Heatmap matrix: a multidimensional data visualization technique

Miguel Mechi Naves Rocha
School of Technology
University of Campinas
Limeira, SP 13484-332
Email: m156778@dac.unicamp.br

Celmar Guimarães da Silva
School of Technology
University of Campinas
Limeira, SP 13484-332
Email: celmar@ft.unicamp.br

***Abstract*—Information Visualization literature presents lots of multidimensional data visualization techniques. We propose a new technique, called *heatmap matrix*, which aims to show relationships between pairs of categorical or ordinal variables. A heatmap matrix is composed by a matrix of heatmaps that represent frequency of occurrence of values from all possible pairs of variables of a dataset. In this paper we define this technique and illustrate how it may be used for analyzing some scenarios.**

I. Introduction

Information Visualization (InfoVis) research area has been developing lots of interactive visual representations that aim to help users to understand data [1] [2] [3], even when they are part of datasets with high dimensionality and cardinality.

We highlight the role of InfoVis when the dataset to be analyzed is multidimensional. Multidimensional visualization techniques receive increasing interest and visibility from InfoVis researchers, given that they make available fast and simple ways for visual recognition of patterns and trends, and for observation of dataset's non-obvious characteristics [4].

In our research we are particularly interested on techniques for visualizing multidimensional categorical or ordinal data. We propose *heatmap matrix* as a visualization technique for this kind of datasets. It combines multiple juxtaposed displays of pairs of categorical or ordinal variables, similar to those used by scatterplot matrices [2], but with heatmaps instead of scatterplots. In this paper, we define how this technique works and exemplify its use in some real world scenarios. An extended version of this work is published elsewhere [5].

This paper is organized as follows: Section II presents some related works on multidimensional visualizations; Section III explains our proposed technique; Section IV presents the tool that implements this technique and its use to analyze real-world examples; and Section V concludes our paper and points out future works.

II. Related works

Our technique has some similarities with other classic multidimensional visualization techniques. We highlight dimensional stacking, scatterplot matrix and parallel sets.

