Cluster Analysis of Economic Data

Hana Řezanková[1] | *University of Economics, Prague, Czech Republic*

Abstract

In the paper, some classical and recent approaches to cluster analysis are discussed. Over the last decades researchers focused mainly on categorical data clustering, uncertainty in cluster analysis and clustering large data sets. In this paper some of the recently proposed techniques are introduced, such as similarity measures for data files with nominal variables, algorithms which include uncertainty in clustering, and the method for data files with many objects.

Keywords

Cluster analysis, similarity measures, hierarchical clustering, k-clustering, fuzzy clustering, large data sets

JEL code

C10, C38

INTRODUCTION

Cluster analysis is a strong tool of the multivariate exploratory data analysis. It involves a great amount of techniques, methods and algorithms which can be applied in various fields, including economy. However, in most of research papers containing cluster analysis of economic data the classical basic approaches are only applied. In this paper some clustering algorithms proposed in the last decades are introduced.

The aim of cluster analysis is to identify groups of similar objects (countries, enterprises, households) according to selected variables (unemployment rate of men and women in different countries, deprivation indicators of households, etc.). The basic approaches are hierarchical clustering and k-means clustering. There are many types of these techniques.

Agglomerative hierarchical clustering, which is usually applied, starts with objects regarded as individual clusters. The clusters are stepwise linked until all objects are connected in one cluster. In k-means clustering, objects are assigned to a certain number of clusters. In both methods the analyst needs to have some tools for determining the number of clusters. In hierarchical cluster analysis it can be done intuitively via a dendrogram, in *k*-means clustering the objects are usually assigned to different numbers of clusters and according to selected criteria, see e.g. (Gan et al., 2007), the suitable number is chosen.

The basic term in cluster analysis is a *similarity*. An attempt to formalize the similarity measure and relation between similarity and distance is given in (Chen et al., 2009). Let $\mathbf{x}_i$ be a vector of variable values, which characterizes the ith object. If variables are quantitative then the distance between the ith and jth objects can be calculated e.g. as the Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ (in the following text an object and a representing vector will be considered as synonyms), i.e.

$$d(\mathbf{x}_i, \mathbf{x}_j) = d_{ij} = \sqrt{\sum_{i=1}^{m} (x_{il}, x_{jl})^2} = ||\mathbf{x}_i - \mathbf{x}_j||, \quad (1)$$

[1] Department of Statistics and Probability, University of Economics, Prague, nám. W. Churchilla 4, 130 67 Prague, Czech Republic. E-mail: hana.rezankova@vse.cz, phone (+420)224095483.

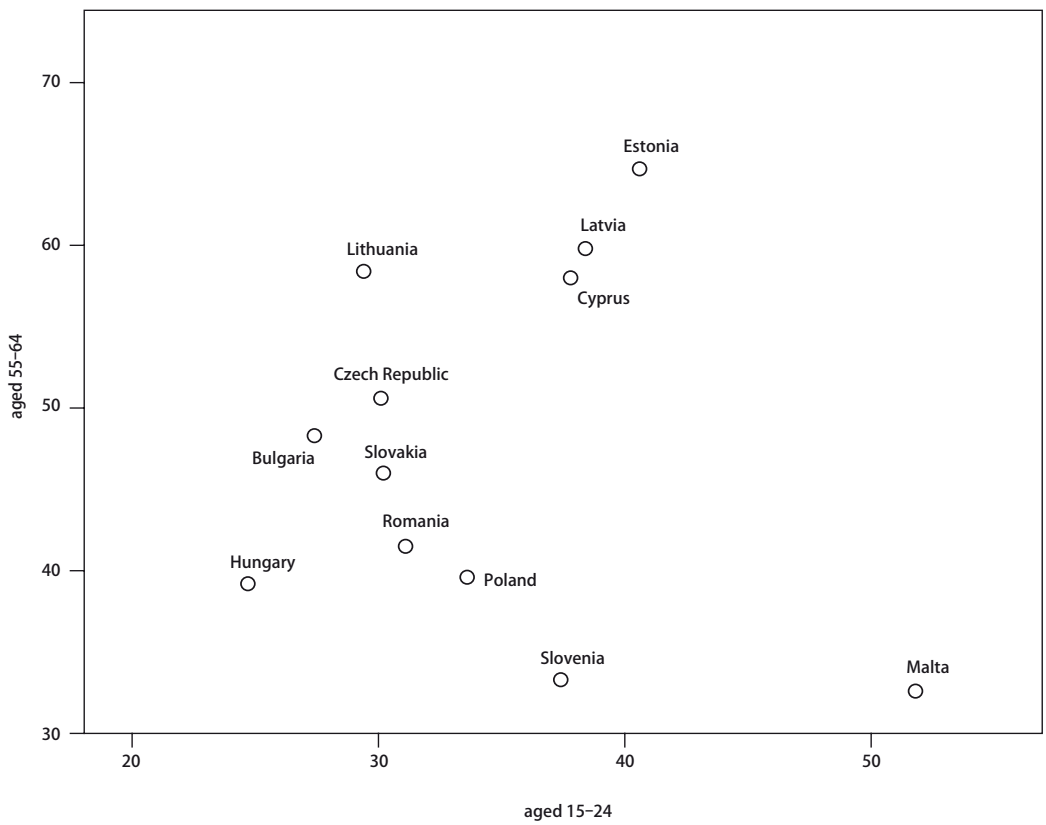
where m is the number of variables (e.g. economic indicators) and x_{il} is the value for the ith object and the lth variable. It is supposed that the data set $\mathbf{X}$ consisting of the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$, where n is the number of objects, should be partitioned into k clusters C_1, C_2…, C_k.

The main tasks for the cluster analysis research of the last decades has been clustering large data sets, clustering data files with categorical variables, fuzzy clustering and other techniques expressing uncertainty. Some related problems are solving, an outlier detection, determining the number of clusters, etc. Although the tasks mentioned above were solved at the beginning of the cluster analysis development, at the end of the 20th century and at the beginning of the 21st century the interest in these methods is growing in connection with the development of data mining techniques. The new algorithms for cluster analysis are proposed not only by statisticians, but also by computer science researchers. In this article the development of selected types of clustering is discussed.

