17th AMSE Applications of Mathematics and Statistics in Economics
International Scientific Conference | Poland • 27-31 August 2014

# EVALUATION OF RECENT SIMILARITY MEASURES FOR CATEGORICAL DATA

**ZDENĚK ŠULC, HANA ŘEZANKOVÁ**
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
email: zdenek.sulc@vse.cz, hana.rezankova@vse.cz

## Abstract

*This paper evaluates recently introduced similarity measures which are based on new approaches and have been proposed for purposes of hierarchical clustering of categorical data. Clustering with these similarity measures is compared to clustering with the simple matching coefficient, and further to alternative methods of categorical data clustering, namely, two-step cluster analysis and latent class analysis. Cluster analysis is applied to economic data. Quality of obtained clusters is evaluated by several indices, which include the normalized Gini coefficient and the normalized entropy, the modified pseudo F indices based on the Gini coefficient and the entropy. The results indicate that clustering with some of recently introduced similarity measures and alternative methods provide better clusters than in case of standardly used simple matching coefficient.*

***Key words:*** *similarity measures, categorical data, cluster analysis.*

DOI: 10.15611/amse.2014.17.27

## 1. Introduction

Cluster analysis is one of important multivariate statistical methods. It is widely used, e.g. in the data analysis from questionnaire surveys, where it helps to identify segments of respondents. Its principle is based on dividing of an examined dataset into several groups according to similarity (or distance) of objects in these groups. When examining a dataset with quantitative variables, the similarity (or distance) measures are well known and they are standardly implemented in statistical software. Different situation occurs when dealing with categorical (nominal) data. There are only several standardly used similarity measures. The best-known is the *simple matching coefficient*, also known as the *overlap* measure, which does not provide good results. There have been proposed many similarity measures for categorical data in recent years; however, none of them was examined properly.

The aim of this paper is to evaluate the clustering with recently introduced similarity measures and compare it to clustering with the *simple matching coefficient* and further to alternative methods for categorical data clustering, namely, *two-step cluster analysis* and *latent class analysis*. Unlike previous reviews, e.g. (Boriah et al., 2008), this paper also compares other than hierarchical methods of clustering. Moreover, different evaluation criteria, which is based on within-cluster variability, are applied for evaluation of clustering results. For the analysis, the data from the EU-SILC survey, which was held in 2011, have been used. Particularly, Czech and Slovak households are going to be compared in their structure. The quality of final clusters is going to be evaluated from a point of view of both within-cluster variability and their economic interpretation.

## 2. Similarity Measures and Methods for Categorical Data Clustering

In this paper, the following similarity measures are going to be evaluated: *Eskin*, *IOF*, *Lin*, *S2* and the *simple matching coefficient*. These measures are applied in hierarchical clustering with the *complete linkage* method. Moreover, the data are analyzed by *two-step cluster analysis* and *latent class analysis*. All formulas in this paper, apart from *S2* formula, are based on data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, ..., n$ ($n$ is the total number of objects); $c = 1, 2, ..., m$ ($m$ is the total number of variables).

### 2.1 Simple Matching Coefficient

The *simple matching coefficient* (hereinafter the *overlap* measure) represents the simplest way for measuring of similarity. When determining the similarity between objects $\mathbf{x}_i$ and $\mathbf{x}_j$, it takes the value 1 for the $c$-th variable in case the objects match and the value 0 otherwise. It is described by the formula

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ 0 \text{ otherwise} \end{cases}. \tag{1}$$

The similarity measure between two objects is then computed as

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}. \tag{2}$$

Every similarity measure can be expressed as a dissimilarity measure. In case of the *overlap* measure, the relationship is given by the expression

$$D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j). \tag{3}$$

The *overlap* measure is wildly used because of its simplicity. On the other hand, this measure neglects important characteristics in a dataset, such a number of categories or frequencies of categories of a given variable. These characteristics can serve for better formulation of similarity between given objects. Recently introduced measures try to deal with this drawback.

