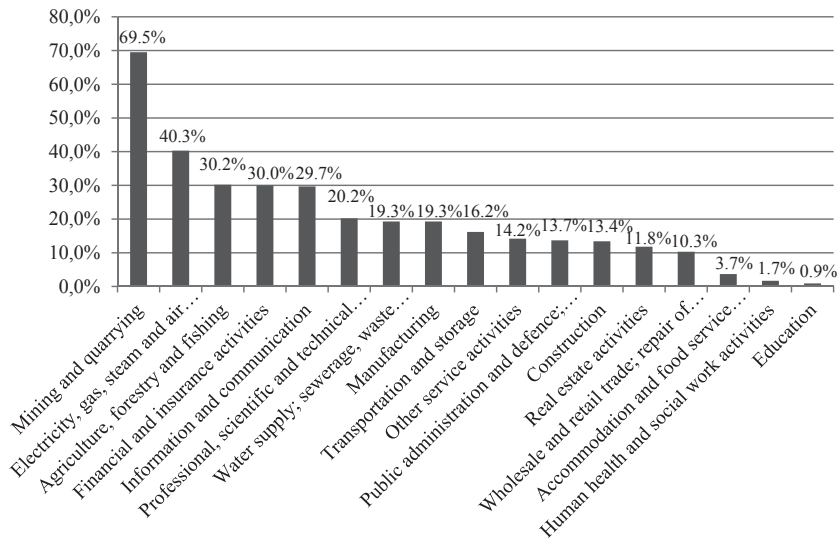

Figure 2. The influence of employer's economic activity on wages in the year 2012

Source: own calculations.

Newell, Socha (2007) found that between 1998 and 2002 the differences between the average wage in population and average wages of managers and those employed in technical professions increased to an considerable extent. In this respect our results confirm the trends in wages observed in earlier studies for Poland. According to our research the best paid group of occupations are managers as the fact of working as chief executive, senior official, or legislator increased wages by 108.5%. Belonging to the other managerial groups also proved beneficial – administrative and commercial managers, production and specialized services managers, and hospitality, retail and other services managers had their wages increased by 76.1%, 67.0%, and 49.4% respectively. Professional, technicians and associated professionals had also considerably increased wages.

Among the worst paid workers were: market-oriented skilled forestry, fishery and hunting workers who had a penalty of –15.9% to their wages and protective services workers who had a penalty of –3.4% to their wages. Working as a cleaner

or helper, sales worker, personal care worker, food preparation assistant increased wages only slightly. Results that brought about similar conclusions obtained Śliwicki (2012). According to his research managers had highest probability to enter the upper decile group. The second most likely group to do so were professionals. However, in comparison to managers, they had a 78.62% lower chance to receive wages from the upper decile group. The worst outcome was achieved by representatives of skilled agricultural, forestry and fisheries workers with a 97.86% lower probability of entering the upper decile group than managers.
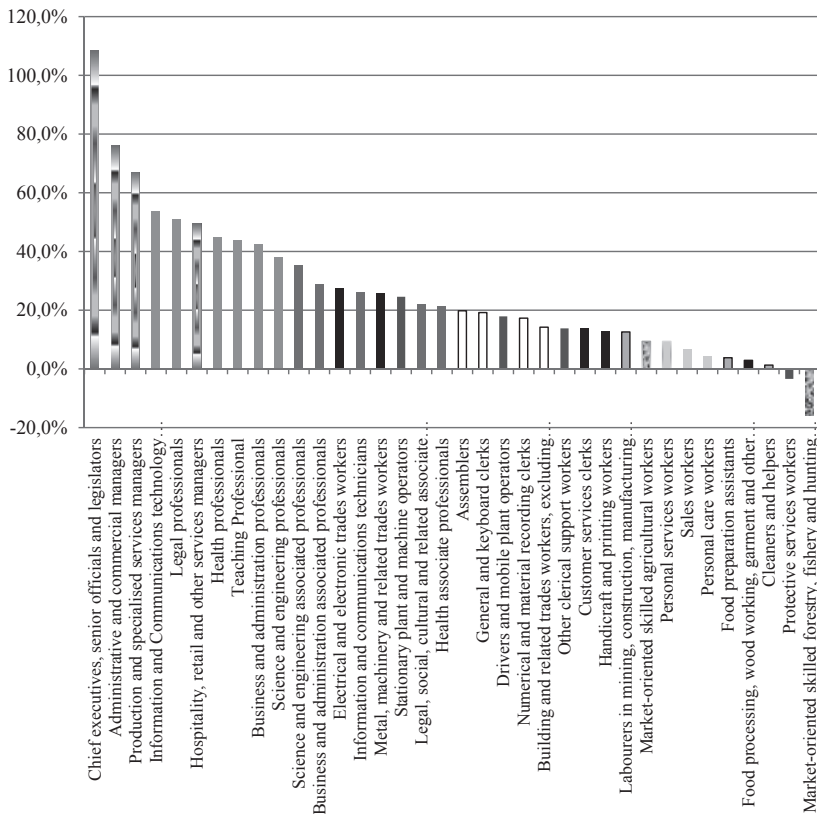