1 HIERARCHICAL CLUSTER ANALYSIS

Probably the most applied method in economy is agglomerative hierarchical cluster analysis. It is based on a proximity matrix which includes the similarity evaluation for all pairs of objects. It means that various similarity or dissimilarity measures for different types of variables (quantitative, qualitative and binary) can be used. Moreover, different approaches for evaluation of the cluster similarity (single linkage, complete linkage, average linkage, Ward's method, etc.) can also be applied.

Figure 1 Scatter plot for countries characterized by economic activity rate in 2011 (IBM SPSS Statistics)



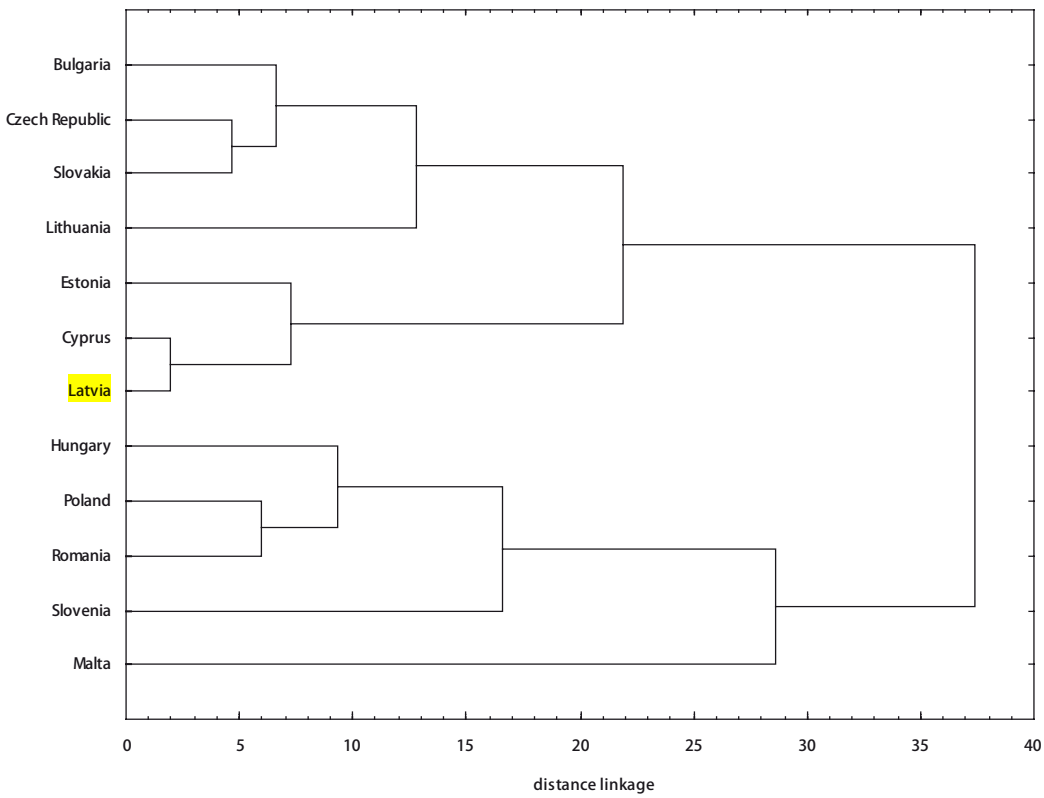
Source: *Slovensko v EÚ 2012 – Trh práce*. Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

Apart from giving a possibility to analyze data files with qualitative variables, the main advantage of this type of analysis is a graphical output in the form of a dendrogram. However, this graph is useful mainly for relatively small data files. In large files (with many objects) individual objects cannot be identified. Another disadvantage is a need to create a proximity matrix in the beginning of the analysis, what may cause a problem for large files.

Hierarchical cluster analysis can be illustrated by grouping selected countries of the European Union (new members of EU from 2004 and 2007 were selected). Let us consider three variables concerning the economic activity rate in 2011 according to the age (aged 15–24, 25–54, 55–64). Two of them can be represented by points in a scatter plot, see Figure 1.

With using the Euclidean distance and the complete linkage method the dendrogram in Figure 2 is obtained. We can see that Cyprus and Latvia are the most similar considering three studied variables (the countries are linked as the first; it is indicated by the smallest distance linkage in the dendrogram), then the Czech Republic and Slovakia are linked, etc. On the basis of the dendrogram we can identify two main clusters, which can be further divided to obtain a larger number of clusters.

Figure 2 Dendrogram for countries characterized by economic activity rate in 2011 (according to the age) obtained by the complete linkage method (STATISTICA)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

For example by cutting the dendrogram according to distance linkage 20 we obtain four clusters. In the first one there are the Czech Republic, Slovakia, Bulgaria and Lithuania, in the second one Cy-

prus, Latvia and Estonia are placed. The third cluster includes Poland, Romania, Hungary and Slovenia. In the fourth cluster there is only Malta. Minimum and maximum values of the analyzed variables characterizing four clusters are in Table 1.

Table 1 Characteristics of four clusters of countries obtained by the complete linkage method according to the economic activity rate in 2011

Cluster number	Aged 15–24		Aged 25–54		Aged 55–64	
	Min	Max	Min	Max	Min	Max
1	27.4	30.2	82.4	90.0	46.0	58.4
2	37.8	40.6	87.6	88.3	58.0	64.7
3	24.7	37.4	79.1	90.1	33.3	41.5
4	51.8	51.8	74.7	74.7	32.6	32.6

Source: Own calculation based on the data from publication: *Slovensko v EÚ 2012 – Trh práce*. Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

Malta has the highest value of the economic activity rate for the aged 15–24 group and the smallest value for the aged 55–64 group. The second cluster is characterized by high values both for the aged 15–24 group and for the aged 55–64 group. The first and the third clusters differ in the values of the aged 55–64 group.

If a data file contains nominal variables, some special measure must be used for similarity evaluation. The basic measure is the *simple matching coefficient*, which is also called the *overlap* measure. Let us denote the similarity of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ as s_{ij}. For its calculation the values in the ith and jth rows of the input matrix are compared for all variables. Evaluation of relationships of the values for the lth variable is denoted as s_{lij}. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = 0$. Similarity s_{ij} is calculated as the arithmetic mean, i.e.

