DE GRUYTER
OPEN

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2018, Vol. 22, No. 2
ISSN 1507-3866; e-ISSN 2449-9994

VISUALIZATION OF CATEGORICAL DATA USING EXTRACAT PACKAGE IN R

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

DOI: 10.15611/eada.2018.2.01
JEL Classification: C30, C31, C4

Abstract: Visualization in research process plays a crucial role. There are several advanced plots for visualizing categorical data, such as mosaic, association, double-decker, sieve or fourfold plot that are based on the graphical presentation of residuals in a contingency table. In this paper we present new methods for visualizing categorical data such as `rmb`, `fluctile` and `scpcp` plot available in `extracat` package in R. This package provides a well-structured representation of categorical data and allows for a detailed presentation of the relationship between categories in terms of proportions. We describe `rmb`, `fluctile` and `cpcp`. Those plots are based on the concept of multiple bar charts, a fluctuation diagram from a multidimensional table and parallel coordinates respectively. Such plots are mostly used for a visualization of a contingency table or a data frame; they can also be used for exploratory analysis and allows for a graphical presentation even for a high number of variables [Pilhöfer, Unwin 2013]. All the calculations and plots are obtained using R software.

Keywords: categorical data, `cpcp plot`, `rmb plot`, `fluctile` plot, R software.

1. Introduction

Over the last decade a modest revolution has been ongoing in the analysis of categorical data, as graphical methods and techniques of data visualization, so commonly used for quantitative data, have begun to be developed for frequency data and discrete data [Friendly 1994; 2000].

In this paper we describe in brief the development of graphical methods for categorical data in R using the new and modern graphs available such as the rmb and cpcp plot in `extracat` package. This package provides interesting and unique graphical tools for displaying categorical data extending the concepts of multiple bar charts and parallel coordinates plots.

In this paper we introduce the `rmb` (*Relative Multiple Bar*) plot based on a crossover of mosaic plots and multiple bar charts to display the frequencies of a data

table split up into conditional relative frequencies of one target variable and the absolute frequencies of the corresponding combinations of the remaining explanatory variables. This plot provides a well-structured representation of the data which is easy to interpret and allows precise comparisons. The graphics can additionally be used as a generalization of spine plots or with bar charts for the conditional relative frequencies. Several options, including ceiling censored zooming, residual shadings and a choice of color palettes, can be provided in R.

The second graph presented in this paper is `cpcp` (*Categorical Parallel Coordinates Plot*).This plot is based on interactive parallel coordinates plots, with sequences of points used to represent each of the variable categories, while ordering algorithms are applied to represent a hierarchical structure in the data and keep the arrangement clear. This type of plot is mostly used in exploratory analysis and allows for a visual interpretation for a higher number of variables and a mixture of categorical and numeric scales.

We also present the `fluctile` plot that creates a fluctuation diagram from a multidimensional table with different shaping variants available in R.

2. Categorical data visualization using `extracat`

Before the analysis is carried out, it is helpful and advisable to make data visualization a priori. The practice of categorical data visualization in comparison to qualitative data is not that common and developed, however due to the development of statistical computer software there are more and more graphs that allow for graphical data presentation.

While advances in software have made it much more simple to fit models to qualitative variables, they can sometimes be difficult to visualize and analyze because there are often no similarities between the levels of the variables. There are many approaches to overcome this difficulty. Basic concepts such as frequency and probability can be extended to introduce a variety of visual methods for analyzing and interpreting categorical data in contingency tables.

In this paper we introduce the advantages of multiple bar charts and classic mosaic plots that can be presented in one display. The R package `vcd` [Meyer, Zeileis, Hornik 2006] provides an implementation of classical mosaic plots and interactive graphics available through the iplots package [Urbanek, Theus 2003]. The main intention of the rmb plots is to precisely display the relative frequencies of a target variable for each combination of explanatory variables divided over a grid-like graphical display and, simultaneously, their corresponding weights. The breakup of absolute frequencies into conditional distributions and weights is a common procedure in many methodologies for categorical data analysis, such as generalized linear models or correspondence analysis, but there seems to be a lack of graphical solutions for exploratory as well as illustrative purposes. In the further parts of this paper, `rmb`, `scpcp` and `fluctile` plots will be described in detail using R software.

