sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2020, Vol. 24, No. 1
ISSN 1507-3866; e-ISSN 2449-9994

THE ANALYSIS OF THE STRUCTURE OF UNIVERSITY POSITIONS IN POLAND USING CLASSIFICATION METHODS

Justyna Brzezińska
University of Economics in Katowice, Katowice, Poland
e-mail: justyna.brzezinska@ue.katowice.pl

ORCID: 0000-0002-1311-1020

© 2020 Justyna Brzezińska
This is an open access article distributed under the Creative Commons Attribution-NonCommercial--NoDerivs license (*http://creativecommons.org/licenses/by-nc-nd/3.0/*)

DOI: 10.15611/eada.2020.1.06
JEL Classification: C30, C31, C38

Abstract: Categorical data analysis is a statistical method that can be successfully applied in different scientific areas, such as: social, medical, psychological and political sciences. Classification and segmentation are statistical methods that usually have been used for large quantitative datasets to identify segments in the data, however if applied for categorical data for contingency tables, one may arrive at impressive results as well. This paper presents the use of classification and segmentation methods for categorical data in a contingency table based on real data from Central Statistics on the number of university positions in Polish voivodeships. The authors compare the results of different approaches and provide graphical results using advanced visualization tools, perceptual map (biplot) and dendrogram. Comparative analysis provides information on corresponding categories of academic positions in different voivodeships. All calculations are conducted in R.

Keywords: categorical data analysis, classification methods, structure of university positions in Poland.

1. Introduction

A contingency table, also known as a cross-classification table, describes the relationships between two or more categorical variables. A table cross-classifying two variables is called a two-way contingency table and forms a rectangular table with rows for the i categories of the X variable and columns for the j categories of a Y variable. Each intersection is called a cell and represents the possible outcomes. The cells contain the frequency of the joint occurrences of the X, Y outcomes.

A contingency table can summarize three probability distributions – joint, marginal, and conditional. The joint distribution describes the proportion of the subjects jointly classified by a category of X and a category of Y. The cells of the contingency table divided by the total provide the joint distribution. The sum

of the joint distribution is 1. The second possible distribution is the marginal distribution which describes the distribution of the X (row) or Y (column) variable alone and the row and column totals of the contingency table provide the marginal distributions. The sum of a marginal distribution using marginal distribution is 1. The third possible distribution is conditional distribution describing the distribution of one variable given the levels of the other variable. The cells of the contingency table divided by the row or column totals provide the conditional distributions. The sum of a conditional distribution is 1.

Contingency tables are mostly used for the measure of association between categorical data, however they can be also used in segmentation research for the following purposes:

- segments characteristics,
- segments profiling,
- evaluation of the correctness of the choice of segmentation variables,
- inference about cause-and-effect relationships.

It is noticeable that methods for the analysis of categorical data have been developed with a certain delay, compared to methods for continuous data. When all d observed attributes in a study are categorical, then the most common way to represent the data is a d-dimensional contingency table, produced by cross-classifying the attributes. The information of a contingency table is traditionally summarized through the appropriate measures of association, which differentiate according to the nature of the underlying classification of nominal or ordinal variables. Association measures, although convenient in computation and interpretation, lead to a major loss of information.

This paper proposes the use of classification and segmentation methods for the analysis of cross-classified categorical data. The empirical study is based on real data from Central Statistics on academic positions in Polish voivodeships. The authors propose applying the correspondence analysis for clustering and classification using graphical methods such as a perceptional map and a dendrogram. The goal of this paper is to provide clusters comprising universities with excellent academic position offers, as well as clusters that do not provide such possibility in term of academic positions. The analysis is based on real data from Statistics Poland. The computer software used for computational analysis is R (www.r-project.org).

2. Methodology: correspondence analysis

Correspondence analysis is a method applicable for analyses of contingency tables used to analyze the relations between two or more categorical variables [Greenacre 1993]. The correspondence analysis is performed in three steps. The first step is to calculate the categorical profiles (i.e. the relative frequencies) and masses (marginal proportions). The next step is to compute the chi-square distances between the points and find the n-dimensional space that best fits the points [Clausen 1998]. The next

part of the paper describes the following steps based on theory [Benzécri 1992; Clausen 1998; Greenacre 1984; Greenacre, Blasius 1994;]. In Polish literature in the field of economics it was published by Stanimir [2005], Bąk [2010] and Brzezińska [2011].