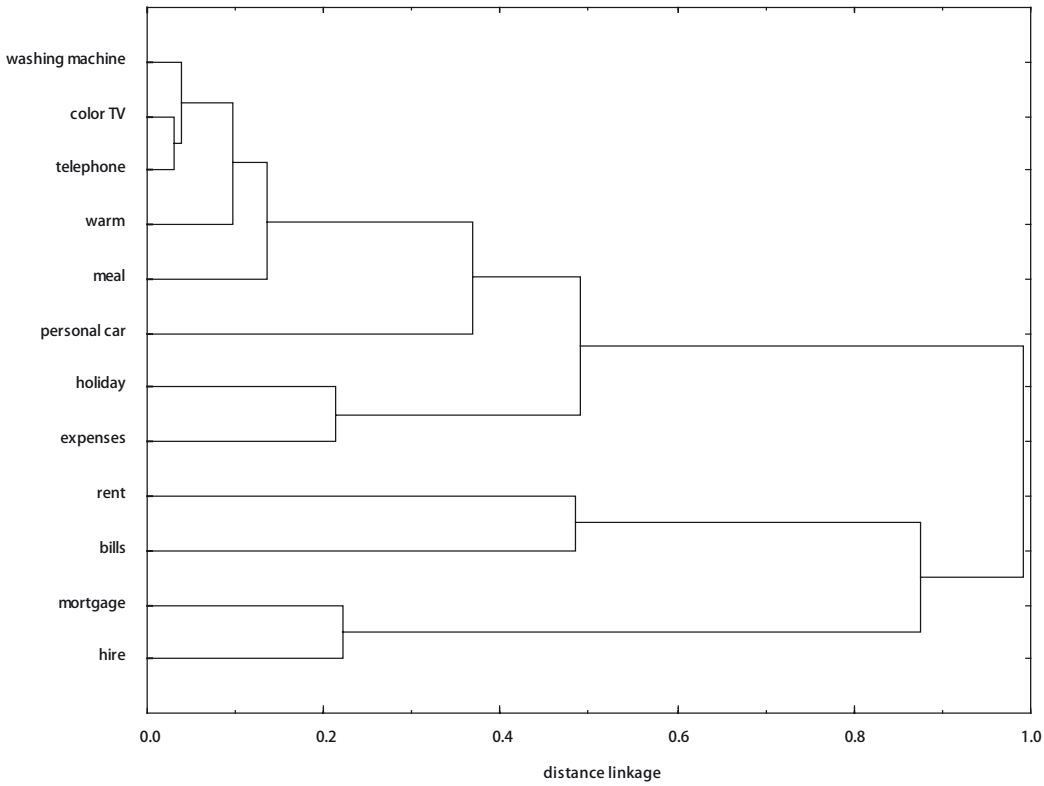
$$s_{ij} = \frac{\sum_{l=1}^{m} s_{lij}}{m}. \tag{2}$$

Hierarchical cluster analysis can be also based directly on a proximity matrix that evaluates the relationship of all pairs of variables. In a dendrogram similarity of variables and groups of variables can be identified.

Clustering of nominal variables will be illustrated by the data from the EU-SILC surveys in the Czech Republic (survey Living Condition 2011, the part "households"). There are nine indicators of material deprivation – eight indicators are answers to questions and the ninth one is composed of four answers. The data set with 12 original variables was analyzed (the number of households was 8 866). The variables contain values indicating whether or not the household can afford: a *washing machine*, a *color TV*, a *telephone*, a *personal car*, keeping the home adequately *warm*, a *meal* with meat, fish or vegetarian equivalent every second day, one week annual *holiday* away from home, coping with unexpected *expenses*, avoiding arrears in *rent*, utility *bills*, *mortgage* and *hire* purchase installments (the name of variables are in italic). These variables are nominal and they have different numbers of categories. The first four variables have three categories, the next four variables have two categories and the last four variables have three categories.

Figure 3 shows that the most similar answers concern a color TV and a telephone. The answers concern a washing machine are also alike (97–98% of the households own these durables). Three separated pairs of variables can be seen: *holiday* and *expenses* (56 and 58% of the households answered *yes*), *rent* and *bills*, and *mortgage* and *hire*.

Figure 3 Dendrogram for indicators of material deprivation of households, survey Living Condition 2011 (STATISTICA)



Source: the EU-SILC 2011 data

The *overlap* measure does not take into account different numbers of categories for individual variables. Recently, several similarity measures for objects characterized by nominal variables were proposed to deal with this problem. In the following text four measures for object similarity evaluation will be introduced. For the first three of them Equation (2) is applied, but the s_{lij} values are calculated differently.

The *Eskin measure* was proposed by Eskin et al. (2002). It assigns higher weights to mismatches which occur on variables with more categories. Let us denote the number of categories of the lth variable as n_l. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = n_l^2/(n_l^2 + 2)$.

The *OF measure* (*Occurrence Frequency*) assigns higher weights to more frequent categories in case of mismatch, see (Boriah et al., 2008). Let us denote the frequency of the category (of the lth variable) equal to the value x_{il} as $f(x_{il})$. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = 1/(1 + \ln(n/f(x_{il}))\cdot\ln(n/f(x_{jl})))$.

The *IOF measure* (*Inverse Occurrence Frequency*), see (Boriah et al., 2008), includes the opposite system of weights to OF. It evaluates mismatches of more frequent categories by lower weights. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = 1/(1 + \ln f(x_{il})\cdot\ln f(x_{jl}))$.

The *Lin measure* (Lin, 1998) assigns higher weights to more frequent categories in case of matches and lower weights to infrequent categories in case of mismatches. Let us denote a relative frequency

of the category equal to the value x_{il} as $p(x_{il})$. If $x_{il} = x_{jl}$, then $s_{lij} = 2 \cdot \ln p(x_{il})$, otherwise $s_{lij} = 2 \cdot \ln(p(x_{il}) + p(x_{jl}))$. The similarity measure for two objects is then computed as

$$s_{ij} = \frac{\sum_{l=1}^{m} s_{lij}}{\sum_{l=1}^{m} (\ln p(x_{il}) + \ln p(x_{jl}))}. \tag{3}$$

The measures mentioned above and some other measures have been reviewed e.g. in (Boriah et al., 2008) and (Chandola et al., 2009). Some other similarity measures have been proposed, e.g. in (Le et Ho, 2005) and (Morlini et Zani, 2012).

