EKONOMETRIA ECONOMETRICS 3(57) • 2017

ISSN 1507-3866
e-ISSN 2449-9994

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

VISUAL TECHNIQUES FOR CATEGORICAL DATA IN R

METODY WIZUALIZACJI DANYCH W PROGRAMIE R

DOI: 10.15611/ekt.2017.3.02
JEL Classification: C35, C 40, C88.

Summary: Visualization is one of the most important parts of statistical analysis. In this paper we present a new method of multiple bar charts to display the frequencies of data tables split up into conditional relative frequencies of one target variable and the absolute frequencies of the corresponding combinations of the remaining explanatory variables. In this paper we present the R package `extracat` allowing for new graphical tools: `rmp` and `cpcp` plot [Pilhoefer, Unwin 2013]. The first plot uses the a crossover of mosaicplots and multiple barcharts to display the frequencies of a contingency table split up into conditional relative frequencies of one target variable and the absolute frequencies of the corresponding combination of the remaining explanatory variables. It provides a well-structured representation of the data with the possibility of easy interpretation. Another plot presented in the paper is the cpcp plot using parallel coordinates. Sequences of points are used to represent each of the variable categories, while ordering algorithms are applied to represent a hierarchical structure in the dataset.

Keywords: visualization, categorical data, multiple bar charts, R software.

Streszczenie: Wizualizacja danych jest jedną z ważniejszych części statystycznej analizy. W niniejszym artykule zaprezentowane zostaną nowoczesne metody wizualizacji z uwzględnieniem wykresów słupkowych w celu prezentacji liczebności, podzielonych na częstości warunkowe pozostałych zmiennych objaśniających. Zaprezentowany zostanie pakiet `extracat` programu R pozwalający na zastosowanie nowoczesnych metod graficznych: wykresu `rmp` oraz `cpcp`. Pierwszy z nich należy do grupy wielowymiarowych wykresów mozaikowych. Drugi natomiast oparty jest na równoległych osiach. Zastosowanie tych wykresów w badaniach pozwala na zwiększenie możliwości interpretacyjnych, a także na przedstawienie struktury analizowanych danych.

Słowa kluczowe: wizualizacja, dane jakościowe, wielowymiarowy wykres słupkowy, program R.

1. Introduction

R.A. Fisher in his crucial work entitled “Statistical Methods for Research Workers” [Fisher 1925] pointed out that “the preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye, they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining conclusions founded upon them". Almost a hundred years later, in the beginning of the 21st century, these words are still very current in the science and they prove that visual presentation is valuable throughout a statistical analysis.

Before any statistical analysis it is always very helpful to create some visualization of the data. This practice is common mainly for quantitative data, and not for qualitative data. This is due to the fact that qualitative data have a sort of measurable scaling. However, graphs and plots can provide preliminary examinations of the data that can help us to decide what type of analysis to carry out in the next stage of the research. On the other hand, visualization is a crucial part of the analyses as a result of the research that have been carried out. Such statistical graphics can also be used to help users to explore the data structure.

In the area of categorical data analysis there are several available plots for visualizing data structure such as the fourfold plot for the two-way table [Fienberg 1975; Friendly 1994, 2000], the sieve plot [Riedwyl, Schüpbach 1983, 1994; Friendly 2000], and the mosaic plot [Hartigan, Kleiner 1981, 1984; Theus 1997; Friendly 2000]. These plots can be done with the use of `vcd` and `vcdExtra` packages. There are also new functions for visualizing data such as `circlize` and `chordDiagram` available in the `circlize` package in R [Gu, Gu, Eils 2014].

This paper presents the theoretical aspects and new knowledge on the visualization and graphical techniques in R software for categorical data. We mainly focus on two new graphical approaches in visualization of categorical data and their implementation in the package `extracat`: `rmb` and `cpcp` plot [Pilhoefer, Unwin 2013]. These plots can be applied for the explanatory analysis of categorical data and allow a visual interpretation. Using these graphical tools we can be provided with a well-structured and organized representation of the data that allow for interpretation and precise comparison. In this paper we present the application of these graphical tools using data from the Central Statistical Office in Poland on economic activity.

2. Visual plots for categorical data

In this section we discuss the theory underlying the visualization of categorical data using R software. We then present major techniques based on this theory, each accompanied by an example in the next section of this paper.