Correspondence analysis is an exploratory statistical method providing a geometric method to represent a complex categorical dataset in a straightforward and low-dimensional, often two-dimensional space so that the inner structure of the original dataset can be visually displayed and one can detect and explain the relationship not only among row or column variables, but also between row and column variables in a very large matrix. This statistical method is similar to principal component analysis, but it is better suited for analyzing categorical data. Principal component analysis, on the other hand, is better suited for continuous data. Another difference between the two techniques is how the data matrix is decomposed. While the total Chi-square value is decomposed in correspondence analysis, total variance is decomposed in principal component analysis. Correspondence analysis also involves mapping a chi-square distance into a particular Euclidean distance. Since no underlying model existence is necessary to conduct the analysis, it has been used mostly for exploratory data analysis.

The correspondence analysis is based on the correspondence matrix, defined as the matrix of elements of **N** divided by the grand total of **N**: $\mathbf{P} = \left[\frac{n_{hj}}{n}\right]$. The aim of the correspondence analysis is the geometrical display of two or more categorical variables by representing the categories of the variables as points in a low--dimensional space [Greenacre 1993]. Having defined the probability matrix, one can go to the next step which is counting the row and column profiles.

The vector of row and column sums of **P** are denoted by **r** and **c** defined as:

$$\mathbf{r} = \left[\frac{n_{h.}}{n}\right] = \left[p_{h.}\right], \tag{1}$$

$$\mathbf{c} = \left[\frac{n_{.j}}{n}\right] = \left[p_{.j}\right]. \tag{2}$$

The row and column profiles of **P** are defined as the vector of rows and columns of **P** divided by their respective sums. Row profiles are defined as:

$$\mathbf{R} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P} = \left[\frac{n_{hj}}{n_{h.}}\right] = \left[\frac{p_{hj}}{p_{h.}}\right] = \begin{bmatrix} \widetilde{\mathbf{r}}_1^{\,\prime} \\ \vdots \\ \widetilde{\mathbf{r}}_H^{\,\prime} \end{bmatrix}, \tag{3}$$

where $\widetilde{\mathbf{r}}_H$ is each respective row profile, and the column profile are defined as:

$$\mathbf{C} = \mathbf{D}_\mathbf{c}^{-1}\mathbf{P}^\mathbf{T} = \left[\frac{n_{hj}}{n_{.j}}\right] = \left[\frac{p_{hj}}{p_{.j}}\right] = \begin{bmatrix} \widetilde{\mathbf{c}}_1' \\ \vdots \\ \widetilde{\mathbf{c}}_J' \end{bmatrix}, \tag{4}$$

where $\widetilde{\mathbf{c}}_J$ is each respective column profile.

The row and column profiles define two clouds of points in respective H and J dimensional weighted Euclidean space. Both the row profiles and column profiles are written in the rows of **R** and **C** respectively. The row profiles and column profiles are a very important element of correspondence analysis, as they have impact on the principal axes. Each row and column profile can be presented as a point in multidimensional space. Thus a profile will tend to lie closer to vertices for which it has higher values. Each row and column profile has a unique weight associated with it, called mass, which is proportional to the row and column sum in the cross-tabulation. The average row and column profile is then the centroid of the row and column profiles, where each profile is weighted by its mass in the averaging process.

The chi-square distance between row profiles, referred to points of categories of first variable in J-dimensional space, is defined as:

$$d(h,h^*) = \sqrt{\sum_{j=1}^{J} \frac{\left(\frac{p_{hj}}{p_{h\bullet}} - \frac{p_{h^*j}}{p_{h^*\bullet}}\right)^2}{p_{\bullet j}}}, \tag{5}$$

where: $h, h^* = 1, ..., H,\ h \neq h^*$ are two different categories of row variable.

The chi-square distance between column profiles, referred to points of categories of second variable in H-dimensional space, is defined as:

$$d(j,j^*) = \sqrt{\sum_{h=1}^{H} \frac{\left(\frac{p_{hj}}{p_{\bullet j}} - \frac{p_{hj^*}}{p_{\bullet j^*}}\right)^2}{p_{h\bullet}}}, \tag{6}$$

where: $j, j^* = 1,...,J,\ j \neq j^*$ are two different categories of column variable.

