Hierarchical clustering of EU countries based on HDI and EPI index

Radka Repiská[1], Nora Grisáková[1], and Peter Štetka[1]

[1] University of Economics in Bratislava, Faculty of Business Management, Department of Business Economy, Dolnozemská cesta 1/A, Bratislava, Slovakia

radka.repiska@euba.sk

https://doi.org/10.53465/EDAMBA.2021.9788022549301.420-430

Abstract.
The aim of this paper is to evaluate the global competitiveness regarding the environmental economics model, considering all three levels: economic, social, and environmental. We measure the socio-economic dimension using HDI (Human Development Index) according to the health and education areas, then we measure the environmental dimension using EPI (Environmental Performance Index), which monitors the behaviour of countries in the field of human health protection and ecosystem protection. This paper focuses on the possibility to group countries by the cluster method in terms of assessing the sustainable competitiveness of European countries. The question is whether there is an appropriate classification for the development of these countries that could help to reduce the differences between the average countries and the EU 27 average. The approach to this topic began with the question whether these countries, which have high values of economic growth, have a high level of EPI or HDI. The intention is to look for the possible existence of a gradual rapprochement of countries belonging to the same group.

Keywords: cluster analysis, HDI index, EPI index

JEL classification: *C 38, P 28, Q 56*

1 Introduction

The aim of this paper is to evaluate the global competitiveness regarding the environmental economics model, considering all three levels: economic, social, and environmental. We measure the socio-economic dimension using HDI according to the health and education areas, then we measure the environmental dimension using EPI, which monitors the behaviour of countries in the field of human health protection and

[1] Corresponding author: Radka Repiská radka.repiska@euba.sk

ecosystem protection. This paper focuses on the possibility to group countries by the cluster method in terms of assessing the sustainable competitiveness of European countries, especially Slovakia and the Netherlands. The question is whether there is an appropriate classification for the development of these countries that could help to reduce the differences between the average countries and the EU 27 average. The approach to this topic began with the question whether these countries, which have high values of economic growth, have a high level of EPI or HDI. The intention is to look for the possible existence of a gradual rapprochement of countries belonging to the same group.

In 1939, Robert Choate Tryon first used the term from noise analysis [1]. Cluster analysis is a classification procedure that groups objects into distinct subgroups that are similar within but different than objects included in other subgroups. The resulting branching diagram is a classification that provides a sequence of clusters (subgroups) according to which a group of objects is divided. For instance, if several ecological units are examined, this analysis is suitable for showing species composition patterns between these units. Cluster analysis essentially creates a dendrogram or tree, the branches of which represent each of the ecological units, and the data on the species composition of these places determine the structure of the branch. Merged branches represent groups or clusters of sites with a similar species composition and the length of a branch before merging is inversely proportional to the degree of similarity of the species composition.

There is a wide range of cluster analyses, we focused on hierarchical, agglomerative, where each object is considered a cluster. The choice of an appropriate method is crucial because it determines (partially) a classification derived from species composition data. Like many multidimensional statistical analyses, cluster analysis attempts to represent complex relationships between objects, in our case between countries, in a simple one-dimensional way. We processed the application of cluster analysis using a comparison of 3 classifications on a set of 15 EU countries. The status of all acquired variables reflects the observed period of the most recently obtained data at the end of 2018, which represents the full coverage of the variables HDI (Human Development Index) and EPI (Environmental Performance Index) for all monitored countries.

2 Methods and methodology

Cluster analysis of a multidimensional data set aims to divide a large set of data into meaningful subgroups of subjects. In cluster analysis, many methods are available to classify objects based on their (un) similarity [2]. Dasgupta [3] framed similarity-based hierarchical clustering as a combinatorial optimization problem, where a “good” hierarchical clustering is one that minimizes a particular cost function. Murlag and Contreas [4] made a survey of agglomerative hierarchical clustering algorithms and discussed efficient implementations that are available in R and other software environments. They look at hierarchical self-organizing maps, and mixture classifications reviewed grid-based clustering, focusing on hierarchical density-based

approaches. Jafarzadegan at all proposes a novel method of combining hierarchical clustering approaches based on principle component analysis (PCA). PCA as an aggregator allows considering all elements of the descriptor matrices. In their approach, basic clusters were made and transformed to descriptor matrices. Then, a final matrix was extracted from the descriptor matrices using PCA and dendrogram were constructed from the matrix that was used to summarize the results of the diverse clustering [5].

