EFFECT OF ORDINAL VARIABLE TRANSFORMATIONS ON HIERARCHICAL CLUSTERING RESULTS: A CASE STUDY ON THE BIG DATA PHENOMENON

HANA ŘEZANKOVÁ[a,*], RICHARD NOVÁK[b]

hana.rezankova@vse.cz, xnovr900@vse.cz

[a] University of Economics in Prague, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Sq. 4, Prague, Czech Republic

[b] University of Economics in Prague, Faculty of Informatics and Statistics, Department of Systems Analysis, W. Churchill Sq. 4, Prague, Czech Republic

Abstract

The aim of the paper is to show some possible transformations of ordinal variables in cluster analysis and discuss their effect on hierarchical clustering results. Although several papers comparing different approaches to clustering objects characterized by ordinal variables have been published, the comparisons are not complete and include also variables other than ordinal variables (e.g. nominal variables). The following possibilities are considered in this paper to capture ordinal variables in clustering: "original" values (from one to the number of categories), standardized values, transformed values based on the range, ranks of the original values (averaged in case of ties), standardized ranks, and transformed ranks based on the range (usually recommended). The results of the complete linkage method obtained by the Manhattan and Euclidean distances for different numbers of clusters are compared. Moreover, these results are compared with the results obtained by the TwoStep algorithm. The case study is based on the answers of 481 respondents concerning the awareness of problems related to the "Big Data Phenomenon" and "New Digital Divide".

Key words

cluster analysis, ordinal variables, hierarchical clustering, data transformation, distance measures, Big Data phenomenon, New Digital Divide

JEL classification

C38, C63, C88

1. Introduction

Cluster analysis is a statistical multivariate method, which includes many techniques and algorithms. During an analysis of this type, data vectors are compared and grouped into clusters. These data vectors usually represent certain objects and the elements of vectors are the values of variables by which the objects are characterized. In the case of quantitative variables when object are characterized by numbers, the clustering algorithms and their features are well known. The techniques for binary, nominal and ordinal variables (and variables of mixed types) have been also proposed and some of them implemented in software packages. However, there is still considerable space for research in this area, the possibility of new similarity measures application, comparison of different measures, data transformation and techniques.

In this paper, we focus on clustering objects characterized by discrete ordinal variables. We compare cluster analysis results obtained on the basis of both original and transformed values. We apply hierarchical cluster analysis because the user obtains an unambiguous solution independent of the order of objects by this technique. In this paper, we present results

obtained by only one linkage method. We evaluate the effects of the transformation type and the distance measure type on clustering results.

The following possibilities are considered in this paper: "original" values (from one to the number of categories), standardized values, transformed values based on the range, ranks of the original values (with the average in case of ties), standardized ranks, and transformed ranks based on the range. Two distance measures are applied: Manhattan and Euclidean. We choose the complete linkage method for the research presentation. The Ward method can be used only in combination with the squared Euclidean measure and the influence of different distance measures could not be investigated. From other linkage methods, the complete and average linkages usually give clusters which are not one-element, except for the existence of outlying objects. For analyzed data, the complete linkage method provided clusters with sizes not too different (in selected numbers of clusters).

In addition, the comparison with results obtained by the TwoStep algorithm with the log-likelihood measure is mentioned. This method analyzes only standardized data, so clustering based on standardized original data and standardized ranks is performed. The log-likelihood measure was chosen for the reason that clusters with acceptable sizes are usually created in comparison with the results obtained using the Euclidean distance within this algorithm.

2. Clustering methods for objects characterized by ordinal variables

The most known technique for the treatment of ordinal variables consists in the transformation of value ranks according to the following formula

$$\frac{r_{il}-1}{M_l-1}, \tag{1}$$

where r_{il} is the rank of the i-th value within the l-th variable and M_l is the maximum rank for the l-th variable. The ranks are calculated from the original values in such a way that if ties of values exist, the ranks are averages of their possible orders. The Manhattan distance is recommended as the dissimilarity measure (see Kaufman and Rousseeuw, 2005). In case of the minimum rank is 1, the difference of two values transformed by (1) corresponds with the dissimilarity of non-transformed i-th and j-th values according to the formula

$$\frac{r_{il}-r_{jl}}{M_l-1}, \tag{2}$$

which is proposed by Podani (1999) for ordinal variables without ties as a part of Gower's general coefficient of similarity designed originally for datasets with nominal and quantitative variables (Gower, 1971). The denominator expresses the range of the l-th variable.

