Review of Visualization Methods for Categorical Data in Cluster Analysis

Jana Cibulková[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Barbora Kupková[2] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 25.1.2022 (revision received 9.8.2022), **Accepted** (reviewed) 1.9.2022, **Published** 16.12.2022

Abstract

The paper focuses on visualization methods suitable for outcomes of cluster analysis of categorical data (nominal data, specifically). Since nominal data have no inherent order, their graphical representation is often challenging or very limited. This paper aims to provide a list of common visualization methods in the domain of cluster analysis of objects characterized by nominal variables. Firstly, the various plot types (such as clustering scatter plot, dendrogram, icicle plot) for cluster analysis are presented, and their suitability for presenting clusters of nominal data is discussed. Then, we study approaches of sorting nominal values on chart axes in such a way that would improve visualization of the data. Lastly, we introduce a simple alternative to cluster scatter plot for nominal data, that makes the final visualization of clustering solution more efficient since the pattern and groups in data are now more apparent. The suggested method is demonstrated in illustrative examples.

Keywords	***DOI***	***JEL code***
Cluster analysis, nominal data, hierarchical clustering, visualization	*https://doi.org/10.54694/stat.2022.4*	*C38, C18*

INTRODUCTION

Cluster analysis belongs to a group of unsupervised learning methods. Usually, its objective is to divide a set of objects into groups, called clusters. The aim is to define clusters in such a way that objects would be homogeneous within one cluster and heterogeneous among different clusters. There have been many clustering algorithms developed in very different fields: artificial intelligence, information technology, image processing, biology, psychology, marketing, etc. In this paper we focus mainly on hierarchical cluster analysis (HCA) and explain how its clustering solutions may be visualized. However, many visualization methods (including newly proposed ones) do not limit themselves to HCA only.

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: jana.cibulkova@vse.cz.
[2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: kupb01@vse.cz.

Often, it is required to use cluster analysis on categorical data, especially nominal data. This situation might occur in the field of biology (taxonomic category), in medicine (the listing of drugs), in marketing (marketing personas), etc. Nominal data clustering has not yet been studied to the same extent as quantitative data clustering has been, which also applies to its visualization. In this paper, we focus on visualization methods suitable for outcomes of cluster analysis of nominal data since visualization is becoming increasingly important for the analysis and research of multidimensional data (Ma and Hellerstein, 1999).

Working with nominal data has its specifics: it is not possible to sort nominal values in any objective way; we can only state whether the values are identical or not. This causes problems with the visualization procedures commonly used for quantitative data as there is no clear right way to place variables on axes of these graphs. Technically, any order is correct. However, as shown in this paper, there are certain ways of sorting these variables on the axes, which can significantly increase the quality of the information that the graph provides.

Several authors introduced new methods and approaches suitable for the visualization of categorical data in cluster analysis. Andrews (1972) presented a smoothed version of a parallel coordinate plot. Chernoff (1973) introduced a method that produces different images of human faces. Later, more advanced techniques were proposed. Hofmann and Buhmann (1995) proposed three strategies derived in the maximum entropy framework for visualizing data structures. Pölzlbauer et al. (2006) and Vesanto (1999) presented visualization methods using self-organizing maps, while the use of a minimum spanning tree for visualizing of HCA is discussed by Kim et al. (2000). Itoh et al. (2004) proposed an algorithm that can provide overviews of structures and the content of the hierarchical data. Chang and Ding (2005) proposed a method for visualizing clustered categorical data based on three-dimensional space.

Other authors, such as Ma and Hellerstein (1999) or Rosario et al. (2004) explored possibilities of ordering nominal data in order to improve visualization. These methods may be useful if one wishes to apply a well-known visualization technique that is commonly used for quantitative data on nominal data.

This paper aims to provide an overview of various available visualization methods in the domain of cluster analysis of objects characterized by nominal variables. Firstly, the most common visualization methods for cluster analysis are presented, and their suitability for presenting clusters of nominal data is discussed. Then, we study approaches of sorting nominal values.

Lastly, we present our alternative of cluster scatter plot for nominal data, which makes the final visualization of clustering solution more efficient since the pattern and groups in data are now more apparent. The suggested methods are demonstrated in illustrative examples, and all the computations and graphs, unless stated otherwise, were prepared by authors in R (R Core Team, 2021); HCA of nominal data was done using package *nomclust* (Šulc et al. 2021).

1 OVERVIEW OF VISUALIZATION METHODS FOR HCA

This section contains an overview of visualization methods suitable for cluster analysis. Each method is briefly described, a simple example is provided, and a method's suitability for presenting clusters of nominal data is discussed.