Figure 3. The influence of occupation on wages in the year 2012

Source: own calculations.

Empirical results indicate that occupations at the ISCO 2-digit level have a meaningful and highly variable influence on wages. There could be many reasons explaining this phenomena, like: the diversified level of responsibility, the need to have appropriate education, qualifications, skills, also some other demand factors or more generally human capital factors are of considerable importance. As wage differences at the ISCO 2 digit level are considerable, we assume that models with ISCO 2

digit are superior to those with only ISCO 1 digit as they more accurately capture the role of occupations. To spot that, we perform an exercise and estimate the very same model, however with ISCO 1 digit occupations (estimated coefficients of the model are reported in appendix 2). The general overview is that except for estimated occupation coefficients, the results of both models are quite similar. It is especially visible when comparing the impact of the education, NUTS 2 regions (however, with noticeable exception of Dolnośląskie voivodship) and the kind of job agreement – however in ISCO 1 digit model the bonus for unlimited-term employment contract is 2.6 p.p. higher. Similarly, the fact of being a man increases the wages by 1.9 p.p. higher in ISCO 1 digit model in comparison to the ISCO 2 digit model. In ISCO 1 digit model the estimated coefficients of economic activity of the employer also tend to be somehow higher than in ISCO 2 digit model. The possible explanation could be that not including the wider range of occupations removes the impact of the better paid occupations in each section and thus the premia for being in particular economic activity rises.

Nevertheless, especially meaningful differences between the two models appear when comparing the impact of occupations as ISCO 1 digit model seems not to capture their role as well as ISCO 2 digit model. In ISCO 1 digit model all elementary occupations are the reference value. However, according to ISCO 2 digit model the impact of particular occupations within this group is diversified as cleaners and helpers coefficient amounts to 0.013 while the coefficient for labourers in mining, construction, manufacturing and transport is almost ten times higher: 0.126. The ISCO 1 digit model seems not to capture the role of occupations particularly well in the group of managers. In this group according to ISCO 1 digit model wages are increased on average by 0.71%, however more detailed estimation with ISCO 2 digit variables shows that in the group of managers mostly rewarded are chief executives, senior officials and legislators (109%) while hospitality, retail and other services managers have a bonus more than twice lower (49%). Moreover, group of skilled agricultural, forestry and fishery workers in ISCO 1 digit model turned out to be statistically insignificant, while as this may be true for the whole group, more detailed estimation with ISCO 2 digits reveals that in this group both market-oriented skilled agricultural workers (coefficient: 0.094) and market-oriented skilled forestry, fishery and hunting workers' (coefficient: –0.159) coefficients are statistically significant and additionally the difference between the impact on wages of these two groups in absolute value amounts to more than 25 p.p. A clearly better explanatory properties of ISCO 2 digit model are visible also in the group of service and sales workers with the estimated coefficient of 0.013 in ISCO 1 digit model. ISCO 2 digit model reveals that this group is also not homogenous. Mostly rewarded in this group are personal services workers (0.094) while protective services workers have a penalty to their wages of "minus" 0.034. Also in other ISCO 1 digit groups the differences between both models are noticeable (compare with appendix 2). The least discrepancy between both models appears to be in the group of professionals with ISCO 1 digit model suggesting that this group has on