Some authors tried to compare other possibilities of ordinal variable treatment in cluster analysis. Žiberna et al. (2004) compared three ways: treating data as interval, using ranks, and treating data as nominal. In the case of the first two ways, the squared Euclidean distance was applied; for the third way, the simple matching coefficient was used. The ordinal data were prepared by cutting the generated interval data into five categories in four different ways. The Ward linkage method was applied in hierarchical clustering, and for the result comparisons, the corrected Rand index was used. The authors confirmed the expectations that treating data as nominal is an inappropriate approach. Significant differences between treatments of ordinal data as interval and as rank were not found.

Ranalli and Rocci (2015) compared several interesting approaches dividing into two groups: in one of them, ordinal variables are treated as interval, in the second one, variables are treating as ordinal. In the first group, the k-means algorithm and three finite mixture of Gaussians approaches, which are the model-based analog of the k-means algorithm, are

evaluated. In the second group, two-step approaches are included. In the first step, principal component analysis for qualitative data is applied; in the second step, the *k*-means algorithm is used. Two variants (with two and three factors) are considered for the comparison. Moreover, two variants of the latent mixture of Gaussians (LMG) approach is contained in the second group. The dataset for the comparison was prepared on the basis of Fisher's Iris quantitative data by their categorization. The results of clustering were compared using the adjusted Rand index. The best results were obtained by the LMG approach.

Giordan and Diana (2011) proposed a new clustering method for discrete ordinal data. Objects are grouped using a multinomial model, a cluster tree and a pruning strategy. This approach also solves a problem with determining the suitable number of clusters. The authors evaluate the proposed method by four external criteria, including the adjusted Rand index. For this purpose, they generate data sets with different features (two and three variables) and different numbers of objects. In addition, the authors use the same external criteria when different four (real) data sets are analyzed by known methods PAM (the *k*-medoids algorithm) and FANNY (the fuzzy clustering algorithm), see (Kaufman and Rousseeuw, 2005). The data were transformed in the way described in the first paragraph of this section. It is obvious, that results obtained by analyses of different datasets (moreover in one case generated and in the second case real) cannot be compared.

A more complex comparison was done by Ruff (2014). He evaluates four clustering methods: the *k*-means algorithm with the Euclidean distance (treating data as interval), the *k*-median algorithm with the Manhattan distance, hierarchical clustering with the Ward linkage method (treating data as interval), and latent class analysis (LCA) with the expectation-maximization (EM) algorithm (a method developed for ordinal variables). Ten experiments were performed; in every experiment 30 datasets were prepared with four variables (with 7 categories) and two clusters. Two variants of ten experiments differing in the sizes of clusters are presented. The methods are compared by the mean and the relative standard deviation of the numbers of correct classifications. The *k*-means algorithm and the LCA approach were evaluated as the best – with similar results.

Walesiak and Dudek (2010) applied a special distance measure for data with ordinal variables. It is called the general distance measure (GDM). The authors compare different clustering algorithms (the *k*-medoids algorithm, seven hierarchical agglomerative algorithms and hierarchical divisive method DIANA) and eight internal evaluation criteria for determining the number of clusters on the basis of adjusted Rand index mean values. Data sets were generated in nine different scenarios which differ in the number of variables, the number of categories for each variable, the density, and shape of clusters, the number of clusters, and the number of noisy (irrelevant) variables. The average linkage method was evaluated as the best and the single linkage method as the worst. However, the results obtained by the GDM measure are not compared with any results which could be obtained using some other distance measure or data transformation.

