DIFFERENT APPROACHES TO THE SILHOUETTE COEFFICIENT CALCULATION IN CLUSTER EVALUATION

HANA ŘEZANKOVÁ

University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill sq. 4, Prague, Czech Republic
email: hana.rezankova@vse.cz

Abstract

Cluster analysis is a useful statistical tool for data exploration. It can help to identify groups of similar objects (e.g. countries) according to selected variables (e.g. economic indicators). The created groups (clusters) can be characterized based either on the variables used in clustering or on some other variables. The problem with using the methods of cluster analysis is that the analyst can obtain different results (assignments of objects into clusters) by different methods, and moreover, he needs to determine the number of clusters. Many coefficients for solving this problem have been proposed until now. However, the suggestion obtained by a certain coefficient can differ from another suggestion obtained by another coefficient. In addition, values of a certain coefficient can differ depending on the implementation in the software product. In this contribution, different approaches to the silhouette coefficient calculation are discussed. They concern the implementations of calculations in the R language and IBM SPSS Statistics system. It is analyzed how the different results influence decisions of an analyst in terms of both the choice of the suitable assignment of objects into clusters obtained by different algorithms, and the determination of the suitable number of clusters. The studied problems are illustrated using selected methods of cluster analysis applied to the EU countries characterized by the gender indicators.

***Keywords**: cluster analysis, number of clusters, silhouette coefficient*

***JEL Codes**: C38, C63, C88*

1. Introduction

The important part of cluster analysis is comparing results (i.e. the assignments of objects into clusters) obtained by different methods and determining the suitable number of clusters. Each object is represented by a vector of the values of input variables. Many coefficients have been proposed for these purposes (Gan *et al.*, 2007). Moreover, other special approaches were described in the literature, e.g. (Erilli *et al.*, 2011; Mur *et al.*, 2016). One of the often applied indices is the silhouette coefficient. Despite the fact that the formula for its calculation is well-known, there are differences in computations implemented in statistical software packages.

The aim of this paper is to discover why values of the silhouette coefficient differ for the same clustering results using two software systems with a long tradition. The R language (R Core Team, 2018), which is based on the S language (as well as the S-PLUS system), and IBM SPSS Statistics (version 24) are compared.

The Euclidean distance between objects both for clustering and for the result evaluation will be considered. However, other distances can be applied as well. The relationship between the selected distance measure and the selected evaluation criterion is investigated e.g. by Löster (2016).

2. Clustering Methods and Evaluating Coefficients

In this paper there will be considered one method of agglomerative hierarchical clustering (for the reason to illustrate a development of the values of the silhouette coefficient) and two partitioning methods. The complete linkage algorithm, the partitioning around medoids algorithm and the FANNY algorithm for fuzzy analysis are applied. All three methods provide the results for hard clustering (closest to the hard clustering in the case of the FANNY method) for the defined number of clusters. All the methods can be applied in the R language. They are included in the "cluster" package (Maechler *et al.*, 2017), which is based on the algorithms proposed for the S-PLUS system in (Kaufman and Rousseeuw, 2005).

In this section, three approaches to the silhouette coefficient calculation will be explained: the original approach implemented in the R language in the "cluster" package (the *silhouette()* function), the calculation available by the STATS CLUS SIL command in IBM SPSS Statistics (the installation of Python is required), and the calculation according to the IBM SPSS documentation (IBM, 2016) for the cluster evaluation algorithms (Goodness Measures).

2.1 Cluster Analysis Methods Used for Object Grouping

The *complete linkage algorithm* (further *COMPLETE*) is a method of hierarchical cluster analysis based on the proximity matrix with values D_{ij}, which evaluate relationships between the *i*-th and *j*-th objects. At the beginning of the clustering, each object is an individual cluster. In the first step, the two most similar clusters are joined. Then, the proximity matrix is recalculated and the next two most similar clusters are joined. In the complete linkage algorithm, the distance between clusters C_g and C_h is defined as the maximal distance of objects $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\mathbf{x}_i$ is an element of C_g and $\mathbf{x}_j$ is an element of C_h, i.e.

$$D_{\text{COMPLETE}}(C_g, C_h) = \max_{\mathbf{x}_i \in C_g;\, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j). \tag{1}$$

The process of clustering finishes when one cluster containing all objects is created.