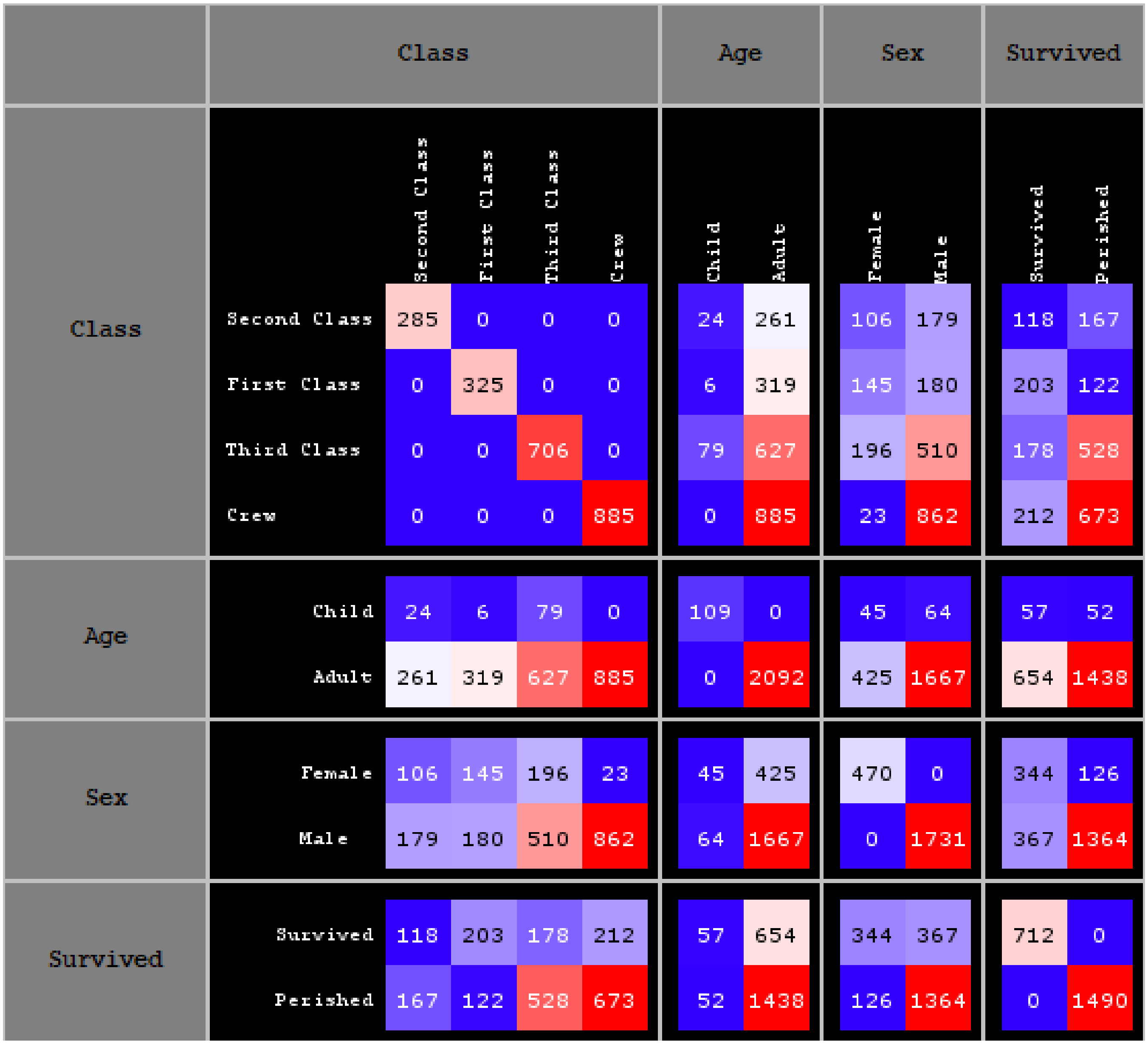
Dimensional stacking [6] is a 2D or 3D projection of a dataset, with embedded groups of dimensions. We focus on 2D dimensional stacking, which splits the plane in rectangles. Given I_1, I_2, I_3 and I_4 as independent variables, and D as a dependent one, the external X and Y axes are mapped to I_1 and I_2, respectively. Each cell of this main graphic has an internal graphic that maps its X and Y axes to I_3 and I_4, respectively. The cells of the internal graphics are painted according to D. When a dependent variable does not exist, a default color is used. Scenarios with more than four independent variables should use other groups of internal graphics inside the $I_3 \times I_4$ graphics. Similarly to dimensional stacking, our technique uses the concept of matrices inside a matrix, but with a different visual mapping, as presented in the following section.

A scatterplot matrix (or SPLOM) [7] shows simultaneously multiple views of a dataset. This matrix represents a dataset with n variables in a $n \times n$ table. Each cell of this table has a scatterplot whose axes map two variables of the dataset. The use of brushing technique enables users to select a set of points in one scatterplot, which highlights related points in the other views. Our technique also uses a $n \times n$ table. However, SPLOM is focused on quantitative variables, whereas our technique focuses on categorical and ordinal variables.

Parallel sets technique [8] is inspired by parallel coordinates plot (PCP) [9]. PCP represents each tuple of a dataset as a polygonal line. On the other hand, parallel sets technique represents frequency of occurrence of values in tuples. It fits a multidimensional scenario with categorical variables. Due to value aggregation, it is able to process datasets with high cardinality without occlusion problems that are recurrent in PCP. Parallel sets technique provides one axis for each categorical variable, and parallelograms represent the relationship between the categories of each pair of neighbor axes, given by frequency values. This technique is similar to our own because it presents frequency of occurrence of values. However, the presented frequencies are related only to pairs of variables that are neighbor in the visual mapping. On the other hand, our technique presents frequencies related to all possible pairs of variables, as depicted in the following sections.

III. Heatmap matrix technique

A heatmap matrix is a visual structure that represents a multidimensional dataset. It is inspired on scatterplot matrix and heatmap techniques.



