A CASE STUDY OF CUSTOMER SEGMENTATION WITH THE USE OF HIERARCHICAL CLUSTER ANALYSIS OF CATEGORICAL DATA

JANA CIBULKOVÁ, ZDENĚK ŠULC
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill sq. 4, Prague, Czech Republic
email: jana.cibulkova@vse.cz, zdenek.sulc@vse.cz

Abstract

Cluster analysis is a multivariate statistical method with a wide range of possible applications. It is especially useful for market segmentation, in which objects are divided into homogenous segments (clusters) which are further analyzed to obtain segment-specific insights. This contribution presents an application of cluster analysis on multivariate data provided by a company from a field of tourism. The dataset contains information about 5,755 travels of its customers, such as the number of passengers traveling together, age, nationality, route details, price, number of destinations booked, etc. The goal of the analysis is to divide the customers into several distinct segments according to their profiles, while demonstrating the importance of distance measure selection and linkage method selection. The results of the analysis will help to develop a targeted marketing program for the company. Since the dataset contains categorical or categorized variables, hierarchical cluster analysis for categorical data is applied to perform the market segmentation. Due to the fact, that clustering process is always strongly dependent on a similarity measure used and also on a linkage method, the optimal cluster assignment is being chosen among five similarity measures for categorical data and three linkage methods. Clustering solutions corresponding to a specific similarity measure and a specific linkage method are compared and evaluated by internal evaluation indices, which allow finding the optimal number of segments, evaluate their internal consistency and determine the best clustering solution possible.

***Keywords**: hierarchical clustering, categorical data, customer segmentation*

1. Introduction

Every company should understand the market they are working in and how to better target their customers. Identifying a target market helps a company to develop effective marketing communication strategies and focus marketing programs on customers who are most likely to purchase the product. Once a company knows its target group, it is much easier to make decisions about media allocations. For content marketing purposes, marketing personas are needed to help delivering a content that will be most relevant to company's target audience. A marketing persona is a composite sketch of a key segment of target audience including customer demographics, behavior patterns, motivations, and goals. It is a semi-fictional representation of company's average customer based on market research and real data about existing customers. If a company provides a wide range of products or its customers are not "homogeneous", it might be useful to create several marketing personas, each persona representing a part of the company's target audience. In this scenario, cluster

analysis could be used to divide customers into more "homogeneous" groups, while a marketing persona for each group would be created.

Cluster analysis has become a common tool in the field of marketing, especially for market segmentation. Smith (1956) has introduced market segmentation to marketing and it became one of the most fundamental strategic marketing concepts. Market segmentation is the division of the market or population into subgroups with similar motivations. The segmentation is often based on geographic differences, personality differences, demographic differences, use of product differences, and psychographic differences. Many authors (e.g. Arimond and Elfessi, 2001; Dolnicar, 2003; Ernst and Dolnicar, 2017; Ibrahim *et al*., 2017; Malone and Lusk, 2017; Müllensiefen *et al*., 2017; Myers, 1996) focus on this field of cluster analysis application, suggest appropriate approaches, create new methods, solve specific problems, point out the most common mistakes, etc., which indicates the importance of the research in this field.

This contribution analyzes a dataset from a company operating in the tourism industry in Europe, which wants to create several marketing personas. Therefore, the goal of the study is to identify distinct groups of customers using cluster analysis on multivariate categorical data provided by this company. The result of the analysis is a segmentation of company's customers into several groups based on the customer's basic characteristics. The company will use this knowledge to develop targeted marketing programs by creating a marketing persona for each group. Second goal of this contribution is to demonstrate the importance of distance measure selection and linkage method selection, that leads to different clustering solutions. The third aim of the study is to emphasize the importance of the evaluation of clustering solution using internal evaluation criteria, that measure certain relevant features of created clusters and can be used to determine the optimal number of clusters and to compare clustering solutions. In order to achieve these goals, distance-based clustering methods are used. To be specific, hierarchical clustering methods (*single-linkage, average-linkage* and *complete-linkage*) are applied to the dataset in combination with five distance measures (*ES, IOF, LIN, VE* and *SM*). These distance-based methods were chosen for the analysis because they are widely used and very popular due to their simplicity and ease of implementation in a wide variety of scenarios.