2 K-CLUSTERING

In *k*-clustering the set of objects is divided to a certain number (*k*) clusters. We can distinguish different approaches from different points of view. The first classification is for hard and fuzzy clustering. In the first one, an object is assigned exactly to one cluster. The result is a membership matrix for objects and clusters with ones (the object is assigned to the cluster) and zeroes (the object is not assigned to the cluster). In the second approach membership degrees are calculated for all cluster-object pairs. Moreover, some other approaches to expressing uncertainty in cluster analysis have been proposed, see below.

The second classification is for *k*-centroid and *k*-medoids clustering. In the former, the center of a cluster is represented by a vector of variable characteristics (e.g. vector of means for quantitative variables). In the latter, the center of a cluster is represented by a selected object (by a vector from the input matrix).

The most applied *k*-centroid technique is the *k*-means (also HCM – *hard c-means*) algorithm (MacQueen, 1967), which analyzes the data set with the aim to minimize the objective function

$$J_{HCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{hard,\,ih} d_{ih}^2, \tag{4}$$

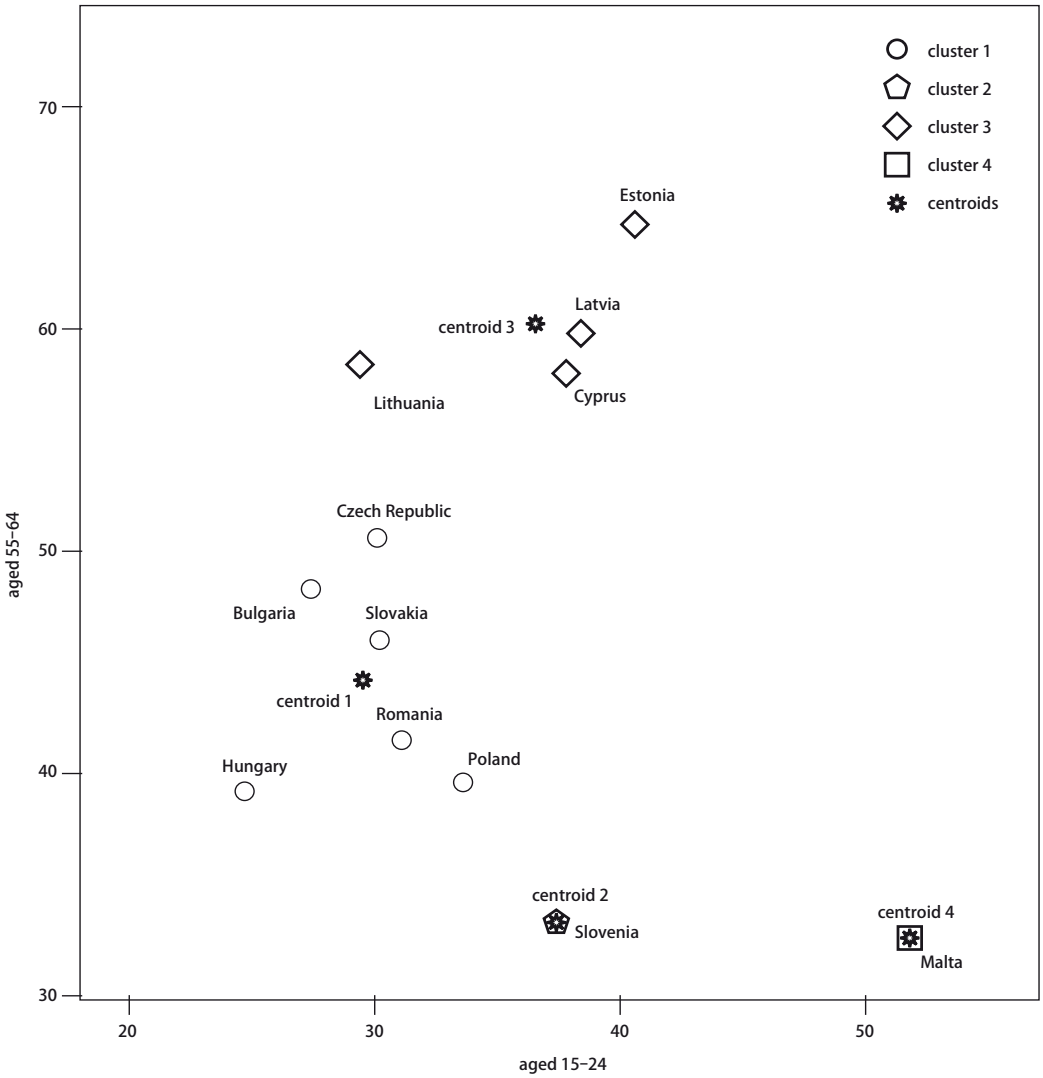
where the elements $u_{hard,ih} \in \{0, 1\}$ indicate the assignment of object vectors to clusters (1 means the assignment) and d_{ih} is the Euclidean distance between the *j*th object and the center (a vector of means) of the *h*th cluster. Further, the following conditions have to be satisfied:

$$\sum_{h=1}^{k} u_{hard,\,ih} = 1 \text{ for } i = 1, 2, ..., n \text{ and } \sum_{i=1}^{n} u_{hard,\,ih} > 0 \text{ for } h = 1, 2, ..., k.$$

Let us analyze the data file with three variables concerning the economic activity rate in 2011 according to age (see Section 1, Figure 1). With using the k-means algorithm for clustering countries to four clusters we obtain two one-element clusters (Slovenia and Malta). All four clusters and their centroids are presented in Figure 4 and the obtained clusters are characterized in Table 2. They differ from results obtained by the complete linkage method (Figure 2) but they are consistent with the results obtained by the average linkage method (these results are not presented in this paper).