Unusually ways were also compared, e.g., Coroiu et al. (2016) evaluated the Slink and Naive agglomerative clustering algorithms in software ELKI based on seven linkage methods by means of eight both internal and external criteria. However, it is not specified which distance measure was applied. The authors analyzed two real datasets, both with ordinal variables with three categories (either original values or transformed variables based on interval scale). They conclude that the Ward linkage method gives more accurate results in terms of cluster validity than the other linkage methods.

This contribution does not provide a complex comparison of all possible transformations, clustering methods and distance measures based on large amount datasets. However, it can serve as a design of how some experiments could be prepared. An important part of the result

evaluation is a description of obtained clusters, including investigation of dependency of analyzed variables.

3. Applied way of result comparison

When effects of both transformation and distance measure are compared using a distance between objects, several problems should be solved: which data matrix should be a basis for comparison (original or transformed) and which distance measure should be applied. All result solutions (assignments of objects into clusters) should be evaluated in the same way. However, it is difficult to decide which way is the best one. For the reason mentioned above, we evaluate result solutions by means of within-cluster variability.

A variability measure proposed for discrete ordinal variables is the discrete ordinal variance *dorvar*. It takes only cumulative relative frequency into account. It means that the orders of categories are important, not their values. In case of a constant, there is not any variability and this measure has the zero value. The greatest variability for the known number of categories is in the case in which respondents choose only answers corresponding to the lowest and the highest categories in the same ratio. This measure can be normalized to the interval from 0 to 1 when we use the following formula

$$\text{normalized dorvar} = \frac{4\sum_{c=1}^{C-1}(P_c(1-P_c))}{C-1}, \tag{3}$$

where P_c is the cumulative frequency of the cth category and C is the number of categories. The dorvar measure is not implemented in software packages and in case of a great number of variables and categories the computation could be only obtained by programming.

When we take the *variance* (proposed for quantitative variables) as a basis for variability evaluation, we cannot interpret the obtained values but we can compare for which case the variability is lower and for which case higher than for others. Comparing variables with the different numbers of categories it is useful to express a variability as a number from the interval $\langle 0, 1\rangle$. It can be achieved using the *proportional coefficient of differentiation* expressed as

$$P_D = \frac{4s^2}{R^2}, \tag{4}$$

where s^2 is the variance and R is the range (it means $C-1$ for the ordinal variable). The interpretation of this coefficient is the same as in the case of the dorvar measure. The greatest variability is for the case when respondents choose only answers corresponding to the lowest and the highest categories in the same ratio.

4. Questionnaire survey on the Big Data problem

To illustrate how to compare clustering results, we use selected variables from the dataset based on a questionnaire survey. The survey concerned the awareness of problems related to a "big data phenomenon" and "new digital divide" (see e.g. Andrejevic, 2014; Boyd and Crawford, 2012). Thus, in this paper, big data is not the term referring to an investigated dataset but the term that refers to the content of the survey focused on awareness of big data phenomenon impacting our society, economy and many other disciplines. At the World Economic Forum in 2012 in Davos, Switzerland, Big Data was declared: "A new class of economic asset, like currency or gold." (Lohr, 2012).

The full questionnaire has the following structure. Firstly, we investigate the attitudes of respondents to the main question focused on the conflict of equality vs digital divide.

Secondly, we investigate the attitudes of respondents to 16 big data issues, called hereinafter as variables. Finally, the attitudes of respondents to eight basic human values are investigated. All together 25 meritorious questions were answered by respondents.

The respondents were IT students and IT professionals mainly from Czechia and Slovakia and their attitudes were investigated via an online questionnaire from October to December 2018. There were 518 respondents but some of them did not answer some questions. For the purpose of this paper, the valid answers of 481 respondents are analyzed.

This statistical case study focuses precisely on a small part of the "big data" and "digital divide phenomenon". Five questions from 25 meritorious ones were chosen for the purpose of cluster analysis. The selection criterion was to select five "big data" variables with the weak dependence among these variables. We are aware that the "big data phenomenon" covers much more complex areas; however, the main focus of this paper is transformations of ordinal variables in cluster analysis. For the deep dive into actual big data topics such as legal regulation or the perspective of different data sources, see e.g. (Novák, 2014; Pavlíček and Novák, 2015).

