18th APPLICATIONS OF MATHEMATICS AND STATISTICS IN ECONOMICS AMSE
International Scientific Conference | Jindřichův Hradec, 2-6 September 2015

# DIMENSIONALITY REDUCTION OF CATEGORICAL DATA: COMPARISON OF HCA AND CATPCA APPROACHES

**ZDENĚK ŠULC, HANA ŘEZANKOVÁ**
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill sq. 4, Prague, Czech Republic
email: zdenek.sulc@vse.cz
hana.rezankova@vse.cz

**Abstract**
*This paper compares two approaches to dimensionality reduction in datasets containing categorical variables: hierarchical cluster analysis (HCA) with different similarity measures for categorical data, and categorical principal component analysis (CATPCA), which represents a model based approach with latent variables. Each approach has its pros and cons. HCA is easier to perform, but most of similarity measures require the same number of categories in all examined variables; CATPCA can deal with ordinal data, but it has worse interpretation of results. In this paper, several similarity measures for HCA are applied and resulting clusters are further compared with the ones gotten by CATPCA. Their quality is determined by the within-cluster mutability coefficient, and by judgments of researchers. For the comparison of approaches mentioned above, datasets from economic and sociological researches conducted in the Czech Republic in 2014 were used. Thus, the conclusion of this paper not only summarizes findings about approaches to reduction of a dataset, but also the opinions of the Czech citizens to actual issues.*

***Key words:*** *dimensionality reduction, categorical data, cluster analysis.*

## 1. Introduction

Questionnaire surveys represent an important tool in economic, marketing and sociological researches. The acquired datasets, based on those surveys, contain mainly categorical variables, both nominal and ordinal ones. The number of these variables is usually large, and thus, dimensionality reduction is a very important part of the analysis, see (Bartholomew et al., 2002). Its principle is based on finding groups of similar variables, which can be interpreted as factors characterizing studying objects (e.g. respondents). From each group, one can choose a variable as a representative of the whole group. Thus, dimensionality reduction leads to saving the computation time in subsequent analyses.

The aim of this paper is to compare the results obtained by hierarchical cluster analysis (HCA) using different similarity measures for categorical data with results obtained by categorical principal component analysis (CATPCA).

For the comparison, two real datasets based on questionnaire surveys conducted by the Institute of Sociology of the Czech Academy of Sciences in 2014 are used. In the surveys, the respondents answered the questions concerning the actual political, economic and social situation.

The paper is organized as follows. Section 2 introduces methods for grouping categorical variables. Section 3 describes evaluation criteria of cluster quality. The application to real datasets is presented in Section 4. The final results are summarized in Section 5.

## 2. Methods for Grouping Categorical Variables

In this paper, the emphasis is put on comparing two approaches for grouping categorical variables, namely, HCA with several similarity measures, and CATPCA. There are other methods for grouping categorical variables as well, e.g. latent class analysis, see (Vermunt and Magidson, 2005).

### 2.1 Complete Linkage Method of HCA and Selected Similarity Measures

HCA is based on a proximity matrix, which contains dissimilarities of analyzed objects (variables in this paper) taken pairwise. At the beginning, each variable is a cluster of its own. Then, in each step, two nearest clusters are merged into a new one. Therefore, the definition of distance between clusters is a very important part of the analysis. Based on our previous research, see (Šulc and Řezanková, 2014b), the *complete linkage* method was chosen. In this method, dissimilarity between the furthest variables from two different clusters is considered as the distance between these clusters.

Dissimilarity of two variables is usually derived from similarity. Based on our previous research, see (Šulc, 2014), several similarity measures for variable clustering were selected. These measures can be used only in case all variables have the same number categories with the same meaning, because it would have no sense to determine the agreement between two nominal variables with different categories. Thus, questionnaire surveys appear to be an ideal application. In this paper, following similarity measures are used: Eskin, Goodall 3, Goodall 4, IOF, Lin, Lin 1, and overlap.

Let us denote the total number of analyzed variables as $m$ and the total number of surveyed objects as $n$. The elements of the input data matrix are $x_{ic}$, where $i = 1, 2, ..., n$ and $c = 1, 2, ..., m$.

