HEATMAP MATRIX: USING REORDERING, DISCRETIZATION AND FILTERING RESOURCES TO ASSIST MULTIDIMENSIONAL DATA ANALYSIS

Miguel Mechi Naves Rocha and Celmar Guimarães da Silva
School of Technology, University of Campinas
Limeira, São Paulo, Brazil

ABSTRACT
Visualization can help users to analyze multidimensional datasets and make better decisions related to them. The current literature presents a set of multidimensional data visualization techniques that deal with discrete variables, continuous ones, or both. Recently, the heatmap matrix was proposed as an alternative for visualizing multidimensional datasets with discrete variables. It comprises a matrix of heatmaps, in which each heatmap shows the frequency of the values of two variables in the records of a dataset. Despite its proposed utility, it only presents a static representation of the dataset, which may limit users understanding about the dataset under analysis. Therefore, we propose to add interactive resources to this visualization, aiming to provide distinct views of a dataset. To guide the choice of new features, we compared the resources of a set of multidimensional visualizations and selected the characteristics that are suitable to enhance the heatmap matrix. We added to it resources for: (a) reordering values (rows and columns) of variables in its heatmaps; (b) filtering variables and values based on correlations of variables and on association rules; (c) discretizing continuous variables to be represented in the visual structure; among others. We present two case studies in which we show the potential of the heatmap matrix, enhanced by the resources proposed in this work, to assist the analysis of multidimensional data.

KEYWORDS
Multidimensional Data Visualization, Matrix Visualization, Reorderable Matrix, Data Mining, Association Rules, Discretization.

1. INTRODUCTION
The current visualization literature provides many techniques to support user analysis of multidimensional datasets. A classical example of these techniques is the scatterplot matrix (a.k.a. SPLOM) (Carr et al., 1987). It is predominantly applicable to datasets with continuous variables, given that scatterplots are not particularly suitable to depict discrete data, as superposition of graphical marks are likely to occur. A more recent example for scenarios with discrete data is the heatmap matrix (Rocha and Silva, 2018). It is similar to a SPLOM, but it is based on reorderable heatmaps instead of scatterplots. In other words, given a dataset \( D \), with \( n \) discrete variables, a heatmap matrix is a matrix of \( n \times n \) heatmaps, where a heatmap \( H_{ij} \) maps discrete variables \( V_i \) and \( V_j \) into its axes \( X \) and \( Y \). The value of a cell \( h_{p,q}^{ij} \) from \( H_{ij} \) indicates the number of records in which \( V_i = p \) and \( V_j = q \) in \( D \). Therefore, a heatmap matrix can be considered as a matrix with bidimensional histograms of all combinations of pairs of variables in a dataset. Figure 1 is an example of a heatmap matrix visualizing a dataset with six variables in a \( 6 \times 6 \) matrix of heatmaps (details about this dataset will be discussed at Section 4).

A restriction that may hamper the exploratory data analysis in the heatmap matrix is the absence of interactive resources that could help users to tailor the visualization to their ongoing analysis. Therefore, in this work, we aimed to select interactive features that could surpass this limitation and enhance this visualization technique. We surveyed a set of features from related visualization techniques (e.g. GPLOM, generalized pairs plot, and mosaic plot) and compared them with the heatmap matrix. Based on this comparison, we proposed to add to the heatmap matrix a set of interactive resources related to matrix reordering, as well as a support to continuous data through discretization procedures. We also included filtering capabilities based on association
rules and variable correlation. The results obtained in two case studies provided positive feedback regarding its potential to aid the analysis of multidimensional datasets.

The remainder of this paper is organized as follows. In Section 2, we compare a set of multidimensional visualizations (including the heatmap matrix), and list features that are absent in the heatmap matrix. Section 3 presents the features we selected to insert into the technique and how we proposed this insertion. We present two case studies in Section 4, in which we discuss the potential of the enhanced heatmap matrix as a tool for multidimensional dataset analysis. Section 5 concludes our paper and points out future work.

![Figure 1. Heatmap matrix for the “Sleep study” dataset. This and other visualizations in this paper uses the red-white-blue diverging color scale, where red, white, and blue stand for minimal, intermediate, and maximal values, respectively.](image)

## 2. RELATED WORK

In our literature review, we selected a set of multidimensional visualizations that support discrete variables. Our aim was to better understand the similarities among them and the heatmap matrix, and to list possible features of these visualizations that could be inserted into the heatmap matrix to improve it.

