Metodološki zvezki, Vol. 14, No. 1, 2017, 37–48

Evaluation of the Gower Coefficient Modifications in Hierarchical Clustering

Zdeněk Šulc[1], Martin Matějka[2], Jiří Procházka[3], Hana Řezanková[4]

Abstract

This paper thoroughly examines three recently introduced modifications of the Gower coefficient, which were determined for data with mixed-type variables in hierarchical clustering. On the contrary to the original Gower coefficient, which only recognizes if two categories match or not in the case of nominal variables, the examined modifications offer three different approaches to measuring the similarity between categories. The examined dissimilarity measures are compared and evaluated regarding the quality of their clusters measured by three internal indices (Dunn, silhouette, McClain) and regarding their classification abilities measured by the Rand index. The comparison is performed on 810 generated datasets. In the analysis, the performance of the similarity measures is evaluated by different data characteristics (the number of variables, the number of categories, the distance of clusters, etc.) and by different hierarchical clustering methods (average, complete, McQuitty and single linkage methods). As a result, two modifications are recommended for the use in practice.

1 Introduction

In various real-life fields where cluster analysis is commonly used, such as biology, social sciences, or marketing surveys, datasets with both quantitative and categorical variables are often applied. This type of data is referred as *mixed data*. However, the majority of multivariate methods including cluster analysis does not allow using the data with different scales as an input for the analysis. According to Podani (1999), there are three possible ways to solve this problem. The first one lies in downscaling (reducing some information of selected variables) or upscaling (incorporating some additional information to selected variables) all variables to the same level, see (Anderberg, 1973). The second way is to analyze variables of each scale separately and to synthesize the results, see (Gordon, 1981). The third way is based on a (dis)similarity measure which can deal both

[1] Department of Statistics and Probability, University of Economics, Prague, Czech Republic; zdenek.sulc@vse.cz

[2] Department of Statistics and Probability, University of Economics, Prague, Czech Republic; martin.matejka@vse.cz

[3] Department of Statistics and Probability, University of Economics, Prague, Czech Republic; jiri.prochazka@vse.cz

[4] Department of Statistics and Probability, University of Economics, Prague, Czech Republic; hana.rezankova@vse.cz

with quantitative and categorical variables. Such a measure can be used in a proximity matrix in hierarchical cluster analysis (HCA) or other multivariate methods, e.g. multidimensional scaling. Since this approach does not need a transformation of a dataset nor handling with the input data, it became very popular.

There have also been introduced several partitioning clustering approaches to mixed-type datasets. For instance, Huang (1998) introduced k-prototype algorithm as an extension to the original k-modes algorithm. There have been introduced many fuzzy clustering approaches, e.g. fuzzy prototype k-means algorithm, see (Ahmad and Dey, 2007) or symbolic dissimilarity approach introduced in (Yang et al., 2004).

This paper focuses on dissimilarity measures for mixed data in HCA. The best-known dissimilarity measure of this type is the *Gower coefficient* (Gower, 1971). However, since its introduction, new approaches to (dis)similarity determination, especially, concerning the nominal variables, were developed. The part of the Gower coefficient which deals with the nominal variables treats the similarity between two categories of a certain variable by one if the categories match and zero otherwise. This is a very simplistic approach. Therefore, the aim of the paper is an evaluation of three Gower coefficient modifications introduced in (Šulc et al., 2016), which were developed to improve dissimilarity determination between categories of nominal variables. All the dissimilarity measures are compared and evaluated from their clustering performance in hierarchical cluster analysis (HCA) with four different clustering methods. The analysis is performed on generated datasets from aspects of internal and external evaluation criteria. The generated data enable to examine the influence of different dataset characteristics on clustering performance of the examined similarity measures.

The paper is organized as follows. Section 2 presents all the examined dissimilarity measures and Section 3 the methods of HCA; Section 4 demonstrates used evaluation criteria. An experimental part of the paper occurs in Section 5, and the results are summarized in Conclusion.