The *Eskin* measure was proposed by Eskin et al. (2002). It assigns higher weights to mismatches by objects with occurrence of higher number of categories. Let us denote the number of categories of the $i$-th object as $n_i$. The similarity $S_{icd}$ between the $c$-th and $d$-th variables for the $i$-th object is expressed in the following way. If $x_{ic} = x_{id}$, then $S_{icd} = 1$, otherwise $S_{icd} = n_i^2/(n_i^2 + 2)$. The total similarity $S_{cd}$ between the $c$-th and $d$-th variables across all objects is calculated as the arithmetic mean of the $S_{icd}$ values for $i = 1, 2, ..., n$. The dissimilarity measure $D_{cd}$ is computed as $(1/S_{cd} - 1)$.

The *Goodall 3* measure, (Boriah et al. 2008), was proposed to assign higher similarity if the infrequent categories match regardless on frequencies of other categories. Similarity between two variables by the $i$-th object is expressed in the following way. Let us denote the relative frequency of the category $x_{ic}$ occurring in the $i$-th object as $p_i(x_{ic})$. If $x_{ic} = x_{id}$, then $S_{icd} = 1 - p_i^2(x_{ic})$, otherwise $S_{icd} = 0$. The total similarity $S_{cd}$ between the $c$-th and $d$-th variables is computed as the arithmetic mean. Since the range of possible values does not exceed the interval from zero to one, the transformation of a similarity measure into a dissimilarity measure can be expressed as a complement to the value 1, i.e. $D_{cd} = 1 - S_{cd}$.

The *Goodall 4* measure, proposed by Boriah et al. (2008), assigns higher similarity if the frequent categories match. Actually, when measuring similarity between two variables, this measure provides complement results of Goodall 3 to one, i.e. if $x_{ic} = x_{id}$, then $S_{icd} = p_i^2(x_{ic})$, otherwise $S_{icd} = 0$. The range of possible values takes values from $2/(m(m - 1))$ to 1. The similarity $S_{cd}$ is computed as the arithmetic mean and the dissimilarity $D_{cd}$ is expressed as a complement to 1.

The *IOF* (Inverse Occurrence Frequency) measure comes from an information retrieval, see Sparck-Jones (1972, 2002). The measure was constructed to assign higher weights to mismatches on less frequent values and lower weights to mismatches on more frequent values. Let us denote the frequency of the category $x_{ic}$ occurring in the $i$-th object as $f_i(x_{ic})$. When determining similarity between the $c$-the and $d$-th variables for the $i$-th object, $S_{icd} = 1$ if $x_{ic} = x_{id}$, otherwise $S_{icd} = 1/(1 + \ln f_i(x_{ic}) \cdot \ln f_i(x_{ic}))$. Similarity $S_{cd}$ is computed as the arithmetic mean and dissimilarity as $D_{cd} = (1/S_{cd} - 1)$.

The *Lin* measure, which was introduced by Lin (1998), represents an information-theoretic definition of similarity based on relative frequencies. It assigns higher weights to more frequent categories in case of match and lower weights to less frequent categories in case of mismatch: $S_{icd} = 2 \cdot \ln p_i(x_{ic})$ if $x_{ic} = x_{id}$, otherwise $S_{icd} = 2 \cdot \ln(p_i(x_{ic}) + p_i(x_{ic}))$. Total similarity between the $c$-th and $d$-th variables is computed as

$$S_{cd} = \frac{\sum_{i=1}^{n} S_{icd}}{\sum_{i=1}^{n} (\ln p_i(x_{ic}) + \ln p_i(x_{id}))}. \qquad (1)$$

Dissimilarity measure is defined as $D_{cd} = (1/S_{cd} - 1)$.