		Class				Age		Sex		Survived	
		Second Class	First Class	Third Class	Crew	Child	Adult	Female	Male	Survived	Perished
Class	Second Class	285	0	0	0	24	261	106	179	118	167
	First Class	0	325	0	0	6	319	145	180	203	122
	Third Class	0	0	706	0	79	627	196	510	178	528
	Crew	0	0	0	885	0	885	23	862	212	673
Age	Child	24	6	79	0	109	0	45	64	57	52
	Adult	261	319	627	885	0	2092	425	1667	654	1438
Sex	Female	106	145	196	23	45	425	470	0	344	126
	Male	179	180	510	862	64	1667	0	1731	367	1364
Survived	Survived	118	203	178	212	57	654	344	367	712	0
	Perished	167	122	528	673	52	1438	126	1364	0	1490

Fig. 1. Heatmap matrix technique representing the Titanic dataset.

This structure's visual mapping is as follows. A heatmap matrix that represents a n-variable dataset has $n \times n$ cells, which are heatmaps. The column labels of each heatmap contain distinct values of a variable V_i. Similarly, the row labels of each heatmap have distinct values of a variable V_j. Each cell belonging to a heatmap represents the frequency of occurrence of the pair of values (v_i, v_j) in the dataset, where $v_i \in V_i$ and $v_j \in V_j$. Colors or numerical values represent the frequency values.

IV. Results

Fig. 1[1] exemplifies our technique with the Titanic dataset[2]. This dataset reffers to the tragic accident of the Titanic passenger ship in April 14, 1912. The dataset has 2202 tuples, and its variables are: passenger's age (child or adult), sex (male or female), class (first-class, second-class, third-class, or crew) and if the passenger survived or not. A red-white-blue palette indicates high, mean and low frequency values. For example, the cell in the row Sex and column Survived at the external

[1] These figures correspond to the most up-to-date versions of the heatmap matrix. Rocha elaborated them after completing his undergraduate course, taking into account modifications suggested by the reviewers of this paper.

[2] Available at https://www.jasondavies.com/parallel-sets/titanic.csv (August 23, 2018)

matrix indicates that 344 women survived and 126 perished in the accident; also, 367 men survived and 1364 died. In the same example, the heatmap in the cell "Class × Age" indicates that there are more adults than children in all classes. It also shows that most adults were in third class and crew. Similarly, the cell "Class × Sex" reveals that there were more men than women in all classes, specially third class and crew. Besides, the cell "Class × Survived" shows that more people died in the third class and crew than in the first and second classes. Given these analyses, a user may conclude that most of the died male adults were from third class and crew.

As a second example, Fig. 3 shows a dataset[3] that includes 638,455 tuples of murders from the FBI's Supplementary Homicide Report from 1976 to the present, and from the Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. This dataset has 24 variables, such as age, race, sex, ethnicity of victims and perpetrators, the relationship between the victim and perpetrator, and the weapon used. For this example we used only four of those variables: "Crime Type", "Crime Solved", "Victim Sex", and "Perpetrator Sex".

There are at least two ways of interpreting this heatmap matrix: a user may analyze a single heatmap, or a row (or column) of the heatmap matrix. In both cases, the user may try to find patterns. As an example, a user that analyzes the row "Crime Type" may note that the value "Murder or Manslaughter" has a high-frequency value associated to it in all the heatmaps. Therefore he may hypothesize that "Murder or Manslaughter" was the crime type that was most solved (value "Yes"), whose most victims were women ("Female") and whose authors were men ("Male").

A. *Heatmap Matrix Tool*

We developed a tool that implements our technique (Figs. 2 and 3). It works as follows. First, the user must indicate a CSV file in the data path section (top-left corner of the tool, see Fig. 2b) and press button "Generate heatmap matrix". The tool draws a heatmap matrix of the input dataset in the center of the screen. Another option is the "Original Data" tab, which shows the loaded dataset as it is.

B. *Limitations*

The proposed technique has some limitations. A first limitation is the number of variables to be plotted in the same screen (SPLOM has a similar limitation). A user that wants to visualize a dataset with high dimensionality in this technique must choose to show only a subset of the dataset variables. Besides, the total number of distinct values of all dataset variables impacts the size of the visualization or its resolution. Both limitations may hamper user understanding.

Two other limitations are the absence of a color legend to improve readability, and the absence of support to quantitative variables.