The *partitioning around medoids* (*PAM*) *algorithm* (the *k*-medoids method) is based on the principle that in each created cluster a representative (a medoid) is determined – it is an object from the cluster for which the sum of the distances from all other objects of this cluster is minimal. It means that the function

$$f_{\text{PAM}} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{ih} \left\| \mathbf{x}_i - \mathbf{m}_h \right\| \tag{2}$$

is minimized, where $\mathbf{x}_i$ is the *i*-th object, $\mathbf{m}_h$ is a medoid of the *h*-th cluster, symbols $\| \, \|$ denote the Euclidean distance and u_{ih} is either 0 or 1 (the *i*-th object is an element of the *h*-th cluster).

The third method used for the problem illustration is the *fuzzy analysis* (*FANNY*) *algorithm* for fuzzy clustering. The memberships degrees u_{ih} for the *i*-th object and the *h*-th cluster and u_{jh} for the *j*-th object and the *h*-th cluster are defined by minimizing the function

$$f_{\text{FANNY}} = \sum_{h=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ih}^2 u_{jh}^2 D_{ij}}{2 \sum_{j=1}^{n} u_{jh}^2}, \tag{3}$$

where D_{ij} is the Euclidean distance between the the *i*-th and *j*-th objects.

2.2 Approaches to Calculation of the Silhouette Coefficient

For the calculation of the silhouette coefficient, for each *i*-th object, which is an element of the cluster C_g, there is computed the value (the width of the rectangle in the silhouette plot)

$$SW_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \tag{4}$$

where according to the *original approach* (Rousseeuw, 1987) implemented in the *R language*

$$a_i = \frac{\sum_{j \in C_g (j \neq i)} D_{ij}}{n_g - 1}, \qquad b_i = \min_{h \neq g} \left(\frac{\sum_{j \in C_h} D_{ij}}{n_h} \right), \tag{5}$$

n_g (n_h) is the number of objects in the *g*-th (*h*-th) cluster. If the *i*-th object is alone in a certain cluster, the value of a_i cannot be computed and SW_i is zero in this case. The silhouette coefficient is the average of values SW_i, i.e.

$$SC = \frac{\sum_{i=1}^{n} SW_i}{n}. \tag{6}$$

The silhouette coefficient takes on values from the interval ⟨−1, 1⟩. The higher value means the better assignment of objects into clusters. The suitable number of clusters can be determined on the basis of the highest value (usually within a certain interval of the numbers).

The values SW_i obtained by the *Cluster Silhouette procedure* in the *IBM SPSS Statistics* software (the *STATS CLUS SIL command* implemented in Python[1,2]) correspond to the calculation in which

$$a_i = \frac{\sum_{j \in C_g} D_{ij}}{n_g}, \tag{7}$$

including the distance D_{ii}. This type of the calculation is not described in the literature and it was discovered in an experimental way by the author of this contribution. If the *i*-th object is alone in a certain cluster, the value of a_i is 0 and SW_i is 1, because according to Eq. (4)

$$SW_i = \frac{b_i - 0}{\max\{0, b_i\}} = 1.$$

The values SW_i which are obtained according to the *IBM SPSS documentation* (IBM, 2016) are based on the values a_i and b_i computed as (for the reason that the original coefficient is computationally expensive, especially within TwoStep Cluster procedure for large data sets)

[1] IBM® SPSS® Statistics – Essentials for Python, which is installed by default with the IBM SPSS Statistics product, includes a set of extension commands that are implemented in Python and that provide capabilities beyond what is available with built-in SPSS Statistics procedures.

