MULTIVARIATE METHODS FOR SURVEY EVALUATION: A CASE STUDY OF BIG DATA AND THE NEW DIGITAL DIVIDE

JANA CIBULKOVÁ[a,*], RICHARD NOVÁK[b], ZDENĚK ŠULC[a]

jana.cibulkova@vse.cz, xnovr900@vse.cz, zdenek.sulc@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 paper evaluates the results of a survey regarding challenges connected to the Big Data phenomenon and the New Digital Divide. Following the authors' previous research, these challenges consist of issues such as business advantage, privacy intrusion, or big brother effect. Perception of these challenges by respondents (IT students and IT professionals) is studied in the paper, and multivariate methods are applied to the survey data to obtain a deeper insight and to discover hidden patterns in the data. The goal of the study is to identify groups of challenges based on respondents' judgment. For this purpose, cluster analysis and factor analysis are used. Results obtained using these methods are compared and evaluated.

Key words

multivariate methods, cluster analysis, factor analysis, survey, Big Data, New Digital Divide

JEL classification

C38, C63, C88

1. Introduction

A survey is a method of gathering information from a sample of people intending to generalize the results to a larger population. Surveys can provide insight for nearly everyone engaged in the information economy, from businesses and the media to government and academics. To analyze a survey's responses, multivariate methods, such as cluster analysis or factor analysis, may be used. Multivariate methods are used to perform studies across multiple dimensions while taking into consideration the effects of all variables on the responses of interest, which is often desired when analyzing a survey's outcome. Both cluster analysis and factor analysis are types of unsupervised learning methods that can be used for data reduction. However, these two methods differ in their objectives. While cluster analysis standardly attempts to group objects, factors analysis attempts to group variables.

In this paper, both methods are applied to survey data to obtain valuable information. The dataset comes from the survey, that is a part of the research focused on a Big Data phenomenon and New Digital Divide in Novák (2019). The paper focus on the specific part of the survey regarding challenges connected to the Big Data phenomenon, also described by other authors such as Boyd and Crawford (2012) for the field of "Big Data challenges" and Andrejevic (2014) for the area of "Digital Divide".

A digital divide represents an economic and social inequality in the access to, use of, or impact of information and communication technologies. The term "New Digital Divide" shifts the focus from technologies and skills towards the focus on identification who benefits from outcomes of new technology introduction. Big Data can be considered as a new technology causing a similar divide issue as the Internet in the eighties. The paper studies a perception of challenges connected to the Big Data phenomenon by respondents. These challenges consist of

issues connected to Big Data, e.g., business advantage, privacy intrusion, big brother effect, etc. and New Digital Divide.

The goal of this paper is to analyze and process data from the survey and to discover hidden patterns in the data, to categorize the challenges based on respondents' perceptions and, if it is possible, to assign them into groups proposed in Novák (2019).

Firstly, the methodology used in the paper is presented. Two multivariate methods were chosen for the case study – factor analysis and cluster analysis. The third section provides information about the study design. Insight into the problematics of the Big Data phenomenon, its challenges, and dataset and survey description are presented here. Results of the analyses are visualized and described in the fourth section. Finally, the conclusion summarizes the findings, explains their importance for the whole study.

2. Methodology and theoretical background

In the study, factor analysis and cluster analysis are applied to the data. Although both methods fall into multivariate methods category, they differ in their objectives. For factor analysis, the common objective is to explain the correlation within a dataset and to find groups of variables expressing a certain construct. The aim of cluster analysis is to identify groups of objects that are similar to each other but different from objects in other groups. In this section, both methods are described.