1.1 Dendrogram

A *dendrogram* is a diagram, that illustrates clusters creation in the process of HCA. Figure 1 shows an example of simple dendrogram on a dataset of 17 observations. According to Sibson (1973), we define a dendrogram to be function:

$$c\colon [0,\infty] \to E(D), \tag{1}$$

where D is the dataset, $E(D)$ is the set of equivalence relations on D, δ represents distance between objects, and c satisfies these conditions:

$$h \leq h' \text{ implies } c(h) \subseteq c(h'), \tag{2}$$

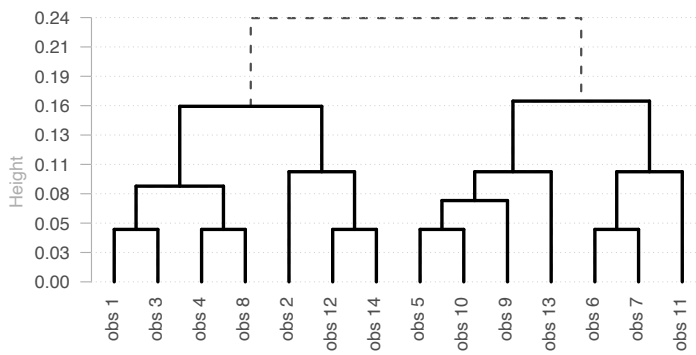
$$c(h) \text{ is eventually } D \times D, \tag{3}$$

$$c(h + \delta) = c(h) \text{ for all small enough } \delta > 0. \tag{4}$$

Hence, a dendrogram is a nested sequence of partitions with associated numerical levels. A dendrogram is usually represented as a tree diagram, but there is a great deal of freedom, and various alterations exist.

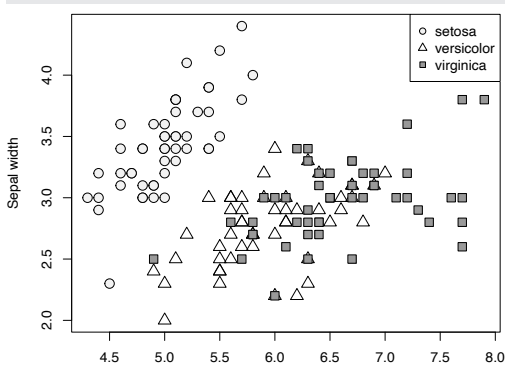
As mentioned above, the dendrogram represents the HCA process. Therefore, dataset with any type (including nominal) may be visualized, which is a big advantage. The disadvantage of this visualization method is that with an increasing number of observations, the dendrogram becomes difficult to read.

Figure 1 Example of a simple dendrogram



Source: Authors

Figure 2 Example of a simple cluster scatter plot on dataset iris



Source: Authors

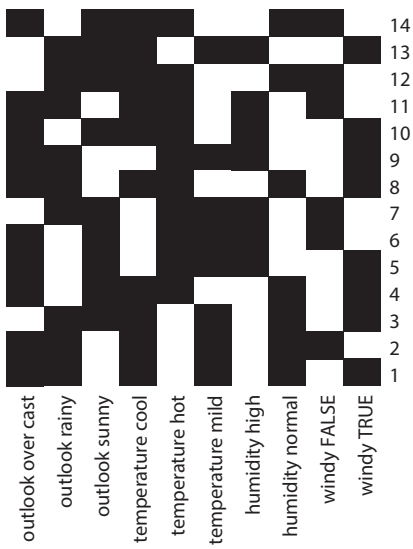
1.2 Cluster scatter plot

A *scatter plot* uses dots to represent values for two different numeric variables. The position of each dot relative to the horizontal and vertical axes indicates values for an individual data point. Scatter plots are used to observe relationships between variables. Up to four variables can be plotted in a scatter plot: two numerical variables on the *x*- and *y*-axis, a numerical or ordinal variable for the definition of point size, and a nominal variable for color definition (Rahlf, 2019).

Cluster scatter plot usually represents three variables: two chosen numerical variables on the *x*- and *y*-axis a nominal variable for

a color definition of an assigned cluster. This is a powerful visualization tool since it shows objects assignation into a cluster with respect to two chosen variables; it shows how well separated the clusters are and how many observations are within each cluster. Figure 2 shows such a cluster scatter plot for dataset iris (Anderson, 1935; Fisher, 1936). However, there is no clear right way to place variables on axes of a scatter plot when dealing with nominal data. Moreover, nominal data usually takes value from a relatively small number of categories; hence problem with overlapping points most likely occur even if we know how to place nominal variables on axes.

Figure 3 Example of a heat map



Source: Authors

1.3 Heatmap

A *heat map* is a two-dimensional matrix in which the cells are colored depending on their value. When visualizing outcomes of cluster analysis, color represents an assignation into a given cluster. It may be a table with individual data or aggregated values (Rahlf, 2019).

Heatmaps can be used for the visualization of all data types and their assignations into clusters. Before constructing a heatmap of clustering solution of nominal data, all nominal attributes are usually transformed into binary variables that are then treated as numeric. Hence, if the nominal attribute has k possible values, it is replaced by $k - 1$ synthetic binary variable, the i-th being 0 if the value is one of the first i in the ordering and 1 otherwise (Witten, 2011). Heatmap becomes hard to read easily when there are many variables with numerous categories, and binary transformation is performed, see Figure 3.