The *Lin 1* measure represents a variation of the Lin measure derived by Boriah et al. (2008). Its weight system is similar to the original Lin measure, but it is much more complex. Similarity between two variables by the $i$-th object is treated according to the expression

$$S_{icd} = \begin{cases} \sum_{q \in Q} \ln p_i(q) \text{ if } x_{ic} = x_{id} \\ 2 \cdot \ln \sum_{q \in Q} p_i(q) \text{otherwise} \end{cases}, \qquad (2)$$

$Q \subseteq \mathbf{x}_i : \forall q \in Q, p_i(x_{ic}) \leq p_i(q) \leq p_i(x_{id})$ assuming that $p_i(x_{ic}) \leq p_i(x_{id})$, where $q$ represents a subset of categories of the $i$-th object whose relative frequencies are in between relative frequencies of $p(x_{ic})$ and $p(x_{id})$. The range of possible values of the Lin 1 measure is from $-\ln(m/2)$ to 0. The similarity measure across all variables is computed according to Eq. (1) and the dissimilarity measure is defined as $D_{cd} = (1/S_{cd} - 1)$.

Clustering with the above mentioned measures is compared with results obtained using the *overlap* measure (the simple matching coefficient), which takes into account only the information whether two observations match or not. When determining similarity between the $c$-th and $d$-th variables for the $i$-th object, $S_{icd} = 1$ if $x_{ic} = x_{id}$, otherwise $S_{icd} = 0$. Similarity $S_{cd}$ across all objects is computed as the arithmetic mean and Dissimilarity $D_{cd}$ is expressed as a complement to 1. Unlike the previously mentioned similarity measures, the overlap measure does not take into account frequency distribution of categories of a given object, which could serve as an important factor for determining similarity between variables.

The comparison of some of the above mentioned measures applied for object clustering with respect to the within cluster variability is described by Šulc and Řezanková (2014a).

### 2.2 Categorical Principal Component Analysis

Principal component analysis (PCA) and CATPCA are methods that reduce observed variables to a number of uncorrelated latent variables, principal components. CATPCA presents a more general approach, which enables to deal with some limitation of traditional PCA. Most importantly, it enables to deal with nominal and ordinal variables, and furthermore, it can discover nonlinear relationships between variables. When working with categorical data in an analysis, categories have to be quantified in order to compute a correlation matrix, and more generally to work with variance as a concept, e.g. for the VAF (Variance Accounted For) computation, which serves as a model diagnostic. This process is called optimal scaling. The choice of a scale level for a given variable is very important, because it influences the structure of correlation matrix. Thus, it is up to a researcher to choose an appropriate scale level for each variable. For more details about choosing a scale level, see (Linting et al., 2007).

The aim of both methods is to create several principal components containing as much variability in a dataset as possible. The proportion of variability expressed by principal components is VAF. On the contrary to PCA, where latent variables are computed directly from a correlation matrix, CATPCA creates a correlation matrix simultaneously with model creation. It is an iterative process, which converges to a stationary point, see (De Leeuw et al., 1976). If all variables are numeric, the iterative process ends up with the same result as original PCA. The main output of PCA and CATPCA are component loadings coordinates; in two-dimensional solution, they can be displayed in a form of a chart.

When determining the number of principal components, a low number is always preferred. If all variables are highly correlated, one principal component is satisfactory. Usually, two principal components are used. In case better differentiation of entry variables is needed, three- or more-component solution is chosen. For a decision, one can use e.g. VAF, a scree plot, or a component loadings plot. CATPCA, on the contrary to PCA, does not provide nested solutions, i.e. each component solution is stand-alone. Thus, one cannot compare the quality of lower- or higher-component solution based on the computed one; those solutions must be computed as well. The method is available in two major commercial statistical packages: the PRINQUAL procedure in SAS, and the CATPCA procedure in IBM SPSS.

## 3. Approaches Used for Evaluation of Obtained Results

In this paper, quality of final clusters is evaluated from aspects of the WCM (Within-Cluster Mutability) coefficient, proposed by Řezanková et al. (2011) which is an important indicator of cluster quality. With an increasing number of clusters, the within-cluster variability decreases, so the clusters become more homogenous. It takes values from zero to one, where one indicates maximal possible variability. It is expressed by the equation

$$WCM(k) = \frac{1}{n}\frac{h}{h-1}\sum_{g=1}^{k}\frac{m_g}{m}\sum_{i=1}^{n}\left(1-\sum_{u=1}^{h}\left(\frac{m_{giu}}{m_g}\right)^2\right), \qquad (3)$$

where $k$ is the number of clusters, $m_g$ is the number of variables in the $g$-th cluster ($g = 1, 2, ..., k$), $m_{giu}$ is the number of variables in the $g$-th cluster by the $i$-th object with the $u$-th category ($u = 1, 2, ..., h$; $h$ is the number of categories).