### 2.2 Recent Similarity Measures

The *Eskin* measure was proposed by Eskin et al. (2002). Its basic idea is to assign higher weights to mismatches by variables with the higher number of categories. The similarity between two objects for the $c$-th variable is then expressed as

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ \dfrac{n_c^2}{n_c^2 + 2} \text{ otherwise} \end{cases}, \tag{4}$$

where $n_c$ is a number of categories of the $c$-th variable. Equation (2) can be used for computation of the similarity between two objects. The dissimilarity measure is computed as

$$D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{S(\mathbf{x}_i, \mathbf{x}_j)} - 1. \tag{5}$$

The *inverse occurrence frequency* (*IOF*) treats mismatches of more frequent categories by lower weights, i.e.

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{jc})} \text{ otherwise} \end{cases}, \tag{6}$$

where $f(x_{ic})$ expresses a frequency of the value $x_{ic}$ of the $c$-th variable. The similarity and dissimilarity measure can be computed by using Equations (2) and (5).

The *Lin* measure, which was introduced by Lin (1998), assigns higher weights to more frequent categories in case of matches and lower weights to less frequent categories in case of mismatches, i.e.

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 2 \cdot \ln p(x_{ic}) \text{ if } x_{ic} = x_{jc} \\ 2 \cdot \ln(p(x_{ic}) + p(x_{jc})) \text{ otherwise} \end{cases}, \tag{7}$$

where $p(x_{ic})$ expresses a relative frequency of the value $x_{ic}$ of the $c$-th variable. The similarity measure between two objects is computed as

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m} (\ln p(x_{ic}) + \ln p(x_{jc}))} \tag{8}$$

and the dissimilarity measure according to Equation (5).

The *S2* measure was proposed by Morlini and Zani (2012) and it is based on a different approach than the previous measures. For its computation a transformation of the initial matrix $\mathbf{X} = [x_{ic}]$ is needed. If the $c$-th variable has at least two categories, $K_c \geq 2$, then for each category one dummy variable is created. The total number of variables in the newly arisen matrix can be computed as a sum of $K_c$ over $m$ original variables. The similarity measure can be computed by using the formula

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} \sum_{u=1}^{K_c} \tau(i,j)_{cu} \ln\left(\frac{1}{f_{cu}^2}\right)}{\sum_{c=1}^{m} \sum_{u=1}^{K_c} \tau(i,j)_{cu} \ln\left(\frac{1}{f_{cu}^2}\right) + \sum_{c=1}^{m} \phi(i,j)_c \sum_{u=1}^{K_c} f_{cu} \ln\left(\frac{1}{f_{cu}^2}\right)}, \tag{9}$$

where $u = 1, 2, \ldots K_c$ is an index of the $u$-th dummy variable of the $c$-th original variable, $f_{cu}^2$ is the second power of the frequency of the $u$-th category of the original $c$-th variable. When using Equation (9), the following conditions have to be held:

$$\tau(i,j)_{cu} = 1 \text{ if } x_{icu} = 1 \text{ and } x_{jcu} = 1,$$
$$\tau(i,j)_{cu} = 0 \text{ otherwise,}$$

$$\phi(i,j)_c = 1 \text{ if } x_{icu} = 1 \cap x_{jct} = 1 \text{ and } x_{ict} = 1 \cap x_{jcu} = 0, \text{ for } u \neq t,$$
$$\phi(i,j)_c = 0 \text{ otherwise.}$$

The *S2* measure takes values from 0 to 1, where 1 indicates maximal similarity, i.e. both objects are identical. The system of weights, which is based on the information theory, assigns higher importance to rarely observed values.

### 2.3 Methods of Cluster Analysis

The similarity measures, which have been introduced above, can be applied in hierarchical cluster analysis. It is based on a proximity matrix, which contains distances among all objects in the dataset. At the beginning, each case 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. For the purpose of this paper, the *complete linkage* method has been used. In this method, distance between two clusters is defined as the distance between two furthest objects from the considered clusters.

An alternative way to categorical data clustering is based on the *log-likelihood distance*. On the contrary to previously mentioned measures, this measure is used in *two-step cluster analysis* (*2STEP*), which is implemented in the statistical software IBM SPSS Statistics. Its algorithm consists of two steps. In the first one, preliminary clusters are created sequentially. In the second step, the clustering algorithms are applied to the preliminary clusters. The full description of *two-step cluster analysis* can be found in (SPSS, Inc., 2014). The advantages of this method include the ability to cluster both quantitative and categorical data. Also, this method is much less time consuming in a comparison to standard hierarchical analysis, especially in case of larger datasets. On the other hand, the created clusters depend on the initial object order. Therefore, it is recommended to order objects randomly before the start of an analysis.

