sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2018, Vol. 22, No. 3
ISSN 1507-3866; e-ISSN 2449-9994

ANALYSIS OF INNOVATIONS IN THE EUROPEAN UNION VIA ENSEMBLE SYMBOLIC DENSITY CLUSTERING

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

ORCID: 0000-0002-2225-5229

DOI: 10.15611/eada.2018.3.06
JEL Classification: C01, C38, O52, O30

Abstract: Innovations play a very important role in the modern economy. They are the key to a higher quality of life, better jobs and economy and sustainable development. The innovation policy is a key element of both national and European Union strategy. The main aim of this paper is to present an ensemble clustering of European Union countries (member states) considering their innovativeness. In the empirical section, symbolic density-based ensemble clustering is used to obtain the co-occurrence matrix. The paper uses `symbolicDA`, `clusterSim` and `dbscan` packages of R software for all calculations. Four different clusters where obtained in the result of clustering. Cluster 1 contains high-innovative countries (innovation leaders). This cluster is also the least homogenous. Cluster 2 contains post-communist countries mainly from central Europe. These countries can be seen as rather mid-low innovative (they try to "catch up" with innovation leaders). Cluster 3 contains moderate innovators. Cluster 4 contains two countries that are also mid-innovative.

Keywords: innovations, European Union, symbolic data analysis, ensemble clustering.

1. Introduction

The term "innovation" is widely used in many economic papers and it is not always clear what it means. In general innovation in the literature is defined in two ways. The first group of definitions focuses on the fact that innovations lead to changes in the company organization, product or services improvement, etc., and the final goal is the success of the company on the market place (see for example [Baregheh et al. 2009, p. 1334; Bessant et al. 2005, p. 1366; Boer, During 2001, p. 84; Lafley, Charan 2008, p. 21; Dosi 1988, p. 222; Hobday 2005, p. 122; Damanpour 1991, p. 556; Trott 2012, p. 15; Van de Ven et al. 1999; Bledow et al. 2009, p. 305; Khan 2012, p. 454]).

The second group of definitions focuses on the changes that are made in the company, but also it is the process that turns an idea into value for the customer

and results in sustainable profit for the enterprise [Carlson, Wilmot 2006, p. 4; O'Sullivan, Dooley 2009, p. 5; Silverstein et al. 2009, p. XVIII].

Since the mid-1990s a new approach in the meaning of science, technology and innovation has appeared. Not that science and technology have vanished, but innovation is understood as a broader and more flexible term that embraces other meaningful issues such as intellectual property rights, education, training, organizational change, institutional framework, standards etc. Innovation has expanded the agenda of EU action at a time when the member states wanted the EU to be the most competitive knowledge-based economy in the world [Borrás 2003, pp. 1-2].

Usually innovation of European Union member countries is measured by the summary innovation index (*SII*). This allows to order countries from the best to the worst, and analyse changes over time in a quite simple way. However the synthetic innovation index does not show patterns – i.e. which countries have similar values of variables that describe innovation.

This paper presents how symbolic density-based ensemble clustering can be applied in the case of the analysis of innovation in European Union members. This article is organized as follows. Section 2 presents the symbolic data and distance measures for this type of data. Section 3 presents European Union innovation policies and funding for innovation. Section 4 presents a general framework for density-based clustering. Section 5 presents a general framework for symbolic data ensemble where a density-based scan is used. Section 6 presents the analysis of innovation with the application of a symbolic data analysis ensemble.

2. Symbolic data and distance measures for symbolic data

In cluster analysis, objects (patterns) are usually described by single-valued variables. This allows to present each object as a vector of qualitative or quantitative measurements, where each column represents a variable. This kind of data representation is too restrictive to represent more 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 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 [Billard, Diday 2006; Bock, Diday (eds.) 2000]).

Table 1 presents examples and the main types of symbolic variables (see [Bock, Diday (eds.) 2000, p. 2]).

Although symbolic data allows to describe objects in more detailed way, it requires special distance measures, methods and algorithms that can deal with this type of data. More about symbolic data, symbolic variables, symbolic objects can be found in

Table 1. Example of symbolic variables and their realizations

Symbolic variable	Realizations	Variable type
Price of a car [in PLN]	<27 000, 38 000>; <35 000, 50 000>; <30 000, 45 000>	symbolic interval-valued (non-disjoint intervals)
Engine's ccm	<1000; 1200>, <1200; 1400>, <1400, 1800>	symbolic interval-valued (disjoint variable)
Preferred car colours	{blue, red, green}, {black}, {yellow, green, grey}	categorical multi-valued
Preferred car brands	{Toyota (0.4), VW (0.4), Skoda (0.2)}, {Audi (1.0)}	categorical modal
Distance travelled daily [km]	{<1, 5> (0.7); <5, 10> (0.3)}, {<1, 5> (0.5); <5, 10> (0.4); <10, 15> (0.1)}	histogram variable
Sex of a person	{M}; {F}	categorical single-valued
Number of children in a household	1,2, 3, …	numerical single-valued

Source: own elaboration.

e.g. Bock, Diday [2000]; Billard, Diday [2006]; Diday, Noirhomme-Fraiture [2008]; Noirhomme-Fraiture, Brito [2011].

