sciendo

Folia Oeconomica Stetinensia
Volume 19 (2019) Issue 2
DOI: 10.2478/foli-2019-0017

WYDZIAŁ NAUK EKONOMICZNYCH
I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

ASSESSMENT OF THE DEVELOPMENT OF THE EUROPEAN OECD COUNTRIES WITH THE APPLICATION OF LINEAR ORDERING AND ENSEMBLE CLUSTERING OF SYMBOLIC DATA

Marcin Pełka, Ph.D.

Wrocław University of Economics
Faculty of Economics, Management and Tourism
Department of Econometrics and Computer Sciences
e-mail: macin.pelka@ue.wroc.pl
ORCID: 0000-0002-2225-5229

Received 13 November 2018, Accepted 15 October 2019

Abstract

The **research background** of the paper covers the development of a country, that can be measured in various ways. Simple indicators, like GDP and also complex indicators such as HDI (*human development index*), can be used to measure country development. However, usually countries are divided into groups via setting some arbitrary levels of final measure. What is more, the composite (complex) indices have some problems and errors.

The main purpose of the paper is the assessment of the development of the selected European OECD countries with the application of the linear ordering and ensemble clustering of symbolic data as well as comparison of the ensemble clustering with a single model.

Research methodology covers linear ordering with the application of multidimensional scaling for a visualisation of results and ensemble clustering for symbolic data.

The results are compared according to adjusted Rand and silhouette indices. The obtained results show that ensemble clustering for symbolic data can be a useful tool in country development analysis and allows reaching better results than a single model.

The novelty of the proposed approach is to use a cluster analysis to obtain the clusters of countries with similar variables' values (indicators of development) and the application of multidimensional scaling for symbolic data in order to visualise linear ordering results.

Keywords: country development, symbolic data analysis, OECD countries, ensemble clustering

JEL classification: C01, C38, O10, O30, O57

Introduction

Recent papers show that there is a deep need for a more comprehensive way to measure the development of a country. A. Sen sees the development as the concept that goes far beyond the accumulation of wealth, measured by gross national product or similar indicators. In his opinion development should also enhance people's lives (Sen, 1999, p. 14). So the measurement of development should take into account different areas of people's lives – social, ecological, political and economic.

Many different papers deal with the problem of development and the comparison of development – see for example S. Vachon and Z. Mao (2008), S. Voigt (2009), A. Demirgüç-Kunt and R. Levine (2004), P.H. Hsu, X. Tian and Y. Xu (2014), S. Dasgupta et al. (1999).

When considering country development, cluster analysis and symbolic data analysis only two papers by M. Pełka (2017 and 2018) present an analysis of innovation which is an important element of sustainable development in the European Union using different ensemble clustering approaches.

A paper by D.B. Alonso et al. (2016) analyzed the how global crisis (during 2008–2012) had an impact on macroeconomic and social factors in the EU member states. The research findings prove the sensitivity and vulnerability of European countries during the crisis and could help policy makers identify effective measures for strengthening the protective capacity of their states in the event of a future economic and social crisis. C H. Ketels and O. Memedovic (2008) present how clusters can be leveraged for economic policy and what the role of different stakeholders in this process is. However their application part of the paper focuses on the concept to resource-rich, oil-dependent economies. K. Liapis et al. (2013) analyses the clusters of similarities among EU member states before and during the recent financial and debt crisis. Using Euclidian Distance and average linkage between groups two clusters were obtained. The first group consists of two subgroups, including the Netherlands, the UK, Luxembourg and Germany, characterised by developed financial sector's and balanced fiscal policy. The second subgroup consists of two subgroups: one subgroup consists of Finland and Sweden and the other one consists of Austria and Denmark. Finally the countries which are faced with several financial or debt problems, Belgium and Ireland are connected with the second group. The second largest group consists of Southern European countries, such as Italy, Spain, France, Greece and Portugal, and are characterised by high deficit and high government debt, low gross wage revenues and low Bank's Assets to GDP, low or medium total taxations performance, and decreases or deficits of current accounts and balance of payments. B. Mercan and D. Goktas (2011) focused their paper on the innovation of ecosystems.