Another approach to categorical data clustering represents *latent class analysis* (*LCA*), which is based on a latent variable consisting of discrete and mutually exclusive latent classes. This latent variable can only be measured indirectly by using two or more observable (also manifest) categorical variables. For each object, the probability of belonging to each latent class is then computed. The object is assigned into the latent class with the highest probability.

*Two-step cluster analysis* and *latent class analysis* are complex methods. Both enable to adjust a lot of parameters, so their final clusters might differ. Because these methods are not the primary aim of the research, they serve for reference purposes only, the standard setting of all their parameters were used.

## 3. Evaluation Criteria of Clusters

The quality of final clusters is going to be evaluated by using indices based on within-cluster variability and by the modified *pseudo F indices*, which are explained e.g. in (Řezanková et al., 2011).

The within-cluster variability is an important indicator of cluster quality. With the increasing number of clusters, the within-cluster variability decreases. Clustering with a certain similarity measure with the greatest degrease is considered to be the best, because its clusters are the most homogenous. The *normalized Gini coefficient* and the *normalized entropy* have been chosen for purposes of determining the within-cluster variability.

The *Gini coefficient*, also known as the *nomvar*, measures the variability of nominal variables. For the *c*-th variable in the *g*-th cluster ($g = 1, 2, ..., k$) it can be expressed as

$$G_{gc} = 1 - \sum_{u=1}^{K_c} \left( \frac{n_{gcu}}{n_g} \right)^2, \tag{10}$$

where $n_g$ is a number of objects in the *g*-th cluster, $n_{gcu}$ is a number of objects in the *g*-th cluster by the *c*-th variable with the *u*-th category ($u = 1, 2, .., K_c$; $K_c$ is a number of categories of the *c*-th variable). The normalized within-cluster variability for *k* cluster solution based on the *Gini coefficient* for *m* variables can be expressed as

$$Gnorm(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{K_c}{K_c - 1} G_{gc}, \tag{11}$$

which takes values from 0 to 1, where lower values indicate more homogenous clusters.

An alternative way to express the variability is the *entropy*, which can be expressed (for the *c*-th variable in the *g*-th cluster, $n_{gcu} > 0$) as

$$H_{gc} = -\sum_{u=1}^{K_c} \left( \frac{n_{gcu}}{n_g} \ln \frac{n_{gcu}}{n_g} \right). \tag{12}$$

To get within-cluster variability for *k* clusters, the *normalized entropy* can be used:

$$Hnorm(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{H_{gc}}{\ln K_c}. \tag{13}$$

The *pseudo F indices* were developed to determine an optimal number of clusters. They are based on the F statistic. A cluster solution with the highest value of this statistics is considered to be the best.

For *k*-cluster solution, the *pseudo F index based on the Gini coefficient* has the formula

$$PSFG(k) = \frac{(n-k)(G(1) - G(k))}{(k-1)G(k)}, \tag{14}$$

where $G(1)$ expresses variability in the whole dataset and $G(k)$ evaluates within-cluster variability in the *k*-cluster solution, both based on the *Gini coefficient*.

The *pseudo F coefficient based on the entropy* is defined by the formula

$$PSFH(k) = \frac{(n-k)(H(1) - H(k))}{(k-1)H(k)}, \tag{15}$$

where $H(1)$ is the *entropy* in the whole dataset and $H(k)$ the *within-cluster entropy* in the *k*-cluster solution.