Table 2 presents all suitable distance measures for symbolic interval-valued data available in R software (all of them, except `C_1` which is defined for symbolic hierarchical variables, will be used in the ensemble).

Table 2. Distance measures for Boolean symbolic objects available in R software

distType & distance name	Elements of distance measure	Distance measure $d(A_i, A_k)$
1	2	3
U_2 Ichino-Yaguchi	$\varphi(v_{ij}, v_{kj}) = \lvert v_{ij} \oplus v_{kj} \rvert - \lvert v_{ij} \otimes v_{kj} \rvert + \gamma\left(2 \cdot \lvert v_{ij} \oplus v_{kj} \rvert - \lvert v_{ij} \rvert - \lvert v_{kj} \rvert\right)$	$\sqrt[q]{\sum_{j=1}^{m} \varphi(v_{ij}, v_{kj})^q}$
U_3 normalized Ichino-Yaguchi	$\psi(v_{ij}, v_{kj}) = \frac{\varphi(v_{ij}, v_{kj})}{\lvert V_j \rvert}$	$\sqrt[q]{\sum_{j=1}^{m} \psi(v_{ij}, v_{kj})^q}$
U_4 weighted and normalized Ichino-Yaguchi	$\varphi(v_{ij}, v_{kj})$ same as in U_2	$\sqrt[q]{\sum_{j=1}^{m} w_j \psi(v_{ij}, v_{kj})^q}$

1	2	3
SO_2 de Carvalho	$\psi\left(v_{ij},v_{kj}\right)=\dfrac{\varphi\left(v_{ij},v_{kj}\right)}{\mu\left(v_{ij}\oplus v_{kj}\right)}$ $\varphi\left(v_{ij},v_{kj}\right)$ same as in U_2	$\sqrt[q]{\sum_{j=1}^{m}\dfrac{1}{m}\left[\psi\left(v_{ij},v_{kj}\right)\right]^{q}}$
SO_1 de Carvalho	$\alpha=\mu\left(v_{ij}\cap v_{kj}\right),$ $\beta=\mu\left[v_{ij}\cap c\left(v_{kj}\right)\right],$ $\chi=\mu\left[c\left(v_{ij}\right)\cap v_{kj}\right],$ $\delta=\mu\left[c\left(v_{ij}\right)\cap c\left(v_{kj}\right)\right]$ $d_1=\dfrac{\alpha}{\alpha+\beta+\chi},$ $d_2=\dfrac{2\alpha}{2\alpha+\beta+\chi},$ $d_3=\dfrac{\alpha}{\alpha+2\left(\beta+\chi\right)},$ $d_4=\dfrac{1}{2}\left[\dfrac{\alpha}{\alpha+\beta}+\dfrac{\alpha}{\alpha+\chi}\right],$ $d_5=\dfrac{\alpha}{\sqrt{\left(\alpha+\beta\right)+\left(\alpha+\chi\right)}}$ $d_f\left(v_{ij},v_{kj}\right)=1-D_f$ $f=1,\ldots,5$	$\sqrt[q]{\sum_{j=1}^{m}\left[w_j d_f\left(v_{ij},v_{kj}\right)\right]^{q}}$
C_1 de Carvalho for hierarchical or logical dependent variables		$\sqrt[q]{\dfrac{\sum_{j=1}^{m}\left[w_j d_f\left(v_{ij},v_{kj}\right)\right]^{q}}{\sum_{j=1}^{m}\delta\left(V_j\right)}}$
SO_3 de Carvalho	–	$\left[\pi\left(A_i\oplus A_k\right)-\pi\left(A_i\otimes A_k\right)+\gamma\left(2\pi\left(A_i\oplus A_k\right)-\pi\left(A_i\right)-\pi\left(A_k\right)\right)\right]$
SO_4 normalized de Carvalho	–	$\left[\pi\left(A_i\oplus A_k\right)-\pi\left(A_i\otimes A_k\right)+\gamma\left(2\pi\left(A_i\oplus A_k\right)-\pi\left(A_i\right)-\pi\left(A_k\right)\right)\right]/\pi\left(A^E\right)$
SO_5 normalized de Carvalho	–	$\left[\pi\left(A_i\oplus A_k\right)-\pi\left(A_i\otimes A_k\right)+\gamma\left(2\pi\left(A_i\oplus A_k\right)-\pi\left(A_i\right)-\pi\left(A_k\right)\right)\right]/\pi\left(A_i\oplus A_k\right)$
H Hausdorff	–	$\left[\sum_{j=1}^{m}\left(\max\left\{\left\lvert\overline{v}_{ij}-\overline{v}_{kj}\right\rvert,\left\lvert\underline{v}_{ij}-\underline{v}_{kj}\right\rvert\right\}\right)^{2}\right]^{\frac{1}{2}}$

Table 2, cont.