Aiming to represent a given dataset $D$ with $n$ variables ($V_1, ..., V_n$), the mosaic plot (Hartigan and Kleiner, 1981) subdivides a rectangular space for each discrete value $v_{ij}^1, j \in [1, ..., \text{numvalues}_i]$ of a variable $V_i, i \in [1, ..., n]$; $\text{numvalues}_i$ is the number of distinct values of $V_i$. The proportions of the areas $a_i^1, a_i^2, ..., a_i^{\text{numvalues}_i}$ follow the respective proportions of the frequency of the values $v_{ij}^1, v_{ij}^2, ..., v_{ij}^{\text{numvalues}_i}$ in $D$. This subdivision starts with $V_1$ and is done recursively for all other variables $V_2, ..., V_n$. The mosaic matrix (Friendly, 1999) is a matrix of mosaic plots of each possible pair of variables $(V_i, V_j), i \neq j$, from $D$; other properties from $D$ may be represented by rectangle colors and dashed or solid rectangle outlines.

The generalized pairs plot (Emerson et al., 2012) is an extension of a SPLOM for dealing with quantitative and categorical variables. As its ancestor, it defines a $n \times n$ matrix of visualizations. A set of visualizations is available according to the classes of the two variables of each cell. For example, a cell related to two quantitative variables (QQ) may have a scatterplot matrix; when the cell is related to two categorical variables (CC), a mosaic plot can represent them; and box plots can be used to visualize data from a quantitative variable and a categorical one (QC). The generalized plot matrix (Im et al., 2013) can be considered as a variation of a generalized pairs plot. It uses scatterplots, heatmaps and bar charts to the QQ, CC and QC situations, respectively.
The *table lens* (Rao and Card, 1994) uses a matrix-like display, with \( n \) columns. Each record is represented by a matrix row. Each column has a starfield display (if the variable it represents is ordinal or categorical) or a bar chart (if the column represents a quantitative variable). Sorting variables in ascending or descending order is an interactive resource to reorder rows and to reveal possible correlations among variables.

The *dimensional stacking* (Grinstein and Trutschl, 2001; Ward et al., 2015) recursively subdivides the display in matrices. The X and Y axes of each matrix represent a pair of variables of \( D \). In a recursive subdivision of the space, each cell of this matrix has another matrix with another pair of variables of \( D \). In the last level, the matrices are heatmaps that reveal frequencies of tuples of \( D \) or the value of one specific variable.

The *parallel sets* technique (Kosara et al., 2006) does not have a tabular display. It is similar to parallel coordinates plots, but has a focus on discrete variables. Each categorical variable \( V_i \) is represented as an axis. Each value \( v_{ij} \) of \( V_i \) is represented as a rectangle, whose size is proportional to the frequency of the value \( v_{ij} \) in \( D \). “Ribbons” are used to connect parallel axes instead of lines, aiming to represent the frequency of pairs of values. Users can reorder the axes as well as the values of each axis.

*Pixel-oriented techniques* (Keim, 1996) dedicate one part (window) of the display for each variable. In a window \( i \), given the vector of values from \( V_i \) in \( D \), these techniques represent each value of this vector as a colored pixel in this window. The layout of the pixels may follow space-filling curves. When the dataset is sorted according to a given variable, correlations among the variables may be revealed through the similarity among some windows of the visualization.

The heatmap at Table 1 summarizes the features we observed on these techniques. Features #1 to #4 deal with the exhibition of relationships and associations among two or more variables. Features #5 and #6 indicate the visualizations that show values of the dataset, frequencies of values, or both. Features #7 and #8 consider interaction aspects such as filtering, brushing and reordering. Features #9 to #12 point out visual mapping aspects related to which graphical property is used to show values and frequencies. Lastly, feature #13 is related to the supported types of variables (categorical, ordinal or quantitative).

Table 1. A comparison heatmap of multidimensional visualization techniques according to a set of features. A blue cell \((x,y)\) indicates that the feature \( y \) is present in the technique \( x \); otherwise, the cell is red.
This heatmap also compares the analyzed techniques and the features they have. Three groups of techniques can be observed in this figure. Group 1 comprises mosaic plot ("M.Plot"), mosaic matrix ("M.Matrix"), parallel sets ("P. Sets"), and heatmap matrix ("H.Matrix"). These techniques cannot represent directly the values of the records (feature #5). However, they show frequencies related to values of two variables (features #2 and #6).