With the increasing number of clusters, the value of WCM always decreases. Thus, there arises a question, how to determine the optimal cluster solution, see e.g. (Löster and Pavelka, 2013). By variable clustering, it is usually sufficient to determine the optimal number of clusters from dendrogram analysis. The other possibility is to use the *pseudo F index based on the mutability*:

$$PSFM(k) = \frac{(m-k)(WCM(1) - WCM(k))}{(k-1)WCM(k)}, \tag{4}$$

where $WCM(1)$ expresses the variability in the whole dataset and $WCM(k)$ evaluates the within-cluster variability in the $k$-cluster solution. The highest value should indicate the best solution.

## 4. Practical Application

In this section, comparison of the HCA and CATPCA approaches is performed using two real datasets, which come from the archive of the Institute of Sociology of the Czech Academy of Sciences (http://archiv.soc.cas.cz). Both datasets were prepared on a basis of questionnaire surveys where respondents answered to questions on actual issues.

From the first dataset, the battery of 13 questions concerning security threats for the Czech Republic, such as terrorists, refugees or natural disasters, was chosen. There were four possible answers to all questions: high thread, low thread, no thread, and does not know. On the whole, 1,084 respondents were surveyed. This dataset was named as Threats.

The second dataset comprises 34 questions concerning importance of various life values, such as helping people, having a family or living healthy. All the questions have three possible categories: disagree, agree, and does not know. In this survey, 1,048 respondents answered to all 34 questions. This dataset was named as Life Values.

### 4.1 Dataset Threats

Table 1 presents values of the WCM coefficient for one- to five-cluster solutions using all similarity measures mentioned in Section 2 applied to the Threats dataset. The within-cluster variability in the one-cluster solution, i.e. in the whole dataset, is the same for all the measures; with the increasing number of clusters it decreases. As it is apparent from the table, the examined similarity measures do not differ very much. Thus, similarly to the previous research (Šulc, 2014), it was proven that the clustering performance of the examined similarity measures could not be fully differentiated when there were only a small number of variables. It is interesting though that all the examined measures have the exactly same four-cluster solution; this solution has the highest value of the pseudo F coefficient as well. That means that the highest relative decrease of the within-cluster variability occurs in this cluster solution.

The four-cluster solution proves to be the best one from an aspect of substantive interpretation as well. The analyzed variables can be divided into four groups. The first one comprises foreigners in the Czech Republic (foreigners in CZ, refugees), the second one crime organizations (terrorists, int. organized crime, radical religious movements), the third one political danger (foreign intelligence services, extreme left wing, extreme right wing), and the last one external factors (wars, epidemics, natural disasters, resource crisis, global economic crisis).

Table 1. Values of WCM for HCA using examined similarity measures (the Threats dataset)

| | WCM(1) | WCM(2) | WCM(3) | WCM(4) | WCM(5) |
|---|---|---|---|---|---|
| Eskin | 0.668 | 0.567 | 0.468 | 0.386 | 0.331 |
| Goodall 3 | 0.668 | 0.567 | 0.468 | 0.386 | 0.326 |
| Goodall 4 | 0.668 | 0.564 | 0.465 | 0.386 | 0.331 |
| IOF | 0.668 | 0.568 | 0.468 | 0.386 | 0.326 |
| Lin | 0.668 | 0.567 | 0.468 | 0.386 | 0.326 |
| Lin 1 | 0.668 | 0.564 | 0.465 | 0.386 | 0.331 |
| Overlap | 0.668 | 0.564 | 0.465 | 0.386 | 0.324 |

Source: Own calculation.