2. Dataset description

The company provided a dataset containing 18 different variables with the basic information of 5,747 trips made by passengers from all around the world organized by the company. The overview of the variables is in Table 1.

Table 1: Original dataset's variables - overview

Variable name	Variable type	Variable description
Passenger Number	numerical	Number of the passengers traveling on the trip.
Leader Age	numerical	Age of the passengers' leader.
Leader Region	categorical	Region, from where the passengers' leader comes from.
Children Traveling	binary	Was there a child traveling?
Trip Season	categorical	Season of the year, when the trip was made.
Departure Hour	numerical	Hour of departure on the trip.
Booking in Advance	categorical	How much in advance was the trip booked?
Payment Method	categorical	What payment method was used?
Trip Distance	categorical	How big distance does the trip cover?

Sightseeing	binary	Did the trip include sightseeing?
Origin Location	categorical	The town of departure.
Origin Country	categorical	The country of departure.
Destination Location	categorical	The destination town.
Destination Country	categorical	The destination country.
Origin is Airport	binary	Was the origin location an airport?
Origin is Capital	binary	Was the origin location a capital?
Destination is Airport	binary	Was the destination location an airport?
Destination is Capital	binary	Was the destination location a capital?

Source: The authors' work.

The variable *Passenger Number* was initially quantitative, so it was necessary to transform it before the actual analysis is performed and create four categories instead – Single, Couple, Family, Large Group.

The variables *Origin Country* and *Destination Country* contained a large number of categories whose over-fragmentation would not be beneficial for the cluster analysis. Moreover, the company considers an information of knowing in what (larger) part of Europe are the customers traveling more beneficial than focusing on a specific country. Hence, the variables *Origin Country* and *Destination Country* were transformed into new variables *Origin Region* and *Destination Region* with three categories – Eastern Europe, Central Europe, Western Europe.

In cluster analysis, variables selection is crucial. It needs to be done with the purpose of the study in mind. The purpose of the study is to divide the customers into clusters, based on their behavior and trip preferences. Private personal information about customers (such as age, gender, preferred payment method, nationality, etc.) will be relevant later, in the process of creating marketing personas by marketing department. After thoughtful consideration, variables in Table 2 were chosen as input variables for the cluster analysis. These variables contain basic information about passengers and their trips with the company.

Table 2: An overview of the analyzed variables

Variable name	Variable type	Variable description
Passenger Number	categorical	Group size and type is the customers.
Children Traveling	binary	Was there a child traveling?
Trip Season	categorical	Season of the year, when the trip was made.
Booking in Advance	categorical	How much in advance was the trip booked?
Sightseeing	binary	Did the trip include sightseeing?
Origin Region	categorical	The region of departure.
Destination Region	categorical	The destination region.

Source: The authors' work.

Also, the variables should not be strongly correlated, as this would increase their weight in the analysis, which could lead to misleading results. Before the variables enter the analysis, their correlations were found using the algorithm CATPCA (non-linear PCA method). This method was chosen since it transforms categorical data into quantitative ones and calculate correlation matrix containing Pearson correlation coefficients. Table 3 shows the correlations between the pairs of the analyzed variables. The variables *Origin Region* and *Destination Region* are almost linearly dependent. It means that customers are very rarely organizing trips

between regions. To prevent distortion of the cluster analysis, only the variable *Origin Region* will be used in the analysis.

Table 3: Correlations among chosen variables

	Passenger Number	Children Traveling	Trip Season	Booking in Advance	Sightseeing	Origin Region	Destination Region
Passenger Number	1.000						
Children Traveling	-0.218	1.000					
Trip Season	-0.124	0.094	1.000				
Booking in Advance	-0.103	-0.010	0.252	1.000			
Sightseeing	-0.045	0.005	0.027	0.132	1.000		
Origin Region	0.064	-0.003	-0.102	0.041	0.044	1.000	
Destination Region	0.061	-0.016	-0.102	0.042	0.044	0.988	1.000

Source: The authors' work.

3. Design of the experiment and methods used

The final dataset consists of six categorical variables, namely *Passenger Number, Children Traveling, Booking in Advance, Sightseeing, Trip Season, Origin Region*. The three chosen hierarchical clustering methods are used in combination with five distance measures. The results of clustering process are evaluated and compared using the *within-cluster mutability* (WCM) and *the pseudo F index based on the mutability* (PSFM). Based on the values of internal evaluation indices the most suitable solution is selected and used for further analysis by marketing department, that is going to use additional information about the customers and create marketing personas.