average a bonus to wages of 39%. However, even here particular occupations are differently rewarded. Mostly rewarded are information and communications technology professionals (0.54%) and least rewarded are science and engineering professionals (38%). In sum, it might be concluded, that models of wages using ISCO 1 digit data might be not enough to properly capture the role of occupations.

According to the estimations, NUTS 2 regions also influenced wages, which is in line with other studies, which indicate for different countries significant inter-regional wage differences among similarly skilled workers. For example, Simón et al. (2006) explain them by both competitive and non-competitive factors, such as an insufficient competition in product markets and industry-level collective bargaining, while Groot et al. (2011) point out to the importance of population density and the total size of the regional labor market to have statistically significant and positive effect on wages. We can observe similar patterns for Poland as in voivodships with the weakest influence on wages we find voivodships with highest unemployment rate, smallest size of regional markets, relatively weaker competition and smallest population density. Mazowieckie voivodship stands out from all of the sixteen voivodships – working in this NUTS 2 region increases wages by 13.3%. Other voivodships have considerably weaker influence on wages. A moderate penalty to wages can be noticed in the eastern voivodships: Podkarpackie, Lubelskie, Podlaskie, and Warmińsko-Mazurskie, where wages are reduced by 12.7%, 9.9%, 9.7%, and 4.7% respectively. Empirical results confirm the existence of differences in the labour market situation between eastern and western voivodships and a relatively strong, positive influence on wages of the Mazowieckie voivodship. The conclusions are thus similar to those of Cieślik, Rokicki (2013b) who obtained that wages decrease as one moves away from the Mazowieckie voivodship[6], as well as from the border with Germany. The results confirm that regional potential might have meaningful impact on wages.

5. CONCLUDING REMARKS

Empirical results indicate that occupations at the ISCO 2-digit level are important factors determining wages. It might be concluded that models of wages basing on ISCO 1 digit data might be not enough to properly capture the role of occupations. The analysis of the econometric model with occupations at ISCO 2 digits (with comparison to analogous model with ISCO 1 digit classification) reveals that the impact of particular occupations on wages within ISCO 1 digit groups may be very diversified. Thus, estimating the impact of occupations on wages only at ISCO 1 digit level may lead to conclusions, which does not have to be necessarily correct for all the detailed occupations within the analyzed ISCO 1 digit group.

---

[6] Authors find that a 10% increase in distance from Warsaw leads to a 0.6% decrease in the relative average regional wage.

Additionally, in the paper were confirmed many earlier empirical results of the wages' determinants on the example of Poland. The positive influence of occupation on wages seems to increase with the level of work-related responsibility and the unique knowledge of the employee. The occupations with the lowest impact on wages include: forestry, fishery, hunting, and protective services workers, cleaners and helpers as well as food processing workers. The estimated coefficients for food preparation assistants, personal care and sales workers show that these occupations also guarantee little positive influence on wages. As the responsibility grows, wages increase. Those most rewarded are managers, professionals, technicians and associated professionals. It can be noticed that also employees with a high level of work-related responsibility and unique skills, like electrical, electronic, metal, and machinery trades workers as well as stationary plant and machine operators can expect relatively higher wages due to their occupation. The influence of occupation on wages was estimated to vary from −15.9% (for market-oriented skilled forestry, fishery, and hunting workers) to 108.5% (for chief executives, senior officials, and legislators).