For the application of the CATPCA procedure in IBM SPSS, the two-dimensional solution was sufficient. The variable principal normalization method was chosen. The important setting in CATPCA is the choice of a scale level of analyzed variables. The nominal scale was chosen because of the category "does not know", which is not ordinal as the rest of the dataset. The nominal scale has other advantages as well; e.g. in (Linting et al., 2007), there is stated that CATPCA has the most freedom in quantifying a variable, when nominal analysis level is set.

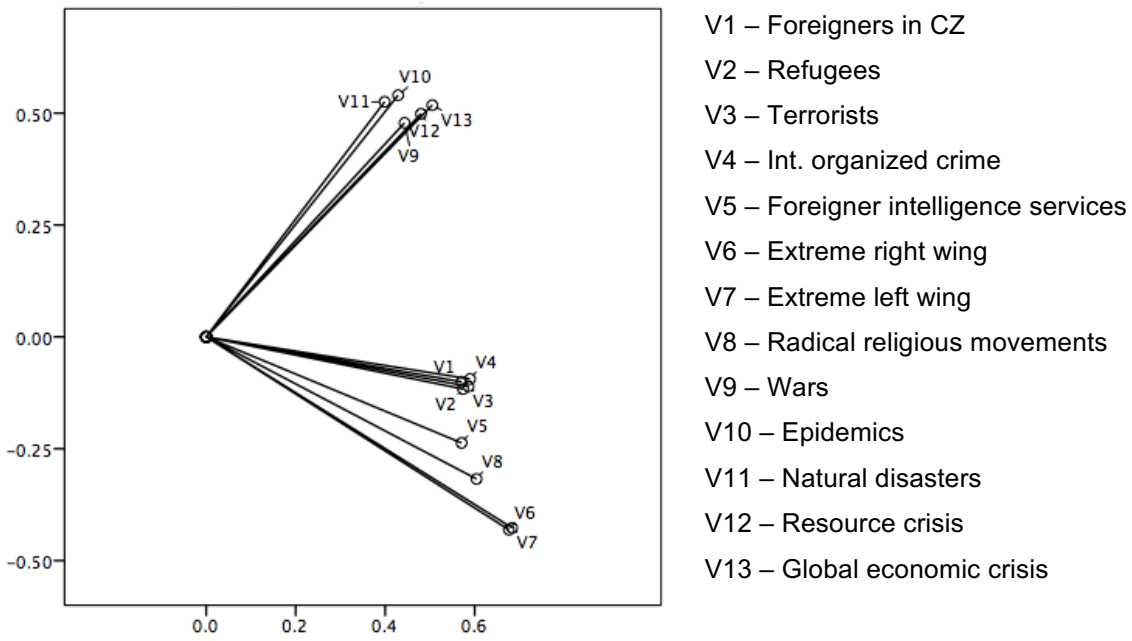


Figure 1. Component loadings plot for the Threats dataset
Source: Own calculation.

The most important output of CATPCA for dimension reduction is a component loadings coordinates table that can be displayed in a form of a plot, which can be eligible particularly when dealing with a lower number of variables. Figure 1 presents such a plot for the Threats dataset. The variables that are situated closely together are closely related as well. The variables which

are in angle approximately 180° are closely negatively related. There is no such variable. Variables in angle approximately 90° are not related. Such situation can be observed e.g. by variables V10 (Epidemics) and V6 or V7 (Extreme left and right wing). In Figure 1, there can be observed several groups of related variables. It depends on a researcher's decision how many clusters to determine. Taking into account the main purpose of the analysis, dimension reduction, the three-cluster solution was chosen. The first one comprises external factors (wars, epidemics, natural disasters, resource crisis, global economic crisis), the second one foreigners (foreigners in CZ, refugees, terrorists, int. organized crime), and the last one political danger (foreign intelligence services, extreme left wing, extreme right wing, radical religious movements).