1	2	3
L_1	a) interval-valued variables $L_1(v_{ij}, v_{kj}) = \lvert\overline{v}_{ij} - \overline{v}_{kj}\rvert + \lvert\underline{v}_{ij} - \underline{v}_{kj}\rvert$ $L_2(v_{ij}, v_{kj}) = \lvert\overline{v}_{ij} - \overline{v}_{kj}\rvert^2 + \lvert\underline{v}_{ij} - \underline{v}_{kj}\rvert^2$	$\sqrt[q]{\sum_{j=1}^{m}\left(L_q(v_{ij}, v_{kj})\right)^q}$ $q = 1$ for L_1 $q = 2$ for L_2
L_2	b) multi-valued variables: $L_1(v_{ij}, v_{kj}) = \sum_{y_f}\lvert q_i v_j(y_f) - q_k v_j(y_f)\rvert$ $L_2(v_{ij}, v_{kj}) = \sum_{y_f}\lvert q_i v_j(y_f) - q_k v_j(y_f)\rvert^2$	

where: $v_{ij}v_{kj}$ – realizations of symbolic variables (interval-valued or multi-valued), $A_i = (v_{i1}, v_{i2}, \ldots, v_{im})$ 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 $q = 2$), $\lvert\ \rvert$ – 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, $\delta(V_j)$ – 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 multi-valued variables: q

Source: [Gatnar, Walesiak (eds.) 2011, pp. 20-23].

3. European Union innovation policies and funding for innovation and results of other studies on EU innovation

In order to accelerate the innovations in different sectors of the EU industry, the European Commission has developed several policies that help with the commercialization of innovations and support the innovation process in the EU mainly through the Horizon 2020 programme.

There are five main innovation policies that support innovation: social innovation, design for innovation, demand-side innovation policies, workplace innovation and public sector innovation. The European Commission promotes commercialization and the uptake of innovation through the Horizon 2020 programme and the European Structural and Investment Funds (ESIF).

Horizon 2020 is the biggest European Union Research and Innovation programme ever with nearly EUR 80 billion of European Union funding available over seven years (2014 to 2020). This programme is coupling research and innovation. The main goal is to ensure Europe produces world-class science, removes barriers to

innovation and makes it easier for the public and private sectors to work together in delivering innovation [European Commission, *Horizon 2020 in brief*; Regulation (EU) No 1291/2013 of the European Parliament and of the Council].

European Structural and Investment Funds (ESIF) dedicates around EUR 110 billion to innovation activities, small and medium-sized enterprises competitiveness, and also a low-carbon based economy. The structural funds will help regions to develop and store smart specialization strategies [see https://ec.europa.eu/growth/industry/innovation/funding/esif_en].

The main goal of the **European Fund for Strategic Investments** is to revive investment in strategic projects around Europe in order to ensure that money will reach the real economy [see https://ec.europa.eu/growth/industry/innovation/funding/efsi_en].

The analysis of innovation within the European Union, ratings and rankings that compare different countries are widely presented in the literature. The papers by Stec [2009] and Nowak [2012] present a comparison study where the innovation level of Poland is compared to European Union countries. A similar paper by Wojtas [2013] analyses the level of Polish innovation compared to other European Union countries. Quite similar work was published by Rynardowska-Kurzbauer [2015] where selected central and eastern European countries (including Poland) where compared according to their level of innovation. The article by Mikołajczyk [2013] compares levels of innovations in the enterprises using the EUROSTAT data. Similar works were suggested by Radicic et al. [2016] and Harrison et al. [2014].

In general, most papers use the summary innovation index (*SII*). This allows to order countries from the best to the worst, and analyse changes over time in a quite simple way. However the synthetic innovation index does not show patterns – that is which countries have similar values of variables that describe innovation.

4. Density-based clustering for symbolic data

Concerning classical data, Ester et al. proposed a density-based algorithm for discovering clusters (DBSCAN, see: [Ester et al. 1996]). This algorithm groups together points that have many neighbours and also makes outliers alone in low-density subregions. In 1998 Sander et. al. proposed a generalized version of DBSCAN (called GDBSCAN). They generalize the DBSCAN in two important directions. The generalized algorithm can cluster point objects as well as spatially extended objects according to both their spatial and non-spatial attributes [Sander et. al. 1998].

In 2013 Compello et. al. proposed a hierarchical version of the DBSCAN algorithm [Compello et al. 2013]. Ankrest et. al. in 1999 proposed the OPTICS algorithm which extends the ideas of DBSCAN [Ankrest et al. 1999]. Also SUBCLU [Kailing et al. 2004] and PreDeCon [Jahirabadkar, Kulkarni 2013] algorithms use similar subspace clustering ideas to those of DBSCAN. WaveCluster is another density-based clustering algorithm. It uses wavelet transform to the dimension space [Sheikholeslami et al. 1998]. However, this method is applicable only to

low-dimensional datasets. DenClue [Hinneburg, Keim 1998] is another efficient algorithm that uses information about density.