2.1. Rmb plot

Rmb (*Relative Multiple Bar*) plots allows for visualization contingency tables for the relative frequencies of some target categories within each combination of the explanatory variables. The weights of these combinations that are absolute frequencies are represented in the total with the corresponding bar chart. This plot available in extracat package in R offers different visualizations of conditional distributions and their corresponding weights (frequencies) of a target variable and the corresponding weights of the explanatory variables including multiple bar charts, spine plots and pie charts, with different ways of displaying the residuals from statistical models.

Rmb plots are a mixture of classic bar charts and mosaic plot, widely used and well-known graphical methods for categorical data. Let us assume we analyze a three-way contingency table with variables X_1, X_2 and X_3, the frequency n_{hjk} is the number of observations falling into *h*-th row, *j*-th column and *k*-th layer. The frequency is split into conditional relative frequency $p_{i|jk}$ of one variable and weights corresponding to the other variables according to the formula [Pilhöefer, Unwin 2013]:

$$n_{hjk} = p_{h|jk} \cdot n_{\bullet jk} = p_{h|jk} \cdot p_{\bullet jk} \cdot n, \tag{1}$$

where: $p_{\bullet jk} = \frac{n_{\bullet jk}}{n}$, $n_{\bullet jk} = \sum_i n_{hjk}$, *n*-total number of observations.

The variable that is represented by the conditional relative frequencies $p_{i|jk}$ is referred to as the target variable. The other variables will be explanatory variables and their combinations are represented by: $n_{\bullet jk} = p_{\bullet jk} \cdot n$.

Next we present the historical background for the rmb plot and its relationship to mosaic plots [Friendly 1994; 2000; Hartigan, Kleiner 1981]. Classic mosaic plots are based on the calculation of $p_{h|jk}$ and $n_{\bullet jk}$, however it is more difficult to describe the relationship and association between categorical data with every additional variable. Comparing the proportions of a target category in different combinations of explanatory variables is only possible in a qualitative manner, because the corresponding rectangles neither share a common axis nor have a common scale [Pilhoefer, Unwin 2013].

In the cpcp plot we consider a set of *m* categorical variables including one target variable. The basis of the plot is a multiple bar chart of the $m - 1$ explanatory variables displaying the observed frequencies $n_{\bullet jk}$ of their combinations. The plot uses horizontal bars which means that all bars have an equal height and their widths are proportional to the ratios $\frac{n_{\bullet jk}}{\max(n_{\bullet jk})}$. Having defined these elements we can define conditional distributions for target categories as probabilities $p_{h|jk}$ displayed inside the bars.

Relative multiple bar plots are a very useful tool for interpretation, however they are a still member of the mosaic plot family and thus should not be applied for a large number of variables and categories.

Next we apply the `rmb` function for the analysis of the categorical dataset on the extent of economic poverty in Poland published by the Central Statistical Office of Poland (www.stat.gov.pl).

In the example we use a dataset on poverty (limit of extreme poverty, relative poverty level, statutory poverty level) due to the type of the household (1-person household, no children, 1 child, 2 children, 3 children, 4 and more children, mother or father with child) in 2015. Firstly, we present the relative multiple bar in horizontal layout.