The techniques from Group 2 – table lens ("T.Lens") and pixel-oriented techniques ("Pixel") – provide support to show discrete or continuous record values (features #5 and #13), and use color or size to represent them (features #9 and #11). However, they do not support the presentation of frequencies (feature #6).

Group 3 has three techniques: GPLOM, generalized pairs plot ("G.P.Plot"), and dimensional stacking ("D.Stacking"). They enable the users to see the record values, and also the frequencies of values (features #5 and #6). Besides GPLOM and generalized pairs plot have almost all the features we analyzed.

It is worth noting that most of the analyzed techniques enable filtering, brushing or reordering (features #7 and #8). Besides, the heatmap matrix is the only technique among the analyzed ones that uses color to map frequency of values (feature #10), while the techniques that represent frequencies use sizes of rectangles (bars and boxes) (feature #11).

3. FEATURE SELECTION

After analyzing the visualizations presented at Section 2, we observed that the heatmap matrix does not have seven of the features listed at Table 1 (#3, #5, #7, #8, #9, #11, and #13). From these ones, we identified that features #7 (reordering), #8 (interactivity/filtering), and #13 (support to categorical, ordinal, or quantitative variables) could be added to the visualization, as we will present in this section. Features #5 and #9 are out of the scope of this technique, which is focused on showing frequency values instead of record values. We did not consider to add feature #11 (mapping format or size of the mark to some variable) because all variables are used in the columns and rows of the heatmap matrix. We also chose not to add feature #3 to avoid potential cognitive overload of the user. Therefore, we propose to add to the heatmap matrix the following resources:

a) Heatmap reordering. Matrix reordering algorithms search for row and column permutations of a matrix to easy the understanding of its data (Behrisch et al., 2016; Silva, 2021). We added to the heatmap matrix a resource for reordering the rows and columns of its heatmaps, in order to reveal patterns in the dataset. We proposed four ways to reorder these heatmaps – we called these ways “reordering focuses”. The proposed focuses are:

1. “Sort all heatmaps”: Individual reordering of each heatmap of the visual structure, aiming to reveal patterns regarding each pair of variables.

2. “Sort heatmap and spread”: Reordering a given heatmap $V_i \times V_j$. The new order of the values of $V_i$ (or $V_j$) is also applied to each heatmap that has $V_i$ (or $V_j$) in one of its axes. Users must select which variables will be considered as $V_i$ and $V_j$. The aim is to reveal patterns regarding a single pair of variables.

3. “Sort selected variables”: Reordering each variable $V_i$ selected by the user. To define the new order of the values of $V_i$, the reordering algorithm should consider a matrix composed by the concatenation of all heatmaps that have $V_i$ in the horizontal axis (except $V_i \times V_i$), as the example of Figure 2. The goal is to reveal patterns regarding the relationship of each selected variable and all other variables.

---

![Figure 2. Reordering with “Sort selected variables” focus applied to variable V1. (a) Heatmap matrix with variables V1, V2 and V3. (b) Matrix that concatenates V1xV2 and V1xV3. (c) Reordered version of the matrix (b); (d) Heatmap matrix reordered according to the new order of V1. Cells that were not drawn in (a) and (d) are not affected by this reordering](image-url)
4. “Sort all variables”. Similar to the previous item, but it considers all variables of the heatmap instead of a subset of them.

Before reordering the heatmap matrix, it is necessary to choose one of these focuses, a set of variables related to the selected focus (when applicable), and a matrix reordering method. It is also worth to consider if the variables to be reordered are nominal, so that they really can be reordered without semantic problems.

b) Filtering by correlation and by association rules. Aiming to focus the users on possibly relevant relationships between quantitative variables, we provided a filtering capability based on correlation between pairs of variables. To use this feature, the user must provide a filtering value \( \rho_{\text{threshold}} \) between 0 and 1. For each pair of quantitative variables \( V_i \) and \( V_j \) from a dataset, consider \( \rho_{i,j} \) as the Spearman correlation coefficient between \( V_i \) and \( V_j \). If \( |\rho_{i,j}| > \rho_{\text{threshold}} \), the heatmap remains visible; otherwise, it is replaced with a label with the value of \( \rho_{i,j} \).