In order to find a cluster, the DBSCAN algorithm requires two parameters: ε (EPS) and the minimum number of points to form a dense region (minPts). All data points (objects to be clustered) can be classified as core points, density-reachable points and outliers according to DBCAN methodology as follows:

1. A data point p is a core point if at least minPts other points are within the distance ε of it, and those points are said to be **directly reachable** from p.

2. A data point k is reachable from p if there is a path p_1, ..., p_n with $p_1 = p$ and $p_n = k$, where each p_{i+1} is directly reachable from p_i (all data points on the way must be core points, with the possible exception of the last point k).

3. All points not reachable from any other points are outliers.

DBSCAN starts from arbitrary data point p and retrieves all points density-reachable from p wrt. ε and minPts. If this point's ε-neighbourhood contains a sufficient number of points (determined by minPts) a cluster is started. Otherwise, p is labelled as noise (outlier). It is important that noise points may later be found in a sufficiently sized ε-neighbourhood of another point. As global values ε and minPts are used in clustering, DBSCAN can merge two clusters into one cluster if two clusters of different density are "close" to each other.

DBSCAN requires three elements to run:

1. minPts – this value can be derived from the number of dimensions in the data set (D) as $minPTS \geq D+1$. For minPts equal to one every data point will be a cluster. If minPts is set to two the results will be the same as for hierarchical clustering with the single metric, with the dendrogram cut at height ε.

2. ε can be found using k-distance graph and plotting the distance to the k = minPts. If ε values are too small, a large part of the data set will be not clustered. The large values will cluster almost all data points.

3. Distance measure – in the case of classical data usually the Euclidean distance is used.

The adaptation of DBSCAN to the symbolic data case involves using one of the distance measures for symbolic data (see Table 2). The other steps are the same as in the case of the DBSCAN algorithm for classical data.

5. Ensemble clustering for symbolic data

Ensemble techniques based on aggregating information (results) from different models have been applied with success in the context of supervised learning (discrimination and regression). The ensemble techniques are applied in order to improve the accuracy and stability of classification algorithms.

Ensemble clustering means combining many different (N) base clustering results (models) P_1, ..., P_N into one aggregated model P^* with k^* clusters. The idea of ensemble clustering can be quite easily applied in the case of symbolic data (see for example [Pełka 2012; 2013; 2016]).

There are several ways to obtain different base partitions:

- use different clustering algorithms,
- use the same clustering algorithm with different initial parameters (different DBSCAN's with different minPts, ε and distance measures). Such an approach is also a solution for selecting initial parameters,
- use subsets of variables,
- receive different partitions by resampling the data.

There are many different ways to combine (aggregate) the results of clustering (see for example [Pełka 2012; 2013; 2016]). In this paper the co-occurrence (co-association) matrix will be used to obtain final partition.

The main idea of this matrix is that objects belonging to the same cluster ("natural cluster") are likely to be co-located in the same clusters among different partitions. The elements of co-association matrix are defined as follows [Fred, Jain 2005, p. 848]:

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

where: i, j – object (pattern) numbers; n_{ij} – number of times patterns (i,j) are clustered together among N different partitions; N – total number of partitions.

The algorithm of the ensemble that uses the co-clustering matrix can be described as follows:

1. Obtain different base partitions.

1. Build the co-association matrix (see equation 1).

2. Apply the co-association matrix as the data matrix for some classical clustering method – e.g. k-means, pam, etc.

3. Choose the best partition. Fred and Jain (see [Fred, Jain 2005]) suggest to apply the "lifetime" criterion in the case of hierarchical clustering methods. The "lifetime" is defined as the value of threshold value on the dendrogram that leads to identifying clusters (e.g. the longest link between two clusters). Besides this proposal, all the well-known cluster quality indices can be applied (e.g. Baker, Hubert, Russeeuw silhouette, Hubert and Levine, etc.) (see for example [Rendón et al. 2011]).

The co-association matrix can be also used as the data matrix for other classical data analysis methods – such as clustering, multidimensional scaling, etc.

6. Analysis of innovation with the application of the SDA ensemble

For purposes of innovation analysis within the European Union countries, data from the Regional Innovation Scoreboard was used. This scoreboard is a regional extension of the European Innovation Scoreboard, assessing the innovation performance of European regions on a limited number of indicators. It covers 220 regions across

22 EU countries, plus Norway, Serbia and Switzerland. In addition, Cyprus, Estonia, Latvia, Lithuania, Luxembourg, and Malta are included at country level.

Regional innovation data from the year 2017 were aggregated into country level using contemporary aggregation of the classical data (see for example [Bock, Diday (eds.) 2000; Billard, Diday 2006]) – symbolic second-order objects were obtained. After data aggregation the following symbolic interval-valued variables (see Table 3) were considered in the analysis (normalized scores per indicator) [Hollanders, Es-Sadki (eds.) 2017a; Hollanders, Es-Sadki (eds.) 2017b].