[2] IBM. Python Reference Guide for IBM SPSS Statistics, p. 260. [cit. 11-06-2018] ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/Python_Reference_Guide_for_IBM_SPSS_Statistics.pdf

$$a_i = D(\mathbf{x}_i, \mathbf{c}_g),\ b_i = \min_{h \neq g} D(\mathbf{x}_i, \mathbf{c}_h), \tag{8}$$

where $\mathbf{c}_g$ ($\mathbf{c}_h$) is the centroids (a vector of the average values of input variables) of the g-th (h-th) cluster. If the i-th object is alone in a certain cluster, SW_i is 1 as in the previous case.

2.3 Illustration of Differences in Calculation of the Silhouette Coefficient

Let us have four objects characterized by two variables (X and Y) with values included in Table 1. In this table there are also displayed the Euclidean distances for all pairs of the object. Let us suppose that we can obtain either two clusters (the first one with objects 1–3 and the second one with object 4) or three clusters (objects 1 and 2, object 3 and object 4).

Table 1: Input values and Euclidean distances for four objects

	Input values		Euclidean distances			
Object	Variable X	Variable Y	Object 1	Object 2	Object 3	Object 4
1	10	20	0	3.162	8.246	18.682
2	11	23	3.162	0	5.099	18.788
3	12	28	8.246	5.099	0	20.616
4	28	15	18.682	18.788	20.616	0

Source: Own calculation.

In the *R language*, the calculation of the silhouette coefficient for two clusters is performed according to Eqs. (5), (4) and (6) as follows:

$$a_1 = \frac{D_{12} + D_{13}}{3-1} = \frac{3.162 + 8.246}{2} = 5.704,\ b_1 = D_{14} = 18.682,$$

$$SW_1 = \frac{18.682 - 5.704}{18.682} = 0.695,$$

$$a_2 = \frac{D_{21} + D_{23}}{3-1} = \frac{3.162 + 5.099}{2} = 4.131,\ b_2 = D_{24} = 18.788,$$

$$SW_2 = \frac{18.788 - 4.131}{18.788} = 0.780,$$

$$a_3 = \frac{D_{31} + D_{32}}{3-1} = \frac{8.246 + 5.099}{2} = 6.673,\ b_3 = D_{34} = 20.616,$$

$$SW_3 = \frac{20.616 - 6.673}{20.616} = 0.676,$$

$$SW_4 = 0,$$

$$SC = \frac{0.695 + 0.78 + 0.676}{4} = 0.538.$$

With the *STATS CLUS SIL command* in *IBM SPSS Statistics*, the average of the SW_i values is different, because the a_i values are computed according to Eq. (7), and thus

$$a_1 = \frac{3.162 + 8.246}{3} = 3.803, \qquad b_1 = 18.682, \qquad SW_1 = \frac{18.682 - 3.803}{18.682} = 0.796,$$

$$a_2 = \frac{3.162 + 5.099}{3} = 2.754, \qquad b_2 = 18.788, \qquad SW_2 = \frac{18.788 - 2.754}{18.788} = 0.853,$$

$$a_3 = \frac{8.246 + 5.099}{3} = 4.448, \qquad b_3 = 20.616, \qquad SW_3 = \frac{20.616 - 4.448}{20.616} = 0.784,$$

$$a_4 = 0 \qquad b_4 = \frac{18.682 + 18.788 + 20.616}{3} = 19.362 \qquad SW_4 = \frac{19.362 - 0}{19.362} = 1,$$

$$SC = \frac{0.796 + 0.853 + 0.784 + 1}{4} = 0.859.$$

According to the cluster evaluation algorithms (*Goodness Measures*) description in *IBM SPSS Statistics*, the centroids of the individual clusters are computed first:

$$\bar{x}_1 = \frac{10 + 11 + 12}{3} = 11, \qquad \bar{x}_2 = 28,$$

$$\bar{y}_1 = \frac{20 + 23 + 28}{3} = 23.667, \qquad \bar{y}_2 = 15.$$

Then, according to Eqs. (8), (4) and (6),

$$a_1 = \sqrt{(10 - 11)^2 + (20 - 23.667)^2} = 3.801, \qquad b_1 = 18.682,$$

$$SW_1 = \frac{18.682 - 3.801}{18.682} = 0.797,$$

$$a_2 = \sqrt{(11 - 11)^2 + (23 - 23.667)^2} = 0.667, \qquad b_2 = 18.788,$$

$$SW_2 = \frac{18.788 - 0.667}{18.788} = 0.965,$$

$$a_3 = \sqrt{(12 - 11)^2 + (28 - 23.667)^2} = 4.447, \qquad b_3 = 20.616,$$

$$SW_3 = \frac{20.616 - 4.447}{20.616} = 0.784,$$

$$a_4 = 0 \qquad b_4 = \sqrt{(28 - 11)^2 + (15 - 23.667)^2} = 19.082 \qquad SW_4 = \frac{19.082 - 0}{19.082} = 1,$$

$$SC = \frac{0.797 + 0.965 + 0.784 + 1}{4} = 0.886.$$

If we consider three clusters, in the R language we obtain $SC = 0.249$ and in the IBM SPSS Statistics system we obtain $SC = 0.875$ according to the both ways of calculation (the values differ since the fourth decimal place).

3. Application to the Real Data Set

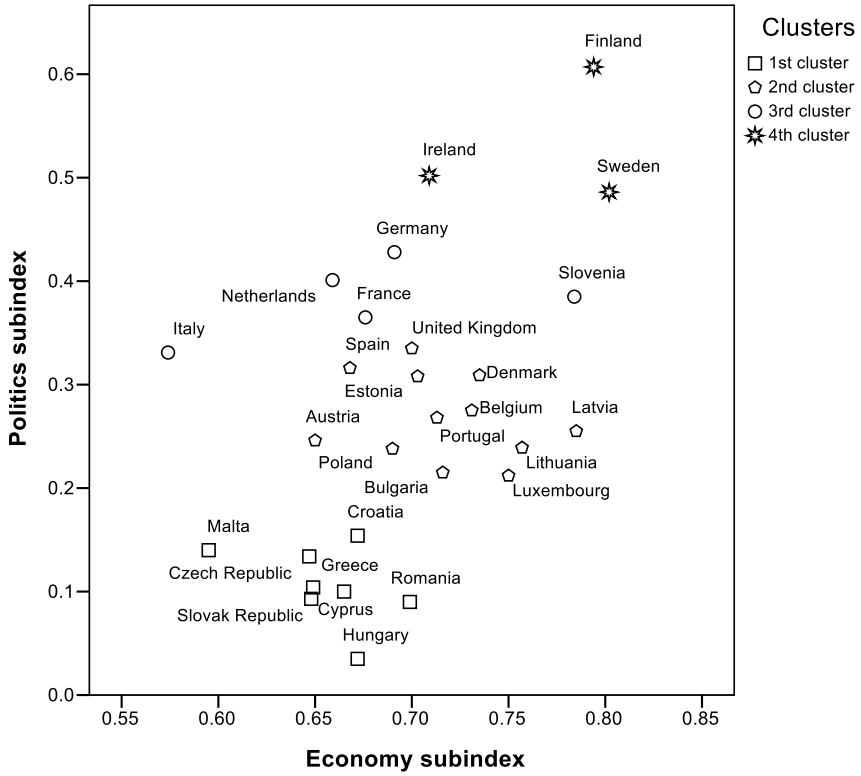
In this section the analyzed data will be described first. Then the values of the silhouette coefficient calculated in different ways will be compared and discussed using the results obtained by three methods of cluster analysis. Because of the second and the third approaches of the silhouette coefficient calculation mentioned in the previous section are similar (the average of distances is replaced by the distance from average values), only two approaches will be compared – the first (original) approach applied in the R language (the R approach) and the

second (modified) approach applied in the Cluster Silhouette procedure in IBM SPSS Statistics (the IBM SPSS approach).