The paper confirms that unlimited-term employment contracts are more profitable than probationary and limited-term employment contracts. The analysis supports also the claim that men in Poland earn on average more than women. The paper presents an estimated influence of economic activities on wages as well. The obtained results confirmed that, generally, the higher the completed level of education, the bigger the chances for high wages. Assuming voivodships as proxies for regional potentials, the results confirm that regional potential has impact on wages, however, its influence seems moderate.

REFERENCES

Adamchik V. A., Bedi A. S., (2001), *Persistence Of The Gender Pay Differential in a Transition Economy*, ISS Working Paper No. 349, Hague: Institute of Social Studies.

Adamchik V. A., King A. E., (2007), Labor Market Efficiency in Poland: A Stochastic Wage Frontier Analysis, *The International Journal of Business and Finance Research*, 1 (2), 41–50.

Bagger J., Fontaine F., Postel-Vinay F., Jean-Marc R., (2014), Tenure, Experience, Human Capital, and Wages: A Tractable Equilibrium Search Model of Wage Dynamics, *American Economic Review*, 104 (6), 1551–96.

Basu S., Estrin S., Svejnar J., (2005), Employment Determination in Enterprises Under Communism And In Transition: Evidence From Ventral Europe, *Industrial and Labor Relations Review*, 58 (3), 353–369.

Brülhart M., Koenig P., (2006), New Economic Geography Meets Comecon Regional Wages And Industry Location In Central Europe, *Economics of Transition*, 14 (2), 245–267.

Cieślik A., Rokicki B., (2013a), Regional Wage Determinants In Poland: The Empirical Verification of the NEG Approach, *Bank i Kredyt*, 44 (2), 159–174.

Cieślik A., Rokicki B., (2013b), Regional Structure of Wages in Poland Over the Period 1995-2009, *Equilibrium*, 8 (3), 65–78.

CSO (2014), *Structure of Wages And Salaries By Occupations in October 2012*, Central Statistical Office, Warsaw, Poland.

Groot S., Groot H., Smit M., (2011), Regional Wage Differences in the Netherlands: Micro-evidence on Agglomeration Externalities, CPB Netherlands Bureau for Economic Policy Analysis Discussion Paper, 184.

Hellerstein J. K., Neumark D., (2004), Production Function and Wage Equation Estimation with Heterogeneous Labor: Evidence From a New Matched Employer-employee Data Set, NBER Working Paper Series, 13, 345–371.

Magda I., Rycx F., Toyerow I., Valsamis D., (2011), Wage Differentials Across Sectors in Europe. An East–west Comparison, *Economics of Transition*, 19 (4), 749–769.

Marcinkowska I, Ruzik A., Strawiński P., Walewski M., (2008), *Badanie struktury i zmian rozkładu wynagrodzeń w Polsce w latach 2000–2006*, Ministerstwo Pracy i Polityki Społecznej, Warszawa.

Mincer J., (1974), *Schooling, Experience and Earnings*, National Bureau of Economic Research, New York.

Mortensen, D. T., (2003), *Wage Dispersion: Why Are Similar Workers Paid Differently?*, MA, MIT Press, Cambridge.

Mouw T., Kalleberg A. L., (2010), Occupations and the Structure of Wage Inequality in the United States, 1980 to 2000s., *American Sociological Review*, 75 (3), 402–31.

Murphy K., Welch F., (1990), Empirical Age Earnings Profiles, *Journal of Labor Economics*, 8 (2), 202–229.

Mysíková M., (2012), Gender Wage Gap In The Czech Republic And Central European Countries, *Prague Economic Papers*, 3, 328–346.

Newell A., Socha M., (2007), The Polish Wage Inequality Explosion, *Economics of Transition*, 15 (4), 733–758.