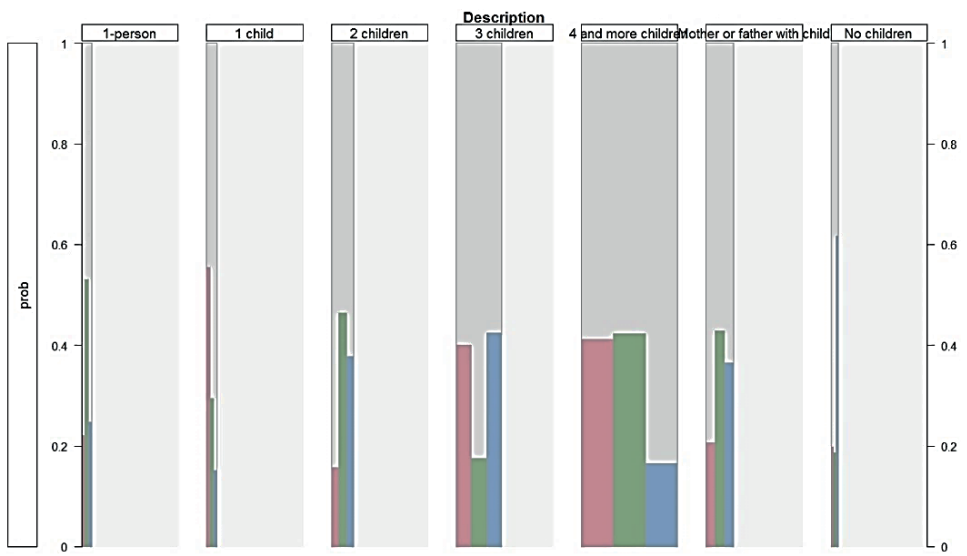

Fig. 1. Rmb plot (horizontal layout) for extent of economic poverty due to the type of household in Poland in 2015

Source: own calculations in R.

From the analysis of Figure 1 we can see the distribution of one variable with the related probability for each of the category of the other variable. We can also analyze the proportions and conclude that the group most endangered by the poverty are households with 4 and more children, with the highest and widest bars. On the opposite end of the scale we can observe the group of households that is least endangered by poverty which is 1-person households.

We can also use the vertical layout for the `rmb` plot presenting distribution and categories for the variable that is analyzed. The interpretations of the bars in Figure 2 is similar to those in Figure 1. Each bar corresponds to the probability of the

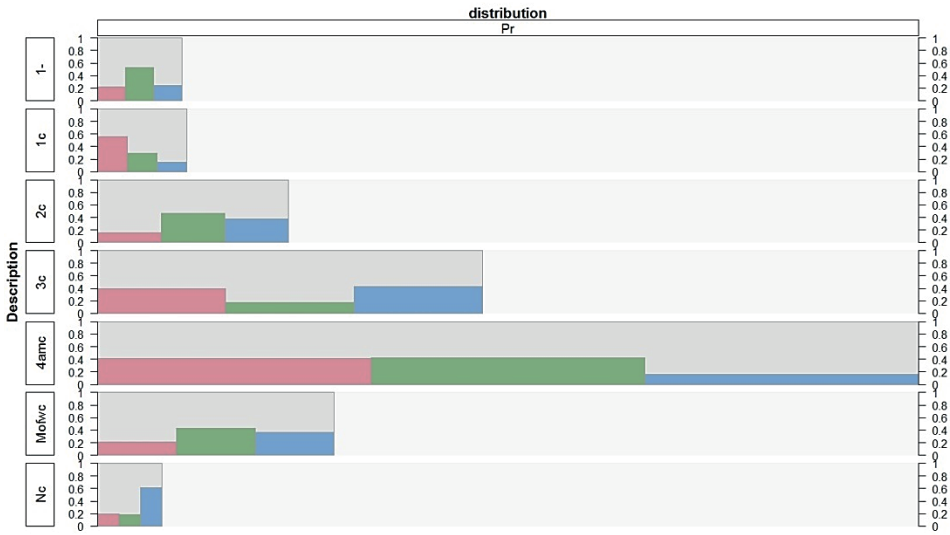

Fig. 2. Rmb plot (vertical layout) for extent of economic poverty due to the type of socio-economic group in Poland in 2015

Source: own calculations in R.

occurrence of one category of the first variable and the colours are related to categories of the other variable.

A different layout available for the relative multiple bar plots in the extracat package in R matches the user`s needs and different requirements.

2.2. Scpcp plot

Scpcp (*Static Categorical Parallel Coordinates Plot*) plots visualize a contingency table or a data frame with categorical variables via interactive parallel coordinates plots. The concept of a parallel coordinates plot (Unwin, Volinsky, and Winkler 2003) is very useful, especially for high-dimension tables including a large number of variables without losing information in the raw data values.