The development of a country can be measured in many various ways. The simplest way is to use well-known wealth indicators such as gross domestic product (GDP), gross national product (GNP), gross national income (GNI). Usually the GNI is the standard way of measuring the level of wealth in a country (Baker, 2011, p. 6). As the measurement of total value of GNI, GDP, GDP can be misleading because it does not take into account the population of the country, the GNI, GDP or GNP are measured per capita – that is the total wealth of a particular country is divided by its population.

There were many efforts done to build rather a composite index that would capture all aspects of development. A major work was done by the United Nations Development Programme (UNDP) that proposed the Human Development Index (HDI). The HDI goes far beyond the measurement of the pure GDP, GNI, GNP and takes into account many different aspects of development (see for example Aziz et. al., 2015; McGillivray, 1991; Stanton, 2007; Sen, 1994).

Many other development indices, besides the HDI, have been proposed for many different purposes – like Quality of Life, Inequality-adjusted Human Development Index, OECD Better Life Index, Gender-related Development Index, Bhutan GNH Index (see for example Magee, Scerri, James, 2012; Durand, 2015; Dijkstra, Hanmer, 2000; Bates, 2009). However even the composite indices have some drawbacks (see Sagar, Najam, 1998; McGillivray, 1991).

For example A.D. Sagar and A. Najam (1998) show that a Human Development Index fails to include any ecological considerations. According to them over the years, the HDRs seem to have become stagnant, repeating the same rhetoric without necessarily increasing the HDI's utility.

According to M. McGillivray (1991) the HDI assesses intercountry development levels on the basis of three so-called deprivation indicators: life expectancy, adult literacy and the logarithm of purchasing power adjusted per capita GDP. Using a simple statistical analysis, his paper questions both the composition of the HDI and its usefulness as a new index of development. This paper concludes that the HDI is both flawed in its composition and, like a number of its predecessors, fails to provide insights into intercountry development level comparisons which pre-existing indicators, including GNP per capita, alone cannot.

What is more the HDI covers long-term changes (e.g. in GDP per capita) and may not respond to recent short-term changes. Many composite indices do not cover wide divergence within countries. A well-known problem is that higher national income may not mean welfare.

It depends on how it is spent. Some countries with high real GNI per capita have high levels of inequality (e.g. Saudi Arabia or Russia).

Thus, there is still a need for development of new indices, development measurement and cross-country development comparisons. The main aim of the paper is to present a symbolic ensemble clustering of the selected European OECD countries considering their development as well as the linear ordering results for this type of data using multidimensional scaling.

The obtained results show that ensemble clustering for symbolic data and linear ordering for this type of data can be a useful tool for a development analysis.

1. Symbolic data analysis

In classical data analysis objects are usually described by a vector of quantitative or qualitative measurements, where each column represents a single variable (a number or a category). However, this kind of data representation is too restrictive to represent more complex data. This type of data takes into account the uncertainty and/or variability to the data, variables must assume a set of categories or intervals even with frequencies or weights. Such data has been mainly studied in Symbolic Data Analysis. It provides suitable methods and algorithms to deal with aggregated or complex data that are described by multi-valued variables, where cells of the data table can contain sets of categories, intervals or weights (probabilities) distributions (see for example Bock, Diday, 2012; Billard, Diday, 2006; Billard, Diday, 2008; Noirhomme-Fraiture, Brito, 2011). Table 1 presents examples of the main types of symbolic variables and their realisations (see Bock, Diday, 2000, p. 2).

Table 1. Example of symbolic variables and their realisations

Symbolic variable	Realisations	Variable type
Money spent monthly on food (PLN)	<100, 200>; <150, 300>; <170, 400>	symbolic interval-valued (non-disjoint intervals)
Distance to work (km)	<0, 5>; <5, 10>; <10, 15>; <15, 20>	symbolic interval-valued (disjoint)
Preferred car make	{Toyota}, {VW, Audi}, {Skoda, Kia}	categorical multi-valued
Preferred laptop brands	{Asus (0.6), Lenovo (0.4)}, {Acer (0.4), Asus (0.3), Dell (0.3)}	categorical modal
Time spent travelling to work weekly (min.)	{<0, 10> (0.6); <10, 20> (0.4)}, {<0, 10> (0.1); <10, 20> (0.9)}, {<0, 10> (0.1); <10, 20> (0.5); <20, 30> (0.4)},	histogram
Gender of a person	{M}, {F}	categorical single-valued
Number of rooms in a flat	1, 2, 3, …	numerical single-valued

Source: own elaboration.