We expand the data matrix X of *pxk* type with *p* objects and k indicators into the set C by means of clustering procedures with all clusters *m*, where the objects of the primary matrix X were grouped. The total number of clusters *m* has the possibility to range from 1 to *p*, while the best situation occurs when we reach the number of clusters smaller than the number of objects (in our case the studied countries) [6].

From the most well-known metrics of distances between objects, we chose the *Euclidean distance of objects* for our analysis, which is set by the following equation [6]:

$$d\left(X_i, X_J\right) = \sqrt{\sum_{s=1}^{k}(x_{is} - x_{js})^2} \quad (1)$$

Where:

x_{is} is the value of the s-th variable for the i-th object.

x_{is} is the value of the s-th variable for the j-th object.

This distance measurement, which generalises the concept of physical distance in two- or three-dimensional space to multidimensional space, is often referred to as the "Pythagorean distance" and forms the basis for Ward's method.

The main types of analysis are hierarchical clustering procedures, which are divided into:

- **agglomerative -** the decomposition process begins with each cluster that contains exactly one object and continues the decomposition by a suitably selected method until all of them are merged into one cluster;
- **divisive** - the opposite procedure begins with one cluster containing all objects and gradually splits into smaller clusters [7].

Next, we will deal with hierarchical clustering procedures, where there are several different methods used to determine which clusters should be combined at each stage, *Nearest-neighbour clustering method, Median method* and *Ward's method* were chosen to collect minimised heterogeneity clusters.

The median method is described by the following two equations [7]

1. Nearest-neighbour clustering method (*"Nearest"*)[2]

[2] The nearest neighbour method uses the distance of the nearest cluster elements C_h and C_r

$$D_1(C_h, C_r) = min\left\{d(X_i, X_j)\right\}$$
$$X_i \in C_h, X_j \in C_r \quad (2)$$

2. Median method (*"Median"*)[3]

$$D_2(C_h, C_r) = d\left(\overline{X_h}, \overline{X_r}\right) \quad where \quad \overline{X_h} = \frac{1}{n_h}\sum_{Xi \in C_h} Xi \;\; , \;\; \overline{X_r} = \frac{1}{n_r}\sum_{Xj \in C_r} Xj \quad (3)$$

Ward's method is a correct hierarchical procedure and makes it possible to determine how many groupings should be considered, and its great advantage is the tendency to remove small clusters and form clusters of roughly the same size. The similarity between 2 clusters is the sum of the squares in the clusters summarised in all variables, the proximity between the 2 clusters being defined as the increase in the square root error resulting from the merging of 2 clusters [8]. In the case of the Ward's method in terms of distance, equation 4 can be formulated in the form of the product of the Euclidean distance of objects between the centre of clusters conditioned to join and the coefficient, based on the size of the cluster[4] [9]:

$$D(C_h, C_r) = \frac{n_h n_r}{n_h + n_r} \times d^2\left(\overline{X_h}, \overline{X_r}\right) \quad (4)$$

The results of hierarchical clustering can be displayed graphically using a tree diagram - "dendrogram", which shows all the steps in a hierarchical process, including distances, where clusters combine.

3 International Sustainability Indices

In this part of the paper, we come to specific variables, sustainability indices. In the case of the HDI index, we used 3 main dimensions and related indicators within the EU countries. We proceeded in a similar way in the case of the EPI index, where we evaluated countries in 24 performance indicators in ten categories of problems related to Environmental Health and Ecosystem Vitality.

[3] The Median clustering method uses the distance between the centroids of the clusters and serves as an improvement to *the Centroid method*

[4] In hierarchical grouping, the sum of squares starts from zero (each point is in its own grouping) and then increases as we merge the clusters. Ward's method keeps this growth as small as possible. Considering two pairs of clusters whose centres are equidistant from each other, the method prefers to merge the smaller ones.