Parallel coordinates plots are made fit for categorical data with the function cpcp, which is based on the ipcp function from package iplots. To reflect the categorical structure of the data, point sequences and orderings are applied to the categories of the different variables. It is possible to use interactive highlighting and linking with other iplots. Furthermore the cpcp function also accepts continuous data.

This plot applies sorted numeric point sequences to the categories which indicate the relative frequencies and allow a sensible interactive highlighting. There are options to change the rule for the gaps between these sequences and to apply an additional ordering algorithm.

Parallel coordinates plot was introduced by d'Ocagne [1885] and Inselberg in 1959 [Inselberg 2009]. The original concept of parallel coordinates plot did not allow for the analysis of categorical variables, which was a great disadvantage. Kosara, Bendix and Hauser [2005] introduced an application designed for categorical data. The `cpcp` plot (categorical parallel coordinates plot) is a new approach which allows for the graphical display of both numeric and categorical variables in the one plot. The `cpcp` plot is based on the R package `iplots` and takes advantage of the interactive capabilities of the package. The application of this plot into categorical data analysis is possible because within every variable, each category is assigned a sequence of equidistant points with one point for each case and a range proportional to each category's relative frequency. The dataset is recursively sorted, starting with the last variable and ending with the first one before assigning points to the cases. This procedure leads to a display which shows a hierarchical splitting structure from left to right. The polylines of cases which are identical in the first m variables are drawn together on the corresponding axes and within each such group they will not cross each other.

Next we apply the scpcp plot for the analysis of data on poverty (limit of extreme poverty, relative poverty level, statutory poverty level) due to the type of socio-economic group (employees, farmers, self-employed, retired, pensioners, other) in Poland in 2015.

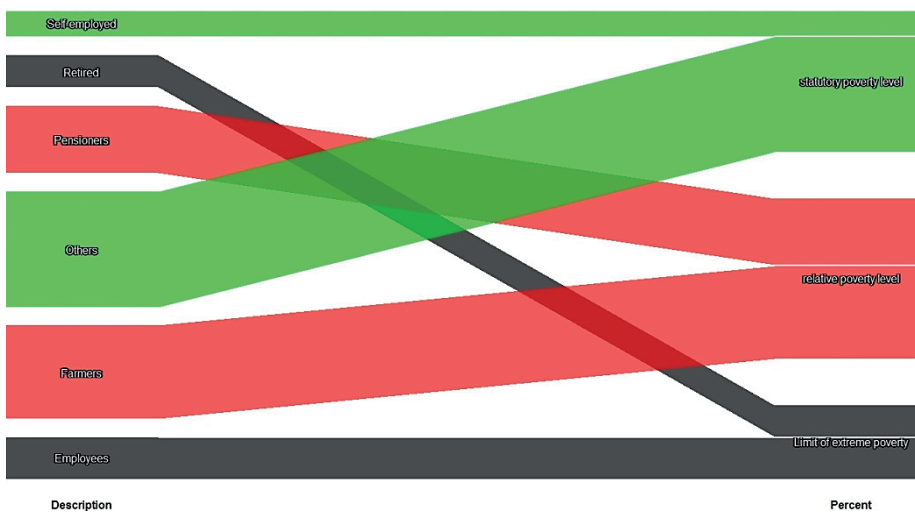

Fig. 3. Cpcp plot for extent of economic poverty due to the type of socio-economic group in Poland in 2015

Source: own calculations in R.

The analysis of the cpcp plot in Figure 3 presents groups of households under different type of poverty risk (in green, grey and red). The height of these lines is related to the frequency showing groups of high and low risk. From Figure 3 we can

see that the highest line is highlighted in green and the group is others related to the statutory poverty level. The second largest group under poverty risk are farmers related to the relative poverty level. We can also make groups under the same type of poverty risk. Self-employed and others belong to the group under the statutory poverty level. The other are retired and employees most endangered by the limit of extreme poverty, and last group are pensioners and farmers mostly related to relative poverty level.

