20th International Scientific Conference AMSE
Applications of Mathematics and Statistics in Economics 2017
Szklarska Poręba, Poland 30 August 2017 – 3 September 2017

COMPARISON OF ECONOMIC AND NON-ECONOMIC DATASETS WITH CATEGORICAL VARIABLES IN HIERARCHICAL CLUSTERING

ZDENEK SULC, JANA CIBULKOVA, JIRI PROCHAZKA, MARTIN MATEJKA
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability, W. Churchill sq. 4, Czechia
email: zdenek.sulc@vse.cz, cibj00@vse.cz, jiri.prochazka@vse.cz, martin.matejka@vse.cz

Abstract
The economic datasets have their specifics; they usually describe human behavior or activity, which are difficult to measure. Thus, in comparison to non-economic datasets, they are less consistent. The paper analyzes differences between categorical economic and non-economic datasets in hierarchical clustering (HCA). To achieve this goal, two analyses based on 25 real-world datasets are carried out. In the first one, groups of economic and non-economic datasets are compared from the point of view of their internal characteristics based on HCA results; in the second one, homogenous groups of datasets are recognized and they are further examined by internal characteristics and graphical outputs. For each group of datasets, the most appropriate similarity measures are identified. The results show substantial differences between economic and non-economic datasets, primarily in terms of the within-cluster variability decrease. We were also successful in classification of the examined datasets into easily interpretable groups, for which suitable similarity measures were identified.

Key words: *economic datasets, categorical data, hierarchical clustering, similarity measures*

JEL Codes: *C38*

DOI: 10.15611/amse.2017.20.38

1. Introduction

In many research tasks, e.g. (Löster and Pavelka, 2013), economic datasets are analyzed. Under this term, datasets from surveys of individuals or households, or financial datasets can be classified. These datasets have their specifics; they usually study human behavior or activity which are difficult to measure and comprehend completely. In comparison to non-economic datasets, it is suspected that they are less consistent. Their different structure may lead to different clustering results.

This paper aims to examine differences between economic and non-economic datasets characterized by categorical variables using two analyses based on results of hierarchical cluster analysis (HCA). In the first one, groups of economic and non-economic datasets are compared from the point of view of their internal characteristics, such as the optimal number of clusters or the within-cluster variability. In the second one, using the HCA, different groups of datasets are recognized, and they are further characterized by selected internal characteristics and graphical outputs. In both the analyses, suitable similarity measures are identified for each group of the datasets.

The paper is organized as follows. Section 2 presents the similarity measures for data with nominal variables and evaluation indices which are used for the analyses. Section 3 describes the used datasets and analyses setting, and the experiments are conducted in Section 4. The results are summarized in Conclusion.

2. Analyses Background

In this section, a background for internal characteristics computation used in the analyses of the paper is presented. First, similarity measures for nominal data are presented; second, internal indices for cluster evaluation are described.

2.1 Similarity measures

Four similarity measures for nominal data were chosen for the analyses because they provided the best clustering results in HCA, (Šulc, 2016). All of them can be applied directly on the data matrix $\boldsymbol{X} = [x_{ic}]$, where $i = 1,2,...,n$ (n is the total number of objects); $c = 1,2,...,m$ (m is the total number of variables). The number of categories of the c-th variable is denoted as K_c, absolute frequency as f, and relative frequency as p. Their overview can be found in Tab. 1, where similarity between two categories by the c-th variable is denoted as $S\left(x_{ic},x_{jc}\right)$ and similarity between the objects $\mathbf{x}_i = [x_{i1},x_{i2},\ldots,x_{im}]$ and $\mathbf{x}_j = [x_{j1},x_{j2},\ldots,x_{jm}]$ as $S\left(\mathbf{x}_i,\mathbf{x}_j\right)$.