According to the studied variables Malta is significantly different from the other countries regardless of the method used. Slovenia is characterized by the low value of the economic activity rate for the age group 55–64 and the highest value for the age group 25–54. The first and the third clusters differ mainly in the values of the age group 55–64.

Figure 4 Scatter plot for countries characterized by economic activity rate in 2011 with centroids of four clusters obtained by *k*-means clustering (IBM SPSS Statistics)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

The advantage of *k*-centroid clustering is a possibility to apply it to large data sets. The disadvantage is its instability; for different orders of object vectors different assignments of objects to clusters can be obtained. Further, the result of clustering depends on a type of initialization (determination of *k* initial centroids), which is the first step of the algorithm. *K*-clustering methods search for the optimal solution, but the optimum can only be local, not global. Despite some negative properties these methods play an important role in the exploratory data analysis.

Table 2 Characteristics of four clusters of countries obtained by k-means clustering according to the economic activity rate in 2011

Cluster number	Aged 15–24		Aged 25–54		Aged 55–64	
	Min	Max	Min	Max	Min	Max
1	24.7	33.6	79.1	88.0	39.2	50.6
2	37.4	37.4	90.1	90.1	33.3	33.3
3	29.4	40.6	87.6	90.0	58.0	64.7
4	51.8	51.8	74.7	74.7	32.6	32.6

Source: Own calculation based on the data from publication: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

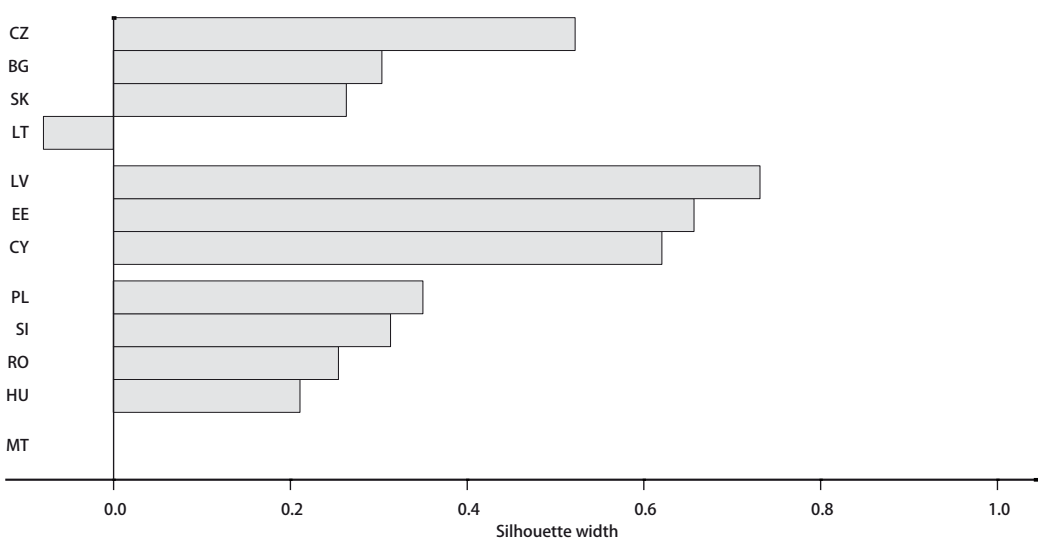
In the hard *k-medoids* (also *PAM – Partitioning Around Medoids*) algorithm, see (Kaufman et Rousseeuw, 2005), the objective function

$$f_{KM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{hard,\,ih} ||\mathbf{x}_i - \mathbf{m}_h|| \tag{5}$$

is minimized, where $\mathbf{m}_h$ is a medoid of the hth cluster and for values $u_{hard,ih}$ the same conditions as in case of k-means clustering must be satisfied.

With using the k-medoids algorithm for clustering countries to four clusters we obtain one one-element cluster (Malta). All four clusters are presented in Figure 5 in the form of a silhouette plot. The silhouette widths are computed on the basis of distances of an object from the other objects from both the same cluster and the other clusters. The first object in the cluster is a medoid. In Figure 5 the medoids are the Czech Republic, Latvia, Poland and Malta. An opposite direction (a negative value) in case of Lithuania (belonged to the first cluster) means that this country is closer the objects from other clusters (according to the special measure).

Figure 5 Silhouette plot for countries characterized by economic activity rate in 2011 for four clusters obtained by the PAM algorithm (S-PLUS)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

From Figure 5 it is obvious that generally some objects can be assigned to two (or more) clusters. In this case an uncertainty can be expressed in results of clustering. One of the approaches how to express an uncertainty in cluster analysis is a fuzzy assignment of objects to clusters. It is applied in fuzzy cluster analysis. This technique is based on the theory of fuzzy sets (Zadeh, 1965). Fuzzy clustering has been studied very intensively in the past decades. A lot of papers have been published in journals, conference proceedings and in some monographs, e.g. (Abonyi et Feil, 2007) and (Höppner et al., 2000). There are many different algorithms used for fuzzy (soft) cluster analysis. Fuzzy *k*-means is one of them, see e.g. (Kruse et al., 2007). It is based on a generalization of the *k-means* (HCM) algorithm.

The *fuzzy k-means* (frequently FCM – *fuzzy c-means*) algorithm (Bezdek, 1981) minimizes the objective function

$$J_{FCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{ih}^{q} d_{ih}^{2}, \tag{6}$$

where the elements $u_{ih} \in \langle 0, 1 \rangle$ are membership degrees, and the parameter q ($q > 1$) is called a fuzzifier or weighting exponent (usually $q = 2$ is chosen). Furthermore, the following conditions have to be satisfied:

$$\sum_{h=1}^{k} u_{ih} = 1 \text{ for } i = 1, 2, ..., n \text{ and } \sum_{i=1}^{n} u_{ih} > 0 \text{ for } h = 1, 2, ..., k.$$

We can again illustrate the application of fuzzy cluster analysis to the data on selected countries of the European Union (see Section 1, Figure 1). Using the FANNY algorithm in the S-PLUS statistical software, see (Kaufman et Rousseeuw, 2005), we obtain the results in Table 3 with the assignment of countries to four clusters.