## 4. Application to Economic Data

In this paper, data from the EU-SILC (*European Union – Statistics on Income and Living Conditions*) survey, held in 2011, were used. The aim of the analysis is to compare the structure of the Czech and Slovak households from a point of view of their financial possibilities and used durables. Studies of the Czech and Slovak households with respect to material deprivation and poverty were published e.g. by Bartošová and Želinský (2013), Želinský (2012). Cluster analyses of the Czech households according to some durables were applied in (Řezanková and Löster, 2013). Six following categorical variables were used for the purpose of this paper; their categories are displayed in brackets: *Capacity to afford paying for one week annual holiday away from home* [*yes, no*], *Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day* [*yes, no*], *Capacity to face unexpected financial expenses* [*yes, no*], *Do you have a computer* [*yes, no – cannot afford, no – other reason*], *Do you have a car* [*yes, no – cannot afford, no – other reason*], *Ability to make ends meet* [*with great difficulty, with difficulty, with some difficulty, fairly easily, easily, very easily*]. In total, the data come from 8866 Czech and 5193 Slovak households.

The analysis consists of two steps. In the first one, the clusters of households are going to be computed and evaluated according to measures and methods, which were introduced in Section 2. In the second step, the final clusters created with selected measures and methods are going to be evaluated from a point of view of their economic interpretation.

The similarity measures, which have been introduced in Section 2, are not part of statistical systems, so they had to be programmed in the *Matlab* software. The proximity matrices based on particular similarity measure have been processed in IBM SPSS Statistics. The *complete linkage* method has been used. The entire *two-step cluster analysis* was performed in IBM SPSS Statistics and *latent class analysis* in the *LatentGold* software. Solutions for two to six clusters have been computed. At the end, the evaluation criteria have been computed using the *Matlab* software for all the measures and the methods.

### 4.1. Czech Household Data

Table 1 summarizes evaluation criteria for clustering with individual similarity measures and for the other methods based on the Czech household data. The quality of a particular cluster solution can be evaluated according to the within-cluster variability. The faster decrease of within-cluster variability, the more homogenous clusters are. The fastest decrease of variability is by the *Eskin* measure; its variability decreases from 0.803 to 0.387 according to the *Gnorm* measure and from 0.815 to 0.376 according to the *Hnorm* measure. Throughout all cluster solutions of the *Eskin* measure, the variability is smaller in a comparison to other approaches of clustering. That means that clustering using this measure provides the best cluster solutions. It is followed by the *IOF* measure, which also performs very well. The *overlap* measure, which is commonly used, ends up on the third place, with a significant distance from the first two measures. It is closely followed by the last two measures *Lin* and *S2*. Both reference methods, *2STEP* and *LCA*, have the results in the middle of the range. They would have been placed behind the *IOF* measure in the following order: *LCA* and *2STEP*.

According to the optimal number of clusters, which is marked boldly, the most of similarity measures and methods prefer the two-cluster solution. There are two exceptions

though; the *overlap* measure prefers five-cluster solution and the *S2* measure three- or four-cluster solutions.

Table 1. Evaluation criteria for clustering with examined similarity measures and for other methods for the Czech household data