2.1 Factor analysis

Factor analysis is a commonly used statistical technique for examining relationships among correlated variables, see e.g. (Child, 2006). It assumes there are hidden latent variables (factors, latent constructs) which cannot be observed directly but are reflected in the analyzed variables. Hence, the transformation of the original set of variables into an equal number of variables such that each new variable is a combination of the current ones in some weight is possible. A typical way to make this transformation is to use eigenvalues and eigenvectors. Data are transformed in the direction of each eigenvector, and the new variables (*factors*) are represented using the eigenvalues. Then, the factors are sorted in decreasing order of the variances they explain. Thus, the first factor is the most influential factor, followed by the second factor, and so on. The variable reduction is performed by removing the last few factors. However, there is a need to read each of the factors and the combination of original features out of which they are made to understand what they represent. The weights of each original variable that were combined to obtain the factor are retained post-transformation. These weights are known as *factor loadings*. The factor analysis model is

$$\boldsymbol{X} = \boldsymbol{\Lambda F} + \boldsymbol{e}, \tag{1}$$

for a $p \times n$ matrix $\boldsymbol{X}$, a $p \times k$ matrix $\boldsymbol{\Lambda}$ of loadings, a $k \times n$ matrix F of scores and a p-element vector $\boldsymbol{e}$ of errors, where $k < p$, p is number of observable random variables, n represents number of observations. None of the components other than $\boldsymbol{X}$ is observed, but the major restriction is that the scores must be uncorrelated and of unit variance, and that the errors must be independent and with variances $\boldsymbol{\Psi} = \mathrm{Cov}(\boldsymbol{e})$. It is also common to scale the observed variables to unit variance. Hence factor analysis is, in essence, a model for the correlation matrix of $\boldsymbol{X}$ in the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Lambda}' + \boldsymbol{\Psi}. \tag{2}$$

There is still some indeterminacy in the model for it is unchanged if $\boldsymbol{\Lambda}$ is replaced by $\boldsymbol{G\Lambda}$ for any orthogonal matrix $\boldsymbol{G}$. Such matrices $\boldsymbol{G}$ are called *rotations.* It is customary to apply rotation to find a set of loadings that fit the observations equally well but can be more easily interpreted.

As it is impossible to examine all such rotations, computer programs carry out rotations satisfying certain criteria. According to Sass and Schmitt (2010), the choice of the rotation criterion is crucial, and researchers tend to choose rotation criteria based solely on a correlation and do not consider other important aspects of their data. In this study, one of the most widely used criteria, the *varimax* criterion, is applied to the data (Kaiser, 1958). It seeks the rotated loadings that maximize the variance of the squared loadings for each factor; the goal is to maximize some of these loadings and the rest minimize in absolute value.

2.2 Hierarchical cluster analysis – Ward's method

The Ward's method is used for the analysis. It is a distance-based hierarchical clustering method. Hierarchical cluster analysis uses a chosen dissimilarity for the objects being clustered. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is one cluster.

There occurs an inconsistency of the method's algorithm in the literature and software, see (Murtagh and Legendre, 2014). In the study, the original procedure introduced by Ward (1963) is used. It seeks to form partitions $P_n, P_{n-1}, \dots, P_1$ in a manner that minimizes the loss of information associated with each merging, quantified in terms of an *error sum of squares* (ESS) criterion. For a given group of data points $\boldsymbol{C}$, the ESS is defined by the formula:

$$\mathrm{ESS}(\boldsymbol{C}) = \sum_{\boldsymbol{x}\in\mathbf{C}} \big(\boldsymbol{x} - \mu(\boldsymbol{C})\big)\big(\boldsymbol{x} - \mu(\boldsymbol{C})\big)^T, \tag{3}$$

where $\mu(\boldsymbol{C}) = \frac{1}{|\boldsymbol{C}|}\sum_{\boldsymbol{x}\in\mathbf{C}}(\boldsymbol{x})$ is the mean of $\boldsymbol{C}$, while $|\boldsymbol{C}|$ is cardinality of $\boldsymbol{C}$.

At each step, the union of every possible pair of clusters is considered, and two clusters whose joining results in the minimum increase in the loss of information are merged. Hence, the ESS criterion minimizes the total within-cluster variance. So, Ward's minimum variance method aims at finding compact, spherical clusters.

To apply a recursive algorithm under this objective function (ESS), the initial distance between individual objects must be (proportional to) squared Euclidean distance. Therefore, the entries d_{ij} of the dissimilarity matrix for a dataset $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ follow the formula:

$$d_{ij} = d_{EUCLID}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T = \sum_{l=1}^{m} (\boldsymbol{x}_{il} - \boldsymbol{x}_{jl})^2, \tag{4}$$

where m is the dimensionality of the dataset $\boldsymbol{X}$.