3.1 Methods of hierarchical clustering

Three hierarchical methods are used in this contribution: the complete linkage, the average linkage, and the single linkage methods.

The complete linkage method defines a dissimilarity between two clusters as the dissimilarity between two farthest observations from both clusters. Let us denote $D_{complete}(C_k, C_l)$ the distance between clusters C_k and C_l and $\mathbf{x}_i$ the i-th data observation. Then, the dissimilarity between two clusters can be expressed by the formula:

$$D_{complete}(C_k, C_l) = \max_{\boldsymbol{x}_i \in C_k;\, \boldsymbol{x}_j \in C_l} D(\mathbf{x}_i, \mathbf{x}_j). \quad (1)$$

The average linkage method takes average pairwise dissimilarity between observations in two different clusters. Let us denote $D_{average}(C_k, C_l)$ the distance between clusters C_k and C_l, with the numbers of observations in k-th and l-th clusters denoted as n_k, n_l. Let $\mathbf{x}_i$ be the i-th data observation. Then, the dissimilarity between two clusters can be expressed by the formula:

$$D_{average}(C_k, C_l) = \frac{\sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_l} D(\mathbf{x}_i, \mathbf{x}_j)}{n_k n_l}. \quad (2)$$

The single linkage method defines a dissimilarity between two clusters as the dissimilarity between two closest observations from both clusters. Let us denote $D_{single}(C_k, C_l)$ the distance between cluster C_k and C_l and $\mathbf{x}_i$ the i-th data observation. Then, the dissimilarity between two clusters can be expressed by the formula:

$$D_{single}(C_k, C_l) = \min_{\boldsymbol{x}_i \in C_k;\, \boldsymbol{x}_j \in C_l} D(\mathbf{x}_i, \mathbf{x}_j). \quad (3)$$

3.2 Distance measures

Due to the nature of the data provided by the company, these five measures were used for the experiment: ES measure (Eskin *et al.*, 2002), IOF measure (Sparck-Jones, 1972), LIN measure (Lin, 1998), VE measure (Šulc, 2016) and SM measure (Sokal and Michener, 1958).

Let us denote the categorical data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, \ldots, n$ and $c = 1, 2, \ldots, m$; n is the total number of objects and 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.

The overview of formulas can be found in Table 4, where the column $S_c(x_{ic} = x_{jc})$ presents similarity computation for matches of categories in the *c*-th variable for the i-th and j-th objects, and the column $S_c(x_{ic} \neq x_{jc})$ represents mismatches of these categories. The last column represents the total similarity $S(\mathbf{x}_i, \mathbf{x}_j)$ between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

Table 4: Similarity measures overview

Measure	$S_c(x_{ic} = x_{jc})$	$S_c(x_{ic} \neq x_{jc})$	$S(\mathbf{x}_i, \mathbf{x}_j)$
ES	1	$\frac{K_c^2}{K_c^2 + 2}$	$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$
IOF	1	$\frac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{jc})}$	$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$
LIN	$2 \ln p(x_{ic})$	$2\ln(p(x_{ic}) + p(x_{jc}))$	$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m} (\ln p(x_{ic}) + \ln p(x_{jc}))}$
VE	$-\frac{1}{\ln K_c} \sum_{u=1}^{K_c} p_u \ln p_u$	0	$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$
SM	1	0	$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$

Source: The authors' work.

In order to compute a proximity matrix, it is necessary to compute dissimilarities $D(\mathbf{x}_i, \mathbf{x}_j)$ between all pairs of objects. To obtain dissimilarities from similarities, transformations inspired by Deza and Deza (2013) are used. For the measures SM and VE, the dissimilarities are calculated using the following formula:

$$D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

For the measures ES, IOF and LIN, the dissimilarities are calculated using this formula:

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

.

3.3 Evaluation of clustering solution

The evaluation of the created clusters is necessary to obtain only one final clustering solution. Řezanková *et al*. (2011) proposed several criteria based on variability measures suitable for categorical data.