Two other filtering resources are based on the analysis of which variables and variable values should be hidden, according to association rules provided by an algorithm. Given a set of association rules, the “Filter Variables” feature keeps in the heatmap matrix only the variables that are mentioned by these rules. The “Filter Labels” feature only shows the values (labels) that are cited by these rules, and their respective variables. Given that we opted by using the Apriori association rule mining algorithm (Agrawal and Srikant, 1994) in this proposal, users that want to use these filters must inform the minimal support (\( \text{minsup} \)) and the minimal confidence (\( \text{minconf} \)) values for it.

When one of these filters is applied, the visualization highlights the cells whose values are related to some rule. In our current implementation, we use a version of the Apriori algorithm that creates rules with only 1 consequent. Considering this, cells related to rules with the same consequent are colored with the same color.

c) Discretization. If a user wants to use a heatmap matrix to analyze a dataset that has quantitative variables, first they must be discretized. Among a large set of discretization methods (see Garcia et al. (2013) for a survey), we chose two unsupervised ones to add to the technique: equal width and equal frequency (Dougherty et al., 1995). Equal width splits the domain of a variable in a set of \( k \) intervals with the same size. On the other hand, equal frequency splits the domain of a variable in a set of \( k \) intervals with approximately the same number of elements. Therefore, the user must inform the desired number \( k \) of intervals (bins) and the discretization method to each variable to be discretized. These methods were selected due to their simplicity, as an initial approach to emphasize our point of view about the relevance of discretization for the heatmap matrix.

d) Global or local color mapping. Users can switch between global or local color mapping. When using local color mapping, the color scale is applied independently to each heatmap, considering its minimal and maximal values. On the other hand, the global color mapping defines that the color scale must consider the minimal and maximal values of the entire set of heatmaps of the visual structure. Therefore, the global color mapping can better reveal outliers in the entire heatmap matrix, while the local color mapping can help to reveal relationships between pairs of variables of a heatmap.

It is also worth to report that we reduced the unused space of the original version of the heatmap matrix technique by representing \( V_i \times V_j \) for every pair \( (V_i, V_j) \) of dataset variables. The version of Rocha and Silva (2018) depicted only \( V_i \times V_j, i > j \). As in the SPLOM, we noted that having a complete row (or column) with all the dataset variables compared to a single variable \( V_j \) is relevant when the user is focused on that specific variable. Therefore, placing \( V_i \) in the same axis in many heatmaps is useful for this kind of comparison.

4. RESULTS AND LIMITATIONS

We present two case studies in which we used a heatmap matrix to assist the interpretation of datasets. Both studies were supported by a Java-based tool that implements all the features presented in the previous section. We also present a brief discussion about limitations of this technique.

The first study comprises a dataset named “Sleep Study”\(^4\), related to a pilot study about sleeping habits of 104 individuals within USA. This dataset has the following six variables (and respective questions and values):

- **Enough**: Do you think that you get enough sleep? (Yes/No);
- **Hours**: On average, how many hours of sleep do you get on a weeknight? (Integer);
- **PhoneReach**: Do you sleep with your phone within arms’ reach? (Yes/No);

\(^4\)“Sleep Study” dataset available at [https://www.kaggle.com/mlomuscio/sleepstudypilot](https://www.kaggle.com/mlomuscio/sleepstudypilot) (February 9\(^{th}\), 2022).
• **PhoneTime**: Do you use your phone within 30 minutes of falling asleep? (Yes/No);
• **Tired**: On a scale from 1 to 5, how tired are you throughout the day? (1: not tired; 5: tired);
• **Breakfast**: Do you typically eat breakfast? (Yes/No).

Figure 1, at the start of the paper, presents a heatmap matrix about this dataset. It uses the global color mapping, with no heatmap reordering and no filtering. All the dataset variables have discrete data; therefore, no discretization was needed. When exploring heatmaps individually in this figure, it is possible to perceive some interesting data. For example, suppose that we want to understand what is the behavior trend of people that do not sleep enough. In the heatmap matrix presented in Figure 1, we may start the analysis with the heatmap **Enough × Enough**. It shows that 68 people (65%) said that they do not sleep enough. Among them, the most common behavior we observed was: sleeping 6 to 7 hours on average per night (**Enough × Hours**); sleeping with cell phone within arms’ reach (49 people, or 72%) (**Enough × PhoneReach**); using the phone for 30 minutes before sleeping (55 people, or 80%) (**Enough × PhoneTime**); having the tiredness level probably between 3 and 4 (46 people, or 67%) (**Enough × Tired**); and having breakfast (38 people, or 55%) (**Enough × Breakfast**).