| | # of clusters | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Overlap | Gnorm | 0.803 | 0.756 | 0.630 | 0.606 | 0.513 | 0.494 |
| | Hnorm | 0.815 | 0.768 | 0.650 | 0.619 | 0.516 | 0.495 |
| | PSFG | | 464.117 | 1002.785 | 814.930 | **1034.185** | 926.993 |
| | PSFU | | 434.004 | 835.242 | 727.061 | **977.649** | 881.846 |
| Eskin | Gnorm | 0.803 | 0.562 | 0.494 | 0.439 | 0.433 | 0.387 |
| | Hnorm | 0.815 | 0.572 | 0.494 | 0.433 | 0.426 | 0.376 |
| | PSFG | | **3023.718** | 2133.429 | 1834.565 | 1408.003 | 1372.450 |
| | PSFH | | **2680.094** | 1930.558 | 1647.784 | 1269.836 | 1222.209 |
| IOF | Gnorm | 0.803 | 0.562 | 0.495 | 0.438 | 0.437 | 0.397 |
| | Hnorm | 0.815 | 0.572 | 0.495 | 0.433 | 0.431 | 0.391 |
| | PSFG | | **3023.72** | 2129.03 | 1837.92 | 1385.53 | 1384.18 |
| | PSFH | | **2680.09** | 1919.72 | 1649.82 | 1248.20 | 1252.12 |
| Lin | Gnorm | 0.803 | 0.691 | 0.680 | 0.587 | 0.534 | 0.509 |
| | Hnorm | 0.815 | 0.691 | 0.675 | 0.586 | 0.523 | 0.499 |
| | PSFG | | **1177.471** | 675.170 | 1009.432 | 1057.670 | 979.813 |
| | PSFH | | **1149.231** | 718.234 | 1013.031 | 1125.325 | 1033.066 |
| S2 | Gnorm | 0.803 | 0.724 | 0.649 | 0.586 | 0.549 | 0.543 |
| | Hnorm | 0.815 | 0.740 | 0.671 | 0.612 | 0.581 | 0.573 |
| | PSFG | | 1082.297 | 1254.480 | **1278.702** | 1227.871 | 1024.126 |
| | PSFH | | 1277.676 | **1455.767** | 1452.600 | 1372.480 | 1157.129 |
| 2STEP | Gnorm | 0.803 | 0.565 | 0.521 | 0.462 | 0.441 | 0.417 |
| | Hnorm | 0.815 | 0.585 | 0.527 | 0.465 | 0.446 | 0.421 |
| | PSFG | | **3074.066** | 2048.426 | 1872.780 | 1629.937 | 1500.663 |
| | PSFH | | **2739.283** | 1969.637 | 1791.560 | 1597.424 | 1477.998 |
| LCA | Gnorm | 0.803 | 0.563 | 0.507 | 0.467 | 0.428 | 0.420 |
| | Hnorm | 0.815 | 0.586 | 0.533 | 0.483 | 0.456 | 0.431 |
| | PSFG | | **3106.425** | 2242.338 | 1885.550 | 1728.027 | 1445.836 |
| | PSFH | | **2750.016** | 1984.105 | 1741.828 | 1493.459 | 1358.503 |

In the second step of the analysis, the final clusters are going to be evaluated from a point of view of the economic interpretation. For this comparison, the best five cluster solutions have been chosen according to their quality in the first step: *Eskin*, *IOF*, *LCA*, *2STEP* and *overlap*.

In the two-cluster solution, clustering with the use of both *Eskin* and *IOF* measures provide the exactly same results. Households are separated into two groups, wealthier (58 %) and poorer (42 %). When creating the clusters, the key importance has the variable *Capacity to face unexpected financial expenses*. All answers *yes* to this question are assigned into the first cluster and all answers *no* into the second one. The three cluster solution is very interesting in case of the *Eskin* measure. The newly created cluster contains households, which cannot afford a holiday, but they have capacity to face unexpected financial expenses. These

households are moderately wealthy; however, they cannot spend their financial reserves on unnecessary expenses.

The two-cluster solution, provided by *LCA*, differentiates pretty well between wealthier (57 %) and poorer (43 %) households. Interesting results are provided by the three-cluster solution as well. The cluster of poorer households has been further separated into other two groups. One describes poor households and the other one the households which are little bit wealthier, but they do not own a car or a computer from other reasons. It would be interesting to describe this group of households in detail; however, the data from the EU-SILC survey does not allow it.

*2STEP* has the most similar clusters according their size. The ratio of wealthier households is 54 % and the poorer ones 46 %. However, their differentiation is worse than by *LCA*. Its three-cluster solution can be interpreted in a similar way as by *LCA*, but again, the differentiation is much poorer.

The clusters, which come from the *overlap* measure, are very unbalanced (wealthier households occupy 85 % of objects and poorer ones 15 %). Moreover, the boundary between the clusters is very fuzzy, which makes them the most inappropriate of all examined ones.

Although the resulting clusters of *LCA* have not been considered as the best in the first step of analysis, they have proven their quality when confronted to their interpretation, which has been the best among all measures and methods, in the second step.

### 4.2. Slovak Household Data

Table 2 contains evaluation criteria for the Slovak household data analyses. Similarly to the results based on Czech data, clustering with the *Eskin* measure provides generally the best results across all other measures and methods. However, the *IOF* measure has the best results in two-cluster solution, which proves to be the optimal one for almost all measures, except for the *Lin* measure. On the whole, when dealing with all measures and methods together, their order is following: *Eskin*, *2STEP*, *IOF*, *LCA*, *Lin*, *overlap* and *S2*.