It is desired to create clusters with the low within-cluster variability. In this paper *the pseudo F index based on the mutability* (PSFM), is used for this purpose, which can be expressed by the following formula:

$$PSFM(k) = \frac{(n-k)[WCM(1) - WCM(k)]}{(k-1)WCM(k)}, \tag{6}$$

where $WCM(1)$ is the variability based on the mutability in the whole dataset with n observations, and $WCM(k)$ is the within-cluster variability in the k-cluster solution. $WCM(k)$ is computed as

$$WCM(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} G_{gc}, \tag{7}$$

where n_g is the number of objects in the g-th cluster ($g = 1, 2, \ldots, k$), m is the total number of variables and G_{gc} is the mutability by the c-th ($c = 1, 2, \ldots, m$) variable in the g-th cluster expressed as

$$G_{gc} = 1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \right)^2, \tag{8}$$

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

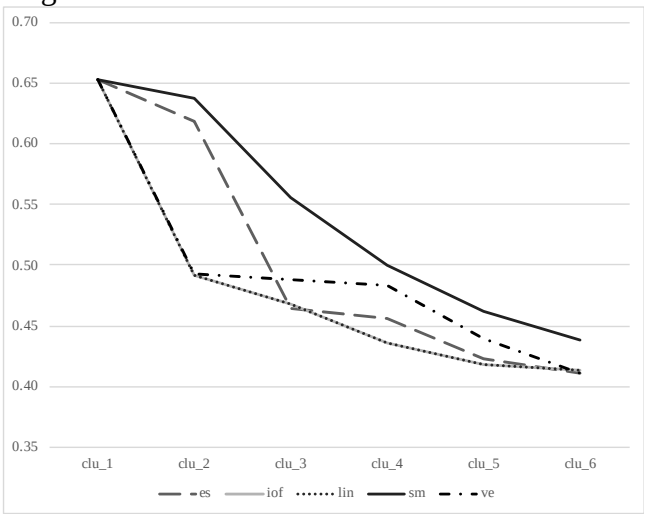
A number of cluster is considered to be the optimal number of clusters, when its value of PFSM index is the highest of PSFM values across all examined cluster solutions. In such a cluster solution, the highest decrease of the within-cluster variability occurs.

4. Experiment and results

The whole analysis was performed in IBM SPSS Statistics 25 (IBM Corp., 2017), using its standard procedures. First, the dataset was modified as described in Section 2. Then, three hierarchical clustering methods introduced in Section 3.1 in combination with five distance measures from Section 3.2 were used in hard clustering process. Lastly, the WCM and PSFM indices for each clustering solution is computed. As the final clustering solution is considered the one, where the highest decrease of WCM occurs and it is also interpretable and makes sense for the company. The optimal number of clusters is determined using the PSFM internal evaluation index.

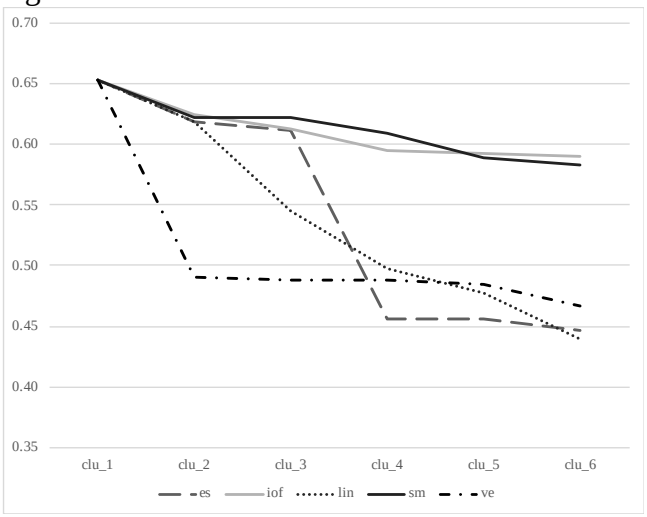
The within-cluster variabilities of clustering solutions with given distance measure, where the number of clusters in the clustering solution varies from one to six, are shown in Figure 1, Figure 2 and Figure 3. Each figure corresponds to one hierarchical clustering method and shows values of WCM for different numbers of clusters.

