ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 3
ISSN 1507-3866; e-ISSN 2449-9994

ANALYSIS OF HAPPINESS IN EU COUNTRIES USING THE MULTI-MODEL CLASSIFICATION BASED ON MODELS OF SYMBOLIC DATA

Marcin Pełka

Wrocław University of Economics, Wrocław, Poland
e-mail: marcin.pelka@ue.wroc.pl

ORCID: 0000-0002-2225-5229

DOI: 10.15611/eada.2019.3.02
JEL Classification: C38; C87; O52

Abstract: The results of happiness analysis are presented in the form of a World Happiness Report that covers 156 countries and 17 different indicators. In the article model-based clustering ensemble is built to determine what selected European countries have similar patterns of happiness. The results are analyzed using multidimensional scaling and a decision tree to find out what factors determine cluster memberships. In the empirical part, three clusters were detected The first contains countries: Austria, Denmark, Finland, Germany, Ireland, Luxembourg, the Netherlands, Norway, Sweden, Switzerland and the United Kingdom. They have the highest values for all the variables, except the negative affect. The second cluster contains seven countries: Bulgaria, Estonia, Hungary, Lithuania, Poland, Romania and Slovakia. This cluster is also the most homogeneous one. The third cluster contains eight countries: Cyprus, the Czech Republic, France, Greece, Italy, Portugal, Slovenia and Spain.

Keywords: happiness, the European Union, symbolic data analysis, ensemble clustering.

1. Introduction

Happiness is a quite new direction in economics. It is used to measure the overall happiness, and life quality in general, in different countries. In order to compare different countries many different happiness indicators were introduced. The results of happiness measurement are presented in the World Happiness Report [Helliwell et al. 2018]. The latest World Happiness Report ranks 156 countries by their total happiness, and 117 countries by the happiness of their immigrants. The data results from Gallup World Poll surveys and show both change and stability of happiness [Helliwell et al. 2018].

The problem of happiness measurement, and the correlation between migrations and happiness are quite often addressed in the literature (see for example [Helliwell et al. 2018; Graham 2012; Diener et al. 2010; Graham 2010; Wallis 2005; Deaton, Stone 2013; Henne et al. 2012; Krok 2016; Rokicka 2014; Machowska-Okrój 2014]).

Usually World Happiness Report indicators are used to build the more or less complex World Happiness Index (a composite index), that uses concepts of popular linear ordering. However, an interesting approach is to find patterns of similar happiness levels among different countries, regions, etc. To obtain such a goal, a cluster analysis can be applied.

The presented paper uses model-based clustering ensembles built with the application of R software (base models are obtained with `mlclust`, `mixture` and `Rmixmod` packages of R software) to determine which selected European countries share similar patterns of happiness.

2. World happiness indicators

The World Happiness Report [Helliwell et al. 2018] contains many different factors that describe different aspects of happiness. The following factors were taken into consideration in the empirical part to cluster countries:

1. logGDP per capita (x_1),

2. social support – the national average of the binary responses to the question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?” (x_2),

3. healthy life expectancy at birth – life expectancy at birth calculated by the World Happiness Report authors on data from the World Health Organization, World Development Indicators and statistics published in articles (x_3),

4. freedom of making choices – the national average of responses to the question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” (x_4),

5. generosity – the residual of regressing the national average of responses to the question “Have you donated money to a charity in the past month?” according to GDP per capita (x_5),

6. perception of corruption – the measure is the national average of the survey responses to two questions “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” The overall perception is just the average of the two 0-or-1 responses. In case the perception of government corruption is missing, the perception of business corruption is used as the overall perception. The corruption perception at national level is simply the average response of the overall perception at individual level (x_6),

7. positive affect – defined as the average of three positive affect measures of happiness, laughter and enjoyment in the Gallup World Poll waves. These measures are the responses to the following three questions, respectively: “Did you experience the following feelings during A LOT OF THE DAY yesterday?, How about Happiness?”, “Did you smile or laugh a lot yesterday?”, and “Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?”.