There is no rule stating how the rows or columns should be arranged (Rahlf, 2019). Rearranging rows or columns in a meaningful way would increase the overall readability of the plot, and it could help to discover hidden patterns in the data.

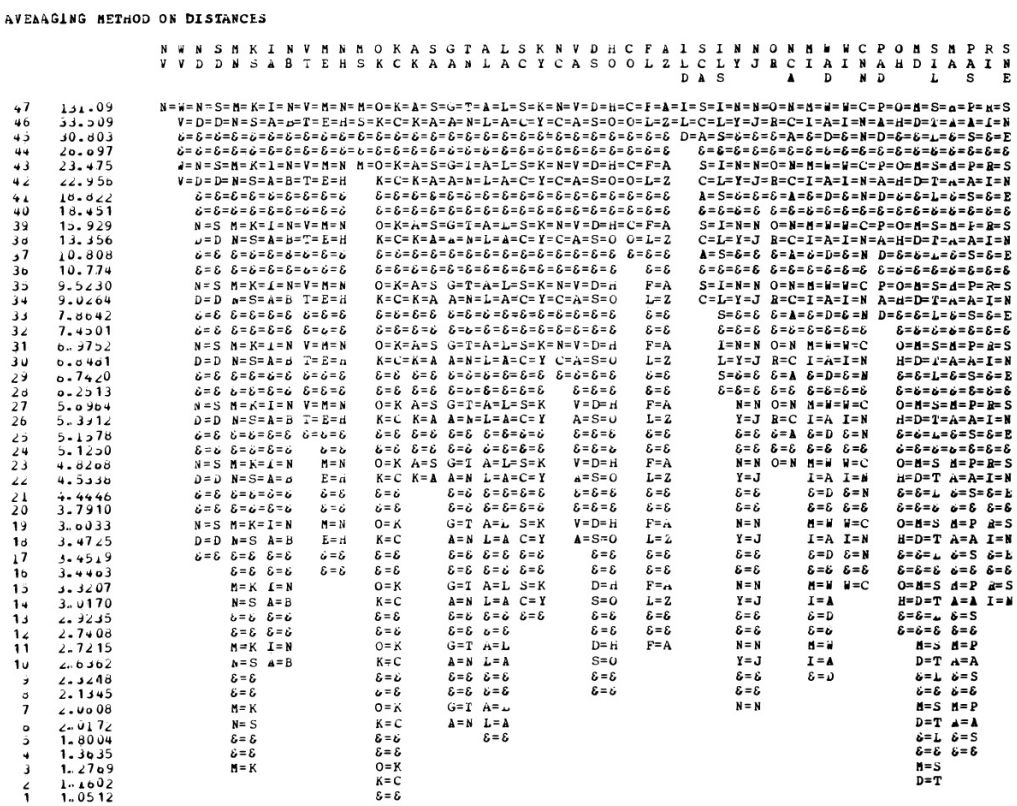
1.4 Icicle plot

According to Kruskal and Landwehr (1983), an *icicle plot* should be easier to read off which object belongs to which clusters and which objects join or drop out from a cluster as we move up or down the levels of hierarchy. However, this method was proposed in 1983, and it took into consideration the very limited capability of printers to print plots, see Figure 4. This plot is basically an upside-down variation of a dendrogram; therefore, it is suitable for visualization of the process of hierarchical clustering of any data type, including nominal one.

1.5 Andrew's plot

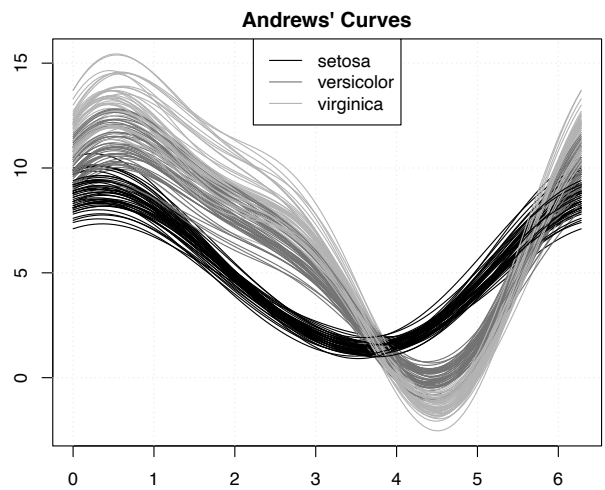
Andrew (1972) introduced a method to plot high-dimensional data with curves. This plot is based on the same principles as a parallel coordinate plot, and it is called *Andrew's plot* or *Andrew's curve*. Each curve represents one object, obtained by using the components of the data vectors as coefficients of orthogonal sinusoids, which are then added together pointwise. Figure 5 shows Andrew's plot for dataset iris (Anderson, 1935; Fisher, 1936). This graph surely can be used for the visualization of clustering solutions. However, it doesn't solve the problem with sorting categories, and overlaps of curves make it difficult to see the inherent structure in the data (especially if clusters are not completely homogeneous).