Figure 1: Complete-linkage method



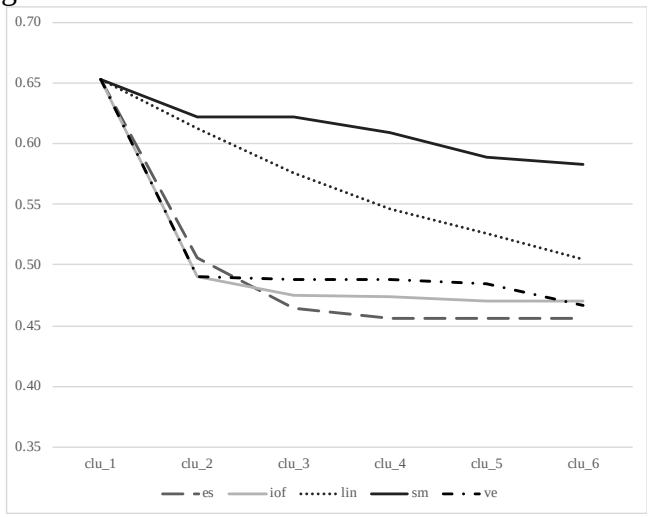
Source: The authors' work.

Figure 2: Average-linkage method



Source: The authors' work.

Figure 3: Single-linkage method



Source: The authors' work.

Table 5 shows the optimal numbers of clusters based on values of the PSFM index. Bold values in the table represent the optimal number of clusters in the clustering solutions corresponding to the highest value of the PSFM index for a given distance measure and a given linkage method.

Table 5: The optimal number of clusters based on PSFM

Measure	Complete-linkage	Average-linkage	Single-linkage
ES	**3**	4	2
IOF	**2**	2	**2**
LIN	**2**	4	2
VE	4	2	2
SM	2	**2**	**2**

Source: The authors' work.

As seen from Figures 1-3 and Table 5, the complete-linkage method provided the best results (based on the values of internal evaluation indices), since it actually decreased the variability of clustering solution with increasing number of clusters by all similarity measures. The average-linkage method gives satisfying results only in combination with the VE measure. A significant decrease of within-cluster variability for the single-linkage method occurs only in the two-clusters solution with measures ES, IOF, LIN; however, the variability does not have decreasing tendency with an increasing number of clusters.

A combination of the complete-linkage method with the ES and IOF measures provides the same results. The VE measure gives the same results combined with the average-linkage method as well as with the single-linkage method.

Based on the values in Table 5, it is also obvious that a distance measure is a crucial parameter of cluster analysis, that determines an outcome of an analysis. For example, the optimal number of clusters for a clustering solution from the complete linkage method in combination with the LIN measure is equal to two, while combination of the complete linkage method and measure VE leads to a solution with four clusters. Also, within-cluster variability can be very different for different distance measures. While clustering using the average linkage method with measures ES, VE and LIN leads to clustering solutions with low within-cluster variability, clustering solutions corresponding to the average linkage method with measures SM and IOF have much higher within-cluster variability.

The most promising clustering solutions appear to be:

- **three-clusters solution with the ES measure and the complete linkage method**
    - Using this combination of hierarchical clustering method and distance measure, a division of the dataset into three clusters was achieved. The first cluster consists of customers traveling for pleasure (with a lot of sightseeing) without children. The second cluster is the cluster of customers traveling without children for purposes other than sightseeing. The third cluster includes all customers traveling with children. This third cluster is small due to a small number of clients with children.

- **two-or-three-clusters solution with IOF or LIN measure and complete linkage method**
    - This two-clusters solution divides customers into two groups using mostly *Sightseeing* variable. The first cluster includes customers interested in sightseeing. The second one includes customers, that are not interested in sightseeing trip.
    - Three-clusters clustering solution is slightly worse than the previous one, however it also gives a relatively meaningful division of customers into three clusters. First cluster contains customers interested in sightseeing. Customers traveling without any stops on their way are divided into two clusters – in the second cluster are the customer who booked the trip on the last-minute and in the third cluster are the ones, who booked the trip sufficiently in advance.
- **two-clusters solution with the ES measure and the average linkage method**
    - This two-cluster solution creates one very large cluster of customers traveling without children and one small cluster with customers traveling with children.

According to internal evaluation criteria, the best customer segmentation seems to be dividing customers into three clusters using the complete-linkage hierarchical clustering and the IOF or LIN distance measure or the complete-linkage hierarchical clustering in the combination with the ES measure.