In the beginning of this section we present the theory of visual plots for categorical data. There are several packages for visualizing categorical data available in R. The most popular packages are: `vcd`, `vcdExtra` that allow for graphical presentation of data in two and multi-way contingency tables. Using these packages we can use the `mosaic, sieve, double-decker`, and `association` plots. They are based on the cells residuals and they provide information on the structure of the data in the contingency table. However they do not provide information on the distribution and probability of the occurring particular category of the categorical variable.

There are plenty of plots available for categorical data allowing for the examination of the relationship between the variables analyzed, however in this paper we introduce new graphical approaches in visualization of categorical data and their implementation with the use of the `extracat` package in R system for statistical computing (R Core Team 2013). This package is available from the Comprehensive R Archive Network at: http://CRAN.R-project.org/package=extracat. It offers a variety of functions that can be used for categorical data analysis. This package offers a variety of functions that can be used for categorical data analysis, or at least deal with categorical data. Among the most interesting features are `rmb`, and `cpcp` presented in detail in this paper.

2.1. Basic buildup of `rmb` plot

A mosaic plot is a graphical display that allows to examine the relationship among two or more categorical variables. The mosaic plot starts as a square with length one [Meyer, Zeileis, Hornik 2006]. The square is divided first into horizontal bars whose widths are proportional to the probabilities associated with the first categorical variable. Then each bar is split vertically into bars that are proportional to the conditional probabilities of the second categorical variable. Additional splits can be made if wanted using a third, fourth variable, etc [Friendly 1994; Hartigan, Kleiner 1981].

The `rmb` plots are a mixture of two members of this family. The `rmb` function basically produces a Multiple Bar Chart for the relative frequencies of some target categories within each combination of the explanatory variables. The weights of those combinations (that is the absolute frequencies) are represented in the total with the corresponding bar chart. The result is a graphic which allows to read the conditional target distributions exactly without losing the information about it. Additionally the rmb function allows to draw spineplots instead of the bar charts within each explanatory combination; On that score it can be seen as a generalization of Spineplots.

The absolute frequencies n_{ijk} of the frequency table are the number of observations in the *i*-th, *j*-th, and *k*-th category of the first, second and third variable respectively. The frequencies are split into conditional relative frequencies $p_{i|jk}$ of one variable and weights corresponding to the other variable according to the equation [Pilhöfer, Unwin 2013]:

$$n_{ijk} = p_{i|jk} \cdot n_{\bullet jk} = p_{i|jk} \cdot p_{\bullet jk} \cdot n, \tag{1}$$

where n is the total number of observations, $n_{\bullet jk} = \sum_i n_{ijk}$ and $p_{\bullet jk} = \frac{n_{\bullet jk}}{n}$. The variable which is represented by the conditional relative frequencies $p_{i|jk}$ will be referred to as the target variable; other variables will be called explanatory variables and their combinations will be denoted by $n_{\bullet jk} = p_{\bullet jk} \cdot n$.

In basic `rmb` plot construction we consider a set of m categorical variables including one target. 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 `rmb` plot uses horizontal bars which means that all bars have an equal height and their widths are proportional to the ratios $\frac{n_{\bullet jk}}{\max\left(n_{\bullet jk}\right)}$.

2.2. Basic buildup of the `cpcp` plot

The categorical parallel coordinates plot (cpcp plot) is a member of the mosaicplot family and thus not capable of displaying a large number of variables and categories that can be visualized in one display. The categorical parallel coordinates plot [Uwin, Volinsky, Winkler 2003] is one of the most useful graphs in which a relatively high number of variables can be presented summarized in one display. It was first introduced by d`Ocagne [1885] and independently developed by Inselberg in 1959. The parallel coordinates plot allows for visualizing multivariate data in one graph without dropping information on the raw data values. The original concept of the cpcp plots did not allow for categorical variables, which was a serious disadvantage. Bendix, Kosara, and Hauser [2005] introduced an application for categorical variables. The cpcp plot (categorical parallel coordinates plot) is a different approach which displays both numeric and categorical variables in the same plot.

The `rmb` function basically produces a multiple bar chart for the relative frequencies of some target categories within each combination of the explanatory variables. The weights of those combinations (that is the absolute frequencies) are represented in the total with the corresponding bar chart. The result is a graphic which allows to read the conditional target distributions exactly from the graphic without losing the information about the importance (in the sense of the number of observations) of the different combinations. Additionally the `rmb` function allows to draw spineplots instead of the bar charts within each explanatory combination. On that score it can be seen as a generalization of spineplots.