The chi-square statistics measures the discrepancy between the observed frequencies in a contingency table and the expected frequencies calculated under a hypothesis of homogeneity of the row and column profiles [Greenacre 1993]. The chi-square statistics measures how far the row and column profiles are from their average profile. The inertia, or total inertia, of the contingency table is the quantity chi-square divided by total n. The inertia for rows is defined as:

$$\lambda_k = \sum_{h=1}^{H} d_h^2 \cdot p_{h.}, \tag{7}$$

where: d_h^2 – the chi-square distance between *h*-th row and the centroid.

The inertia for columns is defined as:

$$\lambda_j = \sum_{j=1}^{J} d_j^2 \cdot p_{.j}, \tag{8}$$

where: d_h^2 – the chi-square distance between *j*-th column and the centroid.

The total inertia is the same as the inertia of rows and the inertia of columns:

$$\lambda = \lambda_h = \lambda_j. \tag{9}$$

When the inertia is low, the row profiles are not dispersed very much, lie close to their average profile and do not extend out to the column vertex points. In this case there is low association, or correlation, between the rows and columns. The higher the inertia, the more the row profiles lie closer to the column vertices, i.e. the higher the row-column association. Zero inertia is attained when all the profiles are identical and lie at the same point. Maximum inertia is reached when all the profiles lie exactly at the vertices of the profile space. The total number of dimensions is defined as: $K = min\{H - 1, J - 1\}$, where *I* and *J* are the number of categories in the variable in the table. Correspondence analysis is also defined as a technique for decomposing the chi-square (i.e. the deviation from independence) for a frequency table. First of all, the total inertia is decomposed into a set of eigenvalues. These eigenvalues express the relative importance of the dimensions or how large a share of the total inertia each of them explain [Clausen 1998; van der Heijden, de Leeuw 1985; Murtagh 2005].

3. Data analysis

The academic structure of Poland resembles the transitional Central Eastern European model characteristic of many countries of the former Soviet bloc. Starting from the 1990s, the government began to introduce the principles of the Bologna Process and to decentralize the system of public higher education, giving universities more autonomy in setting their own principles for recruiting academic staff, including the number of faculty members.

At the same time, public authorities remain responsible for fixing the minimum amount of teaching hours with respect to specific positions although universities are free to increase salaries by drawing on their own resources. The law also fixes the criteria for career advancement.

The higher education system comprises both state and non-state institutions. The latter are created on the basis of the 1990 Higher Education Act. Prior to that, there were only state higher education institutions (with the exception of the Catholic University of Lublin).

Most higher education institutions are under the responsibility of the Ministry of Science and Higher Education. Some, however, are under the control of other relevant Ministries. Apart from the universities, scholars can develop an academic career in the scientific and research-development units, among which are the scientific units of Polish Academy of Science.

This paper presents the analysis carried out on a dataset from the Central Statistical Office (Statistics Poland) available on the website: https://stat.gov.pl on the structure of Polish scientists in Poland in 2017. This shows the structure of academic teachers in a number of existing positions in Polish voivodeships in Table 1.

Table 1. The structure of academic teachers in Poland in 2017 (number of positions)

Voivodeship	Academic teachers	Professors	Assistant professors	Tutors	Assistant lecturers
Dolnośląskie	8 475.8	1 852.1	50.1	3 969.1	983.1
Kujawsko-pomorskie	4 273.8	1 141.3	10.0	1 707.7	664.3
Lubelskie	6 151.8	1 282.8	19.0	2 509.1	1 276.6
Lubuskie	1 245.3	387.8	8.8	518.0	139.8
Łódzkie	6 259.4	1 579.6	15.0	2 560.9	754.0
Małopolskie	12 896.0	2 621.1	23.5	5 643.0	2 103.9
Mazowieckie	17 110.9	4 351.4	150.5	7 138.2	1 958.5
Opolskie	1 521.7	444.0	2.0	588.1	136.0
Podkarpackie	3 071.9	768.3	16.0	888.5	481.4
Podlaskie	2 682.5	593.0	7.0	1 048.0	515.5
Pomorskie	5 855.4	1 524.0	27.0	2 199.0	856.9
Śląskie	8 349.8	1 846.5	24.0	3 827.2	1 092.0
Świętokrzyskie	1 736.2	500.4	8.0	602.0	244.0
Warmińsko-mazurskie	2 308.6	599.1	0.0	945.9	335.2
Wielkopolskie	9 095.2	2 255.3	41.3	3 595.0	855.6
Zachodnio-pomorskie	3 673.1	943.9	1.0	1 434.5	554.3

Source: Statistics Poland.