Figure 4 Example of an icicle plot



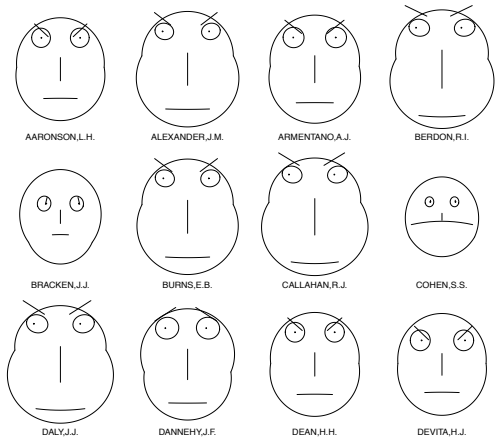
Source: Kruskal and Landwehr (1983)

Figure 5 Example of Andrews plot on dataset iris



Source: Authors

Figure 6 Example of Chernoff faces on dataset USJudgeRatings



Source: Authors

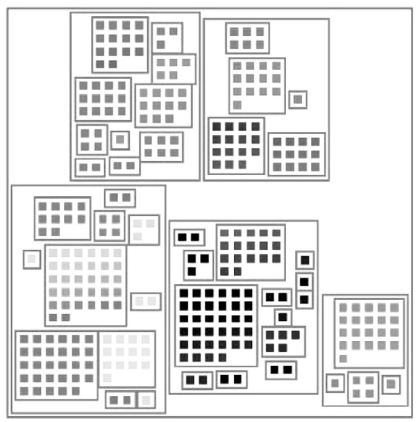
1.6 Chernoff's faces

Chernoff (1973) famously introduced a method to visualize multivariate data with faces. Each point in a d-dimensional space ($d \leq 18$) is represented by a facial caricature whose features such as length of nose, eyebrow position, mouth size, and shape are determined by the value of the corresponding variable. Figure 6 shows *Chernoff's* faces on dataset USJudgeRatings using library TeachingDemos in R (Snow, 2020).

1.7 Itoh's nested rectangles

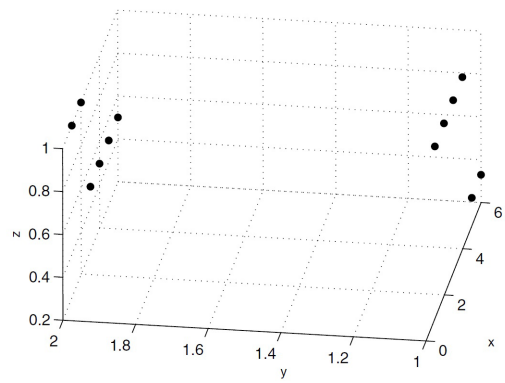
Itoh et al. (2004) presented a technique for representing large-scale hierarchical data by using nested rectangles. It first packs icons or thumbnails of the lowest-level data and then generates rectangular borders that enclose the packed data.

Figure 7 Hierarchical data using Strip Squarified Treemap



Source: Itoh (2004)

Figure 8 Plot of the two clusters



Source: Chang and Ding (2005)

It repeats the process of generating rectangles that enclose the lower-level rectangles until the highest-level rectangles are packed. It provides good overviews of complete structures and the content of the data in one display space. The approach refers to Delaunay triangular meshes connecting the centers of rectangles to find gaps where rectangles can be placed, see Figure 7.

1.8 Chang's and Ding's three-dimensional scatter plot

Chang and Ding (2005) proposed a method for visualizing clustered categorical data. Their method allows users to adjust the clustering parameters based on the visualization, so it is a new visualization method as well as a new clustering method. In this method, a special three-dimensional coordinate system is used

to represent the clustered categorical data. The three-dimensional coordinate system to plot a variable's value is constructed such that the x-axis represents the variables, the y-axis represents the variable's values, and the z-axis represents the probability that the variable's value is in the cluster. To display a set of clusters, the methods construct a coordinate system such that interference among different clusters can be minimized in order to observe closeness.

The visualization of the two categorical clusters using this method is shown in Figure 8. However, this approach doesn't provide information about a cluster's size, and it may be confusing when a large number of variables and their categories are visualized.

2 SORTING VARIABLES

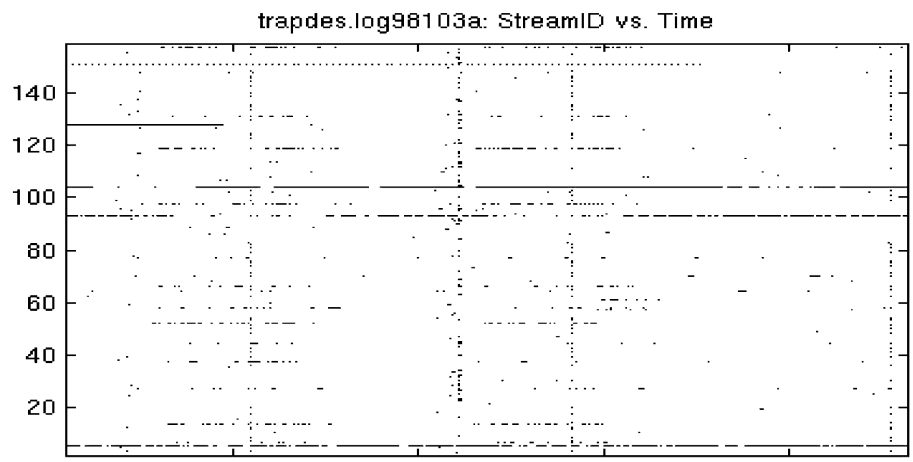
In this section, we focus on reordering categories of nominal variables in such a way that would:

- improve visualization of the data,
- allow us to adjust well-known methods for clusters visualization of quantitative data and to extend them nominal data as well.

Although the importance of sorting categorical values on chart axes is obvious, there are not many approaches to sorting them yet. Among the best known is the manual sorting of categories, which requires an experienced user. Another option to sort values of the nominal variables is to sort by an auxiliary quantitative variable, such as time. This method is not intended for visualizations and generally does not lead to satisfactory results (Ma and Hellerstein, 1999).

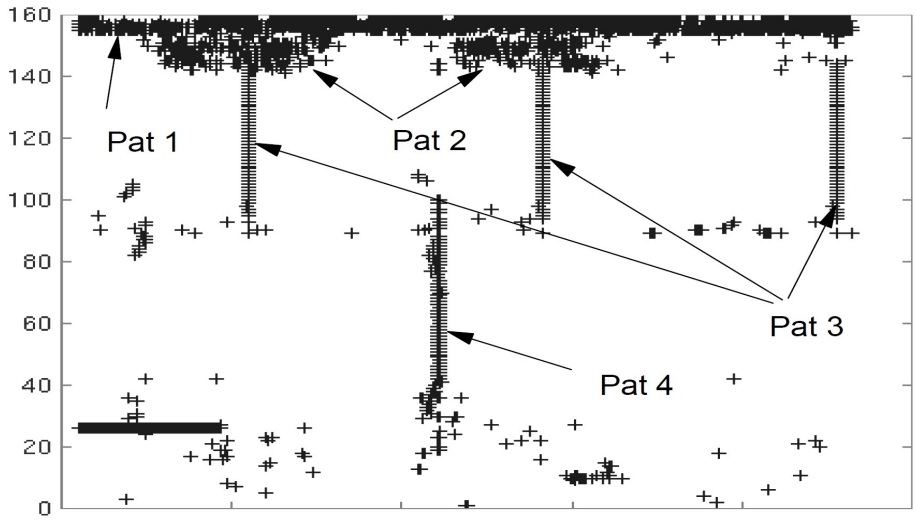
Ma and Hellerstein (1999) proposed an algorithm for ordering nominal data by constructing natural clusters, sequencing these clusters to minimize order conflicts (situations in which the same category must have two or more positions in the ordering) and ordering the values within the clusters to eliminate the above-mentioned order conflicts. The effect of this approach can be seen in Figures 9 and 10, where x-axis represents time and y-axis represents a host name in a production network. We can see which hosts generate events at specific time in Figure 10, while there are no obvious patterns visible in Figure 9. Even though the algorithm seems effective, it is computationally demanding.

Figure 9 Scatter plot example of randomly ordered data



Source: Ma and Hellerstein (1999)

Figure 10 Scatter plot example of data ordered by Ma's and Hellerstein's algorithm



Source: Ma and Hellerstein (1999)

Rosario et al. (2004) proposed another method, called the *Distance-Quantification-Classing approach* (DQC), to preprocess nominal variables before being imported into a visual exploration tool. This method solves the problem of order-and-spacing assignation among nominal values and reduces the number of distinct values to display. It works in three steps:

1 *Distance Step:* We identify a set of independent dimensions that can be used to calculate the distance between nominal values.
2 *Quantification Step:* We use the independent dimensions and the distance information to assign order and spacing among the nominal values.
3 *Classing Step:* We use results from the previous steps to determine which values within the domain of a variable are similar to each other and thus can be grouped together.

Each step in the DQC approach can be accomplished by a variety of techniques.

3 ALTERNATIVE OF CLUSTER SCATTER PLOT FOR NOMINAL DATA

In this section, we present an alternative of cluster scatter plot for nominal data because cluster scatter plot is such an essential visualization tool for data analysis.

The alternative of cluster scatter plot for nominal data shows a relationship between two nominal variables. Let's assume, that we have a dataset with nominal variables var_1, var_2, … var_m. Let's also assume that cat_l is a number of unique categories of a variable var_l; $l = 1, 2 \ldots, m$. Moreover, let's assume that we know an assignation to a cluster of each observation, let's assume k is number of clusters of our final clustering solution, that we wish to visualize on the plot.