2 Modifications of the Gower Coefficient

The Gower coefficient (Gower, 1971) was originally introduced as a similarity measure. However, for purposes of HCA, it is usually expressed in the form of a dissimilarity measure. Let us assume the data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, \ldots, n$ (n is the total number of objects) and $c = 1, 2, \ldots, v$ (v is the total number of variables). Then, the dissimilarity between the objects $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iv}]$ and $\mathbf{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jv}]$, which are characterized by values of mixed-type variables, is expressed using the formula

$$d_G\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{c=1}^{v} d_{ijc}$$

where d_{ijc} is a dissimilarity measure between the i-th and j-th objects by the c-th variable. The formula requires a dataset with excluded objects containing missing values.

If the c-th variable is nominal (or alternative), the dissimilarity between two categories x_{ic} and x_{jc} is treated as zero for matches of categories, and as one otherwise. If the c-th variable is quantitative, dissimilarity is expressed by the formula

$$d_{ijc} = \frac{|x_{ic} - x_{jc}|}{\max(x_c) - \min(x_c)}. \tag{2.1}$$

If the c-th variable is ordinal, all the categories are transformed based on the formula

$$x_{ic} = \frac{r_{ic} - 1}{R_c - 1}, \tag{2.2}$$

where r_{ic} is the rank number of the i-th ordinal category ($r = 1, \dots, R_c$), and the R_c is the maximal rank number of the c-th variable. After this transformation, the outcome values can be used in equation (2.1) for quantitative variables.

In this paper, three modifications of the Gower dissimilarity coefficient are compared: Gower_IOF and Gower_LIN and Gower_VE, which were introduced in (Šulc et al., 2016). All the modifications aspire to improve the part concerning the nominal variables of this coefficient. They are based on similarity measures for nominal data, which use additional characteristics about a nominal variable in comparison to the classic simple matching approach used in the Gower coefficient. The parts for quantitative and ordinal variables remained unchanged by all the modifications, i.e. they are computed using equation (2.1) and equation (2.2).

The first modification, the *Gower_IOF* (G_IOF) dissimilarity measure, is based on the *Inverse Occurrence Frequency* (IOF) measure, which was originally introduced in (Sparck-Jones, 1972) as a similarity measure for information retrieval. Its part for nominal variables assigns higher weights to less frequent mismatches as it is expressed by the formula

$$d_{ijc} = \begin{cases} 0 & \text{if } x_{ic} = x_{jc} \\ 1 - \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{jc})} & \text{otherwise} \end{cases},$$

where $f(x_{ic})$ is an absolute frequency of the value x_{ic} in the c-th variable. The dissimilarity measure takes the value zero in the case of a match of categories, and the values from zero to the number $(1 - 1/(1 + \ln(n/2)^2))$ otherwise. The upper border of this range converges to one with the increasing dataset size n.

The *Gower_LIN* (G_LIN) measure is inspired by the LIN measure, see (Lin, 1998). The part concerning the nominal variables assigns higher weights to mismatches to less frequent categories with the formula

$$d_{ijc} = \begin{cases} 0 & \text{if } x_{ic} = x_{jc} \\ 1 - \dfrac{2 \cdot \ln(p(x_{ic}) + p(x_{jc}))}{\ln p(x_{ic}) + \ln p(x_{jc})} & \text{otherwise} \end{cases},$$

where $p(x_{ic})$ is a relative frequency of the value x_{ic} in the c-th variable. It takes values from zero to one; the value one is obtained if there are only two categories in a variable, and its limit is close to zero if both the observed categories have very small relative frequencies.

The *Gower_VE* (G_VE) measure is based on the *Variable Entropy* (VE) similarity measure introduced in (Šulc and Řezanková, 2015). Its part for nominal variables assigns higher values of the dissimilarities d_{ijc} to the matches in variables with the high variability expressed by the normalized entropy because the matches in such variables are rarer than the matches in the low-variability variables. The measure can be expressed by the formula

$$d_{ijc} = \begin{cases} 1 + \frac{1}{\ln K} \sum_{u=1}^{K_c} p_u \ln p_u & \text{if } x_{ic} = x_{jc} \\ 1 & \text{otherwise} \end{cases},$$

where u ($u = 1, \ldots, K_c$) is a category in the c-th variable, and p_u is a relative frequency of the u-th category in the c-th variable. The dissimilarity measure takes values from zero to one. Values closer to zero indicate lower dissimilarity between the examined categories.

3 Methods of Cluster Analysis

Based on between-cluster distances, four linkage methods of HCA are examined in this paper: the average method, the complete method, McQuitty's method and the single method.