3. Design of the study

The analyzed dataset contains information about the perception of challenges connected to the Big Data phenomenon by respondents in the survey in Novák (2019). Based on the previous research, he differentiates and categorizes the challenges into the following groups:

- *Hot challenges*
  There is almost unanimous agreement that the issue is important among the respondents.
- *Cold challenges*
  A strong majority of the respondents do not consider this issue to be important.
- *Hidden challenges*
  The respondents are expected to be unaware of the importance of hidden factors.

The case study aims to uncover the hidden structure and relations among variables, to categorize the challenges based on respondents' perceptions and, if it is possible, to assign the challenged into proposed categories.

To achieve this goal, firstly, exploratory factor analysis was applied to data to discover hidden patterns in the data and to obtain three factors (ideally corresponding to three groups of challenges defined above). However, interpreting the factors (sets of variables) might be challenging as well as very subjective. This approach has not provided outcomes relevant to the purpose of the study (to divide the challenges into groups based on respondents' perceptions), as shown and explained in Section 4.

Therefore, another approach to discover the desired three groups of challenges was used - cluster analysis. The definition of hot, cold, and hidden groups of challenges (stated at the beginning of this section) was rewritten using only average score and standard deviation of the score from the survey as follows:

- *Hot challenges*
  The "unanimous agreement that the issue is important among the respondents" means that the average score of questions regarding a hot challenge is high, and its standard deviation is low.
- *Cold challenges*
  Definition of cold challenges "majority of the respondents do not consider this issue to be important" assumes, that the average score of questions regarding a cold factor is low. There is no assumption on standard deviation since the definition considers the "majority of the respondents" and several outliers can cause a significant increase of a standard deviation.
- *Hidden challenges*
  This group of issues were defined by "respondents are expected to be unaware of the importance of hidden factors", hence these challenges are expected to have a medium average score, not too low nor too high.

By reformulating the definitions this way, cluster analysis can be applied on averages and standard deviations of scores of each challenge. The dataset was aggregated into a table of 11 rows (each corresponding to one variable/challenge) and two column variables (average score and standard deviation of the score). Ward's method is applied to aggregated data (that had been transformed to 0 – 1 range) to obtain three clusters of issues. This method was chosen for the analysis since this combination leads to the desired compact (spherical) clusters, and it minimizes the total within-cluster variance.

4. Data description

The data is from the survey, that is a part of the research focused on the Big Data phenomenon. The study explores the hidden relationships among 11 variables. Each variable corresponds to one challenge. The survey begins with a brief introduction to a Big Data phenomenon and establishes eleven challenges connected to Big Data. The respondents were asked to assign a score (on a scale 0 – 100) to each challenge to express their perception of the importance of a particular challenge. The score 100 should be assigned to a challenge that is, in their opinion, considered to be a serious problem and score 0 should be assigned to a challenge, which they do not find problematic at all. The overview of the variables is in Table 1.

Data was collected via an online questionnaire and it was collected from 26th November 2018 till 10th January 2019. IT professionals were invited to participate in the survey by direct email invitation while IT students were required to participate during their lessons. The questionnaire was sent to respondents by "an authority" (a teacher/manager), hence the response rate was very high.

The total number of observations (corresponding to a total number of respondents) obtained from the survey is 528. Eight respondents entered the same answer for every question of the questionnaire, which indicates that they were not taking the survey seriously. Hence corresponding eight observations were dropped from the dataset. Also, 36 observations with

missing values were excluded from the dataset as well. Thus, the final number of observations in the dataset is 477.

Since this paper focuses solely on the specific part of the survey regarding the perception of the challenges connected to Big Data phenomenon by respondents, the process of creating a questionnaire (and its methodology), data processing description and exploratory analysis will not be provided here. The in-depth description of the questionnaire and the dataset is available in the original thesis (Novák, 2019).