Symbolic data analysis allows to describe objects in a more detailed, complex, way but is requires a special type of distance measures, clustering algorithms, etc. that can deal with such types of data. More details about symbolic variables, objects can be found in e.g. H. H. Bock and E. Diday (2000), L. Billard and E. Diday (2006), E. Diday and M. Noirhomme-Fraiture (2008), M. Noirhomme-Fraiture and P. Brito (2011).

In the case of symbolic objects two types of data aggregation are used (see Billard, Diday, 2006; Diday, Noirhomme-Fraiture, 2008; Noirhomme-Fraiture, Brito, 2011):

a) temporal data aggregation – where information for classical objects (individuals) are aggregated over time;

b) contemporary data aggregation – where other information than time is used to obtain symbolic objects (e.g. consumers buying the same product).

In the case of symbolic interval-valued variables usually four approaches are used to obtain the intervals:

- minimum values from the data set as lower bounds of intervals and maximum values from the data set as upper bounds of intervals,
- the first quartile from the data set as lower bounds of intervals and the third quartile from the data set as upper bounds of intervals,
- the 10th percentile from the data set as lower bounds of intervals and the 90th percentile from the data set as upper bounds of intervals,
- arbitrary taken values for lower and upper bounds of intervals.

In this paper first and third quartile of the original data values will be used in the empirical part and the contemporary data aggregation was used (data about countries was aggregated over time).

2. Linear ordering and ensemble clustering for symbolic data

The first concepts on pattern and anti-pattern of development and the measurement of development were proposed by Professor Z. Hellwig (see Walesiak, 2017a, p. 2). A two-step procedure that allows visualizing the results of linear ordering was presented by M. Walesiak (see 2016, 2017b, p. 11). This procedure can be applied also for symbolic objects and it involves.

1. Choice of a complex phenomenon that cannot be measured directly. This phenomenon is considered among a set of objects *A*.

2. Choice of objects.

3. Selection of variables and collecting data and the construction of a symbolic data table.

Identification of preferential variables – stimulants, destimulants and nominants. Variable is a stimulant when for every two of its observations v_{ij}^{S}, v_{kj}^{S} for two objects A_i, A_k there is a reference $v_{ij}^{S} > v_{kj}^{S} \Rightarrow A_i \succ A_k$ ($\succ$ means that object A_i dominates over object A_k). Variable is a destimulant for every two of its observations v_{ij}^{D}, v_{kj}^{D} for two objects A_i, A_k there is a reference $v_{ij}^{D} > v_{kj}^{D} \Rightarrow A_i \prec A_k$ ($\prec$ means that object A_k dominates over object A_i). A variable is a unimodal nominant when for every two of its observations v_{ij}^{N}, v_{kj}^{N} for two objects A_i, A_k if $v_{ij}^{N}, v_{kj}^{N} \leq nom_j$ then $v_{ij}^{N} > v_{kj}^{N} \Rightarrow A_i \succ A_k$ and if $v_{ij}^{N}, v_{kj}^{N} > nom_j$ then $v_{ij}^{N} > v_{kj}^{N} \Rightarrow A_i \prec A_k$ (where i, k are object's numbers and j is the variable's number).

4. Transformation of nominants into stimulants. There is no need to transform destimulants into stimulants.

5. Normalisation of variable values.

6. Calculation of distances between objects. For a symbolic data case many different distance measures can be applied. They are chosen upon the type of symbolic variables.

Multidimensional scaling for symbolic data is done. This method allows representing objects from a m-dimensional space on a q-dimensional space (usually $q = 2, 3$). In the case of symbolic data two approaches for multidimensional scaling can be carried out:

a) the symbolic-symbolic approach, where for symbolic data a table containing symbolic interval-valued data only, interval-valued distances are calculated. The interval-valued distances are used to carry out multidimensional scaling with one of three approaches – InterScal, SymScal or I-Scal (see Groenen, Terada 2015; Goenen et al., 2005; Groenen et al., 2006). This approach allows to represent symbolic objects as rectangles, in the case of a 2 dimensional plane, or cuboids (hyperrectangles) in the case of a 3 dimensional plane;

b) symbolic-numeric approach, where for a symbolic data table containing any symbolic variables distances for all objects are calculated. There are many different distance measures for symbolic data. The choice is usually limited to the symbolic variables (see for example Gatnar, Walesiak, 2011). The iterative procedure called **smacof** was used in this paper (Borg, Groenen, 2005, pp. 204–205). The symbolic-numeric approach allows representing symbolic objects as points.