Table 1: An overview of the used similarity measures for nominal data

Measure	$S\left(x_{ic},x_{jc}\right)$	$S\left(\mathbf{x}_i,\mathbf{x}_j\right)$
ES	$= \begin{cases} 1 & \text{if} \quad x_{ic} = x_{jc} \\ \frac{K_c^2}{K_c^2+2} & \text{otherwise} \end{cases}$	$= \frac{\sum_{c=1}^{m} S\left(x_{ic},x_{jc}\right)}{m}$
IOF	$= \begin{cases} 1 & \text{if} \quad x_{ic} = x_{jc} \\ 1+\ln f\left(x_{ic}\right)\cdot\ln f\left(x_{jc}\right) & \text{otherwise} \end{cases}$	$= \frac{\sum_{c=1}^{m} S\left(x_{ic},x_{jc}\right)}{m}$
LIN	$= \begin{cases} 2\ln p\left(x_{ic}\right) & \text{if} \quad x_{ic} = x_{jc} \\ 2\ln\left(p\left(x_{ic}\right)+p\left(x_{jc}\right)\right) & \text{otherwise} \end{cases}$	$= \frac{\sum_{c=1}^{m} S\left(x_{ic},x_{jc}\right)}{\sum_{c=1}^{m}\left(\ln p\left(x_{ic}\right)+\ln p\left(x_{jc}\right)\right)}$
VE	$= \begin{cases} -\frac{1}{\ln K_c}\sum_{u=1}^{K_c} p_u \ln p_u & \text{if} \quad x_{ic} = x_{jc} \\ 0 & \text{otherwise} \end{cases}$	$= \frac{\sum_{c=1}^{m} S\left(x_{ic},x_{jc}\right)}{m}$

Source: The authors.

Each of the measures in Tab. 1 treats similarity between two categories in a different way; the ES measure (Eskin et al., 2002) is based on the number of categories of the c-th variable, whereas the IOF measure (Sparck-Jones, 1972) uses the absolute frequencies of the observed categories x_{ic} and x_{jc}, and the LIN measure (Lin, 1998) uses the relative frequencies instead. The VE measure (Šulc, 2016) is based on the variability of the c-th variable expressed by the entropy.

2.2 Evaluation indices

In the paper, all the examined datasets are evaluated by three internal evaluation criteria. The *within-cluster entropy coefficient* (WCE) expresses the within-cluster variability measured

by the normalized entropy. The coefficient is expressed using the formula

$$WCE(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{1}{\ln K_c} \left(- \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \ln \frac{n_{gcu}}{n_g} \right) \right), \tag{1}$$

where n_g is the number of objects in the g-th cluster ($g = 1, ..., k$), n_{gcu} is the number of objects in the g-th cluster by the c-th variable with the u-th category ($u = 1, ..., K_c$).

Two internal evaluation criteria are able to determine the optimal number of clusters, the PSFE index and the BK index. The PSFE index (Řezanková et al., 2011) with the formula

$$PSFE(k) = \frac{(n-k)(nWCE(1) - nWCE(k))}{(k-1)\, nWCE(k)}, \tag{2}$$

where $nWCE(1)$ is the variability in the whole dataset with n objects, and $nWCE(k)$ the within-cluster variability in the k-cluster solution. $nWCE$ is computed as

$$nWCE(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \left(- \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \ln \frac{n_{gcu}}{n_g} \right) \right).$$

The PSFE index is based on the within-cluster variability measured by the entropy. Its maximal value across several cluster solutions should indicate the optimal number of clusters. In such a cluster solution, the maximal decrease of the within-cluster variability occurs.

The BK index (Chen and Liu, 2009) is an internal entropy-based index which computes the expected entropy H_E for each of the k clusters in a dataset with respect to a cluster membership variable according to the formula

$$H_E(k) = \frac{\sum_{g=1}^{k} \sum_{c=1}^{m} H_{gc} \cdot n_g}{n},$$

where H_{gc} is the normalized entropy in the c-th variable expressed as

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

The expected entropy decreases with increasing cluster solutions. In some point, the decrease slows down. To identify such a point, the second order difference of the incremental expected entropy $I(k)$ is computed. In this point, the BK index takes its maximal value, which indicates the optimal number of clusters k^* as it is stated in the formula

$$k^* = \max_{k \geq 2} BK(k) = \Delta^2 I(k) = (I(k-1) - I(k)) - (I(k) - I(k+1))), \tag{3}$$

where

$$I(k) = H_E(k) - H_E(k+1).$$

3. Experiments and their Evaluation

This section is divided into two parts. In the first one, the analyzed datasets and their adjustments are presented; in the second one, the settings of the analyses are described.