In the second step of the analysis, the same measures and methods as for the Czech households are going to be examined. The two-cluster solution of the *Eskin* measure creates two groups of households, which could be considered as wealthier (64 %) and poorer (36 %). The differentiation is poorer than by the Czech households. The three-cluster solution of this measure behaves in the same way as in the Czech household data, i.e. it contains households, which cannot afford a holiday, but have capacity to face unexpected financial expenses.

The *IOF* measure has very good results in the two-cluster solution. It assigns 56 % of Slovak household to be wealthier and 44 % to be poorer. Thus, the clusters are of similar size and their differentiation is also good.

Despite the good results in the first part of analysis, *2STEP* does not provide very good clusters on this dataset. The two-cluster solution separates the households into wealthier (45 %) and poorer (55 %). The classification is not as good as by the *Eskin* measure or *LCA*.

*LCA* provides very good clusters in the two-cluster solution, they are even better than clusters provided by the *Eskin* measure. There is 56 % of wealthier households and 44 % of poorer in Slovakia. In the same way as in case of the Czech households, the three-cluster solution also separates the households which do not own durables from other reasons.

The final clusters provided by the *overlap* measure are slightly better than in case of the Czech households, but still, they are very insufficient because of their poor differentiation. The ratio of the wealthier households is 65 % and the poorer ones 35 %.

Table 2. Evaluation criteria for clustering with examined similarity measures and for other methods for the Slovak household data

| | # of clusters | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Overlap | Gnorm | 0.846 | 0.669 | 0.642 | 0.563 | 0.534 | 0.500 |
| | Hnorm | 0.855 | 0.682 | 0.657 | 0.581 | 0.556 | 0.518 |
| | PSFG | | **1213.799** | 741.611 | 751.705 | 650.874 | 609.457 |
| | PSFU | | **1123.837** | 685.808 | 669.065 | 567.443 | 530.676 |
| Eskin | Gnorm | 0.846 | 0.630 | 0.516 | 0.448 | 0.419 | 0.387 |
| | Hnorm | 0.855 | 0.640 | 0.517 | 0.449 | 0.418 | 0.384 |
| | PSFG | | **1439.168** | 1296.432 | 1163.495 | 981.570 | 893.556 |
| | PSFH | | **1268.939** | 1152.934 | 1012.660 | 848.886 | 766.809 |
| IOF | Gnorm | 0.846 | 0.605 | 0.566 | 0.548 | 0.457 | 0.453 |
| | Hnorm | 0.855 | 0.612 | 0.566 | 0.548 | 0.457 | 0.453 |
| | PSFG | | **1683.572** | 1028.116 | 767.786 | 858.245 | 699.469 |
| | PSFH | | **1499.943** | 936.213 | 699.588 | 753.724 | 617.284 |
| Lin | Gnorm | 0.846 | 0.785 | 0.635 | 0.610 | 0.578 | 0.484 |
| | Hnorm | 0.855 | 0.795 | 0.649 | 0.620 | 0.584 | 0.477 |
| | PSFG | | 426.171 | **808.045** | 639.165 | 585.180 | 709.009 |
| | PSFH | | 415.071 | **753.809** | 632.024 | 595.899 | 728.655 |
| S2 | Gnorm | 0.846 | 0.752 | 0.732 | 0.629 | 0.592 | 0.587 |
| | Hnorm | 0.855 | 0.762 | 0.740 | 0.653 | 0.626 | 0.621 |
| | PSFG | | **732.199** | 453.244 | 699.282 | 671.351 | 552.808 |
| | PSFH | | **855.585** | 554.863 | 786.693 | 727.244 | 600.372 |
| 2STEP | Gnorm | 0.846 | 0.609 | 0.542 | 0.500 | 0.455 | 0.429 |
| | Hnorm | 0.855 | 0.620 | 0.557 | 0.506 | 0.466 | 0.446 |
| | PSFG | | **1674.963** | 1281.085 | 1072.374 | 1008.041 | 946.544 |
| | PSFH | | **1507.707** | 1154.575 | 1010.648 | 941.556 | 884.983 |
| LCA | Gnorm | 0.846 | 0.611 | 0.556 | 0.496 | 0.475 | 0.454 |
| | Hnorm | 0.855 | 0.631 | 0.578 | 0.517 | 0.498 | 0.475 |
| | PSFG | | **1708.694** | 1220.549 | 1127.542 | 933.846 | 825.988 |
| | PSFH | | **1535.605** | 1109.201 | 1030.682 | 843.284 | 749.679 |