The *average linkage method* takes average pairwise dissimilarity between objects in two different clusters. Distance between the m-th and g-th clusters can be calculated using the formula

$$d(m,g) = \frac{n_h \cdot d(h,g) + n_l \cdot d(l,g)}{n_m},$$

where the m-th cluster was created by merging of the h-th and l-th clusters, and n_h, n_l and n_m are the numbers of objects in the h-th, l-th and m-th clusters.

The *complete linkage method* considers a dissimilarity between two clusters as the dissimilarity between two farthest objects between them. This between-cluster distance usually provides compact clusters with approximately equal diameters. It can be expressed by the formula

$$d(m,g) = \max\left[d(h,g), d(l,g)\right].$$

The *McQuitty's linkage method* defines the distance between two clusters as the arithmetic mean as it is described by the formula

$$d(m,g) = \frac{d(h,g) + d(l,g)}{2}.$$

The *single linkage method* uses dissimilarity between two closest objects of two different clusters. A formula of this algorithm can be expressed as

$$d(m,g) = \min\left[d(h,g), d(l,g)\right].$$

4 Evaluation Criteria

When evaluating clusters produced by cluster analysis, it is more natural to use internal evaluation indices, which use only information from the clustered datasets. This type of indices is usually based either on minimization of the inter-cluster distance, such as the

Dunn index or on the average silhouette width as the silhouette index. In some papers, the (dis)similarity measures are evaluated by the external criteria, which use the class variable to evaluate the quality of classification. In this paper, both internal and external evaluation indices are used.

Most of the internal indices need the original data matrix **X** for their computation. However, when dealing with mixed-type data, where both quantitative and categorical variables occur, it is impossible to use these variables for calculation because they need to be quantitative. There are several internal indices which need only a dissimilarity matrix for the computation, see (Charrad et al., 2014). From these, the Dunn, silhouette and McClain indices are used in this paper. For the external evaluation, the Rand index is used.

The *Dunn index* (Dunn, 1974) is based on an assumption that clusters in a dataset are compact and well separated by maximizing the inter-cluster distance while minimizing the intra-cluster distance, see (Yang, 2012). For the cluster solution with *k* clusters, it can be expressed by the formula

$$\mathit{Dunn}(k) = \min_{1 \le g < h \le k} \left(\frac{d(g,h)}{\max_{1 \le m \le k} \mathit{diam}(m)} \right),$$

where $d(g,h)$ is a distance between the g-th and h-th clusters expressed by one of the between cluster distances, e.g. by the complete linkage method; $\mathit{diam}(m)$ is the maximal distance between two objects in the m-th cluster. As the distance, any known or proposed dissimilarity measure can be used. The Dunn index takes values from zero to infinity, where the higher the values suit for the better cluster solution. The highest value indicates the optimal cluster solution. The Dunn index has one drawback, though; if one cluster is very bad (incompact, indistinct) whereas the others are good, the Dunn index will get very low value in this cluster solution. Thus, the Dunn index always displays the worst result possible.

The *silhouette index* (Rousseeuw, 1986), also known as *average silhouette width*, can be written as

$$\mathit{silhouette}(k) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

where a_i is the average dissimilarity of the i-th object to the other objects in the same cluster, b_i is the minimal average dissimilarity of the i-th object to any cluster not containing i. The silhouette index takes values from -1 to 1. A value close to one indicates well-separated clusters; a value close to minus one suggests badly separated clusters. A value close to zero indicates that the objects in the dataset are often located on the border of two natural clusters.

The *McClain index* (McClain and Rao, 1975) is defined as a ratio of the within-cluster and between-cluster distances

$$\mathit{McClain}(k) = \frac{S_w/n_w}{S_b/n_b} = \frac{S_w n_b}{S_b n_w},$$

where n_w is the number of pairs of objects in the same cluster in the dataset, n_b is the number of pairs of objects not belonging to the same cluster, S_w is the sum of the within-cluster distances between n_w pairs of objects, and S_b is the sum of the between-cluster distances between n_b pairs of objects. Based on the formula, a lower value of the index suits for a better cluster solution.