Table 3. Variables used in the analysis

Variable no. and name	Variable no. and name
x1 – population with tertiary education	x2 – lifelong learning
x3 – scientific co-publications	x4 – most-cited publications
x5 – R&D expenditures by public sector	x6 – R&D expenditures by business sector
x7 – non-R&D innovation expenditures	x8 – product or process innovations
x9 – marketing or organization innovations	x10 – SMEs innovation in-house
x11 – innovative SMEs collaborating with others	x12 – public-private co-publications
x13 – European Patent Office patent applications	x14 – trademark applications
x15 – design applications	x16 – employment in medium and high-tech manufacturing companies and knowledge-intensive services
x17 – exports from medium and high-tech manufacturing companies	x18 – sales of new-to-market and new-to-firm innovations

Source: own elaboration based on [Hollanders, Es-Sadki (eds.) 2017a; Hollanders, Es-Sadki (eds.) 2017b].

Due to lack of data for some or all regions, the following countries were not included in the analysis: Serbia, Estonia, Cyprus, Latvia, Lithuania, Luxembourg, Malta. Due to the lack of some variable data, Switzerland is also not included in the data. In this paper two main approaches are used to obtain base partitions:

1. Subsets of variables are obtained in two different ways – a) set of variables was divided randomly into two groups of equal size without replacement, so each variable was used at least once; b) ten variables were drawn randomly with replacement seven times.

2. The same clustering algorithm – DBSCAN for symbolic data – was applied with different initial parameters (minPts was set to $D + 1$; $D + 2$; $D + 3$ and $D + 4$, ε was selected each time by using k-distance graph) and a different distance measure for symbolic data was used.

`HINoV.SDA` function from `symbolicDA` package of R that is a modification of Carmone, Kara and Maxwell heuristic identification of noisy variables (HINoV) method for symbolic data (see [Walesiak et al. 2018; Carmone et al. 1999]) – see Figure 1.

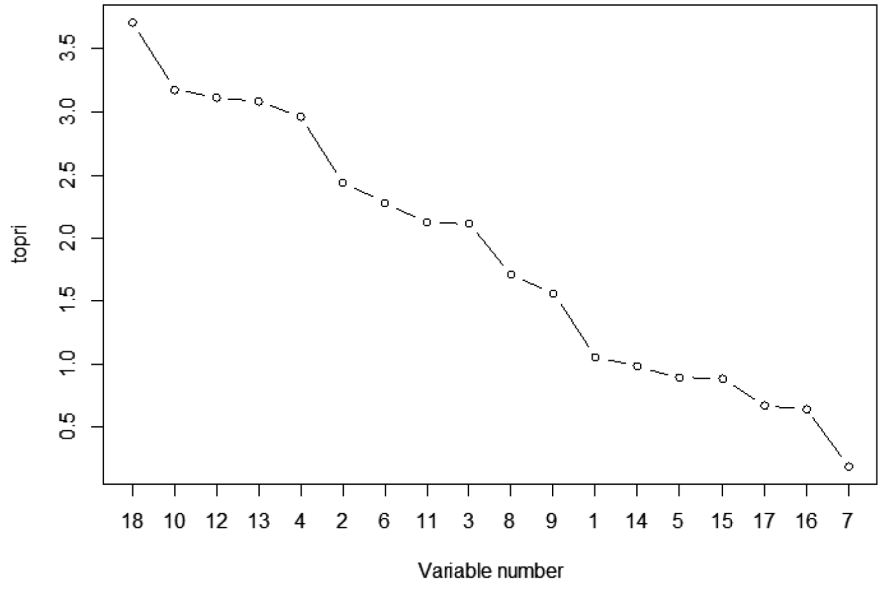

Fig. 1. Topri results for each variable

Source: own elaboration obtained with application of R software.

All the variables were used in further analysis. After the co-association matrix (built upon 504 base models) was obtained it was used as a data matrix in `cluster.Sim` function from `clusterSim` package. This function allows to determine the optimal clustering procedure for a data set by varying all combinations of normalization formulas, distance measures, and clustering methods (see [Walesiak, Dudek 2017]). In this case, Rousseeuw's Silhouette internal cluster quality index was used to find the optimal number of clusters.

In result four clusters were found (Rousseeuw's Silhouette internal cluster quality index was equal to 0.7983418). This result was obtained while using positional standardization, the generalized distance measure for metric data and the single hierarchical clustering method. The cluster validation index (adjusted Rand index equal to 0.766523) suggests that clusters can be seen as quite stable clustering results. In Table 4 the results of clustering are presented.

Cluster 1 contains the following countries: Belgium, Denmark, Germany, Ireland, France, the Netherlands, Austria, Slovenia, Finland, Sweden, the United Kingdom and Norway. These countries are high-innovative ones (innovation leaders). Almost all the variables' spans (ranges) are the largest ones (except the most-cited publications and employment in medium and high-tech manufacturing companies and knowledge-intensive services) for this cluster.