To apply the PCP idea to categorical data it is not sufficient to simply convert the categories into integer values as this would lead to overplotting hiding most of the important information. To avoid this, 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 fact that for any one of these point sequences, the corresponding cases are indistinguishable regarding the corresponding variable can be used to make the display clearer and to display additional information. For this purpose 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.

In R we can use `scpcp` (Static Categorical Parallel Coordinates Plot) in `extracat` library. This function creates a static categorical parallel coordinates plot using base R graphics. The function offers color brush/highlighting and several options for the labels and colors. Efficiency is improved by replacing sets of parallel lines by polygons.

3. Application in R

In this paper we apply the rmb and cpcp functions available in R software. We use the Central Statistical Office report on economic activity of the population in Poland in 2015. In the empirical part of this paper we apply three categorical variables presented in Table 1.

Table 1. Economic activity in Poland in 2015.

Variable	Levels
Voivodeship	łódzkie, mazowieckie, małopolskie, śląskie, lubelskie, podkarpackie, podlaskie, świętokrzyskie, lubuskie, wielkopolskie, zachodniopomorskie, dolnośląskie, opolskie, kujawsko-pomorskie, pomorskie, warmińsko--mazurskie
Sex	male, female
Economic activity	working, unemployed

Source: Central Statistical Office – Local Data Bank [http://stat.gov.pl].

These data form a three-way contingency table (Table 1) presenting the number of people working and unemployed (male and female in 16 voivodeships).

Table 2. Economic activity in Poland in 2015 dataset

Voivodeship	Active working		Active unemployed	
	male	female	male	female
1	2	3	4	5
Łódzkie	657	535	56	47
Mazowieckie	1 447	1 193	105	80
Małopolskie	680	545	51	45
Śląskie	943	767	72	62
Lubelskie	553	434	57	48
Podkarpackie	437	332	55	49

1	2	3	4	5
Podlaskie	260	207	20	16
Świętokrzyskie	323	243	38	28
Lubuskie	223	174	15	13
Wielkopolskie	721	540	40	41
Zachodniopomorskie	297	240	24	20
Dolnośląskie	611	483	45	39
Opolskie	210	169	13	14
Kujawsko-pomorskie	466	356	37	35
Pomorskie	543	414	41	29
Warmińsko-mazurskie	307	233	30	29

Source: Central Statistical Office – Local Data Bank [http://stat.gov.pl].

The data table presented in Table 2 makes a three-way contingency table that will be used in further analysis and visualization. In this paper we present the `extracat` package in R and two modern plots: the `rmb` and the `cpcp` plots. We apply these functions for data on economic activity in Poland in 2016 presented in Table 2.

3.1. `rmb` plot

The first type of plot available in the `extracat` package in R is the `rmb` plot. Figure 1 shows the variable Voivodeship on the x-axis, and Sex on the y-axis, and Activity as the target variable which is by convention on the x-axis.

Below we present two types of rmb plots, each with different variables on the x and y axis (Figures 1 and 2).

The first plot (Figure 1) is an `rmb` plot data on economic activity in Poland in 2015.

The graphic reveals no correlation between sex and target variable – economic activity. The structure of red bars indicating working people is almost equal between males and females; the green bars indicating unemployed males and females are almost equal. We can also see the big difference in the height of the red and green bars. The red bars indicating working people are higher than the green indicating unemployed. This distribution appears almost equal in all voivodeships for males and females.

Using the `rmb` function we can also plot a graph with Sex and Voivodeship on the x-axis, the target variable – Economic activity on the x-axis, and the probability on the y-axis. This type of rmb plot is presented in Figure 2. This plot presents the probability of the occurrence of females and males economically active and unemployed in the voivodeships.