To create a good visualization of clusters of objects represented by nominal variables *var_i* and *var_j*; where $1 \leq i, j \leq m$; $i \neq j$, we need to solve following three problems:

- *Sorting and spacing,*
- *Cardinality,*

- *Clarity of the graph.*

3.1 Sorting and spacing

As mentioned in Section 1.2, a scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. We aim to create an alternative to a scatter plot that uses dots to represent values for two different *nominal* variables var_i and var_j; $1 \leq i, j \leq m$; $i \neq j$. Hence, we need to assign a numerical value to each nominal value. In order to simplify the problem, we assume that distances among values are equal. Hence, we only need to sort values of var_i and var_j on *x*-axis and *y*-axis in such a way that clusters will be the most homogeneous and compact. These are the step we are proposing:

1. We calculate contingency tables of relative frequencies of var_i and var_j for each of k clusters. Let's denote C is an array of k contingency tables.
2. Sorting of var_i values: We calculate the sum of relative frequencies across all var_i values for each cluster. This way a new temporary dataset of k columns and cat_i rows is created. Each row corresponds to a vector of relative frequencies of the corresponding value of var_i of clusters 1 to k. We perform average-linkage HCA with Euclid distance on this temporary dataset. We obtain a dendrogram and from that dendrogram we can easily get its sorted leaves, that correspond to desired sorted values of var_i.
3. Sorting of var_j values, analogously to var_i.
4. We re-arrange contingency tables C according to the obtained new order for values of var_i and var_j.

3.2 Cardinality and clarity of the graph

Rahlf (2019) says that up to four variables can be plotted in a scatter plot: two numerical variables on the x- and y-axis, a numerical or ordinal variable for the definition of dot size, and a nominal variable for color definition. In our case, the first and second variables on the x- and y-axis need to be nominal variables. Color (and dot shape) definition must correspond to the assigned cluster. Dot size must correspond to the overall relative frequency of given values combination of examined two variables. Moreover, the relative frequencies need to be reasonably scaled so all the dots on the plot would be visible.

Hence *min-max normalization* is applied to contingency tables, to set point size within a range from a to b. Normalized contingency tables C' follow this formula:

$$C' = \frac{C - min(C)}{max(C) - min(C)} \times (b - a) + a\,, \tag{5}$$

where C is an array of k contingency tables, $\min(C)$ is minimal relative frequency across all k contingency tables, $\max(C)$ is maximal relative frequency across all k contingency tables. A range of point sizes in the final graph, defined by a and b, can be set based on personal preferences, we suggest $a = 1$ and $b = 6$.

3.3 Illustrative example of newly proposed visualization method applied on data

We demonstrate the newly proposed visualization method on a dataset from Cortez and Silva (2008). The dataset was obtained in a survey of students' math courses in a secondary school. It contains a lot of interesting social, gender, and study information about students. For our analysis, we performed average-linkage HCA with Eskin distance (Eskin et al., 2002) measure on three nominal variables:

- *Mjob* – mother's job;
  nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other',
- *Fjob* – father's job;

nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other',

- *Address* – student's home address type;
  nominal: 'U' – urban or 'R' – rural.

The above mentioned Eskin distance is a distance measure which can express a dissimilarity between two objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ that correspond to rows of given categorical data matrix $X = [\boldsymbol{x}_{ic}]$, where $i = 1, 2, \ldots, n$ and $c = 1, 2, \ldots, m$. It can be calculated as follows:

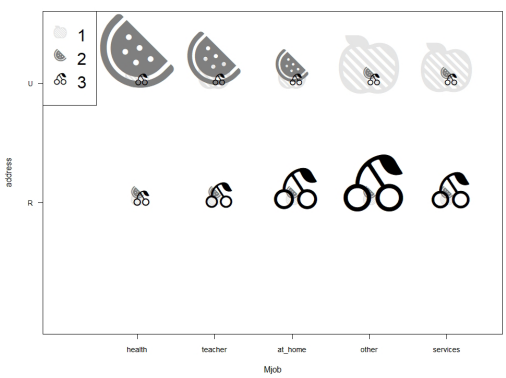
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{S(\boldsymbol{x}_i, \boldsymbol{x}_j)} - 1 , \tag{6}$$

where $S(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is total similarity between the objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ calculated as arithmetic mean of similarities $S(x_{ic}, x_{jc})$ given by formula:

$$S(x_{ic}, x_{jc}) = \begin{cases} 1; \; if \; x_{ic} = x_{jc} \\ \dfrac{K_c^2}{K_c^2 + 2} \;\; if \; x_{ic} \neq x_{jc} \end{cases} , \tag{7}$$

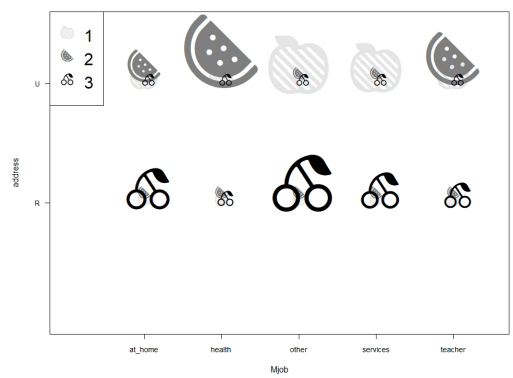
where the number of categories of the c-th variable is denoted as K_c.