Waves 3-7 cover the years 2008 to 2012 and a small number of countries in 2013. For waves 1-2 and those from wave 8 on, positive affect is defined as the average of laughter and enjoyment only, due to the limited availability of happiness (x_7),

8. negative affect – defined as the average of three negative affect measures in the Gallup World Poll. They are worry, sadness and anger (x_8),

9. confidence in national government – the average answer to the question about confidence in the national government (x_9),

10. democratic and delivery quality – the measures of governance that are based on the Worldwide Governance Indicators (x_{10}),

11. The World Bank's estimate of the GINI index (x_{11}),

12. The GINI index of household income reported in Gallup (x_{12}).

3. Symbolic data cluster ensemble

In cluster analysis, the objects (patterns) to be clustered are usually described by single-valued variables. This allows the to be represented as a vector of qualitative measurements, where each column represents a variable.

Nevertheless, this kind of data representation is too restrictive to represent complex data. To take into account the uncertainty and/or variability to the data, variables must assume sets of categories or intervals, including in some cases frequencies or weights.

Such a kind of data have been mainly studied in *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 cells of the data table contain sets of categories, intervals, or weight (probability) distributions (see for example [Bock, Diday (eds.) 2000; Billard, Diday 2006; Noirhomme-Fraiture, Brito 2011]). Table 1 contains examples of symbolic variables.

Table 1. Examples of symbolic variables and their realizations

Symbolic variable	Realizations	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 brand	{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 person	{M}, {F}	categorical single-valued
Number of rooms in the flat/house	1, 2, 3, …	numerical single-valued

Source: own elaboration.

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

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

- temporal data aggregation – where information for classical objects (individuals) are aggregated over time,
- contemporary data aggregation – where information other than time is used to obtain symbolic objects (e.g. consumers buying the same product).

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

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

In this paper, the first and third quartile of the original data values will be used in the empirical part and temporal data aggregation will be used.

Ensemble learning techniques, that combine the results provided by different methods into one single model, are a useful tool for discriminant or regression tasks. However, the same idea of combining different base models (the results of clustering) can also be applied in the case of clustering for symbolic data.

For symbolic data the following ensemble techniques can be applied:

- 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,
- adaptations of the popular bagging procedure for clustering (see [Hornik 2005; Dudoit, Fridlyand 2003; Leisch 1999]).

The algorithm that uses the co-association matrix can be described as follows [Fred, Jain 2005, p. 848]:

- obtain different base partitions (models). This can be done in many ways – such as by using the same clustering algorithm with different initial parameters (e.g. number of clusters, normalization method, distance measure, etc.), using subsets of objects, subsets of variables, and different clustering algorithms. In this paper, spectral model-based clustering techniques will be used,

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

$$C(i,j) = \frac{n_{ij}}{N}, \tag{1}$$

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,

- the obtained co-association matrix is used as the data matrix for some classical clustering method – like *k*-means, pam, etc. In this paper, partitioning around medoids (pam) clustering will be applied,
- choosing the best partitions – e.g. by using cluster quality indices. In the paper, a popular silhouette index will be used.

4. Spectral model-based clustering for symbolic data

As model-based clustering uses EM algorithm for estimation, the direct application of any symbolic data cannot be done. In order to apply symbolic data for model-based clustering, spectral decomposition of the symbolic data table is needed.

The spectral approach is not in fact a new clustering algorithm, but rather a new way to prepare an original data set for some clustering algorithm (like pam, *k*-means, DBSCAN, hierarchical, etc.) [Ng et al. 2001; von Luxburg 2006].

The properties of spectral clustering have been studied from a theoretical point of view in many papers (see for example: [Ng et al. 2001; von Luxburg 2006; Shi, Malik 2000; Karatzoglu 2006)]).