### 4.2 Dataset Life Values

In the Life Values dataset, 34 categorical variables were clustered with the aim to reduce a dataset size. The same seven similarity measures as in the analysis of the Threats dataset were used. Table 2 presents their clustering performance on one- to five-cluster solutions. Apart from the Goodall 4 measure, which performed very poorly, the other similarity measures do not have distinctive differences in their WCM values. In spite of this, one may observe that those similarity measures can be divided into two groups. The first one consists of the measures Eskin, IOF and overlap, where all three measures provided the exactly same clusters. The second one contains Goodall 3, Lin and Lin 1. According to Table 2, it is impossible to determine, which group provides better results. In our previous research (Šulc, 2014), was shown that the substantial interpretation of dendrograms provided by different similarity measures could be a successful strategy.

Table 2. Values of WCM for HCA using examined similarity measures (the Life Values dataset)

| | WCM(1) | WCM(2) | WCM(3) | WCM(4) | WCM(5) |
|---|---|---|---|---|---|
| Eskin | 0.509 | 0.399 | 0.371 | 0.328 | 0.308 |
| Goodall 3 | 0.509 | 0.401 | 0.362 | 0.336 | 0.310 |
| Goodall 4 | 0.509 | 0.472 | 0.435 | 0.404 | 0.372 |
| IOF | 0.509 | 0.399 | 0.371 | 0.328 | 0.308 |
| Lin | 0.509 | 0.399 | 0.355 | 0.336 | 0.310 |
| Lin 1 | 0.509 | 0.402 | 0.355 | 0.336 | 0.310 |
| Overlap | 0.509 | 0.399 | 0.371 | 0.328 | 0.308 |

Source: Own calculation.

When studying the dendrograms along with WCM values in Table 2, it is apparent that measures Goodall 3, Lin, Lin 1 have better cluster allocation than the rest of the measures. Their clusters are more balanced and meaningful. The rest of the measures focus more on variability minimization, which resulted in higher amount of small clusters, which are usually not very helpful for the purpose of dimension reduction. In the end, the five-cluster solution of Goodall 3, Lin and Lin 1 measures was chosen (they are all same). The clusters are as follows. The first cluster contains variables regarding the respondent's active life (improving life in a place of residence, to foster democracy, protection of nature, general overview, to be informed); the second one consists of variables concerning respondent's influence in society (to enforce party policy, to have own business, to manage people, to have nice things, religious principles, an

important position). The third cluster deals with variables connected with a respondent's job (to have any job, an interesting job, to earn a lot of money, a meaningful job, good work team, to work professionally). The fourth cluster, which consists of only two variables, separates the interesting activities from the rest (a job allowing doing interesting things, interesting life). The five cluster accents the private life (time for hobbies, undisturbed privacy, a content family, to help people in need, to help the family, to live according to own liking, to enjoy, healthy environment, nice environment, to live healthy, to have friends who understand, useful friends, to be popular, to have children, own housing).

Generally, it cannot be determined which similarity measure provides the best results. All of them provide similar results, so their clustering performance may vary on every single dataset. Thus, it is always up to researcher to evaluate created clusters and to decide what the most appropriate measure is.