Figure 11 Sorted alternative of cluster scatter plot for nominal data



Source: Authors

Figure 12 Unsorted alternative of cluster scatter plot for nominal data



Source: Authors

The Eskin distance measure is implemented in R package nomclust (Šulc et al., 2021), that was used to perform HCA. Then, observations were divided into three clusters.

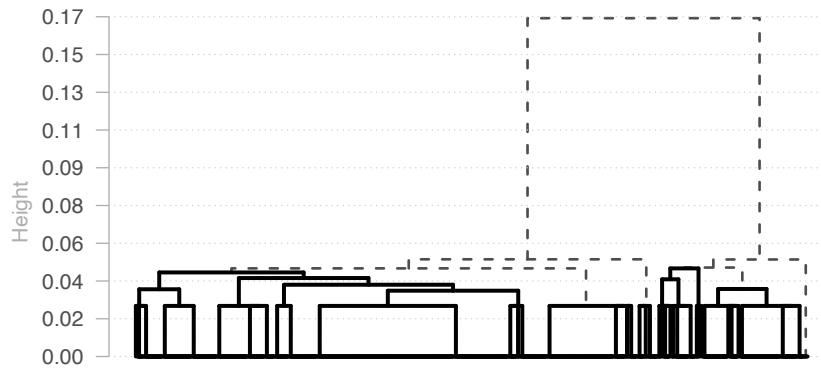
Hence, we can visualize the relationship between variable *Mjob* and *address*. Three contingency tables were calculated, and the new order of values of given variables was obtained. Contingency tables were re-arranged based on the new order and then normalized to range from 1 to 6. Lastly, Figure 11 was created.

The plot re-arranges the nominal values based on frequencies, so it is easier to spot any pattern in the data. For example, we can clearly see, that there are main families living in rural area in the third cluster, while urban area is occupied by families from the first and second cluster. Mothers working in health

care and in education are dominating in the second cluster. Mothers working in services or elsewhere are more prominent in the third cluster. The same interpretation would be harder to spot if we did not perform any sorting; see Figure 12 for comparison.

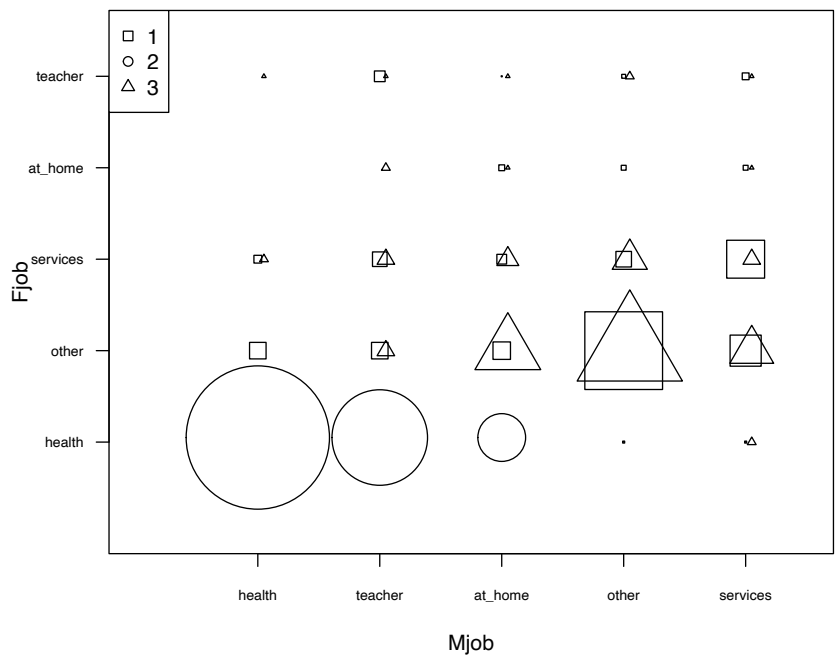
While a dendrogram illustrates a process of HCA, agenesis of clusters, and closeness of clustered objects, it does not provide any information about clusters' internal composition. However, the newly

Figure 13 Dendrogram of HCA of Mjob and Fjob



Source: Authors

Figure 14 Alternative of cluster scatter plot for clusters of Mjob and Fjob



Source: Authors

proposed alternative of cluster scatter plot provides such inside, and thus it is a useful complementary tool to dendrogram. This may be illustrated in the case of clustering objects in the aforementioned dataset from Cortez and Silva (2008). Based on the dendrogram in Figure 13, we can see how three clusters were formed. Figure 11 and Figure 13 show the inner structure of these clusters. Figure 11 shows the relationship between address and mother's job, while Figure 14 shows the relationship between mother's and father's job, with respect to the prevalence of category combinations within clusters. Due to the new visualization method, the patterns within data are visible. Hence it is obvious that the first and second cluster is formed by students who live in an urban area, and the third cluster is formed by students living in rural areas. There are almost no fathers who work as teachers or who stay at home, and very few of them work in civil services. Civil services or other areas are the most dominant for mothers in the first and third clusters. Stay-at-home mothers are the most prominent in the third cluster with students from rural areas, and in the second cluster with students living in an urban area but with a father who works in healthcare. Not all the mothers of students from the second cluster stay at home; they usually work in healthcare or as teachers.