The five chosen variables are: *Equality vs Big Data* (whether new big data technologies distort equality among people), *Privacy Intrusion*, *New Barriers* (division of society on people with benefits and others), *Missing Transparency* (the world is driven by complex mathematical algorithms that I cease to understand) and *Confusion* (I lose confidence in the world and the future, because the big data technologies creates confusion in previously clear contexts and information). The questions were answered with values belonging to five categories: 0 ("I totally disagree" or "completely unimportant"), 25, 50, 75, and 100 ("I completely agree" or "very important") in the survey.

We recoded the original variables into ordinal variables with different numbers of categories so that frequencies of categories are at least 30. In this way, we obtained the main variable *Equality vs Big Data* with three categories. The variables *Privacy Intrusion* and *New Barriers*, expressing the important problems concerning the big data technologies, contain values in four categories. The remaining variables *Missing Transparency* and *Confusion*, expressing the agreement with a certain opinion, contain values in five categories. So, we analyzed the dataset with 481 rows (data vectors based on answers of 481 respondents) and five columns (variables chosen for the analysis). The information concerning recoded variables is in Tables 1 (frequency distributions) and 2 (selected statistics).

Table 1: Frequency distributions for five variables chosen for the analysis

Variable	Category				
	1	2	3	4	5
Equality vs Big Data	106	214	161	–	–
Privacy Intrusion	31	80	120	250	–
New Barriers	78	169	155	79	–
Missing Transparency	39	100	162	135	45
Confusion	71	134	168	78	30

Source: the authors.

As regards Table 2, the average cannot be interpreted but the value provides information about frequency distribution. For example, in the case of two last variables, the median is three for both variables. However, the values of the average are different. For one variable the average is greater than the median, for the second one it is the opposite because the frequencies of the 4th and 5th categories are higher for the first variable from these two ones.

The values of the *normalized dorvar* variability measure and the *proportional coefficient of differentiation* differ but their orders are the same for both variables. Dependence of the variables chosen for the analysis is weak. Kendall's tau-b (for pairs of variables) has values

from 0.097 (p-value 0.013) to 0.327 (dependence of the variables with five categories). It means that correlation is weakly positive for all pairs of variables.

Table 2: Selected statistics for five variables chosen for the analysis

Variable	Median	Average	Normalized dorvar	Proportional coefficient of differentiation
Equality vs Big Data	2.00	2.11	0.789	0.543
Privacy Intrusion	4.00	3.22	0.650	0.398
New Barriers	2.00	2.49	0.697	0.402
Missing Transparency	3.00	3.10	0.599	0.295
Confusion	3.00	2.71	0.603	0.300

Source: the authors.

5. Comparison of analysis results

Before clustering, the objects were displayed within two dimensions obtained by principal component analysis for categorical data. No disjunctive clusters were identified; the highest density of points was in the center of coordinates. A set of objects could be divided into two or four groups according to coordinates. The group sizes should not be too different. We applied different linkage methods for object clustering, and we obtained the acceptable sizes for two and four clusters only by the complete linkage method.

So, we applied the complete linkage method using the Manhattan and Euclidean distances for cluster analysis. First, we use "original" data described in the previous section, when categories of variables are coded from value 1 to the value expressing the number of categories, and five transformations of these data. We used the standardization into the variables with the average equaled 0 and the standard deviation equaled 1 and the transformation into the variables with values from the interval $\langle 0, 1 \rangle$ analogously according to the Eq. (1). Both types of transformations were applied to the original data and to the ranks with the average in case of ties. If original data are transformed, in Eq. (1) the original values are used instead of ranks and the numbers of categories instead of the maximum ranks.