7. Graphical representation and interpretation of the results (multidimensional scaling and linear ordering results). For the multidimensional scaling results a straight line that connects pattern and anti-pattern (so-called axis of the set) is added. Isoquants of development are determined and based on a pattern object – e.g. by dividing the axis of the set into four equal

parts. The objects that are located between isoquants represent a similar level of development that can be archived by different objects.

Normalised distances of the *i*-th object from the pattern of development are calculated as follows (see Hellwig, 1981, p. 62):

$$d_i^+ = \frac{\sqrt{\left(z_{ij} - z_{+j}\right)^2}}{\sqrt{\left(z_{+j} - z_{-j}\right)^2}}, \ d_i^+ \in [0;1] \tag{1}$$

where:

$\sqrt{\left(z_{ij} - z_{+j}\right)^2}$ – Euclidean distance between the *i*-th object and the pattern object,

$\sqrt{\left(z_{+j} - z_{-j}\right)^2}$ – Euclidean distance between pattern and anti-pattern objects,

i – the object's number,

j – the variable's number,

$+$ – means pattern object,

$-$ – means anti-pattern.

The objects are ordered by the increasing values of distance measure (1). The linear ordering results are presented in Figure 1.

When considering clustering methods for symbolic data we can distinguish the following groups of methods:

a) adaptations of the classical clustering method and clustering methods designed strictly for symbolic data (see for example Verde, 2004; Bock, Diday, 2000; Billard, Diday, 2006; Diday, Noirhomme-Fraiture, 2008);

b) density based clustering for symbolic data – an adaptation of a well-known DBSCAN algorithm for symbolic data (see Pełka, 2018);

c) conceptual clustering methods for symbolic data, e.g. pyramids or the adaptation of COBWEB (see Pełka, 2015; Brito, 2002; Brito, 1995).

In general, in ensemble clustering for symbolic data we can use three main groups of methods (see Ghaemi et al., 2009; de Carvalho, Lechevallier, Melo, 2012; Hornik, 2005; Leisch, 1999; Dudoit, Fridlyand, 2003):

1. Clustering based on multiple relational matrices proposed by F.D.A. de Carvalho, Y. Lechevallier and F.M. Melo (2012), where multiple distance matrices (different points of view)

are used to calculate relevance weight vectors. These vectors and distance matrices are used to cluster objects.

2. A clustering ensemble that uses one of the consensus functions, e.g. co-clustering matrix, hypergraph partitioning, mutual information, finite mixture model (Ghaemi et al., 2009). In this paper the co-clustering (co-association) matrix will be used. The algorithm that uses a co-association matrix can be described as follows (Fred, Jain, 2005, p. 848):

a) obtain different base partitions (models). This can be done in many ways – e.g. by using the same clustering algorithm with different initial parameters (e.g. number of clusters, normalisation method, and distance measure, etc.), using subsets of objects, using subsets of variables, and using different clustering algorithms. In the paper different clustering techniques will be used (SClust, DClust, DBSCAN for symbolic data, spectral clustering for symbolic data, single, and complete link clustering) and also these methods will be used with different initial parameters (distance measure, normalisation, number of clusters varying from 2 to 20);

b) use obtained partitions to build a co-clustering (co-association) matrix. The elements of this matrix are defined as follows:

$$C(i,j) = \frac{n_{ij}}{N}, \tag{2}$$

where:

i, j – objects (pattern) number,

n_{ij} – number of times objects i, j were clustered together among N partitions,

N – total number of partitions;

c) the obtained co-association matrix is used as the data matrix for some classical clustering methods – like *k*-means, pam, etc.;

d) choosing the best partitions – e.g. by using cluster quality indices. In the paper a well-known silhouette index will be used (see Kaufman, Rousseeuw, 2009 for further details).