Table 4. Results of clustering

Clusters	**Countries**
Cluster 1	Belgium, Denmark, Germany, Ireland, France, the Netherlands, Austria, Slovenia, Finland, Sweden, the United Kingdom, Norway

Variables (lower and upper bounds of intervals)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
0.23	0.27	0.13	0.48	0.00	0.21	0.12	0.32	0.07	0.20	0.09	0.09	0.14	0.06	0.13	0.29	0.11	0.25
1.00	0.99	0.95	0.93	1.00	0.99	0.86	0.86	0.73	0.75	1.00	0.79	0.85	0.54	0.91	0.98	0.99	0.90

Cluster 2	Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia

Variables (lower and upper bounds of intervals)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
0.16	0.01	0.02	0.09	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.10	0.09	0.16	0.27	0.03
0.87	0.43	0.62	0.48	0.64	0.47	0.65	0.39	0.35	0.39	0.38	0.34	0.24	0.48	0.55	0.91	0.90	0.37

Cluster 3	The Czech Republic, Spain, Italy

Variables (lower and upper bounds of intervals)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
0.15	0.26	0.09	0.40	0.19	0.04	0.06	0.12	0.15	0.12	0.04	0.06	0.06	0.18	0.15	0.19	0.23	0.19
0.82	0.55	0.76	0.74	0.85	0.55	0.58	0.58	0.53	0.62	0.46	0.45	0.49	0.54	0.87	0.78	0.90	0.64

Cluster 4	Greece, Portugal

Variables (lower and upper bounds of intervals)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
0.27	0.03	0.05	0.20	0.22	0.00	0.20	0.34	0.27	0.24	0.04	0.03	0.05	0.15	0.00	0.16	0.00	0.08
0.73	0.54	0.52	0.86	0.73	0.33	0.62	0.68	0.60	0.72	0.66	0.26	0.15	0.52	0.61	0.60	0.56	0.54

Source: own elaboration based on R software.

Besides that, this cluster is quite similar to other clusters when considering the following variables: most-cited publications, product or process innovations, SMEs innovation in-house and trademark applications. This cluster is also the least homogenous.

Cluster 2 contains post-communist countries mainly from Central Europe: Bulgaria, Croatia, Hungary, Poland, Romania and Slovakia. They are rather low-moderate innovators in Europe. This cluster is quite similar to cluster 1 when considering the following variables: population with tertiary education and employment in medium and high-tech manufacturing companies and knowledge-intensive services (for which the variable's span is the highest). For all other variables their spans are not too large. These countries can be seen as rather mid-low innovative (they try to "catch up" with the innovation leaders).

Cluster 3 contains moderate innovators from Spain, Italy and the Czech Republic. Clusters 2 and 3 are very similar to each other. They have the lowest variable spans for lifelong learning, most-cited publications and trademark applications.

Cluster 4 contains two countries, Greece and Portugal. They are also moderate innovators – this cluster is the most homogenous. This cluster has the lowest variable spans for population with tertiary education, scientific co-publications, R&D expenditures by public sector, R&D expenditures by business sector, non-R&D innovation expenditures, product or process innovations, marketing or organization innovations, public-private co-publications, European Patent Office patent applications, employment in medium and high-tech manufacturing companies and knowledge-intensive services, exports from medium and high-tech manufacturing companies and the highest ones for most-cited publications.

7. Final remarks

Symbolic density based clustering allows to use well-known density-based clustering to the symbolic data case. The only limitation is the need of a suitable distance measure for symbolic variables.

Symbolic data ensemble based on density clustering for symbolic data can be done when using the co-occurrence matrix (density clustering is used to build the matrix). This co-occurrence matrix is used as a data matrix for some classical clustering method – e.g. pam, *k*-means, etc. This approach is known as the symbolic-numeric approach. However other ensemble clustering approaches can also be used – like adaptations of bagging that have been proposed by Hornik [2005], Leisch [1999] or Dudoit and Fridlyand [2003].

The symbolic clustering ensemble based on density-based clustering allowed to discover four different clusters. Cluster 1 contains the following countries: Belgium, Denmark, Germany, Ireland, France, the Netherlands, Austria, Slovenia, Finland, Sweden, the United Kingdom, and Norway. These countries are high innovative ones (innovation leaders). Cluster 2 contains post-communist countries mainly from Central Europe: Bulgaria, Croatia, Hungary, Poland, Romania, and Slovakia. They are rather low-moderate innovators in Europe. Cluster 3 contains moderate innovators from Spain, Italy and the Czech Republic. Clusters 2 and 3 are very similar to each other, and have the lowest variable spans for lifelong learning, most-cited publications and trademark applications.

Cluster 4 contains two countries, Greece and Portugal. They are also moderate innovators – this cluster is the most homogenous.

What is important when considering variables most-cited publications, product or process innovations, SMEs innovation in-house and trademark applications, is that all the four clusters are quite similar to each other. This suggests that the European Union should reconsider using these variables in following innovation studies as they do not provide too much discrimination power to the data set.