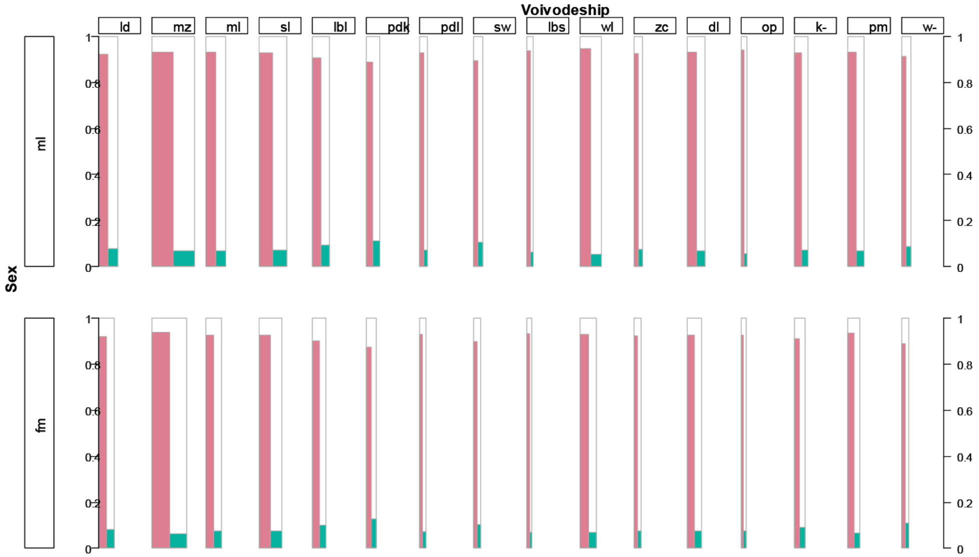

Fig. 1. `rmb` plot for data on economic activity in Poland in 2015

Source: own calculations in R.

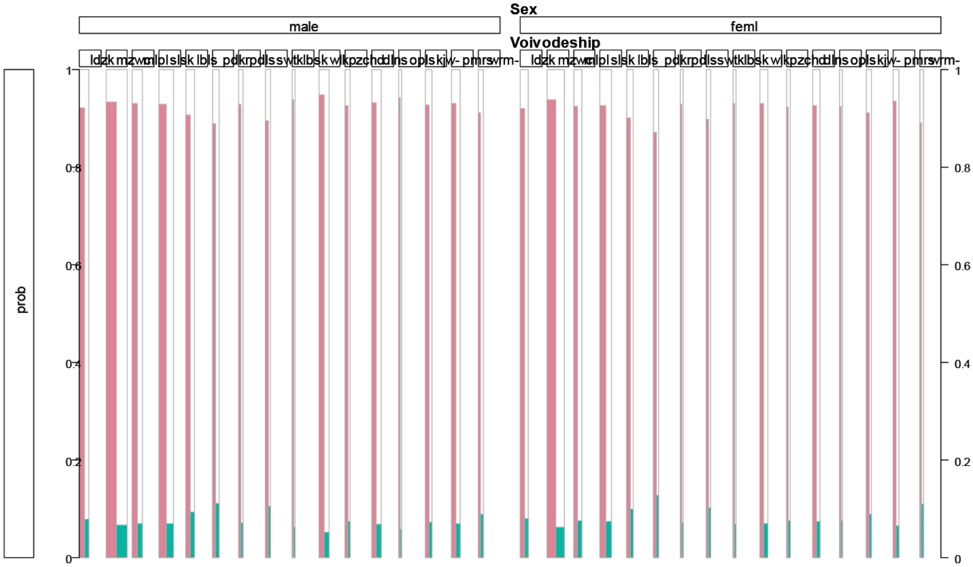

Fig. 2. `rmb` plot for data on economic activity in Poland in 2015 using the options `spine` and `freq.trans`

Source: own calculations in R.

We can observe that there are high bars for economic active with high probability values around 0.9, and low probability (green bars) for unemployed with the values at around 0.1. Looking at the rmb plot in Figure 2 we can see that there is no correlation between Sex and the target variable – Economic activity – in every voivodeship that is analyzed.

3.2. cpcp plot

The second type of the new plot available in the `extracat` package in R is the `cpcp` plot. The concept of parallel coordinates plot (pcp plot) was discovered in the late 19th century by Unvin, Volinsky and Winkler [2003]. These plots are amongst the most useful graphical solutions with which a relatively huge number of variables can be visualized in one display. The original concept did not allow for the analysis of categorical variables, which was a serious disadvantage. Bendix, Kosara and Hauser [2005] developed an application for categorical data which was implemented firstly in Parallel Sets software (ver. 2.1), and later in R software [Pilhöfer, Unwin 2013]. This approach displays both numeric and categorical variables in one plot.