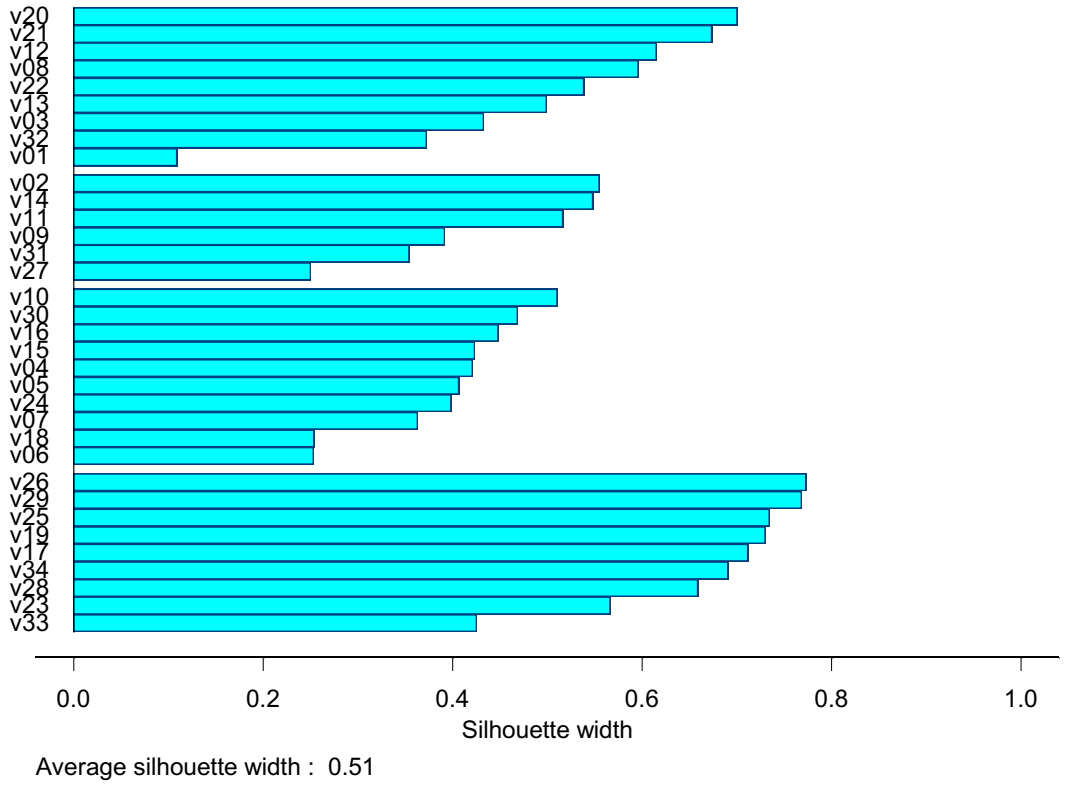


Figure 2. Silhouette plot for the four-cluster solution using fuzzy cluster analysis based on the results of the CATPCA procedure (the Life Values dataset)
Source: Own calculation.

Next, the CATPCA procedure was performed. Similarly as by the Threats dataset, all variables were set to nominal scale level and the variable normalization method to variable principal. The two-dimensional solution proved to be sufficient. The output of CATPCA, component loadings plot, might be confusing to interpret due to the large number of variables. Thus, fuzzy cluster analysis, see (Kaufman and Rousseeuw, 2005), for two- to five-cluster solution was applied to original component loadings coordinates. The resulting clusters are very good; especially, the four-cluster solution has both the interpretable results and the small

number of clusters. The first cluster can be summarized under the term active life (v20 – general overview, v21 – to be informed, v12 – protection of nature, v8 – a job allowing doing interesting things, v22 – interesting life, v13 – to work professionally, v3 – to foster democracy, v32 – to be popular, v1 – improving life in a place of residence). The second cluster exactly matches the corresponding cluster of HCA, which was named respondent's influence in society with variables: v2 – to enforce party policy, v14 – to have nice things, v11 – to manage people, v9 – to have own business, v31 – an important position, and v27 – religious principles. The third cluster deals with variables concerning the respondent's work-life balance (v10 – a good work team, v30 – useful friends, v16 – undisturbed privacy, v15 – time for hobbies, v4 – to have any job, v5 – an interesting job, v24 – to enjoy, v7 – a meaningful job, v18 – to help people in need, v6 – to earn a lot of money). And the fourth cluster is associated with the respondent's family life (v26 – nice environment, v29 – to have friends who understand, v25 – healthy environment, v19 – to help the family, v17 – a content family, v34 – own housing, v28 – to live healthy, v23 – to live according to own liking, v33 – to have children).

Figure 2, a silhouette plot, displays quality of resulting clusters. For each variable, there is a value in the range from –1 to 1. A value close to one indicates a well-clustered variable; a value close to minus one suggests a badly clustered variable, which should be included in a different cluster; a value close to zero shows that a variable is located on the border of two natural clusters, and the value zero marks a one-variable cluster. In Figure 1, most of variables seem to be clustered pretty well, so the final clusters can be considered as good and distinguishable.