Bibliography

Ankrest M., Breunig M., Kriegel H.-P., Sander J., 1999, *OPTICS: Ordering Points to Identify the Clustering Structure*, ACM SIGMOD international conference on Management of data, pp. 49-60.

Baregheh A., Rowley J., Sambrook S., 2009, *Towards a multidisciplinary definition of innovation*, Management Decision, 47, pp. 1323-1339.

Bessant J., Lamming R., Noke H., Phillips W., 2005, *Managing innovation beyond the steady state*, Technovation, 25, pp. 1366-1376.

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

Bledow R., Frese M., Anderson N., Erez M., Farr J., 2009, *A dialectic perspective on innovation: Conflicting demands, multiple pathways, and ambidexterity*, Industrial and Organizational Psychology, 2, pp. 305-337.

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

Boer H., During W.E., 2001, *Innovation, what innovation? A comparison between product, process and organizational innovation*, International Journal of Technology Management, 22, pp. 83-109.

Borrás S., 2003, *The Innovation Policy of the European Union: From Government to Governance*. Edward Elgar Publishing, Massachusetts.

Carlson C.C., Wilmot W.W., 2006, *Innovation: The Five Disciplines for Creating what Customers Want*, Crown Business, New York.

Carmone F.J., Kara A., Maxwell S., 1999, *HINoV: a new method to improve market segment definition by identifying noisy variables*, Journal of Marketing Research, November, vol. 36, pp. 501-509.

Compello R., Moulavi D., Sander J., 2013, *Density-based Clustering based on Hierarchical Density Estimates*, Advances in Knowledge Discovery and Data Mining, pp. 160-172.

Damanpour F., 1991, *Organizational innovation: a meta-analysis of effects of determinants and moderators*, Academy of Management Journal, 34, pp. 555-590.

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

Dosi G., 1988, *The nature of the innovative process*, [in:] G. Dosi, C. Freeman, R. Nelson, G. Silverberg, L. Soete (eds.), *Technical Change and Economic Theory*, Pinter Publishers, London, NY, pp. 221-238.

Dudoit S., Fridlyand J., 2003, *Bagging to improve the accuracy of a clustering procedure*, Bioinformatics, vol. 19, no 9, pp. 1090-1099.

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, *A density-based algorithm for discovering clusters in large spatial databases with noise*, Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining, Portland, pp. 226-231.

European Commission, *Horizon 2020 in brief*, http://ec.europa.eu/programmes/horizon2020/sites/horizon2020/files/H2020_inBrief_EN_FinalBAT.pdf.

Fred A.L.N., Jain K., 2005, *Combining multiple clustering using evidence accumulation*, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 27, pp. 835-850.

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

Hahsler M., 2017, *Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*, www.r-project.org.

Harrison, R., Jaumandreu, J., Mairesse, J., Peters, B., 2014, *Does innovation stimulate employment? A firm-level analysis using comparable micro-data from four European countries*, International Journal of Industrial Organization, 35, pp. 29-43.

Hinneburg A., Keim D., 1998, *An efficient approach to clustering in large multimedia databases with noise*, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 58-65.

Hobday M., 2005, *Firm-level innovation models: Perspectives on research in developed and developing countries*, Technology Analysis & Strategic Management, 17, pp. 121-146.

Hollanders H., Es-Sadki N. (eds.), 2017a, *European Innovation Scoreboard*. *Methodology Report*, http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards.

Hollanders H., Es-Sadki N. (eds.), 2017b, *European Innovation Scoreboard*, http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards.

Hollanders H., Es-Sadki N., Kaberva M. (eds.), 2015, *Innovation Union Scoreboard*, http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards.

Hornik K., 2005, *A CLUE for CLUster ensembles*, Journal of Statistical Software, vol. 14, pp. 65-72.

Jahirabadkar S., Kulkarni P., 2013, *Clustering for high dimensional data: density based subspace clustering algorithms*, International Journal of Computer Applications, vol. 63, issue 20, pp. 29-35.

Kahn K.B., 2012, *The PDMA handbook of new product development*, John Wiley & Sons, Inc., Hoboken, NJ.

Kailing K., Kriegel H.-P., Kröger P., 2004, *Density-connected subspace clustering for high-dimensional data,* Proc. SIAM Int. Conf. on Data Mining, pp. 246-257.

Lafley A.G., Charan R., 2008, *The Game-Changer: How You Can Drive Revenue and Profit Growth with Innovation*, Crown Business, New York.

Leisch F., 1999, *Bagged Clustering, Adaptive Information Systems and Modelling in Economics and Management Science*, Working Papers 1999, SFB, p. 51.

Mikołajczyk B., 2013, *Innowacyjność przedsiębiorstw w krajach UE – pomiar i ocena*, Annales Universitatis Mariae Curie-Skłodowska, Lublin, vol. XLVII, 3, pp. 421-431.

Noirhomme-Fraiture M., Brito, P., 2011, *Far beyond the classical data models: symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, issue 2, pp. 157-170.