3. Adaptations of well-known bagging procedure for clustering (see Hornik, 2005; Leisch, 1999; Dudoit, Fridlyand, 2003). In this paper the F. Leisch's adaptation of bagging for clustering will be used (see Leisch, 1999):

a) the initial data set is divided into *M* subsets, drawn from the initial data set with a replacement – in this paper 20 subsets with 20 objects each will be used;

b) subsets are clustered and in the case of classical data centers of clusters are obtained. In the case of symbolic data medoids will be used;

c) medoids (cluster centers) are used as the data matrix for some clustering algorithms – e.g. *k*-means, ward, complete, DIANA, etc. and final clusters are obtained;

d) for the final clusters new medoids are calculated;

e) all objects are assigned to the nearest medoid.

3. Results of the empirical study

The empirical study uses the statistical data obtained from the World Bank (World Development Indicators available at https://datacatalog.worldbank.org/dataset/world-development-indicators). The data contains information about development indicators for 217 countries. From this data set 30 OECD countries were taken. The evaluation of their development was done using the following variables (where contemporary data aggregation was used – the data about countries was aggregated over time):

v_1 – access to clean fuels and technologies for cooking (percent of population),
v_2 – account in a financial institution (percent of people aged 15+),
v_3 – account in a financial institution (40% of poorest people, aged 15+),
v_4 – adjusted net national income (constant 2010 in USD),
v_5 – adjusted net savings, excluding particulate emission damage (percent of GNI),
v_6 – adolescent fertility rate (births per 1,000 women ages 15–19),
v_7 – age dependency ratio (percent of working population),
v_8 – alternative and nuclear energy (percent of total energy use),
v_9 – annual freshwater withdrawals caused by agriculture (percent of total freshwater withdrawal),
v_{10} – annual freshwater withdrawals caused by households (percent of total freshwater withdrawal),
v_{11} – annual freshwater withdrawals caused by industry (percent of total freshwater withdrawal),
v_{12} – birth rate (per 1,000 people),
v_{13} – total central government debt (percent of GDP),
v_{14} – CO_2 emissions (kt),
v_{15} – cost of business start-up procedure (percent of GNI per capita),
v_{16} – GDP per capita (current USD's),

v_{17} – unemployment with higher education (percent of total labour force with higher education),

v_{18} – unemployment with basic education (percent of total labour force with basic education),

v_{19} – life expectancy at birth.

Variables v_1, v_2, v_3, v_4, v_5, v_7, v_8, v_{12}, v_{16}, v_{19} are stimulants, variables v_6, v_9, v_{10}, v_{11}, v_{13}, v_{14}, v_{15}, v_{17} and v_{18} are destimulants. To find the optimal multidimensional scaling (for a symbolic-numeric approach) the mdsOpt package of R was used (Walesiak, Dudek, 2018).

The best results were obtained for standardisation in terms of normalisation and Ichino-Yaguchi distance measure. Figure 1 presents the results of the multidimensional scaling of 32 objects (30 OECD countries, the pattern and the anti-pattern object). Objects 31 (pattern object) and 32 (ani-pattern object) were connected with a straight line to obtain the so-called axis of a set. Four isoquants were added by dividing the axis into four equal parts.

The distances of each country from a pattern object were calculated in accordance with formula (1). OECD countries were ordered by the growing values of this measure. The results are presented in Table 2.

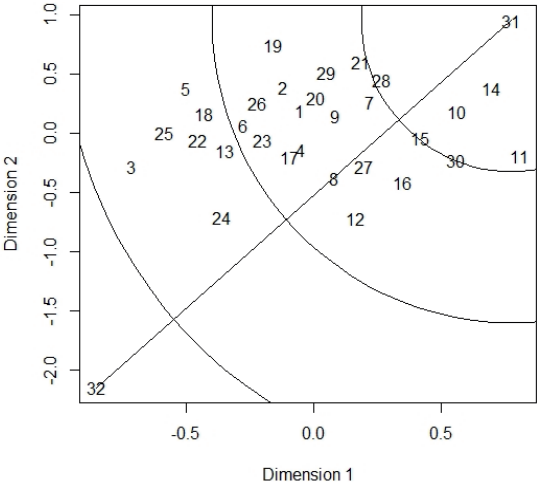

Figure 1. Results of the multidimensional scaling in the two-dimensional space for 32 objects – 30 OECD countries, pattern object (31) and anti-pattern object (32)

Source: own elaboration using R software.