Spectral decomposition for symbolic data table can be started in the following way:

a) let **V** be a symbolic data table with n rows and m columns. Let u be a number of clusters,

b) let **A** = $[A_{ik}]$ be an affinity matrix of objects from **V**. The **A** matrix can be obtained in many different ways. Most often its elements are defined as follows:

$$A_{ik} = \exp(-\sigma \cdot d_{ik}) \text{ for } i \neq k, \tag{2}$$

where: σ – scaling parameter that should minimize the sum of inter-cluster distances for a given number of clusters; d_{ik} – distance measure between *i*-th and *k*-th symbolic object. There are many different distance measures for symbolic data – e.g. the Ichino and Yaduchi distance measure, the normalized de Carvalho measure, etc. (see [Bock, Diday (eds.) 2000; Billard, Diday 2006; Diday, Noirhomme-Fraiture (eds.) 2008] for details on distance measurement for symbolic data),

c) calculation of the Laplacian $\mathbf{L} = \mathbf{D}^{1/2}\mathbf{A}\mathbf{D}^{1/2}$ (**D** – a diagonal weight matrix with sums of each row form **A** matrix on the main diagonal),

d) calculation of eigenvectors and eigenvalues of **L**. The first *u* eigenvectors will create **E** matrix,

e) normalization of the **E** matrix according to $y_{ij} = e_{ij} / \sqrt{\sum_{j=1}^{u} e_{ij}^2}$,

f) the **Y** matrix is then clustered with some usual clustering algorithm (i.e. pam, *k*-means).

In model-based clustering, we assume that the joint distribution is a mixture of *G* components, each of which is multivariate normal with density $f_k(x|\mu_k, \sum_k)$, $k = 1, \ldots, G$ (see for example [Fraley, Raftery 2000; Raftery, Deam 2006]. Then the mixture model can be described as follows:

$$f_k(x|\pi, \mu, \Sigma) = \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x_i|\mu_k, \Sigma), \tag{3}$$

where: π_k – probability that x_i belongs to the *k*-th component ($0 < \pi_k < 1$, $\sum_k \pi_k = 1$).

The popular EM algorithm is used to estimate model parameters.

5. Results of ensemble clustering

The R Statistical Software provided many model-based clustering packages that can be used in cluster analysis. In the paper the following packages will be used:

a) mclust [Scrucca et al. 2016],

b) mixture [Ryan, McNicholas 2014; Celeux, Govaert 1995],

c) Rmixmod [Lebret et al. 2015].

These methods allow to obtain model-based clustering results. In order to build ensembles, different distance measures (namely Ichino and Yaguchi, normalized Ichino and Yaguchi, de Carvalho based on description potential, normalized de Carvalho based on description potential and Hausdorff) will be used, also different *σ* and *u* parameters for thespectral approach were used in the spectral part. As a result 100 different base models were combined.

Finally, three clusters were obtained with a silhouette index equal to 0.5800769. The clustering results are presented in Table 2.

The first cluster contains eleven countries: Austria, Denmark, Finland, Germany, Ireland, Luxembourg, the Netherlands, Norway, Sweden, Switzerland and the United Kingdom. These countries are “the core” of the European Union together with highly-developed countries that are cooperating with the EU. These countries have the highest values for all variables, except the negative affect.

The second cluster contains seven post-communist countries from eastern Europe: Bulgaria, Estonia, Hungary, Lithuania, Poland, Romania and Slovakia. This cluster is also the most homogenous.

The third cluster contains eight countries: Cyprus, the Czech Republic, France, Greece, Italy, Portugal, Slovenia and Spain.

Table 2. Results of clustering

Cluster number	Countries (objects) and their number in the data set
1	Austria (1), Denmark (5), Finland (7), Germany (9), Ireland (12), Luxembourg (15), the Netherlands (16), Norway (17), Sweden (24), Switzerland (25), the United Kingdom (26)
2	Bulgaria (2), Estonia (6), Hungary (11), Lithuania (14), Poland (18), Romania (20), Slovakia (21)
3	Cyprus (3), the Czech Republic (4), France (5), Greece (10), Italy (13), Portugal(19), Slovenia (22), Spain (23)

Source: own computation using R software.