For purposes of this paper, the *Rand index* (Rand, 1971) is used. It considers all possible combinations of each of $n(n-1)/2$ pairs of objects in a dataset to one of four possible states. TP (true positive) is the number of pairs of similar objects assigned to the same cluster, TN (true negative) is the number of pairs of dissimilar objects assigned to different clusters, FP (false positive) is the number of pairs of dissimilar objects assigned to the same cluster, and FN (false negative) is the number of pairs of similar objects assigned to different clusters. Then

$$\textit{Rand} = \frac{TP + TN}{TP + FP + TN + FN}.$$

The Rand index represents a ratio of the number of correctly assigned objects, both positively and negatively, out of the number of all possible pairs. It takes values from zero to one. Values closer to one represent better accuracy in cluster assignment.

5 Experiment

To evaluate the dissimilarity measures presented in Section 2, an experiment on generated datasets is carried out. Its aim is to focus on factors which can influence cluster quality of the examined dissimilarity measures, such as a difficulty structure expressed by numbers of categories of the clustered variables, or a ratio of quantitative and categorical variables. Based on that, it is possible to determine in which situation is useful to use a certain dissimilarity measure.

For the experiment, 810 datasets were generated using the modified `genRandomClust()` function in the `clusterGeneration` R package (Qui and Joe, 2015). The categorical variables generated by this function are in fact ordinal; however, since this paper focuses mostly on an evaluation of the nominal part of the Gower coefficient, all the categories of categorical variables are considered as nominal.

The generated datasets differ by their minimal between-cluster distance (low, middle, high). The low distance was defined by the parameter *sepVal* $= 0.1$ in the `genRandomClust()` function, the middle distance by *sepVal* $= 0.3$, and the high distance by *sepVal* $= 0.5$. By the high distance, the clusters do not intersect in most of datasets. Furthermore, there were examined three different numbers of variables (four, six, eight), three types of difficulty structure (easy, medium, hard) based on the numbers of categories in categorical variables, and all possible combinations of counts of quantitative and categorical variables (including the variant when all variables are categorical). To ensure the robustness of obtained results, each combination was five times replicated. Three types of difficulty structure were generated in the way that categorical variables contain two to four categories in the *easy* structure, three to seven categories in the *medium* structure, and eight to ten categories in the *hard* structure. The numbers of objects in generated datasets were not firmly set; they varied from 300 to 700 cases. All datasets were generated with four clusters.

On each of these datasets, HCA with the three examined dissimilarity measures presented in Section 2 was applied. The average, complete, McQuitty's and the single linkage

Table 1: Mean values of evaluation indices broken down by the dataset difficulty

index	difficulty	Gower	G_IOF	G_LIN	G_VE
Dunn	easy	0.393	0.284	0.365	0.463
	medium	0.388	0.262	0.448	0.458
	hard	0.247	0.154	0.346	0.289
silhouette	easy	0.142	0.009	0.009	0.141
	medium	0.100	−0.021	−0.006	0.097
	hard	0.048	−0.043	0.000	0.044
McClain	easy	0.974	0.573	0.596	1.122
	medium	1.071	0.566	0.585	1.161
	hard	1.168	0.607	0.629	1.188
Rand	easy	0.590	0.343	0.346	0.621
	medium	0.558	0.340	0.344	0.578
	hard	0.529	0.346	0.350	0.534

methods were used. The resulting clusters were evaluated by mean values of the Dunn, silhouette, McClain and Rand indices, which were broken down by the minimal between-cluster distance, linkage method, difficulty structure and combinations of quantitative and categorical variables.

First, the influence of the dataset difficulty was examined. Table 1 demonstrates the mean values of the indices presented in Section 4 broken down by the dataset difficulty. The results are based on analysis of 810 datasets. Each of them was further analyzed by four methods of HCA presented in Section 3. Thus, each value in the table is calculated as the mean of $4 \times 810 = 3240$ clustering results. By all the examined similarity measures, it is apparent that with the increasing difficulty (increasing numbers of categories) the clustering performance decreases. According to the Dunn and Rand indices, the best clusters are produced by the G_VE measure. The silhouette index slightly prefers the original Gower measure to the G_VE measure, and the McClain index favors the G_IOF and G_LIN measures.