Table 2. The ordering of 32 objects regarding the development of the OECD countries

Object no.	Distance	Name
31	0.0000	pattern
14	0.1665	Iceland
21	0.1958	Norway
28	0.2034	Sweden
10	0.2288	France
29	0.2419	Switzerland
7	0.2532	Denmark
19	0.2721	Luxembourg
20	0.2872	The Netherlands
2	0.3008	Belgium
15	0.3010	Ireland
9	0.3021	Finland
1	0.3197	Austria
11	0.3269	Germany
30	0.3422	The United Kingdom
26	0.3469	Slovenia
27	0.3884	Spain
4	0.3893	Croatia
6	0.3917	Czech Republic
5	0.3978	Cyprus
23	0.3995	Portugal
16	0.4107	Italy
18	0.4111	Lithuania
17	0.4132	Latvia
8	0.4288	Estonia
13	0.4484	Hungary
22	0.4540	Poland
25	0.4720	The Slovak Republic
12	0.5087	Greece
3	0.5509	Bulgaria
24	0.5754	Romania
32	1.0000	anti-pattern

Source: own elaboration using R software.

Then for the same data set ensemble symbolic clustering was done. Both Leisch's adaptation of bagging and the co-clustering matrix allowed obtaining the same final clustering results. A two cluster structure was obtained (with silhouette the clustering index equal to 0.87321):

1. Cluster 1 contains the following countries: Denmark, Iceland, Norway, Sweden, France, Germany, Italy, the United Kingdom, Finland and the pattern object. These countries are the most developed ones. They reach the narrowest symbolic interval-valued variables spans (ranges) for all variables. This means that the objects from this cluster are most similar to each other. They have high, but narrow in terms of interval length, values of stimulants (v_1, v_2, v_3, v_4, v_5, v_7, v_8, v_{12}, v_{16}, v_{19}) and low and narrow values for destimulants (v_6, v_9, v_{10}, v_{11}, v_{13}, v_{14}, v_{15}, v_{17}, v_{18}). So it means people living in these countries have very good access to clean fuels and technologies for cooking, both the young and the poorest have accounts in financial institutions. They have a high adjusted net national income (measured via constant 2010 in USD or GDP per capita) also their adjusted net savings damage are high. Countries from this cluster have low adolescent fertility rate, and a high age dependency ratio (% of the working population). People from these countries usually have a choice to use alternative and nuclear energy and they care about annual freshwater withdrawals caused by agriculture, households and industry. Unfortunately the birth rate is usually lower than in other countries. Total central debt and CO_2 emissions are an issue in these countries. The costs of business start-up procedures are usually lower. Both unemployment with higher or basic education are quite low. Life expectancy is quite high.
2. Cluster 2 contains the following countries: Austria, Belgium, Bulgaria, Croatia, Czech Republic, Estonia, Cyprus, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, the Netherlands, Poland, Portugal, Romania, the Slovak Republic, Slovenia, Spain, Switzerland and the anti-pattern object. This cluster contains all other countries with high and mid-high development (according to HDI). Countries from this cluster are the least similar ones. People from these counties have good and very good access to clean fuels and technologies for cooking, usually the youngest members of these countries have accounts in financial institutions. The poorest usually do not have one. Their income varies a lot, when looking at different countries from this cluster. Net savings are not so high as in cluster one. The countries from these clusters also have quite a low fertility rate and age dependency ratio. People from these countries not always have a choice to use nuclear energy and their interest about freshwater withdrawals is not always so clear. Sometimes the governments of these countries have problems when considering CO_2 emissions and CO_2 limits. The costs of business start-up procedures

are higher than in cluster 1. Both unemployment with higher education is lower than in the case of basic education. Generally speaking life expectancy at birth is a bit lower than in cluster one.

When compared the ensemble clustering results with the single clustering (pam method with Ichino-Yaguchi normalised distance measure) the results are the same – also the two cluster structure is reached with the same objects in each cluster. For ensemble and single clustering the results were adjusted and the Rand index was calculated. In both cases we have quite stable clustering results with 0.673212 (for the single model) and 0.772136 (for ensemble clustering).

Conclusions

As the results of the applied research the analysis of development for 30 OECD countries was done. The linear ordering (see the results presented in Table 2) and cluster analysis were conducted for 30 OECD countries using a symbolic-numeric approach for linear ordering visualisation, and single and ensemble clustering for symbolic interval-valued data.

The classifications done by international organisations (e.g. UNDP) were done on the basis of the composite development index where clusters are obtained by selecting some arbitrary values of these indices. This paper has used clustering methodology to obtain two clusters.