From the analysis of Table 2 one can see that there are two main university areas: the Mazowieckie and Małopolskie voivodeships. This is due to the fact that the best

universities in Poland are located there: the University of Warsaw, the Jagiellonian University in Cracow, the Warsaw University of Technology and the University of Science and Technology (AGH) in Cracow. The fact that the best universities are located in Warsaw and Cracow means that the largest number of students study there, and as a result the largest number of academic staff is needed there as well.

Pearson`s chi-square test for independence with 60 degrees of freedom equals 39 223, meaning that there is a strong association between voivodeship and academic position.

```
Principal inertias (eigenvalues):

 dim value % cum% scree plot
 1 0.215964 93.3 93.3 **********************
 2 0.012402 5.4 98.6 *
 3 0.002462 1.1 99.7
 4 0.000743 0.3 100.0
 -------- -----
 Total: 0.231571 100.0

Rows:
 name mass qlt inr k=1 cor ctr k=2 cor ctr
1 | DOLN | 70 991 88 | -438 657 62 | 312 334 548 |
2 | KUJA | 61 978 42 | 386 928 42 | -90 50 40 |
3 | LUBE | 58 908 5 | -139 907 5 | -5 1 0 |
4 | LUBU | 32 979 107 | 814 861 98 | 301 118 235 |
5 | ÓDZK | 50 988 53 | -485 960 55 | -84 29 28 |
6 | MAOP | 111 961 40 | -276 911 39 | -65 50 38 |
7 | MAZO | 172 818 8 | -90 802 6 | -13 17 2 |
8 | OPOL | 66 989 375 | 1139 989 398 | -8 0 0 |
9 | PODK | 27 996 12 | -310 906 12 | -97 89 21 |
10 | PODL | 28 943 1 | -35 102 0 | 99 841 22 |
11 | POMO | 53 979 19 | -287 979 20 | 2 0 0 |
12 | LSKI | 78 975 9 | -163 972 10 | -9 3 1 |
13 | WITO | 42 995 129 | 837 994 137 | -20 1 1 |
14 | WARM | 22 957 3 | -173 956 3 | -6 1 0 |
15 | WIEL | 75 997 64 | -441 975 67 | -65 22 26 |
16 | ZACH | 54 999 43 | 421 952 44 | -94 47 38 |

Columns:
 name mass qlt inr k=1 cor ctr k=2 cor ctr
1 | Acdm | 559 997 239 | -314 996 255 | -7 1 2 |
2 | Prfs | 134 933 47 | -274 926 47 | -22 6 5 |
3 | Assstntp | 2 485 19 | 912 458 9 | 221 27 9 |
4 | Ttrs | 223 1000 565 | 760 983 596 | -99 17 177 |
5 | Assstntl | 82 999 131 | 498 668 94 | 350 330 806 |
```

The summary of the correspondence analysis gives some overall statistics and lists the distribution of the so-called eigenvalues. The eigenvalues in the summary

indicate the explanatory power of each principal axis. They are given in three ways: the first row (value) shows the actual eigenvalues, the second row (%) shows the relative values, and the third row (cum%) shows the cumulative relative values.

Next the authors present the graphical structure of points (categories) in two--dimensional space (Figure 1). The red triangles refer to academic positions (academic teachers, professors, assistant professors, tutors and assistant lecturers), whereas the blue dots refer to the 16 voivodeships.