The `cpcp` function in R software offers color brush/highlighting and several options for the labels and colors. Efficiency is improved by replacing sets of parallel lines by polygons. This function provides a parallel coordinates plot for categorical as well as continuous data based on the `ipcp` function in the `iplots` package. It 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. The `cpcp` function provides a parallel coordinates plot for categorical as well as continuous data based on the `ipcp` function in the `iplots` package. It 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.

We apply the `cpcp` function to data on economic activity in Poland in 2015 (Table 2).

In Figure 3 using the `cpcp` function, no category is highlighted or blended. This plot was done using default parameters. We can see that the result is in grey only. This plot is called a static version of the `cpcp` plot.

We can use another option of the `cpcp` plot with one category highlighted and blended. In Figure 4 we can see that the first category of the first variable is shown in red. This is an interactive `cpcp` plot.

The new interactive `cpcp` plot (Figure 4) shows in red the first category of the first variable. This increases the interpretability of the plot and better reveals the hierarchical splitting from left to right. From Figure 4 we can see that the biggest group are working and economically active females and males shown in grey. The

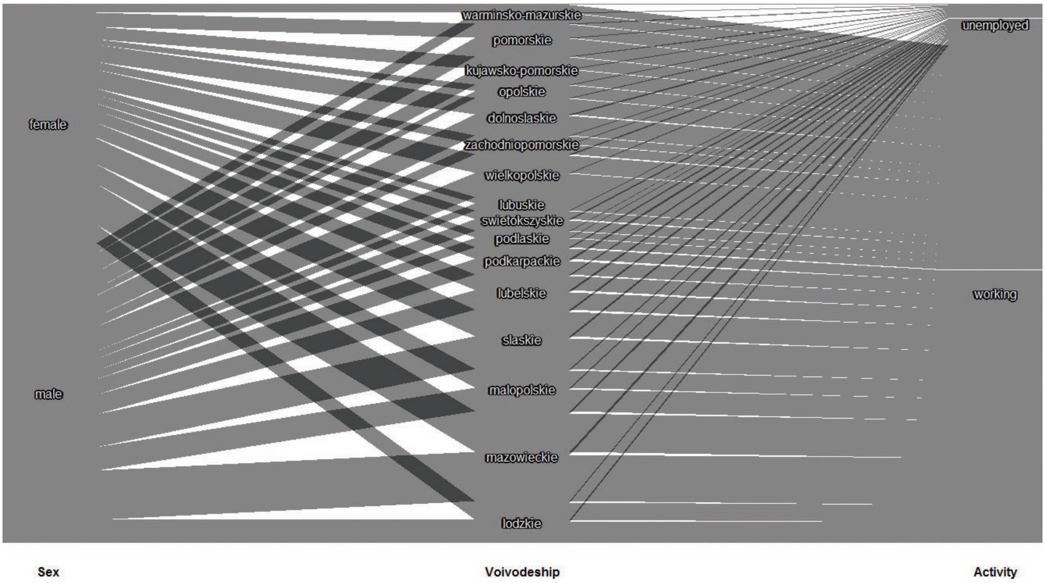

Fig. 3. `cpcp` plot of sex, voivodeship and activity using default parameters

Source: own calculations in R.