Table 1: Overview of variables in the dataset

Variable name	Variable description
Privacy Intrusion	Whether a respondent feels like there is an important impact of Big Data to the privacy of individuals.
New Barriers	Whether a respondent feels like Big Data divide people with advantages from the rest.
Power of All Data	Whether a respondent feels like there are only a few data monopolies (such as Google or Facebook) that can see the global view and predict a future.
Business Advantage	Whether a respondent feels like specific business advantages are available to corporations that actively collect and use Big Data.
New Big Brother Effect	Whether a respondent feels like people are being observed by technologies all the time and their life can be manipulated without their knowledge.
Missing Transparency	Whether a respondent feels that due to complicated Big Data technologies they lose global transparency.
Confusion	Whether a respondent feels like Big Data cause confusion in determining what is right and wrong.
Social Pressure	Whether a respondent feels that there is pressure on people to use new services that are used by others.
Belief in Legislation	Whether a respondent believes that proper legal regulation can solve all Big Data problems.
End of Theory	Whether a respondent feels like it is not important to understand underlying principals but to be able to show results in a lot of figures and graphs.
Data Religion	Whether a respondent feels like the quality of decisions depends only on how much data one is able to collect.

Source: the authors.

5. Implementation and results

First, this section informs briefly about the implementation of the study. Then the results of the study are presented, which is regarded as the core of the section.

5.1. Implementation in R

The complete analysis was done in R language and environment for statistical computing version 3.6.0, see R Core Team (2018). The function *factanal* was used for factor analysis. Hierarchical clustering was done using the base function *hclust*.

5.2. Results of the study

Exploratory factor analysis with various rotations was applied to the data. The interpretable outcomes were obtained when varimax rotation method was used. Figure 1 visualizes the results of the analysis – it shows each challenge and corresponding loadings for each factor. Only the loadings higher than 0.3 are considered to be relevant. Hence, let us establish the cut off 0.3 to improve visibility represented by a black dotted vertical line in Figure 1. The outcomes of the analysis are described as follows:

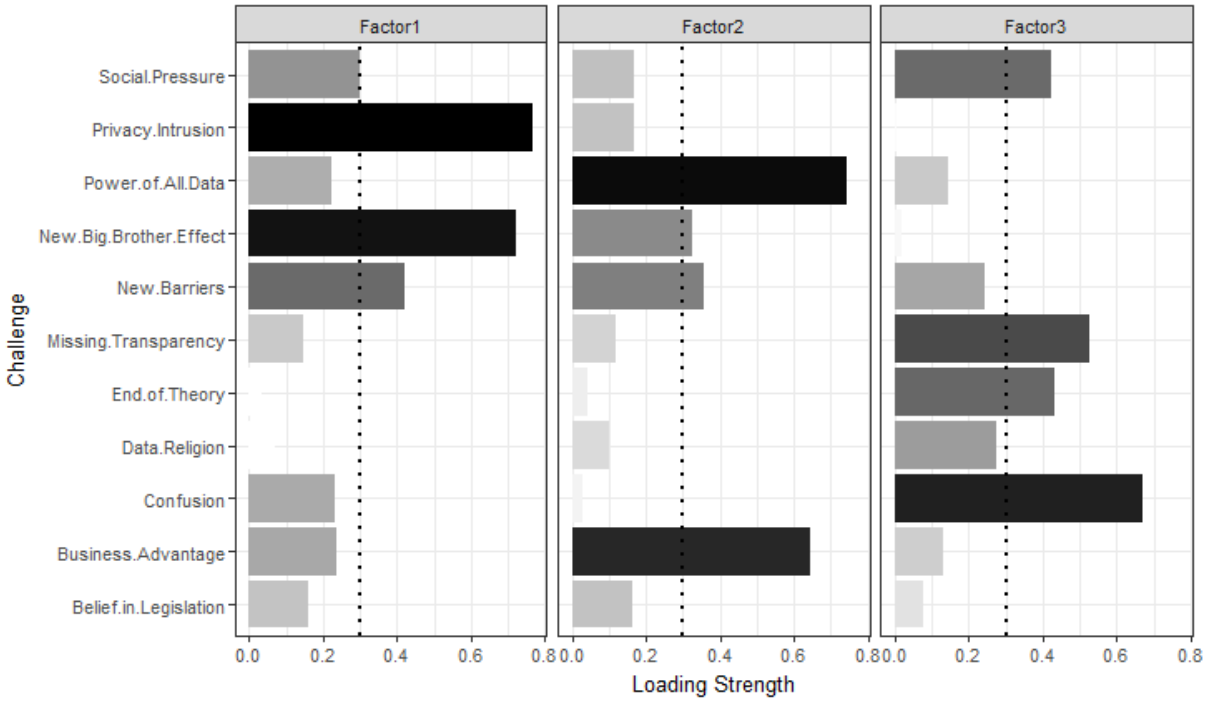
- Challenges “Privacy Intrusion”, “New Big Brother Effect” and “New Barriers”, appear together in one factor, that could be called a factor of hot challenges, and it somehow corresponds with assumptions of hot challenges made by Novák (2019). Despite that, this