## 5. Conclusion

In this paper, six similarity measures for categorical data clustering were evaluated. Clustering performance with these measures was compared to the performance of two alternative methods for categorical data clustering. There were two main fields of comparison. Firstly, the final cluster solutions of measures and methods were compared from a point of view of within-cluster variability; secondly, from a point of view of the economic interpretation. The Czech and Slovak household data from survey EU-SILC 2011 were used.

In both datasets, the best clusters were provided by hierarchical clustering with the *Eskin* measure from a point of view of the within-cluster variability. The order of other measures and methods differs. In the Czech household dataset it is: *IOF*, *LCA*, *2STEP*, *overlap*, *Lin* and *S2*, whereas in the Slovak household dataset it is *2STEP*, *IOF*, *LCA*, *Lin*, *overlap* and *S2*. The order is not surprising; in some our previous researches, e.g. in (Šulc, 2014), the similarity measures *Eskin* and *IOF* performed very well in datasets with the simple structure similar to the one used in this paper.

In the second part of the analysis, the examined measures and methods were evaluated from a point of view of their economic interpretation. In both datasets, *latent class analysis* differentiated the wealthier and poorer households at the best. According to this method, the ratio of wealthier and poorer households in the Czech Republic and in Slovakia is almost the same. Good results were also provided by the *Eskin* and the *IOF* measures. When comes to three-cluster solution, only clusters provided by the *Eskin* measure and *LCA* have the economic interpretation. On a basis of the findings presented in the paper, one might recommend the use of *LCA* for clustering the EU-SILC data and similar surveys. The use of the *complete linkage* method with the *Eskin* measure can be considered as a good alternative.

**Acknowledgements**

This work was supported by the University of Economics, Prague under Grant IGA F4/104/2014 and by the Slovak Scientific Grant Agency as part of the research project VEGA 1/0127/11 Spatial Distribution of Poverty in the European Union. The EU-SILC datasets were made available for the research on the basis of contract no. EU-SILC/2011/33, signed between the European Commission, Eurostat, and the Technical University of Kosice. Eurostat has no responsibility for results and conclusions which are those of the researcher. The authors would like to thank to Adam Mohammad for his advice on the *Matlab* software.

**References**

1. BARTOŠOVÁ, J., ŽELINSKÝ, T. 2013. The extent of poverty in the Czech and Slovak Republic 15 years after the split. In Post-Communist Economies, 2013, vol. 25, iss. 1, pp. 119-131.LE, S. Q., HO, T. B. 2005. An association-based dissimilarity measure for categorical data. In Pattern Recognition Letters, 2005, vol. 26, iss. 16, pp. 2549-2557.
2. BORIAH, S., CHANDOLA, V., and KUMAR, V. 2008. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, pp. 243-254.
3. ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S., V. 2002. A geometric framework for unsupervised anomaly detection. In D. Barbará and S. Jajodia, editors, Applications of Data Mining in Computer Security, pp. 78-100.
4. 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.
5. MORLINI, I., ZANI, S. 2012. A new class of weighted similarity indices using polytomous variables. In Journal of Classification, 2012, vol. 29, iss. 2, pp. 199-226.
6. ŘEZANKOVÁ, H., LÖSTER, T. 2013. Cluster analysis of households characterized by categorical indicators. In Ekonomie a Management, 2013, vol. 16, iss. 3, pp. 139-147.
7. 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.
8. SPSS, Inc. 2014. Help. Chicago, IL : SPSS, Inc., 2014.
9. ŠULC, Z. 2014. Porovnání nových přístupů ve shlukování nominálních dat. In Sborník prací vědeckého semináře doktorského studia FIS VŠE. Praha : Oeconomica, 2014, pp. 214-223.
10. ŽELINSKÝ, T. 2012. Changes in Relative Material Deprivation in Regions of Slovakia and the Czech Republic. In Panoeconomicus, 2012, vol. 59, iss. 3, pp. 335-353.