3.1 Analyzed Data

For the problem illustration, two indicators described in detail in (Řezanková and Křečková Kroupová, 2017) were included into the analysis. There are two of the four subindices of the Global Gender Gap Index published for the EU countries in 2016, i.e. subindex *Economic Participation and Opportunity* (further *economy*) and subindex *Political Empowerment* (further *politics*). The data come from World Economic Forum (2016). The aim of the analysis is to cluster the EU countries according to these two indicators. Relationships among the countries are graphically displayed in Figure 1. One possible result of cluster analysis (the assignment of the countries into four clusters) obtained by the complete linkage method with the Euclidean distance is shown in this figure.

Figure 1: Scatter plot displaying a relationship between economy and politics subindices



Source: Own analysis based on World Economic Forum (2016).

3.2 Experiments and Results

For creating clusters of countries, the complete linkage method of hierarchical cluster analysis was applied first. Seven different assignments of countries into 2–8 clusters were obtained, see Table 2. In this table, the label (number) of a certain cluster does not mean any order. The labels of clusters in case of the assignment of countries into four clusters do not correspond with the labels in Figure 1, in which the higher number denotes the higher values of the politics subindex. In Table 2, splitting of clusters is demonstrated. For example, for obtaining 3 clusters, the first cluster obtained for the case of two clusters (with 25 countries) is divided into two clusters with 17 and 8 countries, etc.

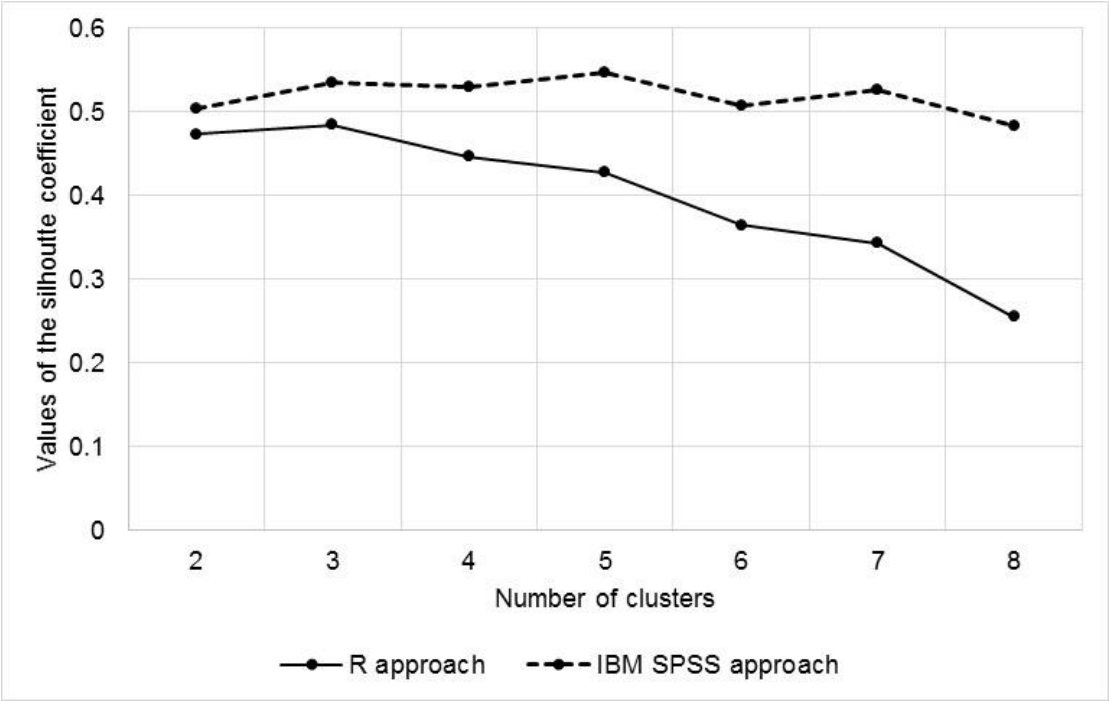
Table 2: Numbers of countries in the clusters as the results of clustering into 2–8 clusters with the complete linkage method

Number of clusters	Cluster 1	2	3	4	5	6	7	8
2	25	3						
3	17	8	3					
4	**12**	**5**	**8**	**3**				
5	12	4	1	8	3			
6	9	3	4	1	8	3		
7	9	3	4	1	8	1	2	
8	9	3	4	1	7	1	1	2

Source: Own calculation.