O'Darchai S., (2011), The Gender Pay Gap In Research: A Comparison Of 23 European Countries, *Brussels Economic Review - Cahiers Economiques De Bruxelles*, 54 (2/3), 237–275.

Pryor F. L., (2013), Research Note: Intraoccupational Wage Dispersion, *WorkingUSA: The Journal of Labor and Society*, Volume 16, 389–394.

Rokicka M., Ruzik A., (2010), The Gender Pay Gap in Informal Employment in Poland, Case Network Studies and Analyses, No. 406, Warsaw.

Ryczkowski M., (2012), Analysis of Selected Unemployment Determinants in Poland and Selected Proposals for the Fight Against, in: Balcerzak A. P., (ed.), *Labor Markets After Global Financial Crisis*, Polish Economic Society Branch in Toruń.

Simón H., Ramos R., Sanromá E., (2006), Collective Bargaining and Regional Wage Differences in Spain: An Empirical Analysis, *Applied Economics*, 38, 1749–1760.

Śliwicki D, Zwara W., (2012), Wpływ wykształcenia na aktywność ekonomiczną i wynagrodzenia, *Polityka edukacyjna wobec rynku pracy*, 115, 255–265.

Śliwicki D., (2012), Czynniki determinujące poziom wynagrodzenia, *Wiadomości Statystyczne*, 10 (617), 1–15.

Śliwicki D., Ryczkowski M., (2014), Gender Pay Gap in the Micro Level – Case of Poland, *Quantitative Methods in Economics*, 15 (1), 159–173.

Stiglitz, J., (1974), Wage Determination and Unemployment in L.D.C.s: the Labor Turnover Model, *Quarterly Journal of Economics*, 88, 194–227.

Strategia rozwoju społeczno-gospodarczego Polski Wschodniej do roku 2020 (2008), dokument przyjęty przez Radę Ministrów w dniu 30 grudnia 2008 roku uchwałą 278/08, Ministerstwo Rozwoju Regionalnego, Warszawa, 25–47.

Van Ours J. C., Stoeldraijer L., (2010), Age, Wage and Productivity, IZA Discussion Paper No. 4765.

White H., (1980), A Heteroscedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity, *Econometrica*, 48, 817–838.

Willis R. J., (1986), *Wage Determinants: A Survey And Reinterpretation Of Human Capital Earnings Functions*, in: Ashenfelter O., Layard R., (eds.), *Handbook of Labor Economics*, Volume L, Elsevier Science Publishers BV.

<div align="center">APPENDIX A</div>

Theoretical monthly wage *TMW* is a sum of TMW_{nom} – theoretical monthly wage paid for nominal time of work (nominal time is the amount of time specified to work according to job agreement, usually 40 hours per week) – (in other words TMW_{nom} is standard remuneration) and payment for overtime work (over-hours) *OH*:

$$TMW = TMW_{nom} + OH.$$

TMW_{nom} is a sum of theoretical wage paid for nominal time of work in October $TMW^{October}_{nom}$ and 1/12 of bonuses, additional annual bonuses for public sector employees, payments due to participation in the profit or budgetary surplus etc.(we will refer to all the additional elements, which increase the basic remuneration as TBN). Moreover, earnings for periods longer than one month are recalculated per a month, i.e. in case of annual bonuses we need to divide *TBN* by 12:

$$TMW_{nom} = TMW^{October}_{nom} + TBN / 12 + OH.$$

Theoretical wage $TMW^{October}_{nom}$ is the actual remuneration that an employee received for the performed job in nominal working hours recalculated per full time job:

$$TMW^{October}_{nom} = MW^{October} \times \frac{4.6 \times WNH^{FT}}{HP^{October}}.$$

In the upper equation 4.6 is the number of weeks in October, WNH^{FT} is the weekly number of hours delivered by particular occupation assuming working full-time without any breaks in work due to, for example, sick leaves (mostly often WNH^{FT} simply equals 40), $HP^{October}$ is the number of hours paid in October by the company to an employee.