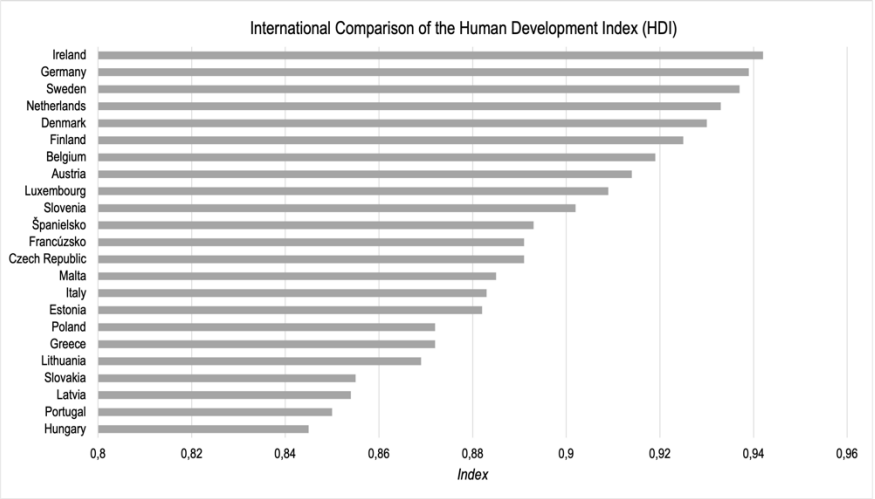

Fig. 1. International Comparison of the Human Development Index (HDI) [5]

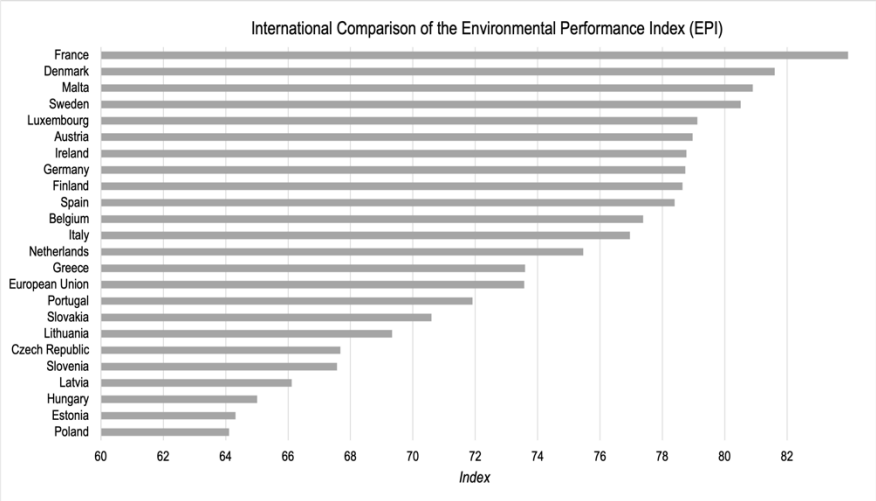

Fig. 2. International Comparison of the Environmental Performance Index [6]

[5] Source: Own processing according to UNDP (United Nations Development Programme: Human Development Index (HDI). Dimension: Composite indices.

[6] Source: Own processing according to Yale Center for Environmental Law & Policy. Center for International Earth Science Information Network. Earth Institute. Columbia University. 2020.

4 Hierarchical clustering procedures

The last presented analysis is a comparison of 3 classifications of cluster analysis on a set of 15 countries of the European Union. Our 2 examined variables were: Human Development Index (HDI) and Environmental Performance Index (EPI) as aggregated indicators, which we described in more detail in the introduction in the first chapter, from a methodological point of view in the third chapter and their application in the last chapter Results.

The characteristics of the raw data was considered in the selection of appropriate hierarchical clustering procedures. In the cluster analysis of our data, we used the statistical software SAS Enterprise Guide 4.2[7], which forms hierarchical clusters of observations containing the coordinates of the data, but also their distances. If the data set contains coordinates, the cluster analysis calculates the Euclidean distance of the objects before the clustering method is applied. The result of hierarchical agglomerative clustering is a graph displayed as a tree diagram - a "dendrogram", which can be displayed in the SAS system in 2 ways, vertically or horizontally. The main use of the dendrogram is to find the best way to assign objects to clusters, and the key to interpretation is to focus on the height at which the two different objects are connected.

4.1 Nearest Neighbor Method

To compare the first cluster analysis classification, we used the *Nearest Neighbour Method* as the first of the hierarchical clustering methods. The principle of the nearest neighbour method is that the algorithm uses a minimum distance to measure the distance between clusters and 2 objects placed in a cluster are separated from each other by the shortest possible distance, gradually adding more clusters to the original objects by creating the 3rd nearest neighbour. After processing the classification using SAS, we constructed a dendrogram.