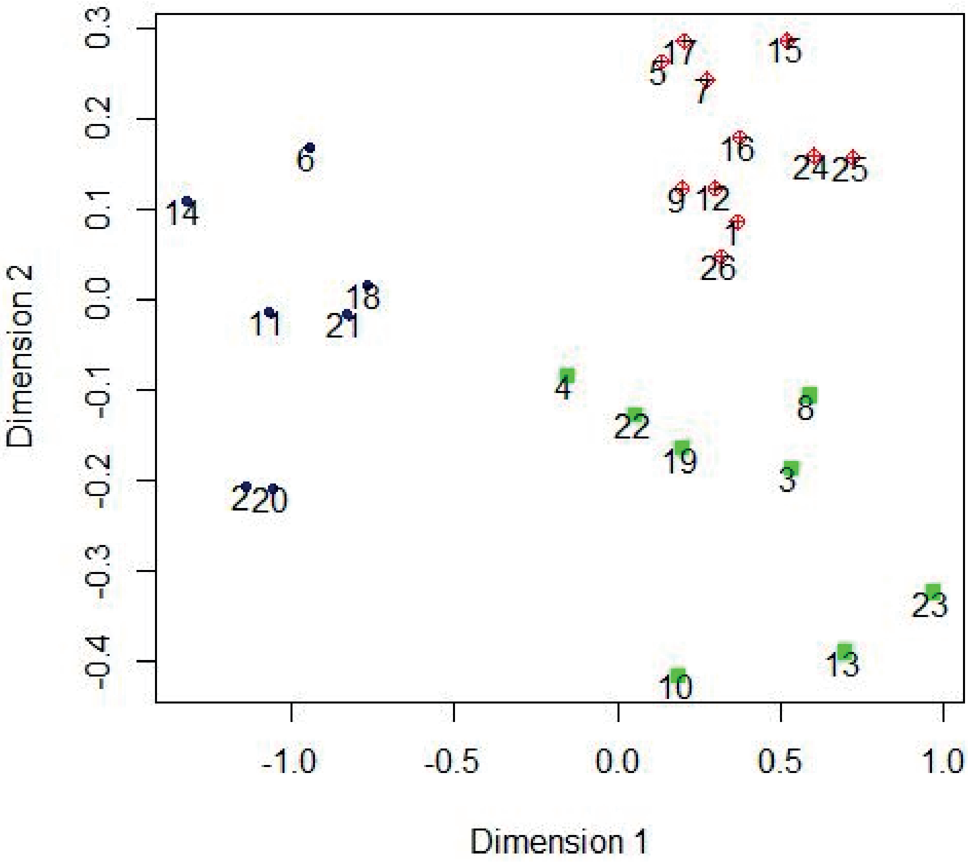

Fig. 1. Results of clustering presented in two-dimensional space

Source: own computation using R software.

To present the results of clustering (see Figure 1) a multidimensional scaling for symbolic interval-valued data was applied where a symbolic-numeric approach was used (symbolic interval-valued data were used to obtain the distance matrix which was used in classical multidimensional scaling – see Groenen et al. 2006 for details on multidimensional scaling for symbolic data).

Red points represent objects from the first cluster, blue objects from the second one and green represent objects from the third one (see Table 2 for clustering results to identify countries).

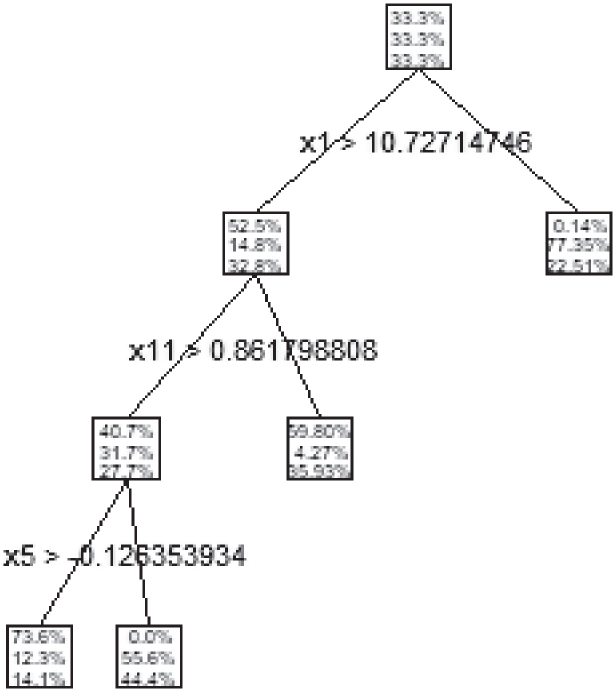

Fig. 2. Decision tree

Source: own computation using R software.