Clustering solution corresponding to a combination of the complete-linkage method and the IOF or LIN distance measures makes it very easy to separate sightseeing tourists from the non-tourists (most likely people on business trips), which are further divided into a group of those who book the trip on time and a group of those who book it on last-minute. This solution might be interesting for the company due to several reasons. The marketing department can use additional information about the customers and create three marketing personas. This will help them develop targeted marketing programs for people who are interested in sightseeing, people who are not interested in sightseeing and book the trip on very last minute (people who urgently need a transport from one place to another) and people who are not interested in sightseeing and book the trip in advance (most likely businessmen).

Another possibility is to follow the customer segmentation corresponding to the clustering solution obtained by the complete-linkage hierarchical clustering in combination with the ES measure. This clustering solution divides customers into three groups. If the marketing department decides to create marketing personas for these three groups, they can focus their marketing on people who are travelling with children, people who are travelling for pleasure (and are interested in sightseeing) and on people who just need a transfer from one place to another.

Both solutions are well interpretable, final clusters have the low within-cluster variability and make sense for the company.

5. Conclusion

This contribution focused on application of cluster analysis in the field of marketing. Cluster analysis of multivariate categorical dataset containing the basic information about customers of one company was performed in order to divide the company's customers into distinct groups. For this purpose, three hierarchical clustering methods were used in

combination with five distance measures, leading to several clustering solutions. The obtained clustering solutions were analyzed, compared, and interpreted. Internal evaluation criteria were used for the optimal number of cluster determination and for evaluation of created clusters. The results of the analysis will be used by the company's marketing department to create targeted marketing programs. This contribution also illustrated the importance of distance measure selection and linkage method selection in hierarchical clustering, as they lead to different clustering solutions.

Acknowledgements

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

References

[1] Arimond, G., Elfessi, A. 2001. A Clustering method for categorical data in tourism market segmentation research. Journal of Travel Research - J TRAVEL RES, vol. 39, pp. 391-397.

[2] Deza, M. M., Deza, E. 2013. Encyclopedia of Distances, Springer. ISBN 978-3-642-30957-1

[3] Dolnicar, S. 2003. Using Cluster Analysis for Market Segmentation – Typical Misconceptions, Established Methodological Weaknesses and Some Recommendations for Improvement, Journal of Marketing Research, vol. 11 (2), pp. 5-12.

[4] Ernst, D., Dolnicar, S. 2017. How to Avoid Random Market Segmentation Solutions. Journal of Travel Research, vol. 57 (1), pp. 69-82. doi:10.1177/0047287516684978

[5] 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.

[6] IBM Corp. 2017. IBM SPSS Statistics, Version 25.0. Armonk, NY: IBM Corp.

[7] Ibrahim, A., Knox, K., Rundle-Thiele, S., Arli, D. 2017. Segmenting a Water Use Market. Social Marketing Quarterly, vol. 24 (1), pp. 3-17. doi:10.1177/1524500417741277

[8] Lin D. 1998. An information-theoretic definition of similarity. In ICML '98: Proceedings of the 15th International Conference on Machine Learning, pp. 296-304. San Francisco: Morgan Kaufmann Publishers Inc.

[9] Malone, T., Lusk, J. L. 2017. If you brew it, who will come? Market segments in the U.S. beer market. Agribusiness, vol. 34 (2), pp. 204-221. doi:10.1002/agr.21511

[10] Müllensiefen, D., Hennig, C., Howells, H. 2017. Using clustering of rankings to explain brand preferences with personality and socio-demographic variables. Journal of Applied Statistics, vol. 45(6), pp.1009-1029. doi:10.1080/02664763.2017.1339025

[11] Myers, J. H. 1996. Segmentation and Positioning for Strategic Marketing Decisions, Chicago: American Marketing Association. ISBN-13: 978-0877572596

[12] Řezanková, H., Löster, T., Húsek, D. 2011. Evaluation of Categorical Data Clustering. Advances in Intelligent and Soft Computing Advances in Intelligent Web Mastering, vol. 3, pp. 173-182. doi:10.1007/978-3-642-18029-3_18

[13] Smith, W. R. 1956. Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing, vol. 21 (1), pp. 3-8.

[14] Sokal, R.R., Michener, C.D. 1958. A Statistical Methods for Evaluating Relationships. University of Kansas Science Bulletin, vol. 38, pp. 1409-1448.

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

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