Table 3 Country membership degrees based on economic activity rate in 2011 for four clusters obtained by the FANNY algorithm (S-PLUS)

Country	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster number
Bulgaria	**0.51**	0.11	0.28	0.10	1
Czech Republic	**0.71**	0.10	0.12	0.07	1
Estonia	0.15	**0.66**	0.11	0.09	2
Cyprus	0.09	**0.82**	0.05	0.04	2
Lithuania	0.35	**0.36**	0.17	0.12	2
Latvia	0.05	**0.89**	0.03	0.02	2
Hungary	0.22	0.09	**0.53**	0.15	3
Malta	0.21	0.18	0.28	**0.34**	4
Poland	0.21	0.08	**0.48**	0.23	3
Romania	0.12	0.05	**0.74**	0.08	3
Slovakia	**0.62**	0.08	0.20	0.09	1
Slovenia	0.03	0.02	0.04	**0.90**	4

Source: Own calculation based on the data from publication: *Slovensko v EÚ 2012 – Trh práce*. Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

According to the highest value of membership degrees over clusters (bold figures in Table 3), the first cluster is created by the Czech Republic, Slovakia and Bulgaria. In the second cluster there are Latvia, Cyprus, Estonia and Lithuania. It can be noticed Lithuania has similar membership degrees to the first and the second clusters (0.35 and 0.36). The third cluster is created by Romania, Hungary and Poland, and the fourth cluster contains Slovenia and Malta. However, membership degrees are very various – it is 0.9 for Slovenia and 0.34 for Malta.

The fuzzy *k*-means algorithm is sensitive to noise and outliers. Let us suppose clustering to two clusters C_h and C_g. If $\mathbf{x}_i$ is equidistant from centroids c_h and c_g then $u_{ih} = u_{ig} = 0.5$, regardless whether the actual distance is large or small. A similar situation can be mentioned in the *fuzzy k-medoids* algorithm (Krishnapuram et al, 2001), in which the objective function

$$f_{FKM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{ih}^{q} \left\| \mathbf{x}_i - \mathbf{m}_h \right\|^2 \tag{7}$$

is minimized under the same condition as in the fuzzy *k*-means algorithm.

For the reason of the negative property of fuzzy clustering, different approaches were proposed later, see (Bodjanova, 2013). One of them is the *possibilistic k-means* (also PCM – *possibilistic c-means*) algorithm (Krishnapuram et Keller, 1993). It minimizes the objective function

$$J_{PCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} w_{ih}^{q} d_{ih}^2 + \sum_{h=1}^{k} \gamma_h \sum_{i=1}^{n} (1 - w_{ih})^q , \tag{8}$$

where the elements $w_{ih} \in \langle 0, 1 \rangle$ are membership degrees, q is a fuzzifier, and the following conditions have to be satisfied

$\sum_{h=1}^{k} u_{ih} = 1$ for $i = 1, 2, ..., n$ and $\sum_{i=1}^{n} u_{ih} > 0$ for $h = 1, 2, ..., k$. and γ_h is a user defined constant (scale parameter). It can be computed e.g. as

$$\gamma_h = \frac{\sum_{i=1}^{m} u_{ih}^{q} d_{ih}^2}{\sum_{i=1}^{m} u_{ih}^{q}} .$$

This algorithm is very sensitive to initialization.

Since both the FCM and the PCM algorithms have some negative properties, the combination of both algorithms has been developped in results of the *FPCM algorithm* (Pal et al., 2005). It minimizes the objective function

$$J_{FPCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} (a u_{ih}^{q_1} + b w_{ih})^{q_2} + \sum_{h=1}^{k} \gamma_h \sum_{i=1}^{n} (1 - w_{ih})^{q_2} , \tag{9}$$

where a, b, q_1, q_2 and γ_h are positive constants. Constants a and b define the relative importance of probabilistic and possibilistic memberships. The *fuzzy-possibilistic c-medoids* algorithm has been also proposed (Maji et Pal, 2007a).

Other approaches which are alternative to hard clustering are techniques based on the *rough set theory* (Pawlak, 1982). The basic technique is the *rough k-means* (or RCM – *rough c-means*) algorithm (Lingras et West, 2004). In this algorithm each cluster C_h is defined by the lower approximation $A_{low}(C_h)$ and the upper approximation $A_{up}(C_h)$. The object $\mathbf{x}_i$ can be a part of most one lower approximation.

If the object $\mathbf{x}_i$ is a part of a certain lower approximation then is also a part of the upper approximation. If $\mathbf{x}_i$ is not a part of any lower approximation then it belongs to two or more upper approximation. A special technique for the cluster mean computation is applied.

In the RCM algorithm two values characterizing the membership for a certain object and a certain cluster are calculated. There are the low membership $u_{low,ih}$ and the up membership $u_{up,ih}$. If $u_{low,ih} = 1$ then the *i*th object certainly belongs to the *h*th cluster. If $u_{low,ih} = 0$ then the assignment of the *i*th object depends on the value of $u_{up,ih}$. If $u_{up,ih} = 1$ then the *i*th object possibly belongs to the *h*th cluster. If $u_{up,ih} = 0$ then the *i*th object does not belong to the *h*th cluster. The following conditions have to be satisfied:

$$u_{low,ih,}\, u_{up,\,ih} \in \{0, 1\}\, u_{low,ih} \leq u_{up,ih,} \text{ and } \sum_{h=1}^{k} u_{low,ih} \leq 1 \text{ for } i = 1, 2, ..., n.$$

The result of a combination of rough and fuzzy approaches is the *rough-fuzzy k-means* (or RFCM – *rough-fuzzy c-means*) algorithm (Mitra et al., 2004). Moreover, Maji and Pal (2007b) proposed the *rough-fuzzy-possibilistic k-means* (or RFPCM – *rough-fuzzy-possibilistic c-means*).

In rough-based techniques the lower approximation of each cluster depends on a fixed threshold *TH* which is defined by the user. For this reason a modification of these approaches based on the *shadowed sets* (Pedrycz, 1998) was proposed by Mitra et al. (2010). This algorithm is called *shadowed k-means* (SCM – *shadowed c-means*). It provides the dynamical evaluation of thresholds for each cluster individually, based on the original data.