The values of the silhouette coefficient for the numbers of clusters from 2 to 8 for two principal different approaches are graphically displayed in Figure 2. While using the R approach dividing the set of the countries into two clusters is identified as the best one, using the IBM SPSS approach it is dividing into five clusters. In this splitting we obtain the first one-element cluster. In such a case, by the IBM SPSS approach the value SW_i for the country which is alone in the cluster (Italy) is 1 and the average value of SW_i for all countries is higher than the silhouette coefficient obtained in R with $SW_i = 0$ for Italy.

Figure 2: Differences between the silhouette coefficient values calculated by two approaches for the numbers of clusters from 2 to 8 (for the results obtained by the complete linkage method)



Source: Own calculation.

In addition, the PAM and FANNY algorithms were applied for assignments of the countries into clusters from 2 to 8. In Table 3 we can see values of the silhouette coefficient obtained by the R environment and the averages of silhouette values computed for each object with the STATS CLUS SIL command for all three clustering algorithms.

Table 3: Evaluation of clustering results obtained by the selected algorithms

Methods and coefficients	Number of clusters						
	2	3	4	5	6	7	8
CL + R	0.47	**0.48**	0.45	0.43	0.36	0.34	0.25
CL + SPSS	0.50	0.53	0.53	**0.55**	0.51	0.53	0.48
PAM + R	**0.47**	0.46	0.45	0.39	0.32	0.29	0.31
PAM + SPSS	0.50	0.52	0.53	0.51	0.47	0.49	**0.54**
FANNY + R	0.42	**0.46**	0.41	0.33	0.28	0.25	0.25
FANNY + SPSS	0.46	**0.52**	0.50	0.45	0.43	0.44	0.48

Notes: CL – the complete linkage algorithm, R – the silhouette coefficient obtained by the silhouette function in R, SPSS – the average of silhouette values computed for each object with the STATS CLUS SIL command in IBM SPSS Statistics.
Source: Own calculation.

In the case of the FANNY algorithm, the dividing the set of countries into 3 clusters was identified as the best one by both approaches. However, we can see that in the case of the R approach the values of the silhouette coefficient with the increasing number of clusters after the maximum value do not increase; by the IBM SPSS approach the value of the silhouette coefficient for 8 clusters is higher than those one for 5 clusters (and also higher than those one for two clusters).

In the case of the PAM algorithm, by the R approach two clusters were identified as the best, while with the IBM SPSS approach, it was dividing into 8 clusters. In this clustering, three one-element clusters were created – in comparison with two one-element clusters obtained by the FANNY algorithm.

Comparing results of all three methods of clustering, the highest value computed by the R environment was gained for 3 clusters obtained by the complete linkage method (0.48). In the case of the IBM SPSS approach it was division into 5 clusters obtained also by the complete linkage method (0.55).

4. Conclusion

In this contribution, the difference between two ways of calculation of the silhouette coefficient was explained. In addition, the third approach was mentioned. From the obtained results it is obvious that the IBM SPSS approach evaluates better results with one-element clusters. However, the aim of clustering is to obtain clusters of objects, not dividing a set of objects into individual objects. For this reason, we can consider the original approach implemented in the R language as the better one. Moreover, it is not correct to use the term *silhouette coefficient* for a measure which is computed in different way than the coefficient well-known more than 20 years. The evaluation coefficients in IBM SPSS Statistics should be named differently.