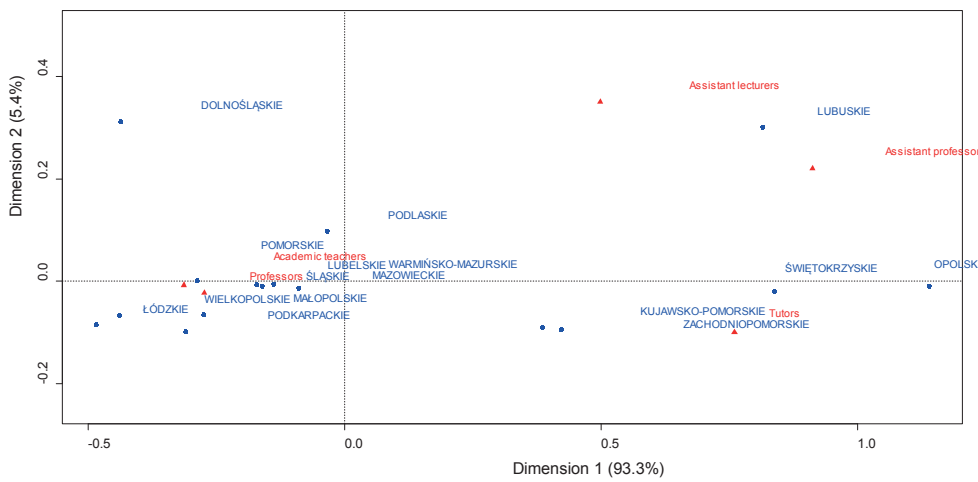

Fig. 1. Perceptual map for correspondence analysis

Source: own calculations in R.

The interpretation of graphical results in correspondence analysis is usually subjective and is difficult to classify objects directly. Looking at Figure 2 one can see that there are three academic positions: assistant lecturers, assistants and assistant professors strongly corresponding to the following group of voivodeships: Kujawsko-pomorskie, Zachodniopomorskie, Świętokrzyskie, Opolskie, and Lubuskie. This cluster represents a group of academic positions for young scientists. The second cluster are academic teachers and professors corresponding strongly to the rest of the voivodeships (Dolnośląskie, Podlaskie, Pomorskie, Łódzkie, Wielkopolskie, Mazowieckie, Małopolskie, Podkarpackie, Śląskie, Lubelskie and Warmińsko--mazurskie). This cluster corresponds to more experienced academics such as professors and academic teachers.

Additional information on the classification of objects can be derived from classification methods. A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters. A dendrogram for categorical data is mostly used for the comparison of results with correspondence analysis.

From the analysis of row data (Table 1), one can see the voivodeships with the largest number of academic positions available, and this is due to the fact that that the Małopolskie and Mazowieckie voivodeships are the two largest academic centres in Poland with the two best rated universities.

Using the classification method the authors applied the Ward approach to obtain a dendrogram (see Figure 2).

Fig. 2. Dendrogram using the Ward method.

Source: own calculations in R.

In this plot (Figure 2) one can see that three main clusters were created. One cluster contains two voivodeships: Małopolskie and Mazowieckie. The other cluster consists of six objects: Lubelskie, Łódzkie, Pomorskie, Dolnośląskie, Śląskie and Wielkopolskie voivodeships, while the last cluster consists of eight, namely Podkarpackie, Podlaskie, Warmińsko-mazurskie, Opolskie, Kujawsko-pomorskie, Zachodniopomorskie, Lubuskie and Świętokrzyskie.

Using median clustering the authors obtained another dendrogram (Figure 3). As previously, there are three clusters, however their structure is different. The first cluster contains the Małopolskie and Mazowieckie voivodeships, as in the first cluster. The second is a one-object cluster consisting of the Opolskie voivodeship, and the last cluster of thirteen other voivodeships (Lubelskie, Łódzkie, Pomorskie, Dolnośląskie, Śląskie, Wielkopolskie, Podkarpackie, Podlaskie, Warmińsko-mazurskie, Kujawsko--pomorskie, Zachodniopomorskie, Lubuskie and Świętokrzyskie).

The outcome does not surprise, especially that the main academic centers (Cracow and Warsaw) are in the same two-object cluster using other methods of classification.

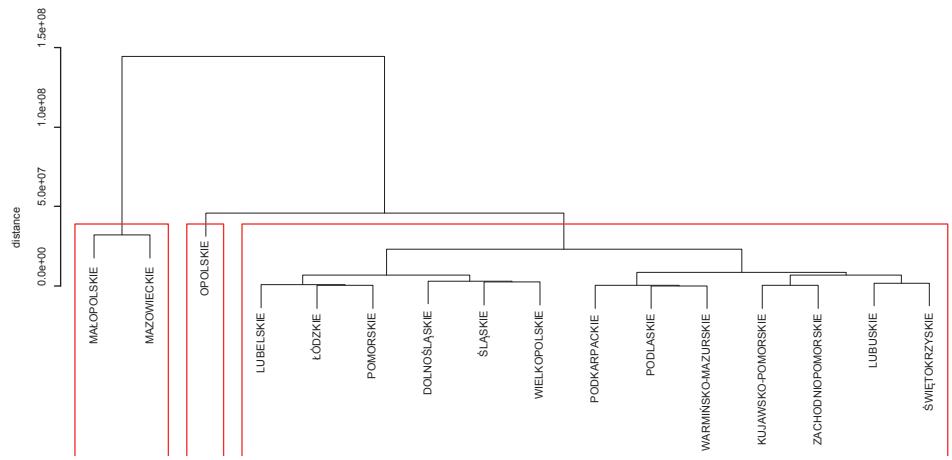

Fig. 3. Dendrogram using the median method.