3.1 Used Datasets

For the analyses, 25 real-world datasets (both economic and non-economic) were collected. Their overview with the selected properties occurs in Tab. 2. The datasets come mostly from the *UCI Machine Learning Repository* (UCI), see (Bache and Lichman, 2013), four datasets come from the book *Analyzing Categorical Data* (CAT), see (Simonoff, 2010), four of them from the *Australian government* webpage[1] (GOV), four of them from the *Kaggle* (KAG) database[2], and the rest of the datasets were created by the authors (OWN) from the economic surveys SILC 2011 and from the Czech Social Science Data Archive[3]. Very large datasets (datasets: 2, 6, 10, 18, 22, 25), which were not suitable for HCA because of their size, were reduced using the simple sampling method. Furthermore, all objects that contain at least one missing value were removed from all the datasets (the *listwise deletion method*). Also, some numeric variables were categorized, or variables with the very large number of categories were recoded to the smaller number of categories for the purposes of the analyses.

The examined datasets were divided into economic (E) and non-economic (N) ones, see column *Type* in Tab. 2. The economic datasets were defined as those, which come from questionnaires surveys (e.g. Sex Survey), deal with economic issues (e.g. Poverty), or express individuals' opinions. On the contrary, the non-economic datasets dealt with the non-economic topic, e.g. medical data or described a given state. Briefly, the economic datasets were created by surveying the respondents, whereas the non-economic ones were produced by observing a given issue.

3.2 Analyses settings

In the paper, there are two types of analyses. The first one deals with a comparison of economic (E) and non-economic (N) datasets whose overview occurs in Tab. 2. The second one deals with the classification of the examined datasets (both economic and non-economic).

Before the evaluation, all datasets were analyzed with HCA with the four similarity measures presented in Tab. 1. The *complete linkage* method is used in this step, since it usually provides the best clustering results for categorical data, see e.g. (Šulc, 2016). Based on the HCA outputs, the evaluation criteria WCE, PSFE and BK are computed for each similarity measure in the two- to six-cluster solutions. WCE and PSFE were computed in the R software using the *nomclust* package (Šulc and Řezanková, 2015). Based on these criteria, six characteristics for internal evaluation of the examined datasets were computed, see Tab. 3.

The characteristic WCE(1) expresses the total variability in the whole dataset measured by the WCE index. The following mean scores were always calculated as the arithmetic mean of the four similarity measures in Tab. 1. WCE_M expresses the arithmetic mean of the WCE decrease across two- to six-cluster solutions measured by the geometric mean. CLU_1 and CLU_2 present the arithmetic means of the recommended number of clusters using either PSFE and the BK index respectively. WCE_O1 and WCE_O2 are calculated as the arithmetic mean of WCE based on the optimal cluster solution expressed by CLU_1 and CLU_2.

In the first analysis, the economic and non-economic datasets are compared from aspects of the characteristics in Tab. 3. Next, a decrease of the WCE coefficient is graphically compared in both the groups and the best similarity measures are identified.

[1] http://data.gov.au/dataset

[2] www.kaggle.com

[3] http://archiv.soc.cas.cz/en

Table 2: An overview of the analyzed real-world datasets

ID	Dataset name	No. of objects	No. of variables	Range of categories	Type	Source
1	Adults	2409	6	2–5	E	UCI
2	Census US	1200	10	2–13	E	UCI
3	Dominica	944	18	2–10	E	GOV
4	Flats	1162	9	2–3	E	UCI
5	Happiness	1517	3	3–5	E	CAT
6	Homicides	1200	9	2-5	E	KAG
7	Hospital	704	13	3–9	E	KAG
8	House Votes	435	16	3	E	UCI
9	Unemployment	1239	11	2-7	E	OWN
10	Nursery	1287	8	2–5	E	UCI
11	Sex Survey	159	9	2–4	E	CAT
12	Poverty	1063	8	2–6	E	OWN
13	Poverty 2	1059	5	3–6	E	OWN
14	Maths	395	22	2–5	E	KAG
15	Accessibility	68	9	2–10	N	GOV
16	Autos	1652	7	2–8	N	KAG
17	Breast Cancer	683	9	9–10	N	UCI
18	Chlamydia	1243	3	2–10	N	CAT
19	CMC	1473	9	2–4	N	UCI
20	Energy Rating	1263	8	2–6	N	GOV
21	Flags	194	27	1–10	N	UCI
22	Mushrooms	792	22	1–9	N	UCI
23	Post-Operative	90	7	2–3	N	UCI
24	SPECT Heart	267	22	2	N	UCI
25	Vehicle Deaths	1320	4	2–6	N	CAT

Source: The authors.