V. Conclusion

We presented a new multidimensional visualization technique called heatmap matrix. It aims to help users to understand datasets with categorical or ordinal variables. It is focused on representing frequency of occurrence of values, grouped by pairs of variables. We exemplified two datasets represented by this technique, and some information that a user could derive from these visualizations.

Future works should include user interaction capabilities in the technique, such as resources for reordering the heatmaps and the whole matrix.

Acknowledgment

This work was supported by São Paulo Research Foundation (FAPESP) (grant number #2015/14854-7).

References

[1] S. K. Card, J. D. Mackinlay, and B. Shneiderman, Eds., *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.

[2] M. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, MA, USA: A. K. Peters, Ltd., 2010.

[3] R. Mazza, *Introduction to Information Visualization*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[4] R. Pillat, "Dynamic coordination of multidimensional data visualizations," Ph.D. dissertation, Universidade Federal do Rio Grande do Sul, 2006.

[5] M. M. N. Rocha, "Multidimensional data visualization with heatmap matrix," Final graduation work. School of Technology, University of Campinas., 2017. [Online]. Available: https://sistemas2.ft.unicamp.br/tcc/upload/monografia/8853Monografia.pdf

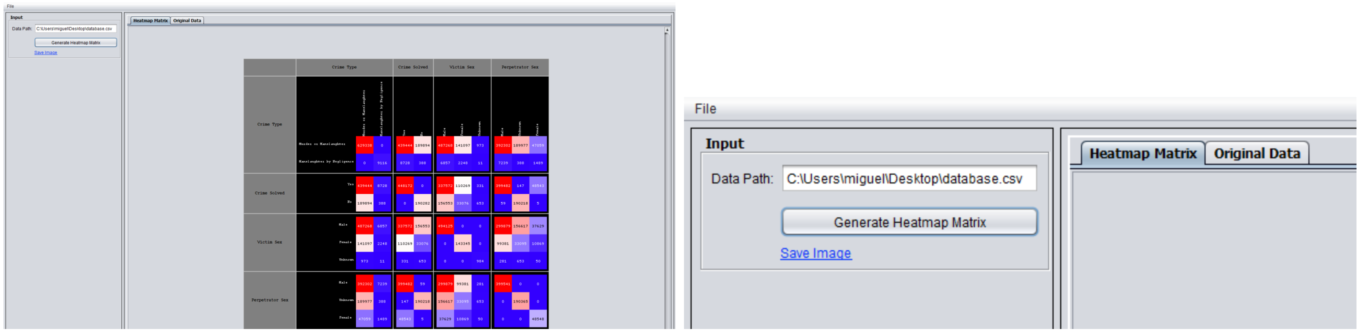
[6] J. LeBlanc, M. O. Ward, and N. Wittels, "Exploring n-dimensional databases," in *Proceedings of the 1st Conference on Visualization '90*, ser. VIS '90. Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, pp. 230–237. [Online]. Available: http://dl.acm.org/citation.cfm?id=949531.949568

[7] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield, "Scatterplot matrix techniques for large n," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 424–436, 1987.

[8] F. Bendix, R. Kosara, and H. Hauser, "Parallel sets: visual analysis of categorical data," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 2005, pp. 133–140.

[9] A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multi-dimensional geometry," in *Proceedings of the 1st Conference on Visualization '90*, ser. VIS '90. Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, pp. 361–378.

[3] Available at https://www.kaggle.com/murderaccountability/homicide-reports (June 15, 2017)