group of challenges could also be called a group of *individual-related* challenges, since all the challenges are concerning a respondent as an individual.

- Challenges "Social Pressure", "Missing Transparency", "End of Theory" and "Confusion" could be considered to be a group of cold challenges. However, "Social Pressure" is balancing between groups of hot and cold challenges, which should not happen. Hence the term *perplexity-related challenges* describes this group better.
- The second factor does not fit the assumed structure; however, it could be well described as a group of *business-related challenges*.
- Moreover, two challenges "Belief in Legislation" and "Data Religion" have become insignificant, when following the 0.3 cut-off rule. Also, the double loadings of challenges "New Big Brother Effect" and "New Barriers" are not welcomed as a result.

To sum up, exploratory factor analysis definitely discovered a hidden structure in the data. However, the structure was more associated to an area of respondents' life that was affected by a challenge. It discovered a factor of challenges related to personal life of an individual, a factor of challenges related to business and a factor of challenges causing perplexity. But it did not uncover a hidden structure of a respondent's perception of given challenges.

Figure 1: Results of the factors analysis



Source: the authors.

Since the study aims to analyze a respondents' perception of given challenges and the outcomes of factor analysis could not be interpreted in such way, the focus was shifted to a definition of the assumed structure of the challenges presented in Section 3.

Hence, cluster analysis was applied on aggregated data – for each variable an average score and a standard deviation of the score was calculated. In the paper, Ward's method is used. Note, that other methods (for example average-linkage or complete linkage) lead to identical clustering solutions. The results of cluster analysis are visualized in Figure 2. This solution is definitely acceptable for further research since the interpretation of groups of challenges is very clean, and it corresponds to the previous research on this topic.

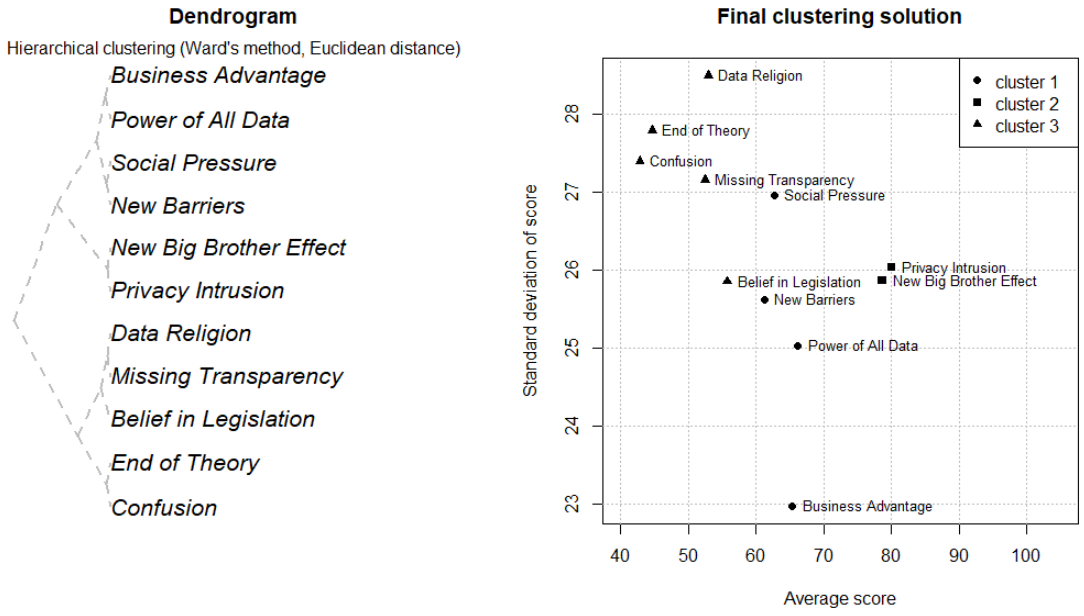
- *Cluster* 2 contains "Privacy Intrusion" and "New Big Brother Effect" (square-shaped points in Figure 2) has a high average and (relatively) low standard deviation of the score. This cluster of challenges can be considered a *cluster of hot challenges* since they fulfill the

definition from Section 3. The fact, that these two challenges were assigned to the cluster of hot challenges actually supports the assumptions about hot challenges made by Novák (2019).