Table 3: Characteristics used for the analyses

Characteristic	Description
WCE(1)	within-cluster variability based on WCE in the dataset
CLU_1	mean optimal cluster solution based on PSFE
CLU_2	mean optimal cluster solution based on BK
WCE_M	mean level of decrease of WCE
WCE_O1	mean WCE in the optimal cluster solution based on CLU_1
WCE_O2	mean WCE in the optimal cluster solution based on CLU_2

Source: The authors.

The second analysis aims to identify and to characterize similar groups of the examined datasets in Tab. 2 using the unsupervised classification, see e.g. (Hartigan, 1975). In the

first step, HCA with the *Ward* method was used on the variables in Tab. 3, since it usually provides the most distinct groups of datasets for quantitative data, see e.g. (Padilla et al., 2007). In the second step, the newly formed groups of datasets are further examined using internal characteristics.

4. Results

In this section, the analyzed datasets are examined in two ways. First, from the point of view of economic and non-economic datasets as defined in Tab. 2; second, from an aspect of the homogenous groups of datasets based on the HCA classification.

4.1 Economic and non-economic datasets

Tab. 4 presents mean values for the examined economic and non-economic datasets for the characteristics from Tab. 3. With respect to the values, general tendencies between the types of datasets can be observed. The economic ones contain slightly more variability than the non-economic ones as it is expressed by WCE(1). According to the variables CLU_1 and CLU_2, the mean optimal number of clusters is lower by economic datasets. This supports the idea that it is more difficult to identify a higher amount of homogenous clusters in the economic datasets than in the non-economic ones. In such the cases, the low-cluster solution is often preferred. Perhaps the clearest differences between economic and non-economic datasets are shown by the WCE_M characteristic which decreases by 6.6% in the economic datasets (1-0.934) on average, whereas the non-economic ones decrease by 11%. According to the Kolmogorov-Smirnov test, all the characteristics are normally distributed, and thus, the independent sample t-tests can be used to test the differences between economic and non-economic datasets. Regarding the p-values in Tab. 4, it is apparent that the WCE_M characteristic significantly differs between economic and non-economic datasets at the standard 5% significance level.

Table 4: Internal characteristics of economic (E) and non-economic datasets (N)

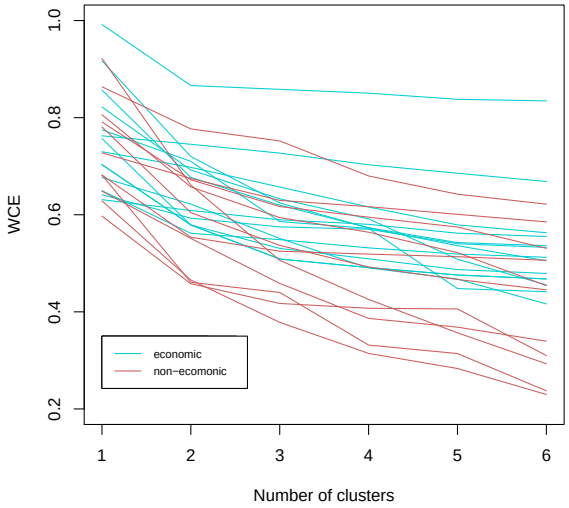
Type	WCE(1)	CLU_1	CLU_2	WCE_M	WCE_O1	WCE_O2
E	0.763	2.596	2.712	0.934	0.635	0.628
N	0.739	2.909	2.864	0.890	0.554	0.553
p-value	0.595	0.317	0.532	0.020	0.112	0.130

Source: The authors.