Objects were clustered to different numbers of clusters – from 2 to 5. For each clustering solution including a certain distance measure and a certain number of clusters, the weighted average of the proportional coefficient of differentiation was calculated. The average is based on values of this coefficient computed for the original values of individual variables in individual clusters. Weights are the numbers of objects in the clusters. Only in four cases, the average value was less for the Euclidean distance than for the Manhattan one. For this reason, we only present the results for the Manhattan distance, see Table 3. We can see that for the original data and ranks, the least within-cluster variability was found for the two-cluster solution. In other cases, the four-cluster solutions were better – the best for the standardized original data. The object assignment into four clusters approximately corresponds with splitting a set of objects according to coordinates, see above.

Table 3: Average values of the proportional coefficient of differentiation (complete linkage method, Manhattan distance)

Data	Number of clusters			
	2	3	4	5
Original	**0.331**	0.355	0.362	0.350
Standardized original	0.317	0.318	**0.288**	0.292
Transformed original	0.385	0.355	**0.333**	0.342
Ranks	**0.318**	0.348	0.376	0.374
Standardized ranks	0.317	0.321	**0.310**	0.330
Transformed ranks	0.334	0.327	**0.319**	0.345

Source: the authors.

The proportional coefficient of differentiation gives the same results both for original variables and for new variables obtained by any linear transformation original variables. Ranks are not a linear transformation of original values. However, the order of variables according to their variability is mostly the same both for the original data and the ranks.

Moreover, we applied the TwoStep cluster analysis using the log-likelihood distance for object clustering. This method was proposed for large data files. In the first step, objects are clustered into sub-clusters characterized by cluster features (CF). In the second step, the sub-clusters are clustered based on the CF-tree (Zhang et al., 1997). Hierarchical clustering is applied in this step. The TwoStep algorithm in IBM SPSS Statistics analyzes only standardized data. That is why we only present clustering solutions for the standardized original data and for the standardized ranks. In Table 4, there are average values of the proportional coefficient for differentiation for numbers of clusters from 2 to 5. The smallest value was obtained for the four-cluster solution when standardized ranks were analyzed.

Table 4: Average values of the proportional coefficient of differentiation (TwoStep method, log-likelihood distance)

Data	Number of clusters			
	2	3	4	5
Standardized original	0.339	**0.300**	0.342	0.331
Standardized ranks	0.314	0.263	**0.250**	0.301

Source: the authors.

The main results of the analyses described above are the following:

- clustering transformed ranks does not guarantee the best clustering solution in terms of within-cluster variability,
- hierarchical cluster analysis with the complete linkage methods usually gives better clustering solutions (in terms of within-cluster variability) with the Manhattan measure,
- the best clustering solution (in terms of within-cluster variability) for the complete linkage methods with the Manhattan measure can be obtained e.g. when standardized discrete ordinal data are analyzed,
- there are some other clustering methods which give better clustering solutions (in terms of within-cluster variability) than hierarchical cluster analysis with the complete linkage methods, e.g. the TwoStep method.

Among performed analyzes described above, the TwoStep method with the log-likelihood distance for standardized ranks gave the best clustering solution for four clusters. When we characterized these clusters by the averages and the values of the proportional coefficient of differentiation, we obtained the values displayed in Table 5. In Cluster 1, the lower values are more frequent. In Cluster 4, the higher values are more frequent.

Table 5: The averages and the values of the proportional coefficient of differentiation (in parentheses) for individual variables in individual clusters (TwoStep method)

Variable	Cluster (number of objects)				
	1 (56)	2 (127)	3 (156)	4 (142)	Total (481)
Equality vs Big Data	1.30 (0.25)	1.61 (0.27)	2.26 (0.35)	2.73 (0.23)	2.11 (0.54)
Privacy Intrusion	2.07 (0.76)	3.84 (0.17)	2.44 (0.18)	3.99 (0.06)	3.22 (0.40)
New Barriers	1.41 (0.17)	2.50 (0.37)	2.28 (0.24)	3.14 (0.29)	2.49 (0.40)
Missing Transparency	2.34 (0.35)	2.79 (0.23)	3.25 (0.21)	3.51 (0.30)	3.10 (0.30)
Confusion	1.48 (0.29)	2.17 (0.17)	2.97 (0.17)	3.40 (0.30)	2.71 (0.30)

Source: the authors.