- "Privacy Intrusion" is a well-known issue thanks to media (GDPR, wiki leaks, etc.) and its relevance to everybody.
- "New Big Brother Effect" is a widely popular issue and people, in general, are very familiar with this topic (George Orwell – 1984, dictatorship regimes, experience with communism, etc.)

- *Cluster 3* that contains "Data Religion", "Missing Transparency", "End of Theory" and "Confusion" (triangle-shaped points in Figure 2) may be considered to be *a cluster of cold challenges*, according to the definition. This group of issues can be described as the issues that are very specific and science-oriented. Hence the general public may not be aware of them.
  - "Data Religion" is investigating the role of data quantity in the decision-making process in a Big Data phenomenon, which might be unknown to the wide audience and therefore underestimated.
  - "Missing Transparency" challenge is dealing with mathematical algorithms, that are distant from respondents' everyday-life problems.
  - "End of Theory" observe the underlined principles of how the world is managed. That may be too philosophical for the respondents to care.
  - "Confusion" represents the fact that Big Data make the world too complicated. This challenge is probably too abstract for the respondents.
- *Cluster 1* could be referred to as *a cluster of hidden challenges* (round-shaped points in Figure 2). It contains challenges "Belief in Legislation", "New Barriers", "Power of All Data", "Social Pressure", "Business Advantage". The average score of challenges in this cluster is medium (as defined in Section 3), the standard deviation of the score varies from very low to high.

Figure 2: Results of cluster analysis



Source: the authors.

6. Conclusion

This study focused on the application of chosen multivariate methods (cluster analysis and factor analysis) in the process of a survey evaluation. Even though these methods are popular widely used unsupervised learning techniques, and they both can be used for data reduction, their objectives differ. Cluster analysis usually groups objects, and factor analysis is usually used for grouping variables. However, we used both the methods mutually in an attempt to distinguish groups of challenges connected to the Big Data phenomenon and New Digital Divide, based on the subjective perception of the challenge and awareness of the issue by respondents.

The study follows the authors' previous research, where significant challenges were determined. Firstly, exploratory factor analysis was applied to data to discover hidden patterns in the data. However, the structure discovered by factor analysis could not be interpreted in such way, that would correspond to a respondent's perception of given challenges. Hence, factor analysis provided the results that were not acceptable for the study. Thanks to a precise definition of the desired groups of challenges in the previous research, it was possible to aggregate the dataset in a way that cluster analysis could be applied to the data. This way, it was possible to assign challenges into distinct groups in a well-interpretable way and to support the partial results and assumptions from the previous research.

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] Child, D. 2006. The essentials of factor analysis. 3rd ed. London : Bloomsbury Academic, 2006. ISBN 978-0826480002.

[4] Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. In Psychometrika, 1958, vol. 23, iss. 3, pp. 187-200.

[5] Murtagh, F., Legendre, P. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? In Journal of Classification, 2014, vol. 31, pp. 274-295.

[6] Novák, R. 2019. Big Data and ethics. Dissertation thesis. Prague : University of Economics in Prague, 2019.

[7] R Core Team 2018. R: A language and environment for statistical computing. Vienna : R Foundation for Statistical Computing, 2018, https://www.R-project.org/.

[8] Sass, D. A., Schmitt, T. A. 2010. A comparative investigation of rotation criteria within exploratory factor analysis. In Multivariate Behavioural Research, 2010, vol. 45, iss. 1, pp. 73-103.

[9] Ward, J. H., Jr. 1963. Hierarchical grouping to optimize an objective function. In Journal of the American Statistical Association, 1963, vol. 58, pp. 236-244.