3 OTHER APPROACHES

Besides of hierarchical clustering and *k*-clustering, there are some other approaches proposed e.g. for large data files (with many objects), for data files with categorical variables, and also for data files of both types. We can mention *two-step cluster analysis* implemented in the IBM SPSS Statistics software as an example of the procedure which can cluster large data sets with both quantitative and qualitative variables. This method is based on the BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) algorithm, see (Zhang et al., 1996).

The algorithm arranges objects of the data set into subclusters, known as cluster features (CFs). These cluster features are then clustered into *k* groups using a traditional hierarchical clustering procedure. A CF represents a set of summary statistics on a subset of the data. The algorithm consists of two phases. In the first one, an initial CF tree is built. In the second one, an arbitrary clustering algorithm is used to cluster the leaf nodes of the CF tree. The disadvantage of this method is its sensitivity to the order of the objects.

In two-step cluster analysis, the user can apply either the Euclidean distance for the quantitative data or the log-likelihood distance measure which is determined for the data files with the combination of quantitative and qualitative variables (Chiu et al., 2001). In the second case the dissimilarity of two clusters is expressed as the difference between a variability of the cluster created by linking of the studied clusters and a sum of the variability of individual clusters. A variability is calculated as a combination of values of the variance (for quantitative variables) and the entropy (for qualitative variables).

The application of this method will be illustrated to the EU-SILC data (Living Condition 2011). After clustering 8 866 households (i.e. large data set for cluster analysis) according to 12 nominal variables analyzed in Section 1, the procedure determines two clusters of households as optimal (the average silhouette width is calculated on the basis of the silhouette widths, see Figure 5).

The output for two clusters indicates that the most important variables for clustering are *personal car, holiday,* and *expenses*. For three clusters, variables *mortgage* and *hire* were added as important. Variables *warm* and *meal* were added as important for four clusters. The relative frequencies of categories for variables mentioned above are in Tables 4 and 5.

Table 4 Relative frequencies of answers in individual clusters obtained by two-step cluster analysis based on indicators of material deprivation, survey Living Condition 2011

Cluster number (size)	Warm		Meal		Holiday		Expenses	
	Yes	No	Yes	No	Yes	No	Yes	No
1/2 (49.9%)	85.9%	14.1%	76.4%	23.6%	29.7%	70.3%	30.9%	69.1%
2/2 (50.1%)	99.8%	0.2%	99.8%	0.2%	83.0%	17.0%	84.2%	15.8%
1/3 (54.8%)	87.5%	12.5%	79.2%	20.8%	32.2%	67.8%	36.1%	63.9%
2/3 (20.1%)	98.8%	1.2%	97.8%	2.2%	68.1%	31.9%	63.6%	36.4%
3/3 (25.1%)	100.0%		100.0%		100.0%		100.0%	
1/4 (37.8%)	99.8%	0.2%	99.6%	0.4%	39.9%	60.1%	43.5%	56.5%
2/4 (19.6%)	99.4%	0.6%	99.5%	0.5%	69.6%	30.4%	54.7%	35.3%
3/4 (17.5%)	60.5%	39.5%	33.7%	66.3%	14.9%	85.1%	19.7%	80.3%
4/4 (25.1%)	100.0%		100.0%		100.0%		100.0%	

Source: the EU-SILC 2011 data

Table 5 Relative frequencies of answers in individual clusters obtained by two-step cluster analysis based on indicators of material deprivation, survey Living Condition 2011

Cluster number	Personal car			Mortgage			Hire		
	Own	Cannot afford	Other	Yes	No	Other	Yes	No	Other
1/2	29.8%	22.5%	47.7%	1.0%	4.4%	94.6%	2.4%	10.8%	86.9%
2/2	98.0%	0.5%	1.5%	0.1%	19.9%	80.0%	0.1%	19.3%	80.7%
1/3	38.1%	19.4%	42.5%	0.8%	1.9%	97.2%	2.1%	5.7%	92.3%
2/3	89.7%	4.2%	6.1%	0.6%	55.4%	44.1%	0.4%	59.4%	40.2%
3/3	100.0%					100.0%			100.0%
1/4	37.5%	14.6%	47.9%		0.1%	99.9%		1.1%	98.9%
2/4	89.1%	3.7%	7.2%	0.6%	54.6%	44.8%	0.3%	58.7%	41.0%
3/4	41.3%	29.9%	28.8%	2.6%	7.9%	89.5%	6.5%	17.7%	75.8%
4/4	100.0%					100.0%			100.0%

Source: the EU-SILC 2011 data

If the households are clustered to two clusters, they can be characterized in the following way. One cluster includes mostly the households that own a personal car and have no problems neither with paying holiday nor with unexpected expenses. The second cluster represents the households which have not a personal car from other reason than financial and have problems to pay holiday and unexpected expenses. Similarly the results of clustering to three and four clusters can be described.

Another application of two-step cluster analysis to the EU-SILC data is described in (Řezanková et Löster, 2013). For the analysis of large data files with quantitative variables, methods *k*-clustering can

be applied, either classical algorithms or their modifications. We can mention the *CLARA* (*Clustering LARge Applications*) algorithm as an example (Kaufman et Rousseeuw, 2005). It is based on the *k*-medoid algorithm and implemented in the S-PLUS system.

The principles of methods proposed for large data files (both with many objects and many variables) are reviewed e.g. in (Kogan, 2007). An example of techniques for clustering in case of high-dimensional data is the R package BCLUST (Partovi Nia et Davison, 2012). The approaches to clustering categorical data are summarized e.g. in (Řezanková, 2009). If a data set contains mixed-type variables, one possibility is to cluster objects according groups of variables of the same type and then combine of individual solutions by cluster ensembles, e.g. by package CLUE for R, see (Hornik, 2005).

CONCLUSION

In the paper selected approaches to cluster analysis were introduced. For cluster analysis of objects which are characterized by values of nominal variables, the analyst can use recently proposed similarity measures. Performed experiments showed (Šulc et al., 2013) that clustering with using some of these measures give better clusters than the overlap measure from the point of view of the within-cluster variability.