Table 1. Clusters according to the nearest neighbour method[8]

CLUSTERS	EU COUNTRIES
1.	Sweden
2.	Hungary, Slovakia, Greece, Italy, Netherlands, Belgium, Luxembourg, Austria, Czech Republic, Finland, Germany, Ireland, France, Denmark

[7] Available on the SAS software website: <https://www.sas.com/sk_sk/trials/software/covid19/form.htm>

[8] Source: Own processing according to data obtained from HDI and EPI index variables

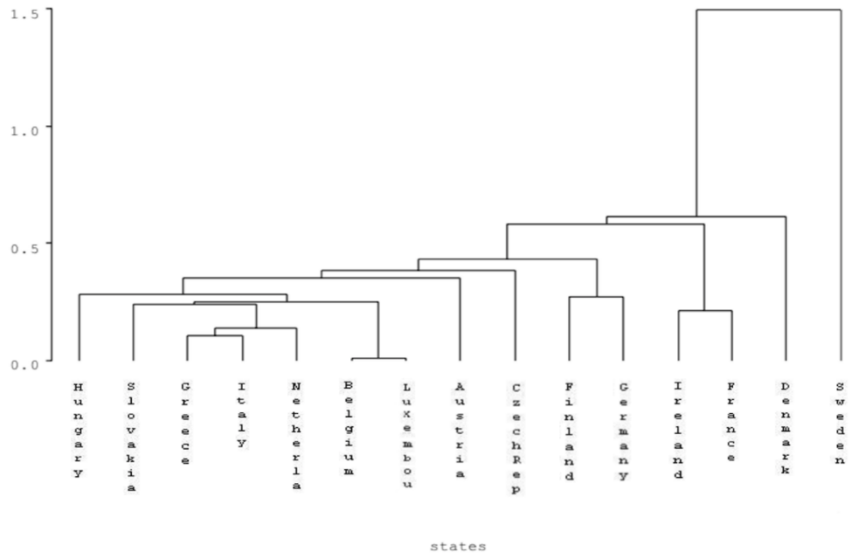

Fig. 3. Cluster created according to the nearest neighbour method[9]

According to the constructed dendrogram (Fig. 2) and from table 1 it follows visually and analytically that we divided the set of 15 countries into 2 clusters. If we take a closer look at the formed clusters, we can state that cluster 2, as a larger group, contains the predominance of 14 developed countries of the European Union. Countries such as Denmark, France, the Netherlands, Italy, Greece, Slovakia, and Hungary used the dendrogram to show a similar level of HDI and EPI indices. Cluster 1 is made up of only one EU country, Sweden, as significantly more advanced in terms of obtaining higher values of HDI and EPI indices.

4.2 Median Method

As the second method of cluster analysis for the comparison of European countries, we chose *the Median method*, which serves as a certain upgrade of the *Centroid method*. We have described the detailed principle of these methods in more detail in the previous chapters. The centroid method uses the distance between the centre of gravity of two clusters to evaluate the overall solution of the cluster, with the centre of gravity representing the centroid of a particular cluster. The distance between two clusters is calculated as the difference between the centres of gravity. The median method is based on the median, which follows from the name itself, and instead of calculating the average for each cluster to determine its centre of gravity, it calculates the mean distance between all pairs of observations or individuals in the clusters. After the data for this classification were processed, we built a dendrogram using SAS software.

[9] Source: Own processing according to data obtained from HDI and EPI index variables

Table 2. Clusters according to the median method[10]

CLUSTERS	EU COUNTRIES
1.	Sweden
2.	Hungary, Slovakia, Greece, Italy, Netherlands, Czech Republic, Ireland, France, Denmark, Austria, Belgium, Luxembourg, Finland, Germany

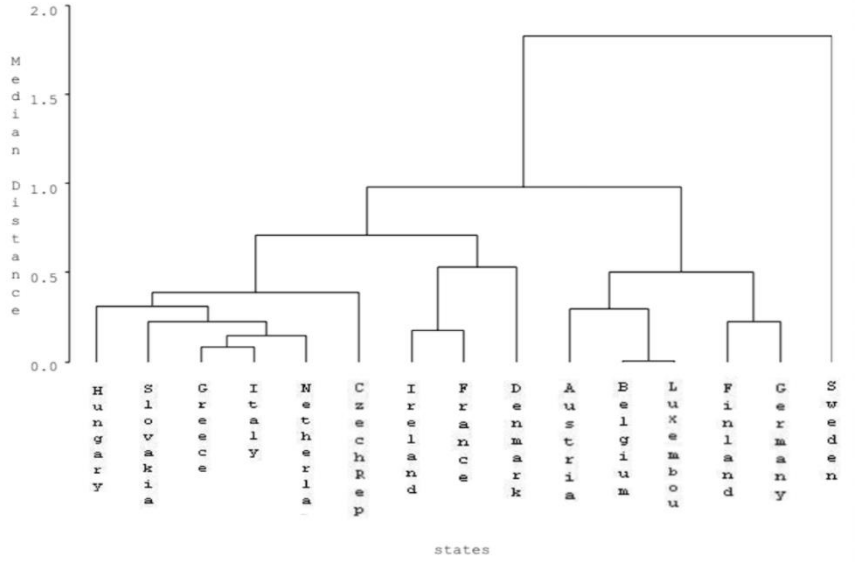

Fig. 4. Cluster according to the median method [11]