The differences between variability decreases of economic and non-economic datasets can also be expressed by visualizing the WCE coefficient values in one- to the six-cluster solution for all the examined datasets, see Fig. 1. Although the results are presented for the ES measure, the other examined similarity measures perform similarly. From the chart, it is apparent that the WCE decrease of the economic datasets is much lower than in the non-economic ones. The change usually comes in the two-cluster solution, where an elbow of the decline for the majority of the economic datasets occurs. Since this point, their within-cluster variability decreases very slowly. The non-economic datasets do not have such an elbow, and thus, their within-cluster variability continues to decline with the increasing number of clusters.

For both types of datasets, it was also found out which similarity measures provide the best cluster assignment regarding the PSFE index, by which the higher value in a certain cluster solution indicates a better clustering. The results are displayed in Tab. 5, which shows the percentages of situations when a given similarity measure ranks as the first one among the others. Regarding the economic datasets, it was the IOF measure, which ranked first in 35.7% of datasets. Concerning the non-economic datasets, it was the ES measure with the share of 36.4% first rank datasets. The lowest share of the first ranks was provided using the LIN measure using both the types of datasets.

Figure 1: WCE values in one- to six-cluster solution for all the examined datasets using the ES measure



Source: The authors

Table 5: Percentage of the best similarity measures for economic (E) and non-economic (N) datasets based on PFSE

Type	ES	IOF	LIN	VE
E	28.6%	35.7%	14.3%	21.4%
N	36.4%	27.3%	9.1%	27.3%

Source: The authors.

4.2 HCA classification

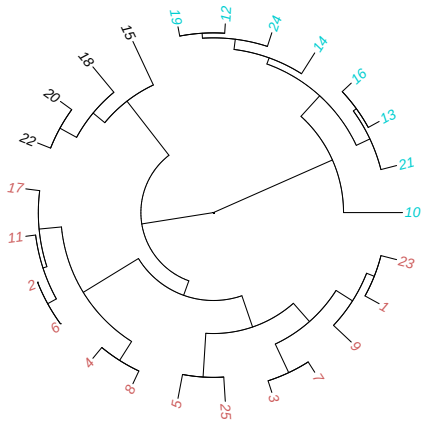
Applying HCA with the *Ward* method on the characteristics in Tab. 3, the three-cluster solution was chosen, see Fig. 2. Tab. 6 contains center, minimal and maximal values of the internal characteristics for these groups. The results are supported by the charts in Fig. 3 in which the economic (range 1–14) and non-economic (range 15–25) datasets are well separated.

Table 6: Center, minimal and maximal values of the internal characteristics for the groups G1-G3

Group	WCE(1)	CLU_1	CLU_2	WCE_M	WCE_O1	WCE_O2
G1	0.86±0.13	2.38±0.38	2.38±0.38	0.94±0.03	0.76±0.14	0.75±0.15
G2	0.64±0.04	3.00±1.00	3.13±0.63	0.85±0.02	0.40±0.06	0.41±0.06
G3	0.78±0.15	3.00±1.00	3.00±1.00	0.90±0.07	0.60±0.06	0.59±0.07

Source: The authors.

Figure 2: Dendrogram of HCA with internal characteristics of the examined datasets (Ward method)