To see what factors have a significant impact on cluster membership, a decision tree for symbolic data was built using R software (see Figure 2). For details on symbolic decision tree construction see for example [Bock, Diday (eds.) 2000; Billard, Diday 2006; Gatnar, Walesiak (eds.) 2011]. As the `decisionTree.SDA` function of the `symbolicDA` package (see [Dudek et al. 2018]) requires to set testSet (objects in the test set) parameter, it was set randomly to contain 33% of the initial data set.

The most important variable for cluster membership is x_1 – logGDP per capita. Countries from the first cluster tend to have high values of the logGDP variable. The second variable is x_{11} – the World Bank's Estimate of the GINI index. This variable allows to distinguish objects from the first and second cluster. The third least important variable is x_5 – generosity. High generosity allows to distinguish objects from the first and second cluster.

6. Final remarks

Model-based clustering can be used for symbolic data only when the initial data pre-processing has been done. The spectral approach allows to conduct such data pre-processing.

In the clustering ensemble, the model-based approach can be used when applying different distance measures, and sigma parameters for the spectral data pre-processing.

In the empirical part of the paper three clusters were obtained. The first one contains "core European Union members" and some other high-developed countries (Austria, Denmark, Finland, Germany, Ireland, Luxembourg, the Netherlands, Norway, Sweden, Switzerland and the United Kingdom). The second cluster contains post-communist countries from Central and Eastern-Europe (Bulgaria, Estonia, Hungary, Lithuania, Poland, Romania and Slovakia). The third cluster contains eight countries (Cyprus, the Czech Republic, France, Greece, Italy, Portugal, Slovenia and Spain).

The most important factor that determines cluster membership is the logGDP, and the second is the World Bank's Estimate of the GINI index. The last variable is generosity.

Bibliography

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, 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.

Celeux G., Govaert G., 1995, *Gaussian parsimonious clustering models*, Pattern Recognition, 28(5), pp. 781-793.

Deaton A., Stone A.A., 2013, *Two happiness puzzles*, American Economic Review, vol. 103(3), pp. 591-597.

Diday E., Noirhomme-Fraiture M. (eds.), 2008, *Symbolic Data Analysis and the SODAS Software*, John Wiley & Sons.

Diener E., Ng. W., Harter J., Arora R., 2010, *Wealth and happiness across the world: Material prosperity predicts life evaluation, whereas psychosocial prosperity predicts positive feeling,* Journal of Personality and Social Psychology, 99(1), pp. 52-61.

Dudek A., Pełka M., Walesiak M., *The symbolic DA package for R software*, www.r-project.org.

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

Fraley C., Raftery A.E., 2002, *Model-based clustering, discriminant analysis and density estimation*, Journal of the American Statistical Association, 97(458), pp. 611-631.

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

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

Ghaemi R., Sulaiman M.N., Ibrahim H., Mustapha N., 2009, *A survey: Clustering ensembles techniques*, World Academy of Science, Engineering and Technology, 50, pp. 636-645.

Graham C., 2005, *The economics of happiness*, World Economics, 6(3), pp. 41-55.

Graham C., 2012, *Happiness around the World: The Paradox of Happy Peasants and Miserable Millionaire,* Oxford University Press.

Graham C., 2019, *Happiness around the World: The Paradox of Happy Peasants and Miserable Millionaires*, Oxford University Press.

Groenen P.J.F., Winsberg S., Rodriguez O., Diday E., 2006, *I-Scal: Multidimensional scaling of interval dissimilarities*, Computational Statistics and Data Analysis, vol. 51, pp. 360-378.