According to Table 2 and the dendrogram (Figure 3), we can observe a very similar situation as with the nearest neighbour method. We redistributed 15 countries into 2 main clusters. Cluster 2 contains again a set of 14 EU countries, whose monitored data of HDI and EPI indices are relatively similar. While Sweden belongs again to the 1st cluster and shows its strength over other countries, especially within the HDI and particularly in the dimension index called the "*Education index*".

4.3 Ward's Method

As a final analysis, we present the most used method in marketing called *the Ward's Minimum Variance method*. Ward's method creates clusters that minimise variance in each cluster. For each cluster, the average for each variable is calculated and, in each cluster, the observations are compared to the average for each variable. The observations or clusters are combined in a way that the variance in the resulting cluster

[10] Source: Own processing according to data obtained from HDI and EPI index variables
[11] Source: Own processing according to data obtained from HDI and EPI index variables

of solutions is minimised as much as possible. Following the summary of the data of our analysis, we prepared a table and constructed a dendrogram using SAS software.

Table 3. Clusters according to the Ward's Minimum Variance method[12]

CLUSTERS	EU COUNTRIES
1.	Hungary, Slovakia, Czech Republic, Greece, Italy, the Netherlands, Austria, Belgium, Luxembourg, Finland, Germany
2.	Ireland, France, Denmark, Sweden

The illustrated dendrogram (Figure 4) illustrates the situation of 2 constructed clusters of countries, which can be very nicely distinguished from the cluster formed by Ward's Minimum Variance method. On the right side of the dendrogram we see cluster 2, which connects the 4 strongest countries in northern Europe. They are the world's richest economies with even income distribution, low unemployment, and highly developed institutionalisation, in terms of human data development index (HDI) and environmental performance index (EPI), what evokes a high level of standard in countries. From the opposite left side of the dendrogram, we can observe developed countries connected by one cluster with relatively similar values of the HDI and EPI indices. Although more significant differences can be seen mainly in countries such as Hungary (left side of the dendrogram) and Germany (closer to the Nordic countries of the dendrogram), where the differences are obvious and Hungary is trying to catch up, but it is not enough yet. Table 3 also clearly shows 2 clusters with a division of countries according to the achieved values of HDI and EPI indices.

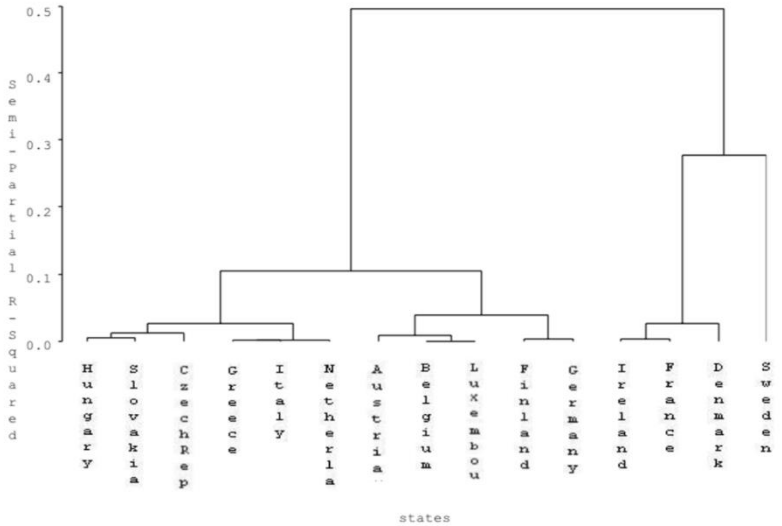

Fig. 5. Cluster according to the Ward's Minimum Variance method [13]

[12] Source: Own processing according to data obtained from HDI and EPI index variables
[13] Source: Own processing according to data obtained from HDI and EPI index variables