Nowak P., 2012, *Poziom innowacyjności polskiej gospodarki na tle krajów UE*, Prace Komisji Geografii Przemysłu, nr 19, pp. 153-168.

O'Sullivan D., Dooley L., 2009, *Applying Innovation*, SAGE Publications, Thousand Oakes, CA.

Pełka M., 2012, *Ensemble approach for clustering of interval-valued symbolic data*, Statistics in Transition, vol. 13, issue 2, pp. 335-342.

Pełka M., 2013, *Clustering of symbolic data with application of ensemble approach*, Acta Universitatis Lodziensis. Folia Oeconomica, vol. 285, pp. 89-95.

Pełka M., 2016, *A comparison study for spectral, ensemble and spectral mean-shift clustering approaches for interval-valued symbolic data*, Analysis of Large and Complex Data, pp. 137-146.

Radicic D., Pugh G., Hollanders H., Wintjes R., Fairburn J., 2016, *The impact of innovation support programs on small and medium enterprises innovation in traditional manufacturing industries: An evaluation for seven European Union regions*, Environment and Planning C: Government and Policy, 34(8), pp. 1425-1452.

Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013, establishing Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020) and repealing Decision No 1982/2006/EC, http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/fp/h2020-eu-establact_en.pdf.

Rendón E., Abundez I., Arizmendi A., Quiroz E., 2011, *Internal versus external cluster quality indexes*, International Journal of Computers and Communications, issue 1, vol. 5, 27-34.

Rynardowska-Kurzbauer J., 2015, *Innowacyjność wybranych krajów Europy Środkowo-Wschodniej*, Zeszyty Naukowe Politechniki Śląskiej, seria „Organizacja i Zarządzanie", nr 86, pp. 93-101.

Sander J., Ester M., Kriegel H.-P., Xu X., 1998, *Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications*, Data Mining and Knowledge Discovery, vol. 2, issue 2, pp. 169-194.

Sheikholeslami G., Chatterjee S., Zhang A., 1998, *Wavecluster: a multi-resolution clustering approach for very large spatial databases*, Proceedings of the 24th VLDB Conference, pp. 428-439.
Silverstein D., Samuel P., DeCarlo N., 2009, *The Innovator's Toolkit: 50+ Techniques for Predictable and Sustainable Organic Growth*, John Wiley & Sons, Hoboken, NJ.
Stec M., 2009, *Innowacyjność krajów Unii Europejskiej*, Gospodarka Narodowa, nr 11-12, pp. 45-65.
Trott P., 2012, *Innovation Management and New Product Development* (5th ed.), FT/Prentice Hall, Harlow, England.
Van de Ven A., Polley D.E., Garud R., Venkataraman S., 1999, *The Innovation Journey*, Oxford University Press, New York.
Walesiak M., Dudek A., 2017, *The clusterSim package for R software*, www.r-project.org.
Walesiak M., Dudek A., Pełka M., 2018, *The symbolicDA package for R software*, www.r-project.org.
Wojtas M., 2013, *Innowacyjność polskiej gospodarki na tle krajów Unii Europejskiej*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, nr 756, seria „Finanse, Rynki Finansowe, Ubezpieczenia", nr 57, pp. 605-617.

ANALIZA INNOWACYJNOŚCI KRAJÓW UNII EUROPEJSKIEJ Z ZASTOSOWANIEM WIELOMODELOWEJ KLASYFIKACJI GĘSTOŚCIOWEJ DANYCH SYMBOLICZNYCH

Streszczenie: Innowacje odgrywają istotną rolę w nowoczesnej gospodarce rynkowej, są kluczem do podnoszenia jakości życia, poprawy warunków pracy i rozwoju gospodarczego kraju. Polityka poprawy innowacyjności jest kluczowym elementem zarówno na poziomie polityk krajowych, jak i polityki całej Unii Europejskiej. Głównym celem artykułu było zaprezentowanie wielomodelowej klasyfikacji gęstościowej krajów Unii Europejskiej ze względu na poziom ich innowacyjności. W części empirycznej zastosowano klasyfikację gęstościową dla danych symbolicznych do budowy macierzy współwystąpień użytej jako macierz danych w klasyfikacji. W artykule wykorzystano pakiety i funkcje programu R pochodzące z pakietów `symbolicDA`, `clusterSim` oraz `dbscan`. Zastosowane podejście pozwoliło zidentyfikować strukturę czterech różnych klas. W klasie pierwszej znalazły się kraje o wysokim poziomie innowacyjności (liderzy innowacji), jest ona jednocześnie najmniej homogeniczną z klas. W klasie drugiej znalazły się kraje Europy Środkowo-Wschodniej, które starają się dorównać liderom innowacyjności. Kraje te można uznać za przeciętnie innowacyjne. W klasie trzeciej znalazły się kraje umiarkowanie innowacyjne, a w czwartej dwa kraje, które również należy uznać za przeciętnie innowacyjne.

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