CONCLUSION

Data visualization is vital in the final step of data mining applications. However, clustering and visualization of nominal data have not been explored to such an extent as clustering and visualization of the quantitative data has been. This paper provided an illustrative overview of available visualization methods in the area of cluster analysis with a focus on the visualization of nominal data. Two methods for sorting nominal data in order to improve visualization were also briefly presented. The goal of this paper was to provide an overview of existing methods but also to identify opportunities to improve the clustering visualization on nominal data. We proposed a new and very simple method to visualize the relationship between two nominal variables and assignation of observations into clusters as an alternative of cluster scatter plot for nominal data.

ACKNOWLEDGMENT

This work was supported by the Prague University of Economics and Business under Grant IGA F4/22/2021.

References

ANDERSON, E. (1935). The Irises of The Gaspe Peninsula [online]. *Bulletin of the American Iris Society*, 59: 2–5. <https://doi.org/10.2307/2394164>.

ANDREWS, D. (1972). Plots of high-dimensional data [online]. *Biometrics*, 28(1): 125–136. <https://doi.org/10.2307/2528964>.

CHANG, C., DING, Z. (2005). Categorical Data Visualization and Clustering Using Subjective Factors [online]. *Data & Knowledge Engineering*, 53(3): 243–262. <https://doi.org/10.1007/978-3-540-30076-2_23>.

CORTEZ, P., SILVA, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *FUture BUsiness TEChnology Conference*, 5: 5–12.

ESKIN, E. et al. (2002). *A Geometric Framework for Unsupervised Anomaly Detection* [online]. Boston: Springer, 77–101. <https://doi.org/10.7916/D8D50TQT>.

FISHER, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems [online]. *Annals of Eugenics.* 7(2): 179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.

HOFMANN, T., BUHMANN, J. (1995). Multidimensional Scaling and Data Clustering [online]. *The MIT Press: Advances in Neural Information Processing Systems*, 7: 459–466. <https://doi.org/10.5555/2998687.2998744>.

ITOH, T. et al. (2004). Hierarchical data visualization using a fast rectangle-packing algorithm [online]. *IEEE Transactions on Visualization and Computer Graphics*, 10(3): 302–313. <https://doi.org/10.1109/TVCG.2004.1272729>.

KIM, S. et al. (2000). Interactive Visualization of Hierarchical Clusters Using MDS and MST [online]. *Metrika*, 51(1): 39–51. <https://doi.org/10.1007/s001840000043>.

KRUSKAL, J. B., LANDWEHR, J. M. (1983). Icicle Plots: Better Displays for Hierarchical Clustering [online]. *The American Statistician*. 37(2): 162–168. <https://doi.org/10.1080/00031305.1983.10482733>.

MA, S., HELLERSTEIN, J. L. (1999). Ordering Categorical Data to Improve Visualization. *IEEE Symposium on Information Visualization*, 15–18.

POLZLBAUER, G. et al. (2006). Advanced visualization of self-organizing maps with vector fields. *Neural Networks*, 19(6–7): 911–922.

R CORE TEAM (2021). *R: a Language and Environment for Statistical Computing* [online]. Vienna, Austria. <http://www.R-project.org>.

RAHLF, T. (2019). *Data Visualisation with R: 111 Examples* [online]. 2nd Ed. Cham: Springer. <https://doi.org/10.1007/978-3-030-28444-2>.

ROSARIO, G. E. et al. (2004). Mapping Nominal Values to Numbers for Effective Visualization [online]. *Information Visualization*. 3(2): 80–95. <https://doi.org/10.1057/palgrave.ivs.9500072>.

SIBSON, R. (1973). SLINK: an Optimally Efficient Algorithm for the Single Link Cluster Method [online]. *The Computer Journal*. 16(1): 30–34. <https://doi.org/10.1093/comjnl/16.1.30>.

SNOW, G. (2020). *TeachingDemos: Demonstrations for Teaching and Learning* [online]. R package Version 2.12. <https://CRAN.R-project.org/package=TeachingDemos>.

ŠULC, Z. et al. (2021). *Nomclust: an R package for hierarchical clustering of objects characterized by nominal variables*. Version 2.5.0.

VESANTO, J. (1999). SOM-based Data Visualization Methods [online]. *Intelligent Data Analysis*, 3(2):111–126. <https://doi.org/10.1016/S1088-467X(99)00013-X>.

WITTEN, I. H. et al. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Ed. Burlington: Morgan Kaufmann publications.