# MULTISET: An Interactive Visual Analysis Tool for Multidimensional Quantitative Data

Liqun Liu and Romain Vuillemot *


## ABSTRACT

Multidimensional quantitative data usually appear in real world, such as the traffic density among different time stamps and students' grades among different subjects. By improving UpSet, we introduce MULTISET, a novel visualization technique for analyzing the distribution of multidimensional quantitative data, which first creates a scale for mapping all the quantitative values into categories and then aggregate the elements into the same groups if they have similar distribution on the scale. Thus, it is able to help users analyze the relations of multidimensional quantitative data. In the future, we will also improve the prototype to support users having a good experience and conduct a user study to evaluate the performance of MULTISET.


## 1 CONTEXT

In many domains, visual analysis of multidimensional data is important to help domain experts understand data. For example, the observation of traffic density over time is a key factor to understand the situation of road segments, as shown in Fig 1. In this dataset, the road segments include multidimensional quantitative values, categorical attributes, and numerical attributes. Therefore, analysis of traffic densities over time is meaningful. For example, **How does the traffic density of a road segment changes over a day?** Let's consider the scenario for evaluating the traffic status of roads. A road might be fluent at a certain time (e. g., morning) but the traffic congestion happens in the evening. If the traffic density of roads is diverse widely from time to time, we can not describe the roads with simple words such as smoothly flowing, heavy or traffic jams.

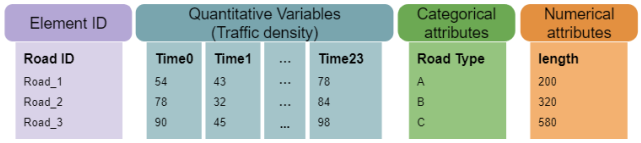


Figure 1: The structure of road segments data.

In this situation, the evaluation for the road situation has clutter since a road might be smooth at 4:00 AM but be heavy at 8:00 AM. In order to understand how roads' traffic density changes over time, we have to observe both the traffic density at both peak time and normal time. Thus, it is necessary to propose a tool for analyzing the distribution of traffic density changing over time.

Statistical visualizations such as histograms [5] or box plots [4], enable analysts to better grasps the frequency distribution of dimensions of the dataset and proceed with modeling and hypothesis tasks. Statistical charts can even be combined together so that all dimensions frequencies are visually displayed at once. Interactions may be added to further support data selection and aggregation [1]. There has been a significant effort to enhance them to support uncertainty and multidimensional data analysis. Moreover, *Mosaic Display* [2] represents the frequency distribution with the vertical and horizontal longs of rectangles. However, those visual techniques have limitations in the scalability of the number of elements and attributes of elements.


*Liqun Liu was with Ecole Centrale de Lyon e-mail: liqun.liu@ec-lyon.fr.

†Romain Vuillemot was with Ecole Centrale de Lyon e-mail: romain.vuillemot@ec-lyon.fr.


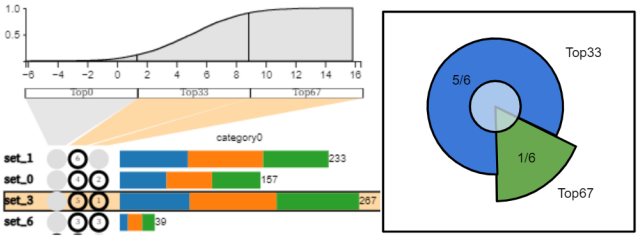


Figure 2: Theory of MULTISET the aggregate process is on the left and an illustration of fuzzy sets is displayed on the right. The elements in selected set (set_3) belonging to Top33 and Top67 (generated from cumulative membership function) simultaneously. There are 5 dimensions being scaled as Top33, 1 attribute being scaled as Top67.

Thus, We introduce MULTISET to analyze the multidimensional quantitative data based on UpSet matrix [3], a matrix-based layout to analyze the intersection of multidimensional categorical data. The novelty is the set-creation part which is inspired by an interactive cumulative distribution function and MULTISET improved the matrix visualization, which not only presents intersections but combines circles and numbers to present how many dimensions that an element has in each category. MULTISET is consists of five separated but linked views: 1) cumulative distribution function view, 2) combination matrix, 3) categorical attributes view, 4) numerical attribute view, and 5) element view. It is illustrated in Fig. 4.

## 2 MULTISET TECHNIQUE

The tool we introduced works as follows. First, the continuous quantitative values are separated into different categories. And then, based on the generated categories, all the elements are aggregated in the same groups if they have the same distribution, e. g., road segments would be aggregated if they have 3 traffic densities mapped into *Low* category and 2 traffic densities mapped into *High* category. Corresponding to aggregated groups, the categorical attributes and the numerical attributes are visualized with stacked bar charts and box plots.