Cluster 1 contains the following countries: Denmark, Iceland, Norway, Sweden, France, Germany, Italy, the United Kingdom, Finland and the pattern object. These countries are the most developed ones. They reach the narrowest symbolic interval-valued variables spans (ranges) for all variables. This means that the objects from this cluster are most similar to each other. People living in these countries have very good access to clean fuels and technologies for cooking; both the youngest and the poorest have accounts in financial institutions. They have a high adjusted net national income (measured via constant 2010 in USD or GDP per capita) also their adjusted net savings damage is high. The countries from this cluster have a low adolescent fertility rate, and high age dependency ratio (percent of the working population). People from these countries usually have a choice to use alternative and nuclear energy and they care about annual freshwater withdrawals caused by agriculture, households and industry. Unfortunately the birth rate is usually lower than in other countries. Total central debt and CO_2 emissions are an issue in these countries. The costs of business start-up procedures are usually lower. Both unemployment with higher or basic education are quite low.

Cluster 2 contains the following countries: Austria, Belgium, Bulgaria, Croatia, Czech Republic, Estonia, Cyprus, Greece, Hungary, Ireland, Latvia, Lithuania, Luxembourg, the

Netherlands, Poland, Portugal, Romania, the Slovak Republic, Slovenia, Spain, Switzerland and the anti-pattern object. This cluster contains all other countries with high and mid-high development (according to HDI). The countries from this cluster are the least similar ones. People from these counties have good and very good access to clean fuels and technologies for cooking, usually the youngest have accounts in financial institutions. The poorest usually do not have one. Their income varies a lot, when looking at the different countries from this cluster. Net savings are not so high as in cluster one. The countries from this cluster also have quite a low fertility rate and age dependency ratio. People from these countries not always have a choice to use nuclear energy and how they care about freshwater withdrawals is not always so clear. Sometimes the governments of these countries have problems when considering CO_2 emissions and CO_2 limits. The costs of business start-up procedures are higher than in cluster 1. Both unemployment with higher education is lower than in the case of basic education.

References

Alonso, D.B., Androniceanu, A., Georgescu, I. (2016). Sensitivity and vulnerability of European countries in time of crisis based on a new approach to data clustering and curvilinear analysis. *Administratie si Management Public*, *27*, 46.

Aziz, S.A., Amin, R.M., Yusof, S.A., Haneef, M.A., Mohamed, M.O., Oziev, G. (2015). A critical analysis of development indices. *Australian Journal of Sustainable Business and Society*, *1* (01).

Baker, B. (2011). *World development: An essential text*. New Internationalist.

Bates, W. (2009). Gross national happiness. *Asian-Pacific Economic Literature*, *23* (2), 1–16.

Bock, H.H., Diday, E. (eds.) (2012). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer Science & Business Media.

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

Brito, P. (2002). Hierarchical and pyramidal clustering for symbolic data. *Journal of the Japanese Society of Computational Statistics*, *15* (2), 231–244.

Brito, P. (1995). Symbolic objects: order structure and pyramidal clustering. *Annals of Operations Research*, *55* (2), 277–297.

Dasgupta, S., Wheeler, D., Mody, A., Roy, S. (1999). *Environmental regulation and development: A cross-country empirical analysis*. The World Bank.

De Carvalho, F.D.A., Lechevallier, Y., De Melo, F.M. (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, *45* (1), 447–464.

Demirgüç-Kunt, A., Levine, R. (eds.) (2004). *Financial structure and economic growth: A cross-country comparison of banks, markets, and development*. MIT press.

Diday, E., Noirhomme-Fraiture, M. (eds.) (2008). *Symbolic data analysis and the SODAS software*. John Wiley & Sons.

Dijkstra, A.G., Hanmer, L.C. (2000). Measuring socio-economic gender inequality: Toward an alternative to the UNDP gender-related development index. *Feminist economics*, *6* (2), 41–75.

Dudoit, S., Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, *19* (9), 1090–1099.

Durand, M. (2015). The OECD better life initiative: How's life? and the measurement of well-being. *Review of Income and Wealth*, *61* (1), 4–17.

Fred, A.L., Jain, A.K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *6*, 835–850.

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

Ghaemi, R., Sulaiman, M.N., Ibrahim, H., Mustapha, N. (2009). A survey: clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, *50*, 636–645.