## 5. Conclusion

In this paper, we compared two approaches for clustering of categorical variables, hierarchical cluster analysis (HCA) and categorical principal component analysis (CATPCA). For the comparison, we used two economic datasets, which differed by the number of variables. In the smaller Threats dataset, HCA provided better results than CATPCA. The clusters had better substantial interpretation, e.g. they separated well the foreigners living in the Czech Republic from much greater danger, such as terrorists. The CATPCA procedure provided quite good results; however, they were much more dependent on researcher's judgment.

The opposite situation occurred when came to dealing with the larger Life Values dataset. All the examined similarity measures, apart from Goodall 4, provided very similar results from a point of view of the within-cluster variability. To determine, which similarity measure provides the best results, it was needed to use their dendrograms and the researchers' qualified opinion. Moreover, with increasing number of analyzed variables, low number cluster solutions usually do not have any meaningful interpretation, and thus, the solution with more clusters has to be chosen. This is in contradiction to the requirement for dimension reduction. On the contrary to the Threats dataset, the component loadings plot in the Life Values dataset became confused, so we suggested a solution based on clustering component loading coordinates by fuzzy cluster analysis. The results of this approach were surprisingly good from an aspect of substantial interpretation.

Generally, we would recommend the use of HCA when analyzing datasets with smaller number of variables. Among the similarity measures there are no significant differences (apart from Goodall 4), so it is upon researcher to choose a right one for a particular dataset. The CATPCA procedure combined with fuzzy cluster analysis presents a powerful tool for dimension reduction in datasets with higher number of analyzed variables. On the contrary to HCA, it is able to create a small number of clusters in such situations.

**References**

1. BARTHOLOMEW, D. J. et al. 2002. The Analysis and Interpretation of Multivariate Data for Social Scientists. Boca Raton : Chapman & Hall/CRC. 2002. ISBN 1-58488-295-6.
2. BORIAH, S., CHANDOLA, V., KUMAR, V. 2008. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 8th SIAM International Conference on Data Mining. SIAM. 2008, pp. 243-254.
3. DE LEEUW, J., YOUNG, F. W., TAKANE, Y. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. In Psychometrika, 1976, vol. 41, pp. 471-503.
4. ESKIN, E. et al. 2002. A geometric framework for unsupervised anomaly detection. In Applications of Data Mining in Computer Security. 2002, pp. 78-100.
5. KAUFMAN, L., ROUSSEEUW, P. 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken : Wiley. 2005. ISBN 0-471-73578-7.
6. LIN, D. 1998. 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.
7. LINTING, M. et al. 2007. Nonlinear principal components analysis: introduction and application. In Psychological Methods, 2007, vol. 12, no. 3, pp. 336-358.
8. LÖSTER, T., PAVELKA, T. 2013. Evaluating of the results of clustering in practical economic tasks. In The 8th International Days of Statistics and Economics. Slaný : Melandrium. 2013, pp. 804–818.
9. ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D. 2011. Evaluation of categorical data clustering. In Advances in Intelligent Web Mastering 3. Berlin : Springer Verlag. 2011, pp. 173-182.
10. SPARCK-JONES, K. 1972. A statistical interpretation of term specificity and its application in retrieval. In Journal of Documentation, 1972, vol. 28, no. 1, pp. 11-21. Later: In Journal of Documentation, 2002, vol. 60, no. 5, pp. 493-502.
11. ŠULC, Z. Similarity Measures for nominal variable clustering. In The 8th International Days of Statistics and Economics. Slaný : Melandrium. 2014, pp. 1536-1545, http://msed.vse.cz/msed_2014/article/275-Sulc-Zdenek-paper.pdf.
12. ŠULC, Z., ŘEZANKOVÁ, H. 2014a. Evaluation of recent similarity measures for categorical data. In Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wroclaw : Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu. 2014, pp. 249-258, http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf.
13. ŠULC, Z., ŘEZANKOVÁ, H. 2014b. Evaluation of selected approaches to clustering categorical variables. In Statistics in Transition new series, 2014, vol. 15, no. 4, pp. 591-610.
14. VERMUNT, J. K., MAGIDSON, J. 2005. Technical Guide for Latent GOLD 4.0: Basic and Advanced. Belmont Massachusetts : Statistical Innovations Inc. 2005, www.statisticalinnovations.com/products/LGtechnical.pdf.