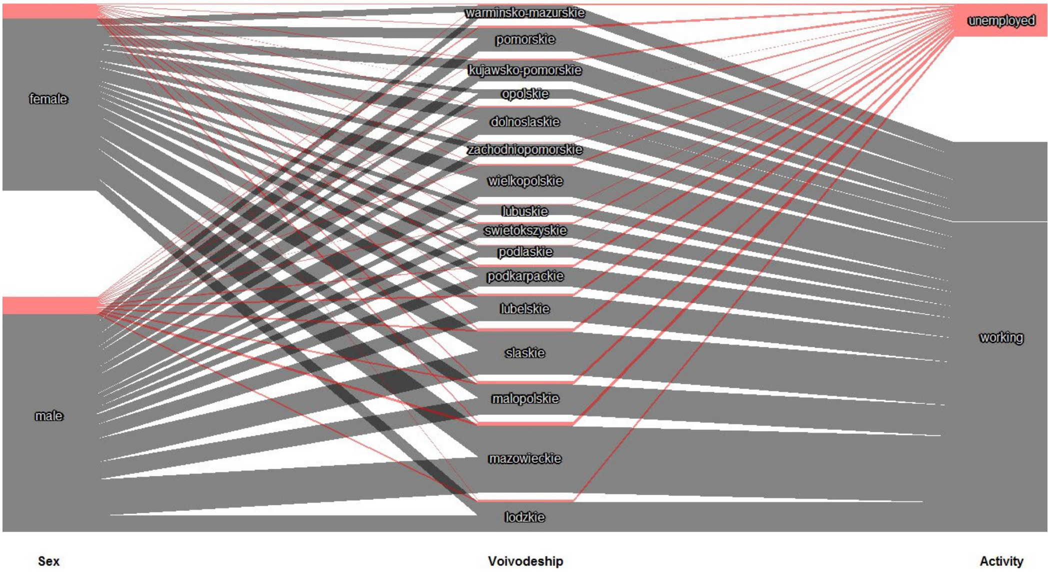

Fig. 4. `cpcp` plot of sex, voivodeship and activity using gap type equal 0

Source: own calculations in R.

larger group is observed for males, however the difference between females and males is very small. The dominance of the economically active over the unemployed is visible in every voivodeship, which can be seen as a great advantage in comparison to other visual plots. The other advantage of this plot is the sorting and ordering option which can be used to minimize the number of lines crossing. The disadvantage of such plots is that it is not possible to read the corresponding proportions of the selected group within the categories from the graphic.

4. Conclusions

In this paper we introduced two extensions of well-known graphics for the visualization of categorical data: the `rmb` and the `cpcp` plots available in the `extracat` package in R.

The rmb plot is a member of the mosaic plot family which displays the natural factorization of absolute frequencies into conditional relative frequencies and their

weights. This is very useful especially for the analysis of target variables. In this plot we have the equal-width option which are the key features for displaying especially small frequencies from the contingency table. Residual shadings are used with log-linear and logistic models and the option to use rmb plots as a generalization of spineplots further increases the flexibility of the graphic. The rmb plot is mostly used for displaying and specifying the data more precisely.

On the contrary, the `cpcp` plot is an attempt to increase the number of displayable categorical variables using the well-established parallel coordinates plot as its basis. There are several possible layouts in R software. 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. One possible way of combining the graphics in a graphical analysis of categorical data is the following: A `cpcp` plot is used for the interactive exploration of the dataset and the `rmb` plots are then used to display any specific findings in the data more precisely. The cpcp plot is mainly used for the interactive exploration of the dataset.

There are several layout options in R which the researchers will find very friendly and flexible to use. Such graphics can be a great tool in modern graphical methods that can be a sort of competitive methods to the well-known mosaic, double-decker, association and sieve plots for contingency tables.

Bibliography

Bendix F., Kosara R., Hauser H., 2005, *Parallel Sets: Visual Analysis of Categorical Data*, Proceedings of the 2005 IEEE Symposium on Information Visualization (InfoVis), pp. 133-140.

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

Fienberg S.E., 1975, *Perspective Canada as a social report*, Social Indicators Research, 2, pp. 153-174.

Fisher R.A., 1925, *Statistical Methods for Research workers*, Originally published in Edinburgh by Oliver and Boyd.

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.

Gu Z., Gu L., Eils R., 2014, *Matthias Schlesner, Benedikt Brors, Circlize Implements and enhances circular visualization in R*, Bioinformatics, Oxford, England.

Hartigan J.A., Kleiner B., 1984, *A mosaic of television ratings*, The American Statistician, 38, pp. 32-35.

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.

Pilhöfer A., Unwin A., 2013, *New approaches in visualization of categorical data: R package extracat*, Journal of Statistical Software, 53(7), pp. 1-25. 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*, Springer-Verlag, New York, pp. 268-273.

Riedwyl H., Schüpbach M., 1983, *Siebdiagramme: Graphische darstellung von kontingenztafeln*, Technical Report 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland.

Riedwyl H., Schüpbach M., 1994, *Parquet diagram to plot contingency tables*, [in:] Faulbaum F. (ed.), Softstat `93: *Advanced In Statistical Software*, Gustav Fischer, New York, pp. 293-299.

Theus M., 1997, *Visualization of categorical data*, Advanced in Statistical Software, Lucius & Lucius, 6, pp. 47-55.

Unwin A., Volinsky C., Winkler S., 2003, *Parallel Coordinates for Exploratory Modelling Analysis*, Computational Statistics & Data Analysis, 43(4), pp. 553-564.