When it comes to cluster properties, it is important to look at the values that countries indicate for the two indicators used for the analysis. In the case of cluster 1, things are clear: we have an economically strong country that seem to be operating under the control of the objectives of the European strategy and the appropriate values for an important environmental factor. In the case of clusters of the 2nd degree, we can observe interesting situations with all 3 analysed methods. In the case of Slovakia, we can see in the first nearest neighbour method how it reworked for the 2nd lowest position, which analyses that of all the countries studied, together with Hungary and Greece, it has the highest average of HDI and EPI indices. On the other hand, the Netherlands is approaching the average values of the indices to Belgium and Luxembourg. As defining features for the country in this grouping, we can say that they have an average employment rate between 70% and 81.1% (except Greece - 64% and Poland 64.60%) as well as high values for greenhouse gas emissions above 102 compared to 1990. These countries are the ones that need to make sustainable efforts to become knowledge-based economies. In the analysis of the median method, we get similar results as in the case of the first method, but the fundamental difference is the distance used between the centre of gravity of the two clusters to evaluate the overall solution of the cluster. However, it is more interesting in the last Ward's method, where the strongest EU countries (Sweden, Denmark, France, and Ireland) separated into a second cluster. The countries in cluster 1 seem to have interesting characteristics: greenhouse gas emissions are less than 71, compared to 1990 at 100, except for Belgium (92) and Sweden (91), and compared to cluster 2, the countries have a higher average of people at risk of poverty and lower average of primary consumption.

Cluster analysis is an important tool for any study to identify possible intentions for convergence in living standards, education, GDP growth, life expectancy and environmental protection to measure overall progress in environmental sustainability.

One of Britain's professors of environmental economics, Paul Ekins, suggested in 2011 that there was a link between environmental performance and measures to improve environmental sustainability. Ideally, these measures would include [10]:

1) development of better measurement and monitoring systems to improve the collection of environmental data, the so-called environmental data;
2) development of environmental policies focused on extremely weak areas;
3) communication of data and statistics at national level to international agencies such as the United Nations (UN);
4) the definition of sub-national metrics and targets for the improvement of environmental performance.

5 Conclusion

In this paper we made a comparison of 3 classifications of cluster analysis on a set of 15 EU countries using 2 examined variables of human development and environmental performance indices as aggregate indicators. During our

multidimensional statistical classification, clusters were designed based on the HDI, EPI indices to evaluate the sustainable performance of EU members, as well as possible convergences between them at EU Member State level. The indicators used in the analysis form different groupings and most of the overlapping occurs in the groupings whose countries came first. This type of behaviour is typical of countries with strong economies, which record performance at all three socio-economic and environmental levels and pursue consistent development policies. Sweden and Denmark are the countries that appear in the first grouping in all analysed cases. Among the EU countries, Sweden appears most often in the leading grouping in all 3 analysed cases. The Czech Republic and Slovakia are ranked the best among the former communist countries and Luxembourg, Belgium, and the Netherlands as the third among the "Benelux" countries.

This paper is outcome of project solution VEGA 1/0646/20 „Diffusion and consequences of green innovations in imperfect competition markets“

References

1. Dasgupta, S., "A cost function for similarity-based hierarchical clustering," Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC'16), pp. 118-127, 2016.
2. Ekins, P., Anandarajah, G., Strachan, N., "Towards a Low-Carbon Economy: Pathways and Policy Requirements," Climate Policy, no. 11, pp. 865-882, 2011.
3. Han, J., Kamber, M., Pei, J., Data Mining: Concepts and Techniques, Waltham: Morgan Kaufmann, 2012, p. 444.
4. Jafarzadegan, J., Safi-Esfahani, F., Beheshti, Z., "Combining hierarchical clustering approaches using the PCA method," Expert Systems with Applications, no. 137, pp. 1-10, 2019.
5. Johnson, S. C., "Hierarchical clustering schemes," Psychometrika, vol. 3, no. 32, pp. 241-254, 1967.
6. Legendre, P., Legendre, L., Numerical Ecology, Elsvier, 2012, p. 990.
7. Murtagh, F., Contreras, P., "Algorithms for hierarchical clustering: an overview," Wiley Interdisciplinary Reviews, vol. 1, no. 2, pp. 86-97, 2011.
8. R. C. Tryon, Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality, Michigan: Edwards brother Inc., 1939, p. 122.
9. Řezanková, H., Húsek, D, Snášel, V., Shluková analýza dat, Pruhonice: Professional Publishing, 2009, p. 156.
10. Vark van, G. N., Howells, W. W., Multivariate Statistical Methods in Physical Anthropology: A Review of Recent Advances and Current Developments, Heidelberg: Springer, 1984, p. 444.