According to the program output, variable *Privacy Intrusion* had the greatest influence on clustering and variable *Missing Transparency* had the least influence. In obtained clusters, variables are either mostly weaker dependent in comparison with the situation in the whole

dataset or negatively dependent (according to Kendall's tau-b). The values of Kendall's tau-b for all pairs of variables in individual clusters are shown in Table 6. The strongest positive correlation (0.34) was found for variables *Missing Transparency* and *Confusion* in Cluster 4, the strongest negative correlation −0.37 for variables *Privacy Intrusion* and *New Barriers* in Cluster 2.

From Tables 5 and 6 we can conclude that respondents assigned to Cluster 1 rather disagree with the claims (or they do not consider them important) but there is a negative correlation between variables *Equality vs Big Data* and *Missing Transparency*. In Cluster 2 there are respondents which consider *Privacy Intrusion* rather important and they assign less importance to other claims; there is a negative correlation between variables *Privacy Intrusion* and *New Barriers*. In Cluster 3, values of variables are rater middle and there is a negative correlation between variables *Equality vs Big Data* and *Confusion*. Cluster 4 can be characterized by respondents considering the claims rather important. There is primarily a negative correlation between variable *Equality vs Big Data* and variables *Missing Transparency* and *Confusion*. The above may be related to that in Clusters 3 and 4 there is a significant positive correlation between variable *Missing Transparency* and *Confusion.*

Table 6: The values of Kendall's tau-b and p-values (in parentheses) for all pairs of variables in individual clusters (TwoStep method)

Variable	Cluster (number of objects)				
	1 (56)	2 (127)	3 (156)	4 (142)	Total (481)
Equality vs Big Data and Privacy Intrusion	0.05 (0.71)	0.12 (0.16)	0.11 (0.13)	−0.07 (0.41)	0.19 (0.00)
Equality vs Big Data and New Barriers	0.07 (0.58)	0.01 (0.89)	0.12 (0.12)	−0.09 (0.24)	0.27 (0.00)
Equality vs Big Data and Missing Transpar.	−0.32 (0.01)	0.08 (0.31)	0.06 (0.45)	−0.25 (0.00)	0.16 (0.00)
Equality vs Big Data and Confusion	0.17 (0.20)	0.08 (0.35)	−0.18 (0.01)	−0.26 (0.00)	0.27 (0.00)
Privacy Intrusion and New Barriers	0.07 (0.57)	−0.37 (0.00)	0.12 (0.10)	−0.13 (0.09)	0.31 (0.00)
Privacy Intrusion and Missing Transpar.	0.06 (0.63)	0.15 (0.07)	0.07 (0.34)	0.01 (0.89)	0.10 (0.01)
Privacy Intrusion and Confusion	0.18 (0.17)	0.14 (0.09)	−0.02 (0.84)	−0.01 (0.92)	0.12 (0.00)
New Barriers and Missing Transpar.	0.06 (0.63)	−0.13 (0.09)	0.14 (0.05)	0.09 (0.22)	0.16 (0.00)
New Barriers and Confusion	0.11 (0.39)	0.08 (0.32)	0.00 (0.98)	0.04 (0.63)	0.23 (0.00)
Missing Transparency and Confusion	0.14 (0.27)	0.07 (0.34)	0.23 (0.00)	0.34 (0.00)	0.33 (0.00)

Source: the authors.

6. Conclusion

If discrete ordinal data are analyzed, it is useful to compare not only results obtained by several clustering methods but also results based on different transformations of original data. The Manhattan distance measure should be considered as a measure which can influence better assignments of objects into clusters in comparison with e.g., the Euclidean distance in hierarchical cluster analysis. In a case study presented in this paper, we obtained the best results (in terms of within-cluster variability) using the TwoStep method with the log-likelihood distance applied to the standardized ranks.