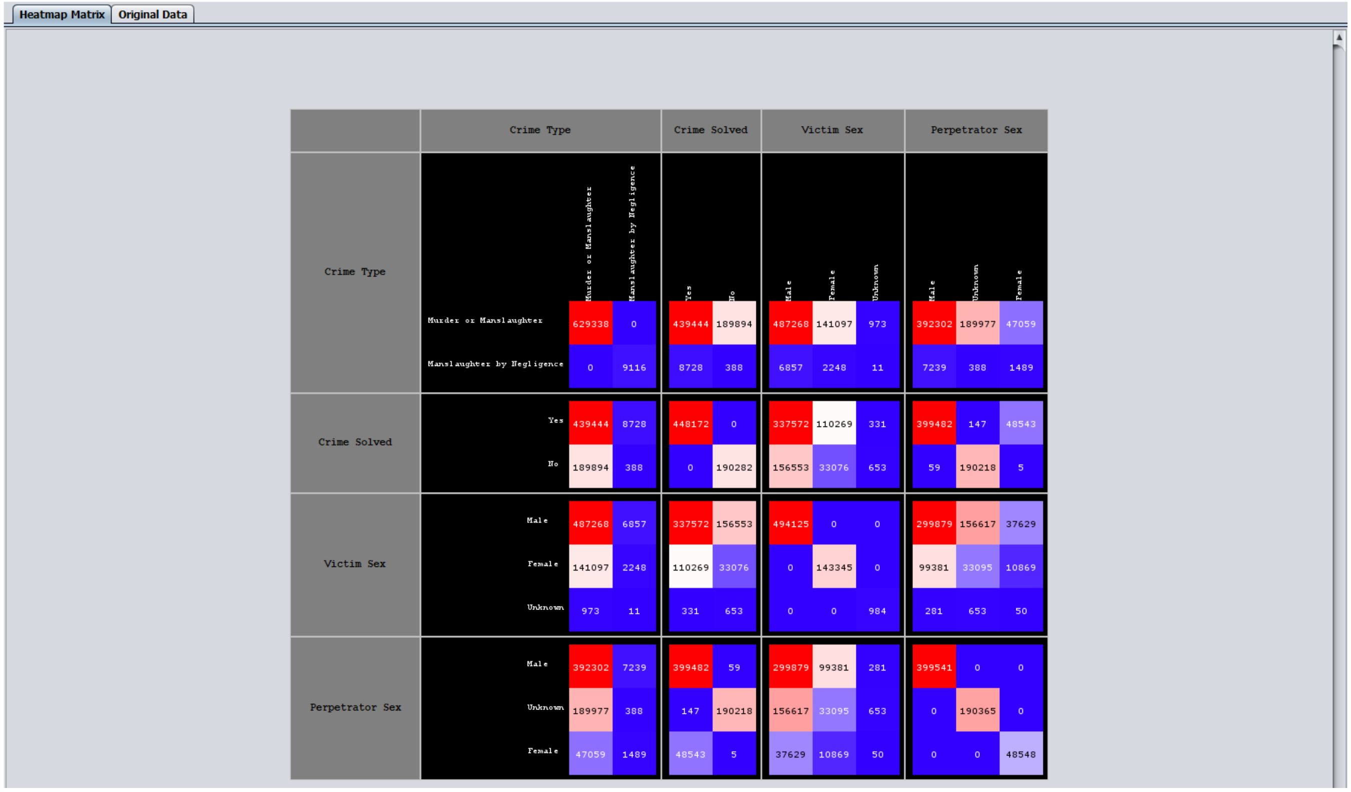


Data Path: C:\Users\miguel\Desktop\database.csv

(a) Overview.

(b) Detailed view of the top-left corner of the tool.

Fig. 2. The Heatmap Matrix Tool.



		Crime Type		Crime Solved		Victim Sex			Perpetrator Sex		
		Murder or Manslaughter	Manslaughter by Negligence	Yes	No	Male	Female	Unknown	Male	Unknown	Female
Crime Type	Murder or Manslaughter	629338	0	439444	189894	487268	141097	973	392302	189977	47059
	Manslaughter by Negligence	0	9116	8728	388	6857	2248	11	7239	388	1489
Crime Solved	Yes	439444	8728	448172	0	337572	110269	331	399482	147	48543
	No	189894	388	0	190282	156553	33076	653	59	190218	5
Victim Sex	Male	487268	6857	337572	156553	494125	0	0	299879	156617	37629
	Female	141097	2248	110269	33076	0	143345	0	99381	33095	10869
	Unknown	973	11	331	653	0	0	984	281	653	50
Perpetrator Sex	Male	392302	7239	399482	59	299879	99381	281	399541	0	0
	Unknown	189977	388	147	190218	156617	33095	653	0	190365	0
	Female	47059	1489	48543	5	37629	10869	50	0	0	48548

Fig. 3. Heatmap Matrix Tool visualization of the homicide dataset.