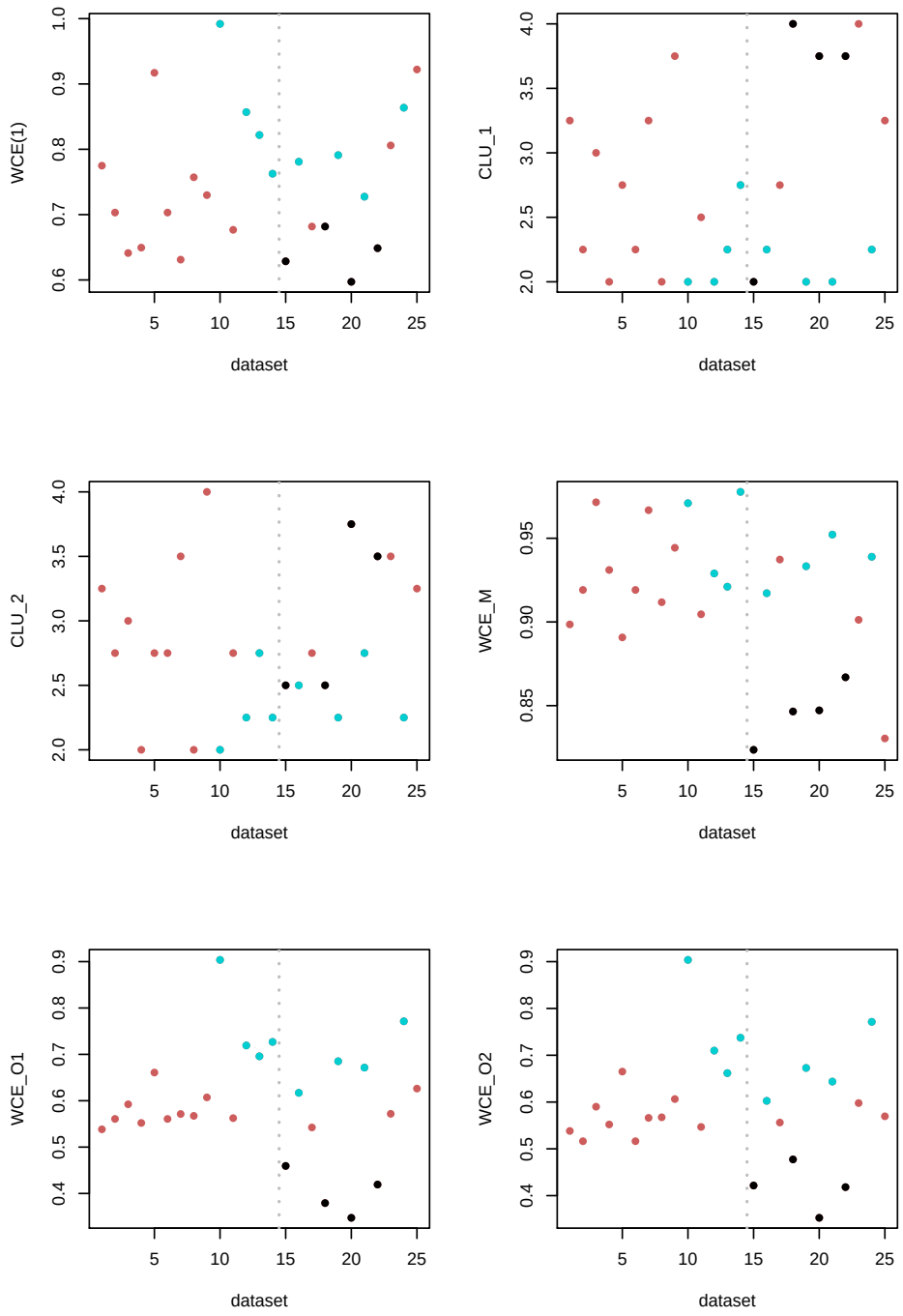


Source: The authors.

Regarding Tab. 6, the resulting groups of datasets can be characterized in the following way. The G1 group contains mostly non-economic datasets with a higher within-cluster variability both in the whole dataset expressed by WCE(1) and also in optimal cluster solutions as it is described by WCE_O1 and WCE_O2. The G2 group is formed by the non-economic datasets with the lowest WCE(1), WCE_O1 and WCE_O2. Also, the within-cluster variability decrease is the highest in this group. Thus, these datasets are easy to cluster. The G3 group comprises mostly the economic datasets. Since the economic datasets are not divided into low- and high-variability ones, their characteristics occur in between the previous two groups.

When looking at Tab. 7, which shows the percentages of first ranks among the four used similarity measures, it is evident that the ES measure provides the best clusters in the first group of datasets (non-economic, high variability). On the contrary, the IOF measure provides the best results in the third group represented in particular by economic datasets. In the second group (non-economic, low variability), the clustering performance of both these measures is at a similar level. The VE measure and especially the LIN measure perform poorly in this comparison of economic and non-economic datasets. However, this does not necessary mean that these are unsuitable measures. For instance, in (Šulc, 2016) has been shown that the VE measure performs well in datasets with the low number of variables, whereas the LIN measure is suitable for datasets with a larger number of variables and with a large number of categories.