Using the cpcp plot in R we can also use different colours as presented in Figure 4. In the following example we use the dataset on poverty (limit of extreme poverty, relative poverty level, statutory poverty level) due to the type of the household (1-person household, no children, 1 child, 2 children, 3 children, 4 and more children, mother or father with child) in 2015.

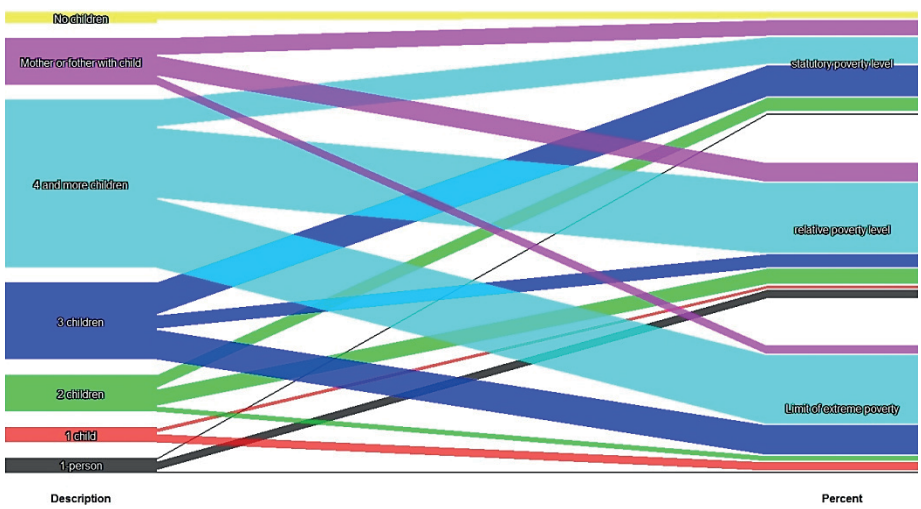

Fig. 4. Cpcp plot for extent of economic poverty due to the type of household in Poland in 2015

Source: own calculations in R.

The interpretation of cpcp in Figure 4 shows on the left groups of household and on the right – different types of poverty level according to the Central Statistical Office levels. We can see that the highest line is light blue for 4 and more children households. This group is mostly related to three types of poverty level showing high proportions of this line on the right side. The group under the least risk of poverty level are 1 person and 1 child households. The heights of the lines related to this group are the lowest.

The cpcp plot is a very powerful visual plot, easy for interpretation showing proportions using colourful lines related to one another. This structure is user-friendly and enables to know the proportions of categories without raw data analysis. Different colouring options make this plot matched to R-users' needs and preferences.

2.3. `Fluctile` plot

In this part of the paper we present the fluctile plot which is a variant of the mosaic plot. The mosaic plot was proposed by Hartigan and Kleiner [Hartigan, Kleiner 1981] as a method for visualizing the counts from contingency tables. In a mosaic plot, a rectangle is drawn for every combination of categories where the area of the rectangle is proportional to the count. To construct a mosaic plot we usually follow three steps: 1. the horizontal axis is divided according to the category counts of the first variable, 2. if there is a second variable, then each vertical column is divided according to the counts of the second variable, 3. if there are more than two variables, repeat steps 1 and 2 according to the counts for each additional variable. That is, each rectangle created in steps 1 and 2 is further sub-divided horizontally and vertically for the third and fourth variables. This subdivision is repeated until all variables have been used. At each grid position, two rectangles are drawn. The first is drawn in a background color and is full size (i.e. the maximum count). A second rectangle is drawn in a foreground color with a height proportional to the count for that particular combination of categories. The background rectangle is drawn to give a sense of scale.

We present a new type of mosaic plot called a fluctuation plot which might be much easier to interpret than the mosaic plot. Although the mosaic and fluctuation plots were developed to visualize counts for categorical data, a data plot can also generate the fluctuation plot for various statistics. To analyse data on poverty (limit of extreme poverty, relative poverty level, statutory poverty level) due to the type of socio-economic group (employees, farmers, self-employed, retired, pensioners, other) in Poland in 2015 we apply the `fluctile` function in the `ectracat` package in R.