Finally, the theoretical measure of all kinds of let's call it "bonuses", i.e. *TBN* are all actually received bonuses by an employee *BN* recalculated per full time job, per uniform number of working days (as they sometimes differ among companies) and adjusted to money paid for economic outage, i.e.:

$$TBN = BN \times AWT^{FT} \times \frac{250}{NWD} / TP^{AN},$$

where AWT^{FT} is the annual number of hours that are to be worked in a given occupation assuming an employee worked full-time, $\frac{250}{NWD}$ is the indicator of the 'number of working days in a year' (mostly often the indicator equals 1 as in most companies number of annual working days *NWD* is 250), TP^{AN} is the annual time for which the employee received remuneration (paid time) – which is a sum of time actually worked in nominal working hours and time not worked paid minus time spent for economic outages (all time is measured in hours). As a result:

$$TMW = MW^{October} \times \frac{4.6 \times WNH^{FT}}{HP^{October}} + (BN \times AWT^{FT} \times \frac{250}{NWD} / TP^{AN})/12 + OH.$$

APPENDIX B

Table 2.

Earnings equations (with ISCO 1 digit occupations) for the year 2012 (model II)

Exogenous variables	Coefficient	Std. error
Intercept	7.149	0.006
Individual characteristics		
Age	-0.001	0.000
Job experience	0.030	0.000
Squared job experience	-0.001	0.000
Cubed job experience	0.000	0.000
Higher education (tertiary studies) with a degree of at least doctor	0.498	0.004
Master's degree, physician's degree or any other degree of equal status	0.369	0.002
Higher education (tertiary studies) with engineer's degree, bachelor, economist with diploma or any other degree of equal status	0.228	0.003
Post-secondary	0.116	0.003
Vocational secondary	0.087	0.002
General secondary	0.099	0.002
Basic vocational	-0.004	0.002
Lower secondary	0.107	0.010
Primary and incomplete primary	Ref	Ref
Male	0.174	0.001
Female	Ref	Ref
Kind of job agreement		
Unlimited-term employment contract	0.112	0.004
Limited-term employment contract	-0.039	0.004
Task oriented employment contract	s.i.	
Probation contract	Ref	Ref
Economic activity of the employer		
NACE Rev.2 A, agriculture, forestry and fishing	0.355	0.005
NACE Rev.2 B, mining and quarrying	0.782	0.003
NACE Rev.2 C, manufacturing	0.255	0.002

Exogenous variables	Coefficient	Std. error
NACE Rev.2 D, Electricity, gas, steam and air conditioning supply	0.483	0.004
NACE Rev.2 E, Water supply; sewerage, waste management and remediation activities	0.241	0.004
NACE Rev.2 F, Construction	0.189	0.003
NACE Rev.2 G, Wholesale and retail trade; repair of motor vehicles and motorcycles	0.146	0.003
NACE Rev.2 H, Transportation and storage	0.202	0.003
NACE Rev.2 I, Accommodation and food service activities	0.0754	0.004
NACE Rev.2 J, Information and communication	0.3698	0.004
NACE Rev.2 K, Financial and insurance activities	0.346	0.003
NACE Rev.2 L, Real estate activities	0.167	0.004
NACE Rev.2 M, Professional, scientific and technical activities	0.252	0.004
NACE Rev.2 N, Administrative and support service activities	Ref	Ref
NACE Rev.2 O, Public administration and defence; compulsory social security	0.193	0.003
NACE Rev.2 P, Education	0.053	0.003
NACE Rev.2 Q, Human health and social work activities	0.060	0.003
NACE Rev.2 R, Arts, entertainment and recreation	s.i.	
Nace rev.2 S, Other service activities	0.165	0.008
Post held within the company by the employee		
Managers	0.708	0.003
Professionals	0.387	0.002