Figure 3: Datasets distribution according to HCA classification (G1 – blue, G2 – black, G3 – red)



Source: The authors.

Table 7: Percentage of the best similarity measures for three groups of datasets based on PFSE

Group	ES	IOF	LIN	VE
G1	62.5%	12.5%	12.5%	12.5%
G2	50.0%	50.0%	0.0%	0.0%
G3	23.1%	38.4%	15.3%	23.1%

Source: The authors.

5. Conclusion

This paper dealt with a comparison of 14 economic and 11 non-economic datasets characterized by categorical variables in hierarchical cluster analysis (HCA). Two types of analyses were performed. In the first one, the definition for economic and non-economic datasets was declared by the authors; in the second one, the datasets were divided according to the results of the HCA classification of their internal characteristics.

In the first analysis, it was discovered that the examined economic datasets have only slightly higher level of the within-cluster variability on average, but due to their more complicated structure, their variability decreases much more slowly compared to the examined non-economical datasets. The faster variability decrease was proofed to be statistically significant.

In the second analysis, economic and non-economic datasets were clearly distinguished by the HCA classification of the examined datasets. Thus, based on a dataset type, one might assume on the ability of a dataset to be clustered. For each dataset type, a particular similarity measure was chosen as well. The IOF measure was chosen for the economic datasets and the ES measure for non-economic ones.

Although the study was performed on a relatively large sample of real-world datasets in comparison to the commonly used number of datasets in other papers dealing with classification evaluation, we are aware that the data generation would help us to focus on this topic more precisely. Therefore, we plan to focus on economic and non-economic datasets generation. There were introduced some interesting approaches using bootstrapping, see (Moreau and Jain, 1987) or (Laan and Pollard, 2003), which could be used as an inspiration for our future research.

Acknowledgements

This work was supported by the University of Economics, Prague under Grant IGA F4/41/2016.

References

[1] Bache, K. and Lichman. 2013. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml/datasets.html`.

[2] Boriah, S. and 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] Chen, K. and Liu, L. 2009. "Best K": critical clustering structures in categorical Datasets. Knowledge Information System, 20, pp. 1–33.

[4] Eskin, E., Arnold, A., Prerau, M. Portnoy, L, Stolfo, S. 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.

[5] Hartigan, J.A. 1975. Clustering algorithms. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley. ISBN=0-471-35645-X

[6] Laan, M. and Pollard, K.S. 2003. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. Journal of Statistical Planning and Inference, vol. 117, iss. 2, pp. 275-303.

[7] 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., pp. 296-304.

[8] Löster, T., Pavelka, T. 2013. Evaluating of the results of clustering in practical economic tasks. In The 8th International Days of Statistics and Economics. Slaný : Melandrium, pp. 804-818.

[9] Moreau, J.V. and Jain, K.J. 1987. The Bootstrap Approach to Clustering. In Proceedings of the Pattern Recognition Theory and Applications, pp. 275-303. ISBN: 978-3-642-83069-3.

[10] Padilla, G., Cartea, M.E., Ordás, A. 2007. Comparison of Several Clustering Methods in Grouping Kale Landraces. Journal of the American Society for Hortucultural Science, vol. 132, iss. 3, pp. 387-395.

[11] Sparck-Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. In Journal of Documentation, 1972, vol. 28, iss. 1, pp. 11-21. Later: In Journal of Documentation, vol. 60, iss. 5, pp. 493-502.

[12] Řezanková, H., Löster, T., Húsek, D. 2011. Evaluation of categorical data clustering. In Advances in Intelligent Web Mastering 3. Berlin : Springer Verlag, pp. 173-182.

[13] Simonoff, J. 2010. Analyzing Categorical Data. Springer Texts in Statistics. Springer New York.

[14] Šulc, Z. and Řezanková, H. 2015. nomclust: An R package for hierarchical clustering of objects characterized by nominal variables. In Proceedings of the 9th International Days of Statistics and Economics. Melandrium, Slaný, pp. 1581-1590.

[15] Šulc, Z. 2016. Similarity measures for nominal data in hierarchical clustering. Dissertation thesis, University of Economics, Prague.