Helliwell J., Layard R., Sachs J., 2018, *World Happiness Report 2018*, Sustainable Development Solutions Network, New York.

Henne K., Jasińaka-Kania A., Skarżyńska K., 2012, *Zadowolenie z życia a zaufanie do ludzi w Polsce i w różnych regionach Europy*, [in:] A. Jasińska-Kania (red.), *Wartości i zmiany. Przemiany postaw Polaków w jednoczącej się Europie*, Wyd. Naukowe Scholar, pp. 78-104.

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

Karatzoglu A., 2006, *Kernel methods. Software, algorithms and applications*, Doctoral thesis, Vienna University of Technology.

Kaufmann L., Rousseeuw P.J., 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Krok E., 2016, *Metody pomiaru szczęścia i jego zależność od dochodu*, Studia Ekonomiczne, no. 286, pp. 43-55.

Lebret R., Iovleff S., Langrognet F., Biernacki Ch., Celeux G., Govaert G., 2015, *Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification mixmod library*, Journal of Statistical Software, vol. 67, issue 6, pp. 1-29.

Leisch F., 1999, *Bagged clustering*, Adaptive Information Systems and Modeling in Economics and Management Science, Working Paper 51.

Machowska-Okrój S., 2014, *Wzrost gospodarczy a dobrobyt ekonomiczno-społeczny w wybranych krajach europejskich*, Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, 35(2), pp. 409-430.

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

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.

Raftery A.E., Dean N., 2006, *Variable selection for model-based clustering*, Journal of the American Statistical Association, 101(473), pp. 168-178.

Rokicka E., 2014, *Rozwój gospodarczy i społeczny a jakość życia. Wybrane kontrowersje teoretyczne i metodologiczne*, Przegląd Socjologiczny, 63(1), pp. 81-107.

Ryan P.B., McNicholas P.D., 2014, *Estimating common principal components in high dimensions*, Advances in Data Analysis and Classification, 8(2), pp. 217-226.

Scrucca L., Fop M., Murphy T.B., Raftery A.E., 2016, *Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models*, The R Journal, 8/1, pp. 205-233.

Shi J., Malik J., 2000, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), pp. 888-905.

von Luxburg U., 2006, *A tutorial on spectral clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.

Wallis C., 2005, *The new science of happiness,* Time Magazine, vol. 22.

ANALIZA ZADOWOLENIA W KRAJACH UNII EUROPEJSKIEJ Z ZASTOSOWANIEM WIELOMODELOWEJ KLASYFIKACJI OPARTEJ NA MODELACH DANYCH SYMBOLICZNYCH

Streszczenie: W analizie zadowolenia stosowane są mierniki zadowolenia, a wyniki mają formę raportów. Dotyczą one 156 krajów, które są opisywane przez 17 zmiennych. W artykule zastosowno podejście wielomodelowe danych symbolicznych, w którym wykorzystano klasyfikację opartą na modelach, aby zidentyfikować, które z wybranych krajów Europy mają podobny poziom zadowolenia. Wyniki zanalizowano z użyciem skalowania wielowymiarowego i drzew klasyfikacyjnych danych symbolicznych. W efekcie zastosowanego podejścia zidentyfikowano strukturę trzech klas. W klasie pierwszej znalazły się: Austria, Dania, Finlandia, Niemcy, Irlandia, Luksemburg, Holandia, Norwegia, Szwajcaria oraz Wielka Brytania. Kraje te mają najwyższe wartości dla większości zmiennych. Klasa druga zawiera: Bułgarię, Estonię, Węgry, Litwę, Polskę, Rumunię, Słowację. Ta klasa jest równocześnie najbardziej homogeniczna. W klasie trzeciej znalazły się z kolei: Cypr, Czechy, Francja, Grecja, Włochy, Portugalia, Słowenia i Hiszpania.

Słowa kluczowe: zadowolenie, Unia Europejska, analiza danych symbolicznych, klasyfikacja wielomodelowa.