Source: own calculations in R.

There is a strong correspondence between the size of the university and the number of academic positions, which results in greater professional opportunities in large cities with higher numbers of students.

4. Conclusions

This paper presents the statistical methods used for the analysis of academic positions in different voivodeships in Poland using data from Statistics Poland available on the website: https://stat.gov.pl on the structure of academia in Poland in 2017. Thee presented methods for the analysis of contingency tables and proposed correspondence analysis for the graphical presentation of the results. The graphical results are shown in a perceptual map where one can observe correspondence and points layout in two-dimensional space. The authors applied hierarchical classification (the Ward and the median method) for the graphical presentation of results on a dendrogram. As a result one can see that using correspondence analysis allows for grouping academic positions in two clusters: firstly positions for young academics, and the other for experienced scholars. Applying cluster analysis (the Ward and the median method) proved that there is a strong dependency between the ranking and the size of the university and the number of academic positions. Using the dendrogram one can see that there are three clusters, one consisting of the Mazowieckie and Małopolskie voivodeships – the two largest in terms of the number of students, universities and the level of education. This paper applied the available libraries in R software: `ca` package for correspondence analysis and `MASS` package for hierarchical clustering methods.

A future goal may be the dependency analysis for multi-way tables using the model-based method such as log-linear models, where one can include interactions between categories, use formal criteria for model selection, and extend the table dimension to a multi-way table. Such analysis would provide information on the relationship with other important variables influencing cell-counts.

References

Bąk I., 2010, *Zastosowanie analizy korespondencji w badaniu aktywności turystycznej emerytów i rencistów*, Metody Ilościowe w Badaniach Ekonomicznych, XI(2), 1-11.

Benzécri J. P., 1992, *Correspondence Analysis Handbook*, Marcel Dekker, New York.

Brzezińska J., 2011, *Analiza korespondencji*, [in:] E. Gatnar, M. Walesiak (ed.), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, PWN, Warszawa, 52-80.

Clausen S. E., 1998, *Applied Correspondence Analysis. An Introduction*, Sage: University Paper, Thousand Oaks, London, New Delhi.

Greenacre M., 1984, *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Greenacre M., 1993, *Correspondence Analysis in Practice*, Academic Press, London.

Greenacre M., Blasius J., 1994, *Correspondence Analysis in Social Science, Recent Developments and Applications*, Academic Press, San Diego.

Stanimir A., 2005, *Analiza korespondencji jako narzędzie do badania zjawisk ekonomicznych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Murtagh F., 2005, *Correspondence analysis and data coding with Java and R*, Computer Science and Data Analysis Series, Chapman & Hall, CRC.

van der Heijden, P.G.M. de Leeuw, J., 1985, *Correspondence analysis used complementary to loglinear analysis*, Psychometrika, 50, 429-447.

www.stat.gov.pl.

ZASTOSOWANIE METOD KLASYFIKACJI I SEGMENTACJI W ANALIZIE DANYCH JAKOŚCIOWYCH

Streszczenie: Analiza danych jakościowych należy do grupy metod statystycznych, która może być z powodzeniem wykorzystywana w wielu obszarach naukowych, takich jak: nauki społeczne, medyczne, psychologiczne oraz polityczne. Metody klasyfikacji i segmentacji są technikami statystycznymi, które wykorzystuje się zazwyczaj do analizy dużych zbiorów danych o charakterze ilościowym w celu identyfikacji segmentów w danych. Zastosowanie tych metod w analizie danych jakościowych może także przynieść zaskakujące wyniki. W niniejszym artykule zaprezentowano metody klasyfikacji i segmentacji do analizy danych jakościowych w analizie tablic kontyngencji. Porównano wyniki i rezultaty różnych podejść, a także zaprezentowano graficznie wyniki analizy. Wszystkie obliczenia przeprowadzono w programie R na danych rzeczywistych pochodzących z Głównego Urzędu Statystycznego.

Słowa kluczowe: analiza danych jakościowych, tablica kontyngencji, klasyfikacja i segmentacja.