Recent methods which include uncertainty are a promising tool to give better results than basic fuzzy cluster analysis. For the data from some surveys, e.g. the EU-SILC, the techniques for large data sets is useful.

It is regrettable that commercial software products react to the recently proposed methods very slowly, or do not react at all. They rarely include measures for nominal variables, fuzzy cluster analysis and methods for large data files. If the software system offers some of these possibilities, it is usually just one of them and the analysts need to use several software products to perform modern analyses.

References

ABONYI, J., FEIL, B. *Cluster Analysis for Data Mining and System Identification*. Berlin: Birkhäuser Verlag AG, 2007.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York: Plenum Press, 1981.

BODJANOVA, S.: Fuzzy sets and rough sets in prototype-based clustering algorithms. In *Olomoucian Days of Applied Mathematics 2013 – Presentations* [online]. Olomouc: Palacky University in Olomouc, 2013. [cit. 11.11.2013]. <http://odam.upol.cz/downloads/presentations/2013/Bodjanova.pdf>.

BORIAH, S., CHANDOLA, V., KUMAR, V. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 8th SIAM International Conference on Data Mining.* SIAM, 2008, pp. 243–254.

CHANDOLA, V., BORIAH, S., KUMAR, V. A framework for exploring categorical data. In *Proceedings of the 9th SIAM International Conference on Data Mining.* SIAM, 2009, pp. 187–198.

CHEN, S., MA, B., ZHANG, K. On the similarity metric and the distance metric. In Formal Languages and Applications: A Collection of Papers in Honor of Sheng Yu. *Theoretical Computer Science*, 2009, 24–25, pp. 2365–2376.

CHIU, T., FANG, D., CHEN, J., WANG, Y., JERIS, C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: ACM, 2001, pp. 263–268.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S. A geometric framework for unsupervised anomaly detection. In BARBARA, D., JAJODIA, S., eds. *Applications of Data Mining in Computer Security*, pp. 78–100. Norwell, MA: Kluwer Academic Publishers, 2002.

GAN, G., MA, C., WU, J. *Data Clustering: Theory, Algorithms, and Applications.* Philadelphia: ASA-SIAM, 2007.

HÖPPNER, F., KLAWON, F., KRUSE, R., RUNKLER, T. *Fuzzy Cluster Analysis. Methods for Classification, Data Analysis and Image Recognition.* New York: John Wiley & Sons, 2000.

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

KAUFMAN, L., ROUSSEEUW, P. *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken: Wiley, 2005.

KOGAN, J. *Introduction to Clustering Large and High-Dimensional Data.* New York: Cambridge University Press, 2007.

KRISHNAPURAM, R., KELLER, J. M. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst.*, 1993, 1(2), pp. 98–110.

KRISHNAPURAM, R., JOSHI, A., YI, L. Fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *IEEE International Conference on Fuzzy Systems 3*. Institute of Electrical and Electronics Engineers Inc., 1999, pp. III-595–III-607.

KRUSE, R., DÖRING, C., LESOT, M.-J. Fundamentals of Fuzzy Clustering. In OLIVEIRA, J. V., PEDRYCZ, W., eds. *Advances in Fuzzy Clustering and Its Applications*. Chichester: John Wiley & Sons, 2007, pp. 3–30.

LE, S. Q., HO, T. B. An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters*, 2005, 26(16), pp. 2549–2557.

LIN, D. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the 15th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

LINGRAS, P., WEST, C. Interval set clustering of web users with rough k-means. *Journal of Intelligent Information Systems*, 2004, 23, pp. 5–16.

MACQUEEN, J. B. Some methods for classification and Analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, 1967, pp. 281–297.

MAJI, P., PAL, S. K. Protein sequence analysis using relational soft clustering algorithms. *International Journal of Computer Mathematics*, 2007a, 84(2), pp. 599–617.

MAJI, P., PAL, S. K. Rough set based generalized fuzzy c-means algorithm and quantitative indices. *IEEE Trans. Syst., Man and Cybernetics Part B*, 2007b, 37(6), pp. 1529–1540.

MITRA, S., BANKA, H., PEDRYCZ, W. Rough-fuzzy collaborative clustering. *IEEE Trans. Syst., Man and Cybernetics, Part B*, 2006, 36(4), pp. 795–805.

MITRA, S., PEDRYCZ, W., BARMAN, B. Shadowed c-means: Integrating fuzzy and rough clustering. *Pattern Recognition*, 2010, 43, pp. 1282–1291.

MORLINI, I., ZANI, S. A new class of weighted similarity indices using polytomous variables. *Journal of Classification,* 2012, 29(2), pp. 199–226.

PAL, N. R., PAL, K., KELLER, J. M., BEZDEK, J. C. A possibilistic fuzzy c-means clustering algorithm. *IEEE Trans. Fuzzy Syst.*, 2005, 13 (4), pp. 517–530.

PARTOVINIA, V., DAVISON, A. High-dimensional Bayesian clustering with variable selection: the R package bclust. *Journal of Statistical Software*, 2012, 47(5), pp. 1–22.

PAWLAK, Z. Rough sets. *International Journal of Computer and Information Sciences*, 1982, 11, pp. 341–356.

PEDRYCZ, W. Shadowed sets: representing and processing fuzzy sets. *IEEE Trans. Syst., Man and Cybernetics,* Part B, 1998, 28(1), pp. 103–109.

ŘEZANKOVÁ, H. Cluster analysis and categorical data. *Statistika*, 2009, 3, pp. 216–232.

ŘEZANKOVÁ, H., LÖSTER, T. Shluková analýza domácností charakterizovaných kategoriálními ukazateli. *E+M Ekonomie a Management*, 2013, 3, pp. 139–147.

ŠULC, Z., ŘEZANKOVÁ, H., MOHAMMAD, A. Comparison of selected approaches to categorical data clustering. In *AMSE 2013*. Banská Bystrica: Univerzita Mateja Bela, 2013, p. 25.

ZADEH, L. A. Fuzzy sets. *Information and Control*, 1965, 8, pp. 338–353.

ZHANG, T., RAMAKRISHNAN, R., LIVNY. M. BIRCH. An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. Montreal: ACM, 1996, pp. 103–114.