This case study only included a small part of wider research focused on the "big data phenomenon" and "new digital divide" problems. Five questions from 25 meritorious ones were chosen for the purpose of cluster analysis. Four clusters of respondents were found as the best assignments of respondents into clusters by most of the performed analyses. We characterized the best cluster solution both by selected descriptive statistics for individual variables and by the association between variables in individual clusters. Although weak positive correlations were found for all pairs of variables in the whole datasets, in individual clusters negative correlations were found for some pairs of variables.

Our research could be considered as a contribution to the exploration of extensive possibilities to analyze ordinal data. Applied transformations and ways of result evaluation

can be also used in case of quantitative data and research of effect of transformations, distance measures, linkage methods, etc. on quality of clustering.

Acknowledgements

The paper was supported by the University of Economics in Prague under the grant scheme IGA No. F4/44/2018.

References

[1] Andrejevic, M. 2014. The big data divide. In International Journal of Communication, 2014, vol. 8, pp. 1673-1689.

[2] Boyd, D., Crawford, K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. In Information, Communication & Society, 2012, vol. 15, iss. 5, pp. 662-679.

[3] Coroiu, A. M., Gaceanu, R. D., Pop, H. F. 2016. Discovering patterns in data using ordinal data analysis. In Studia Universitatis Babes-Bolyai, Informatica, 2016, vol. 61, iss. 1, pp. 78-92.

[4] Giordan, M., Diana, G. 2011. A clustering method for categorical ordinal data. In Communications in Statistics – Theory and Methods, 2011, vol. 40, iss. 7, pp. 1315-1334.

[5] Gower, J. C. 1971. General coefficient of similarity and some of its properties. In Biometrics, 1971, vol. 27, iss. 4, pp. 857-871.

[6] Kaufman, L., Rousseeuw, P. J. 2005. Finding groups in data. Hoboken : Wiley. 2005. ISBN 0-471-73578-7.

[7] Lohr, S. 2012. The age of big data. Sunday review. New York Times, 2012-02-11.

[8] Novák, R. 2014. Big data and legal regulation. In Doucek, P., Chroust, G., Oškrdal, V. (eds.) IDIMT-2014 Networking Societies – Cooperation and Conflict. Linz : Trauner, 2014. ISBN 978-3-99033-340-2.

[9] Pavlíček, A., Novák, R. 2015. „Big data" from the perspective of data sources. In Proceedings of the 11th international conference on Strategic Management and its Support by Information Systems 2015 (SMSIS). Ostrava : VŠB – Technical University of Ostrava, 2015. ISBN 978-80-248-3741-3.

[10] Podani, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. In Taxon, 1999, vol. 48, iss. 2, pp. 331-340.

[11] Ranalli, M., Rocci, R. 2015. Clustering methods for ordinal data: A comparison between standard and new approaches. In Morlini, I., Minerva, T., Vichi, M. (eds.) Advances in statistical models for data analysis. Heidelberg : Springer, 2015. ISBN 978-3-319-17376-4.

[12] Ruff, F. 2014. Clustering methods for ordinal variables. In Karlovitz, J. T. (ed.) Economic questions, issues and problems. Komárno : International Research Institute, 2014. ISBN 978-80-89691-07-4.

[13] Walesiak, M., Dudek, A. 2010. Finding groups in ordinal data: An examination of some clustering procedures. In Locarek-Junge, H., Weihs, C. (eds.) Classification as a tool for researchers. Berlin : Springer, 2010. ISBN 978-364210744-3.

[14] Zhang, T., Ramakrishnan, R., Livny, M. 1997. BIRCH: A new data clustering algorithms and its applications. In Data Mining and Knowledge Discovery, 1997, vol. 1, iss. 2, pp. 141-182.

[15] Žiberna, A., Kejžar, N., Golob, P. 2004. A comparison of different approaches to hierarchical clustering of ordinal data. In Metodološki Zvezki, 2004, vol. 1, iss. 1, pp. 57-73.