It is obvious that the name of the statistical software system does not guarantee a correct calculation in the case of the procedures or extensions added in later years because they are not by many years verified algorithms. The way of the calculation of the silhouette coefficient performed by STATS CLUS SIL command in IBM SPSS Statistics is not described The real way of the calculation was discovered experimentally by the author of this contribution. It was found that the calculation is performed incorrectly.

For the reason the calculation of the silhouette coefficient is based on the distances between objects, it does not depend on the number of variables (for the small number of variables the distance between two objects can be greater than the distance between other objects

characterized by the greater number of variables). However, with an increasing number of objects, the clusters with the larger number of objects can be created. In such a case, the average distance of objects in a certain cluster is influenced by means of calculation (without or with the *i*-th object) less than in the case of the small number of objects in the cluster. With the larger number of objects and small number of one-element clusters, the results of both types of calculation will differ less than in the case of the small number of objects.

We can find other examples of evaluating coefficients, which favor clustering results with one-element clusters, see (Řezanková and Húsek, 2012; Říhová and Pecáková, 2016). Thus, it is needed to interpret the results obtained by statistical software systems with caution.

The silhouette coefficient can be also applied to evaluation of results of cluster analysis in which objects characterized by categorical variables, see (Šulc *et al.*, 2017). The further research could be focused on comparison of values of this coefficient obtained with using different similarity measures proposed for categorical data clustering.

Acknowledgements

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

References

[1] Erilli, N. A. et al. 2011. Determining the most proper number of cluster in fuzzy clustering by using artificial neural networks. Expert Systems with Applications, vol. 38, iss. 3, pp. 2248-2252.

[2] Gan, G., Ma, C., Wu, J. 2007. Data Clustering: Theory, Algorithms, and Applications. Philadelphia: ASA-SIAM. ISBN 978-0-898716-23-8.

[3] IBM. 2016. IBM SPSS Statistics 24 Algorithms, pp. 115-117. [cit. 11-06-2018] ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf.

[4] Kaufman, L., Rousseeuw, P. 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken: Wiley. ISBN 0-471-73578-7.

[5] Löster, T. 2016. Determining the optimal number of clusters in cluster analysis. In: The 10th International Days of Statistics and Economics. Slaný: Melandrium, pp. 1078-1090. ISBN 978-80-87990-10-0.

[6] Maechler, M. et al. 2017. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

[7] Mur, A. et al. 2016. Determination of the optimal number of clusters using a spectral clustering optimization. Expert Systems with Applications, vol. 65, pp. 304-314.

[8] R Core Team. 2018. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, http://www.r-project.org/.

[9] Rousseeuw, P. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65. doi: 10.1016/0377-0427(87)90125-7.

[10] Řezanková, H., Húsek, D. 2012. Fuzzy clustering: Determining the number of clusters. In: 4th International Conference on Computational Aspects of Social Network. IEEE, pp. 277-282. doi: 10.1109/CASoN.2012.6412415.

[11] Řezanková, H., Křečková Kroupová, Z. 2017. Selected gender gap indicators – Comparison of V4 countries within EU context. In: Applications of Mathematics and

Statistics in Economics. Wroclaw: Publishing House of Wrocław University of Economics, pp. 373-386. doi: 10.15611/amse.2017.20.31.

[12] Říhová, E., Pecáková, I. 2016. Determination of the optimal number of clusters. In: 15th Conference on Applied Mathematics. Bratislava: Slovenská technická univerzita, 2016, pp. 947-958. ISBN 978-80-227-4531-4.

[13] Šulc, Z. et al. 2017. Comparison of internal evaluation criteria for categorical data in hierarchical clustering. In: Applied Statistics. Ljubljana: Narodna in univerzitetna knjižnica, p. 21. ISBN 978-961-93547-9-7.

[14] World Economic Forum. 2016. The Global Gender Gap Report 2016. Geneva: World Economic Forum. [cit. 11-06-2018] http://www3.weforum.org/docs/GGGR16/WEF_Global_Gender_Gap_Report_2016.pdf.