Second, the behavior of the examined similarity measures by the three different numbers of variables with different numbers of categorical variables was studied. Table 2 presents the mean values of the Dunn index of the examined dissimilarity measures broken down by different numbers of quantitative and categorical variables. The best clusters are produced by the G_VE measure in all combinations of quantitative and categorical variables. Only in situations, when all the variables are categorical, G_VE is outperformed by the G_LIN measure. Based on the performance of the original Gower measure, a rapid decrease of the Dunn scores is apparent with increasing numbers of categorical variables, which is in contradiction to G_VE, where the decrease is much lower. Since the cluster solution with three categorical variables, the Gower measure is also outperformed by the G_LIN measure.

The silhouette index almost does not change with different combinations of numbers of variables and categories. Concerning the Rand index, if a given dataset contains only

Table 2: Mean values of the Dunn index broken down by the number of variables (# of var.) and the number of categorical variables (# of cat. var.)

# of var.	# of cat. var.	Gower	G_IOF	G_LIN	G_VE
4	1	0.545	0.417	0.507	0.546
	2	0.456	0.271	0.450	0.485
	3	0.339	0.124	0.384	0.422
	4	0.276	0.041	0.370	0.389
6	1	0.578	0.514	0.574	0.582
	2	0.509	0.402	0.507	0.536
	3	0.422	0.304	0.471	0.482
	4	0.343	0.201	0.425	0.420
	5	0.273	0.112	0.369	0.380
	6	0.215	0.052	0.340	0.338
8	1	0.608	0.563	0.604	0.615
	2	0.555	0.484	0.567	0.581
	3	0.496	0.410	0.528	0.546
	4	0.439	0.332	0.489	0.502
	5	0.367	0.259	0.440	0.448
	6	0.315	0.182	0.412	0.418
	7	0.249	0.109	0.365	0.370
	8	0.216	0.060	0.360	0.345

one categorical variable, the original Gower measure provides the highest scores of the Rand index. However, when there are two or more categorical variables in a dataset, the G_VE outperforms the original Gower measure regarding a correct classification. Both these measures also increase their Rand scores with the increasing number of categorical variables, which is in contradiction to the G_IOF and G_LIN measures, whose Rand scores almost do not change. The McClain index is also relatively constant by different combinations of numbers of categorical variables. The only situation, when it is substantially lower, is when all variables are categorical. Based on the facts described in this paragraph, the results for the rest of the used indices (including the Dunn index) are presented in a condensed form in Table 3. From this table, it is apparent that the Dunn and Rand indices favor the G_VE measure, whereas the silhouette prefers the original Gower measure and the McClain index the G_IOF measure.

Table 3: Mean values of the Dunn, silhouette, McClain and Rand indices

Index	Gower	G_IOF	G_LIN	G_VE
Dunn	0.400	0.269	0.453	0.467
silhouette	0.097	−0.018	0.001	0.094
McClain	1.071	0.582	0.603	1.157
Rand	0.559	0.343	0.347	0.578

Third, Dunn index scores decrease by different minimal between-cluster distances and methods of the HCA was examined. The Dunn index was chosen because it is the only examined index whose values decrease monotonically with the increasing number of categorical variables. The results are displayed in Table 4 in the form of ratios between the Dunn scores in datasets with all categorical variables and datasets with one categorical variable. Thus, the higher values suit for more stable results of a given similarity measure by different combinations of quantitative and categorical variables.

Based on the mean ratios in Table 4, it was found out that the produced clusters are more stable by the high minimal between-cluster distance, i.e. they do not worsen so much with the increasing numbers of categorical variables. This is valid for all the examined methods of HCA. Generally, the most stable clusters were provided by the single linkage method. They were closely followed by the complete linkage method, and further by McQuitty's and average linkage methods.

The last row in Table 4 implies that the mean Dunn scores by the G_LIN measure decrease the least (by 35.2%) of all four examined dissimilarity measures. Thus, this measure is the most stable in its clustering performance by different numbers of categorical variables. It is relatively closely followed by the G_VE measure. With a substantial distance, the Gower measure was placed. The worst results were obtained by the G_IOF measure whose clusters worsen strongly with the increasing numbers of categorical variables.