Groenen, P., Terada, Y. (2015). *Symbolic Multidimensional Scaling* (No. EI 2015-15).

Groenen, P.J., Winsberg, S., Rodriguez, O., Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities. *Computational Statistics & Data Analysis*, *51* (1), 360–378.

Groenen, P.J.F., Winsberg, S., Rodriguez, O., Diday, E. (2005). *SymScal: symbolic multidimensional scaling of interval dissimilarities* (No. EI 2005-15). Econometric Institute Research Papers.

Hellwig, Z. (1981). Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych. In: W. Welfe (ed.), *Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną* (pp. 46–68). Warszawa: PWE.

Hornik, K. (2005). A CLUE for CLUster ensembles. *Journal of Statistical Software*, *14* (12), 1–25.

Hsu, P.H., Tian, X., Xu, Y. (2014). Financial development and innovation: Cross-country evidence. *Journal of Financial Economics*, *112* (1), 116–135.

Kaufman, L., Rousseeuw, P.J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Ketels, C.H., Memedovic, O. (2008). From clusters to cluster-based economic development. *International Journal of Technological Learning, Innovation and Development*, *1* (3), 375–392.

Leisch, F. (1999). *Bagged clustering*. Working Paper no. 51. Vienna University of Economicsand Business Administration.

Liapis, K., Rovolis, A., Galanos, C., Thalassinos, E. (2013). The Clusters of Economic Similarities between EU Countries: A View Under Recent Financial and Debt Crisis. *European Research Studies*, *16* (1).

Magee, L., Scerri, A., James, P. (2012). Measuring social sustainability: A community-centred approach. *Applied Research in Quality of Life*, *7* (3), 239–261.

Mercan, B., Goktas, D. (2011). Components of innovation ecosystems: a cross-country study. *International research journal of finance and economics*, *76* (16), 102–112.

McGillivray, M. (1991). The human development index: yet another redundant composite development indicator? *World Development*, *19* (10), 1461–1468.

Nayak, P. (2010). Human development: conceptual and measurement issues. In: P. Nayak (ed.),

Growth and Human Development in North East India (pp. 3–18). New Delhi: Oxford University Press.

Noirhomme-Fraiture, M., Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, *4* (2), 157–170.

Pełka, M. (2017). Klasyfikacja wielomodelowa danych symbolicznych w badaniu innowacyjności krajów Unii Europejskiej. *Ekonometria*, *2* (56), 42–51.

Pełka, M. (2018). Analysis of Innovations in the European Union Via Ensemble Symbolic Density Clustering. *Econometrics*, *22* (3), 84–98.

Pełka, M. (2015). An adaptation of COBWEB for symbolic data case. *Statistica*, *75* (3), 265–273

Sen, A. (1999). Freedom as development. New York: Oxford Univerity Press.

Sagar, A.D., Najam, A. (1998). The human development index: a critical review. *Ecological economics*, *25* (3), 249–264.

Sen, A. (1994). Human Development Index: Methodology and Measurement.

Stanton, E.A. (2007). *The human development index: A history*. PERI Working Papers, 85.

Vachon, S., Mao, Z. (2008). Linking supply chain strength to sustainable development: a country-level analysis. *Journal of Cleaner Production*, *16* (15), 1552–1560.

Verde, R. (2004). Clustering methods in symbolic data analysis. In: *Classification, clustering, and data mining applications* (pp. 299–317). Berlin, Heidelberg: Springer.

Voigt, S. (2009). The effects of competition policy on development–cross-country evidence using four new indicators. *Journal of Development Studies*, *45* (8), 1225–1248.

Walesiak, M. (2016). Visualization of linear ordering results for metric data with the application of multidimensional scaling. *Ekonometria*, *2* (52), 9–21.

Walesiak, M. (2017a). Wizualizacja wyników porządkowania liniowego dla danych porządkowych z wykorzystaniem skalowania wielowymiarowego. *Przegląd Statystyczny*, *64* (1), 5–19.

Walesiak, M. (2017b). The application of multidimensional scaling to measure and assess changes in the level of social cohesion of the Lower Silesia region in the period 2005–2015. *Econometrics/Ekonometria*, *3* (57).

Walesiak, M., Dudek, A. (2018). The mdsOpt package for R software. Retrieved from: www.r-project.org.