Exogenous variables	Coefficient	Std. error
Technicians and associated Professional	0.238	0.002
Clerical support Workers	0.123	0.002
Service and sales Workers	0.013	0.002
Skilled agricultural, forestry and fishery workers	s.i.	
Craft and related trades workers	0.110	0.002
Plant and machine operators, and assemblers	0.142	0.002
Elementary occupations	Ref	Ref
NUTS 2 regions (voivodships)		
Dolnośląskie	0.043	0.003
Kujawsko-Pomorskie	-0.031	0.003
Lubelskie	-0.091	0.003
Lubuskie	-0.036	0.004
Łódzkie	-0.034	0.003
Małopolskie	-0.008	0.003
Mazowieckie	0.134	0.003
Opolskie	-0.025	0.004
Podkarpackie	-0.118	0.003
Podlaskie	-0.094	0.004
Pomorskie	0.049	0.003
Śląskie	0.021	0.003
Świętokrzyskie	-0.099	0.004
Warmińsko-mazurskie	-0.053	0.003
Wielkopolskie	-0.005	0.003
Zachodniopomorskie	Ref	Ref

R-squared: 0.516614, Adjusted R-squared: 0.516578.
Ref stands for reference variable, s.i. stands for statistically insignificant.

EFEKT WYKONYWANEGO ZAWODU
– CZY WYKORZYSTYWANIE KLASYFIKACJI ISCO NA POZIOMIE JEDNEJ CYFRY
WYSTARCZA, ŻEBY MODELOWAĆ WYNAGRODZENIA W POLSCE?

S t r e s z c z e n i e

W przeciwieństwie do neoklasycznych założeń doskonałej konkurencji istnieje konsensus, że do czynników wpływających na wysokość uzyskiwanego wynagrodzenia zaliczyć należy płeć, poziom wykształcenia, wiek, doświadczenie zawodowe, wykonywany zawód, posiadane stanowisko, stopień odpowiedzialności związanej z wykonywaną pracą oraz cały zestaw indywidualnych cech osobowościowych. W artykule posługując się Międzynarodową Klasyfikacją Zawodów i Specjalności (ISCO) na poziomie 2 cyfr uzyskano oceny parametrów strukturalnych modelu, które pozwalają wyjaśnić różnice wynagrodzeń w Polsce. Na podstawie przeprowadzonej analizy stwierdzono, że wykonywany zawód, mierzony na poziomie ISCO 2 cyfry jest ważną i statystycznie istotną zmienną oddziałującą na wysokość uzyskiwanego wynagrodzenia w Polsce. Modelowanie wynagrodzeń wykorzystujące klasyfikację ISCO, ale ograniczoną tylko do poziomu 1 cyfry, może być niewystarczające, aby prawidłowo uchwycić znaczenie wykonywanego zawodu dla uzyskiwanego wynagrodzenia.

Słowa kluczowe: determinanty wynagrodzeń, modelowanie wynagrodzeń, zawody, znaczenie wykonywanego zawodu

EFFECTS OF BEING IN AN OCCUPATION
– IS ISCO 1 DIGIT CLASSIFICATION ENOUGH TO MODEL WAGES IN POLAND?

A b s t r a c t

Contrary to neoclassical assumptions of perfect competition, there is a consensus that factors affecting wages include sex, level of education, age, job experience, occupation, post, work-related responsibility and a whole set of personality traits. The paper presents an econometric model that allows to explain wage differences in Poland and extends analyses of wage determinants in Poland by taking into account occupations broken down in accordance with the 2-digit level of International Standard Classification of Occupations (ISCO). The analysis shows that ISCO 2 digit level is an important and statistically significant determinant of wages in Poland, while models of wages basing on ISCO 1 digit might be not enough to properly capture the role of occupations.

Keywords: determinants of wages, wages modeling, occupations, importance of occupation