Fig. 5. Fluctuation plot for extent of economic poverty due to the type of socio-economic group in Poland in 2015

Source: own calculations in R.

We present several layout options presenting different shapes and colours available in R software (Figures 5, 6 and 7).

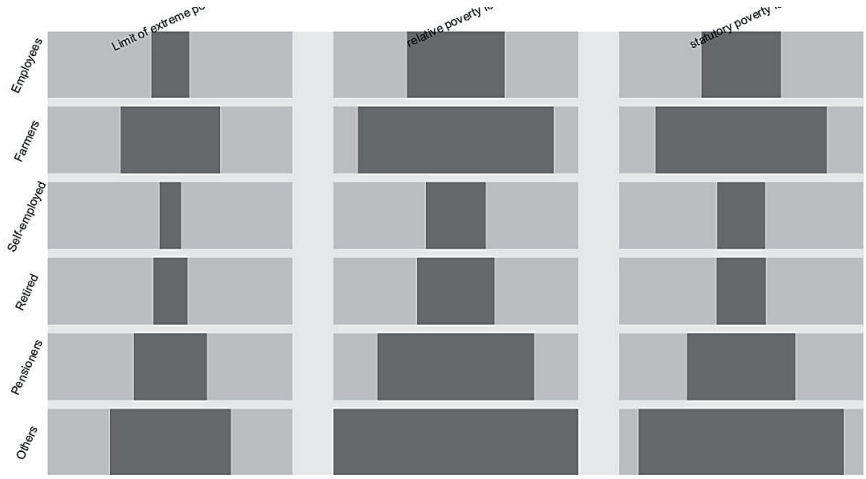

Fig. 6. Fluctuation plot for extent of economic poverty due to the type of socio-economic group in Poland in 2015

Source: own calculations in R.

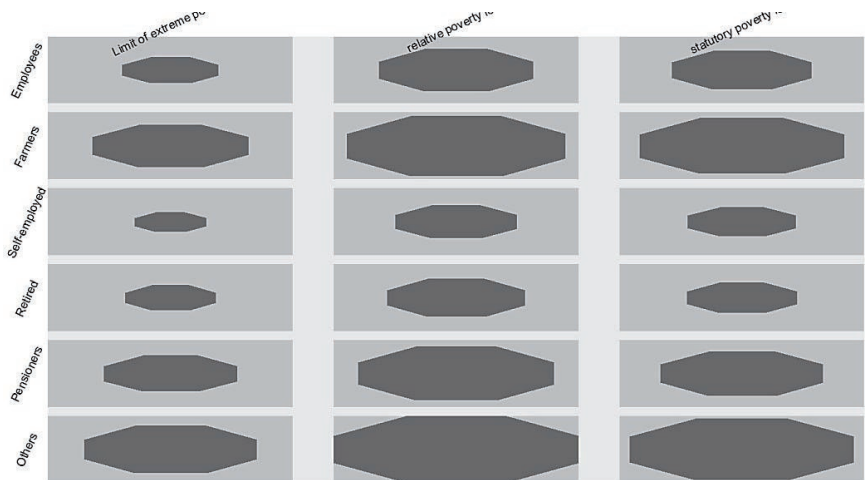

Fig. 7. Fluctuation plot for extent of economic poverty due to the type of socio-economic group in Poland in 2015

Source: own calculations in R.

For a two-dimensional table, this simply puts different dimensions on the *x* and *y* axes, and the area of the rectangles is proportional to the density of observations in that cell of the table.

3. Conclusions

In this paper we introduced interesting modern graphical methods for visualizing categorical data analysis: `rmb`, `cpcp` and `fluctile` plots available in the `extracat` package in R. The rmb plot is a member of the mosaic plot family which are used for the graphical display of the natural factorization of absolute frequencies into conditional relative frequencies and their weights. This property of the rmb plot makes it especially useful for the analysis of target variables. Zooming and the equal-width option are key features for displaying small frequencies. Residual shadings are used with log-linear and logistic models and the option to use rmb plots as a generalization of spine plots further increases the exhibility of the graphic.