We also can filter the visualization by association rules. In this dataset, if we configure the Apriori algorithm to run with 90% of confidence and 10% of support, this algorithm returns the following rules:

- **Rule 1**: 
  6 (Hours)  Yes (PhoneTime) 
  (support: 21.15%; confidence: 91.66%);
- **Rule 2**: 
  5 (Hours)  No (Enough) 
  (support: 11.53%; confidence: 100.0%); and
- **Rule 3**: 
  8 (Hours)  Yes (PhoneTime) 
  (support: 14.42%; confidence: 93.74%).

Based on these rules, we may apply the “Filter Variables” and the “Filter Labels” resources. In the resulting visualizations (Figure 3), note that the filter hid three variables that are not related to the association rules. These rules reinforce the idea that almost all people that sleep 6 or 8 hours per night use cell phone before sleep, and that all people that sleep only 5 hours believe that they did not sleep enough.

---

Figure 3. Filtering by association rules. Left: “Filter Variables” was applied; Right: “Filter Labels” was applied. Cells outlined in green are related to the Rule 2. Cells outlined in yellow are related to the Rules 1 and 3.

In a second study case, we analyzed a dataset with fictitious data about the performance of one thousand students regarding three subjects. This dataset has the following variables: Gender, Race/Ethnicity, Parental Level of Education (“parental educ. lvl.”), Lunch, Test Preparation Course (“test prep”), Math Score, Reading Score, and Writing Score. Given the quantitative nature of the three scores, we need to discretize them.

For our analysis of this dataset, we used a heatmap matrix with global color mapping, equal width discretization, and no filtering (Figure 4). The visualization was reordered by the FVS algorithm (Silva et al., 2017) with the “Sort and Spread” reordering focus, based on the heatmap “Test Preparation × Parental Level of Education”. In practice, it sorted the dataset according to the number of students related to each parental educational level. In this visualization, we observed a better performance of men in math scores, and of women in reading and writing scores (see the heatmaps “Gender × Math Score”, “Gender × Reading Score”, and “Gender × Writing Score”). We also noted that the students’ scores are, in average, close to the interval [60,70],

---

independently of the parental level of education (see the heatmaps “Parental Level of Education × Math Score”, “Parental Level of Education × Reading Score”, and “Parental Level of Education × Writing Score”). Besides, the heatmaps that compare the Lunch variable and the variables related to scores reveal that having a standard lunch (instead of a reduced one, or no lunch) potentially helped the students in the three subjects. Also, the preparatory test seemed to be worthwhile due to the trend of higher scores for the group of students who took it (see the heatmaps related to the “Test Preparation” variable).

Regarding the limitations of this technique, it is worth considering that the use of colors to convey information about frequency of records provides a limited level of precision in the interpretation of the visualization. We tried to minimize this problem using cell labels that may be turned on or off according to users’ needs, and zooming resources. Scalability issues may arise as the number of discrete values per variable or the number of variables increase. We provided filtering capabilities to minimize these potential problems.

**Figure 4.** Heatmap matrix for the “Students Performance in Exam” dataset. Global color mapping, equal width discretization, no filtering. Matrix reordered with the “Sort and Spread” focus applied to the highlighted heatmap

### 5. CONCLUSION

Having a variety of interactive multidimensional data visualization techniques is essential to help users to interpret datasets. In this paper, we added to the heatmap matrix technique a set of interactive features to aid the analysis of multidimensional datasets. The users may use this visualization to analyze datasets with discrete
or continuous variables, with the support of discretization methods. Filters based on correlation of quantitative variables or on the elements of association rules can be used to hide unnecessary data and to focus users on relevant values and patterns. The users may select global color mapping to make a better comparison between heatmaps, or local color mapping to focus on differences inside each heatmap. Matrix reordering capabilities complement this set of features, assisting users to perceive patterns and trends.

Future works related to this technique include: enabling the users to inform the domain and type of each variable of the input dataset; enabling manual and automatic reordering of the variables presented in the visual structure (instead of only providing reordering of values within each variable); and testing the implementation of this technique with users to evaluate how we can better enhance this technique.

ACKNOWLEDGEMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We also thank the grant #2126/22 from FAEPEX/PRP/Unicamp.

REFERENCES