Table 4: Ratios of the mean Dunn index scores broken down by minimal cluster distances and methods of HCA

Method	Distance	Gower	G_IOF	G_LIN	G_VE
average	low	0.324	0.078	0.541	0.553
	middle	0.334	0.084	0.591	0.523
	high	0.413	0.095	0.630	0.606
complete	low	0.387	0.095	0.702	0.628
	middle	0.365	0.097	0.752	0.607
	high	0.476	0.106	0.806	0.750
McQuitty	low	0.330	0.077	0.547	0.559
	middle	0.338	0.081	0.602	0.555
	high	0.444	0.091	0.647	0.680
single	low	0.536	0.137	0.616	0.707
	middle	0.504	0.121	0.650	0.651
	high	0.492	0.157	0.686	0.646
total	average	0.412	0.102	0.648	0.622

6 Conclusion

In this paper, three modifications of the Gower coefficient were compared and evaluated. The comparison was performed on 810 generated datasets with different properties, and the produced clusters were evaluated regarding three internal and one external evaluation criteria.

The examined dissimilarity measures were compared from three main aspects. First, their dependence on dataset difficulty, expressed by different numbers of categories, was examined. Second, the dependence on various combinations of quantitative and categorical variables in datasets was studied. Third, robustness of the dissimilarity measures to a high number of categorical variables was examined.

Generally, the best clusters were provided by the Gower_VE measure, which performed well both in datasets with the low and high number of categorical variables. Also, its classification abilities expressed by the Rand index were also the best ones among the examined measures. Thus, this dissimilarity measure can be recommended for general use. The original Gower measure also provided good results; however, with the increasing number of categorical variables, its clusters worsened substantially. This measure can be recommended for mixed datasets with the low number of categorical variables. The Gower_LIN measure proved to be the most stable concerning the increasing number of categorical variables. Also, its classification performance was good. Thus, this measure can be recommended for datasets with a high number of categorical variables. The clusters produced by the Gower_IOF measure were very poor both from the aspects of increasing number of categorical variables and classification, so this measure cannot be recommended for common use.

In the future research, we are going to compare the examined dissimilarity measures in this paper with some of (dis)similarity measures for mixed data introduced in recent years. We also plan to examine new approaches to evaluation criteria to datasets with mixed-type variables.

Acknowledgement

This work was supported by the University of Economics, Prague under the IGA project No. F4/41/2016.

References

[1] Ahmad, A. and Dey, L. (2007): A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, **63**(2), 503–527.

[2] Anderberg, M.R. (1973): *Cluster Analysis for Applications*. New York: Academic Press.

[3] Charrad M., Ghazzali N., Boiteau V., and Niknafs A. (2014): NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, **61**(6), 1–36.

[4] Dunn, J.C. (1974): Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**(1), 95–104.

[5] Gordon, A.D. (1981): *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Boca Raton: Chapman and Hall.

[6] Gower, J.C. (1971): A general coefficient of similarity and some of its properties. *Biometrics*, **28**(4), 857–871.

[7] Huang, Z. (1998): Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, **2**(3), 283–304.

[8] Kaufman, L. and Rousseeuw, P. (1990): *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.

[9] Lichman, M. (2013): *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California.

[10] Lin, D. (1998): An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning*, 296–304. San Francisco: Morgan Kaufmann.

[11] McClain, J.O. and Rao, V.R. (1975): Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, **12**, 456–460.

[12] Podani, J. (1999): Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*, **48**(2), 331–340.

[13] Rand, W.M. (1971): Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

[14] Rousseeuw, P. (1986): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

[15] Qiu, W. and Joe, H. (2015): *clusterGeneration: Random cluster generation (with specified degree of separation)*. R package version 1.3.4.

[16] Sparck-Jones, K. (1972): A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**(1), 11–21.

[17] Šulc, Z. and Řezanková, H. (2015): Novel similarity measures for categorical data based on mutability and entropy. In: *Conference of the International Federation of Classification Societies*. Bologna: Ospitalia.

[18] Šulc, Z., Matějka, M., and Procházka, J. (2016): Modifications of the Gower similarity coefficient. In: *The 19th Conference of Applications of Mathematics and Statistics in Economics, Banská Bystrica*. Available at: http://amse.umb.sk/proceedings/SulcProchazkaMatejka.pdf.

[19] Yang, M.S., Hwang, P.Y., and Chen, D.H. (2004): Fuzzy clustering algorithms for mixed feature variables. *Fuzzy Sets and Systems*, **141**(2), 301–317.

[20] Yang, R. (2012): *A Hierarchical Clustering and Validity Index for Mixed Data*. Ph.D. thesis of Iowa State University.