Another graph presented in the paper is the scpcp plot. This type of graph allows for the increase of the number of displayable categorical variables with the use of the well-established parallel coordinates plot as its basis. Its strength lies in interactive features like highlighting and the resort-algorithms which make it a powerful tool for exploratory data analysis. Its capability of displaying a mixture of categorical and continuous variables gives it an advantage over alternative plots.

We also introduced a fluctuation plot using the `fluctile` function in R which allows for creatiing a fluctuation diagram from a multidimensional table. The different shape layout allows to adjust the graph to the user`s needs.

The presented plots are an excellent useful tool in the analysis of categorical data. We applied rmb, scpcp and fluctile plots for the analysis of real life data on the poverty risk in Poland in 2015. Graphical analysis allowed for clustering groups of households due to different criteria on groups under the highest and lowest poverty level.

Bibliography

d'Ocagne M., 1885, Coordonnées Parallèeles et Axiales: Méthode de Transformation Géométrique et Procédé Nouveau de Calcul Graphique déduits de la Considération des Coordonnées Parallèlles, Gauthier-Villars, Paris.

Friendly M., 1994, *Mosaic display for multi-way contingency tables*, Journal of the American Statistical Association, 89, pp. 190-200.

Friendly M., 2000, *Visualizing Categorical Data*, SAS Institute, Cary NC.

Hartigan J.A., Kleiner B., 1981, *Mosaics for Contingency Tables*, [in:] W.F. Eddy (ed.), *Computer Science and Statistics*, Proceedings of the 13th Symposium on the Interface, 268-273, Springer-Verlag, New York.

Inselberg A. 2009, Parallel Coordinates, Springer-Verlag, New York.

Kosara R., Bendix F., Hauser H., 2005, *Parallel sets: Interactive exploration and visual analysis of categorical data*, Visualization and Computer Graphics, IEEE Transactions, 12 (4), pp. 558-568.

Meyer D., Zeileis A., Hornik K., 2006, *The strucplot framework: visualizing multi-way contingency tables with vcd*, Journal of Statistical Software, 17(3), pp. 1-48.

Pilhöefer A., Unwin A., 2013, *New Approaches in Visualization of Categorical Data: R Package extracat*, Journal of Statistical Software, 53(7), pp. 1-25.

Unwin A., Volinsky C., Winkler S., 2003, *Parallel coordinates for exploratory modelling analysis*, Computational Statistics & Data Analysis, 43(4), pp. 553-564.

Urbanek S., Theus M., 2003, *iPlots – High Interaction Graphics for R*, [in:] K. Hornik, F. Leisch, A. Zeileis (eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing 2003*, Technische Universität Wien, Vienna, Austria.

WIZUALIZACJA DANYCH JAKOŚCIOWYCH Z WYKORZYSTANIEM PAKIETU EXTRACAT PROGRAMU R

Streszczenie: W procesie decyzyjnym oraz analizie danych kluczową rolę odgrywa wizualizacja wyników. Istnieje wiele zaawansowanych wykresów przeznaczonych dla danych o charakterze nominalnym. Najbardziej znanymi w tym obszarze są: wykres mozaikowy, wykres asocjacji, wykres dwuwarstwowy, wykres sitkowy, wykres czteropolowy. Wykresy te oparte są na graficznej analizie odchyleń liczebności empirycznych od teoretycznych w badanej tablicy kontyngencji. W niniejszej pracy przedstawione zostaną nowoczesne metody wizualizacji danych niemetrycznych za pomocą wykresu częstości względnych (*relative multiple bar*) (funkcja `rmb`), wykresu opartego na koncepcji osi równoległych (*categorical paralel coordinates plot*) (funkcja `cpcp`) oraz wykresu fluktuacji (*fluctuation plot*) (funkcja `fluctile`) dostępnych w pakiecie `extracat` programu R [Pilhöfer, Unwin 2013]. Wszystkie obliczenia prezentowane w niniejszym artykule wykonano w programie R.

Słowa kluczowe: wizualizacja, wykres scpcp, wykres rmb, funkcja fluktuacji, program R.