### 2.1 Concept

In possibility theory, a random variable $X$ is taken from the cumulative distribution function. Besides, there is a value $x$ is calculated with a possibility $P$ by the cumulative distribution function. In this case, it means the random variable $X$ has the $P$ possibility being less than or equal to $x$, the equation is given by:

$$F_X(x) = P(X \leq x) \tag{1}$$

In the cumulative distribution function, it is easy to understand the relative positions of a variable in entire variables. For example,

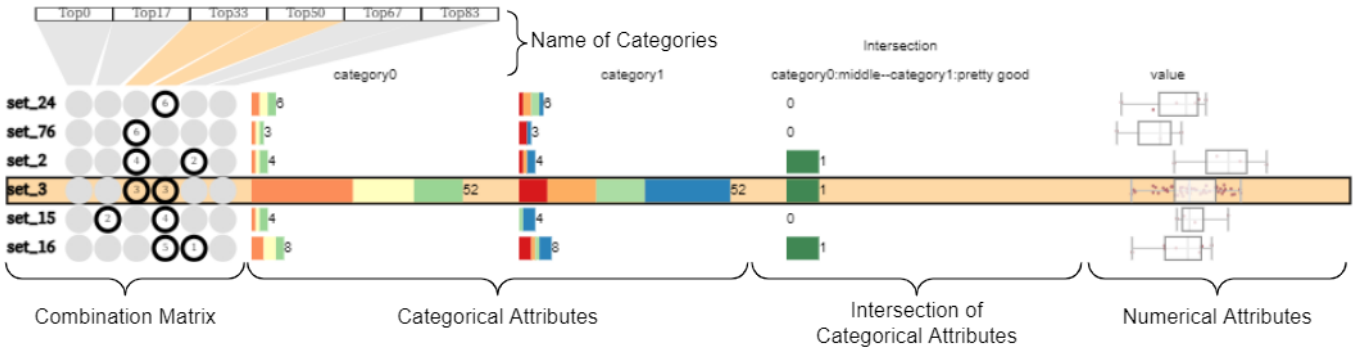


Figure 3: Set-based view: categories from cumulative distribution function is presented as columns. Rows are aggregated sets that have similar distribution of multidimensional quantitative data. The categorical attributes and numerical attributes are presented with stacked bar chart and box plots corresponding to specific sets.

we build a cumulative distribution function for the ages of a people group. If $X$ is a random age, we let it be equal to 30 and the possibility is calculated as 0.5. Thus, in this case, we can say there are 50% of people in this group who are less than 30 years old.

Based on the categories generated from the cumulative distribution function, the values of multidimensional data should be scaled into different categories and the elements should be aggregated if they have the same distributions in different categories. As illustrated in Fig. 2, ⓝ represent the elements in a certain set having $n$ dimensions scaled in the corresponding category. In contrast, ○ means that no dimension of the elements is scaled in the corresponding category. **set_3** is consists of three categories, which means the cumulative distribution function categorizes multidimensional quantitative data into three groups and the count of dimensions in each category is represented with the number in circles. It is able to be explained with the figure on the right. The central circle represents all element in **set_3**. The right figure shows there is no dimension belonging to *Top0*, 5 dimensions belonging to *Top33* represented with the blue sector, and only 1 dimension belonging to *Top67* represented with the green sector.

### 2.2 Split Quantitative Values

We design an interactive mapping function based on the theory of cumulative distribution function, as shown in Fig 4 ①. Through this function, the multidimensional quantitative data are separated into different categories and these categories are named in a specific way that helps users understand the relative positions of the values in the entire values as shown at the bottom of Fig 4 ①.

### 2.3 Set-based View

The set-based view can both address the set-related tasks by the aggregation of elements with similar distribution such as finding membership degrees and members of a specific set. Besides, the set-based view is also able to address the attribute-related tasks by analyzing the categorical attributes and numerical attributes distribution.

#### Combination Matrix

In the combination matrix, as illustrated in Fig 3, all these values are scaled into categories generated from the cumulative distribution function and then we count how many dimensions an element has and how these values distribute in categories. The account of dimensions distributing in a specific category is presented with a combination of circles and a number (ⓝ). If there is no one dimension distributed in a certain category, it is presented with ○.

#### Attributes of Sets

In this section, the attributes of sets are divided into two aspects, one is the categorical attribute whose values do not have inherent order and another one is the numerical attributes whose values are continuous. The categorical attribute is presented with stacked bars, colors of which show the different categorical values in this attribute. As illustrated in Fig. 3, there are two categorical attributes corresponding to sets from the combination matrix. These categorical attributes are visualized with a stacked bar chart, apart from this, there is still an intersection view following with categorical attribute view, which represents the frequency of elements that have specific values in both previous categories. For example, in Fig. 3, the intersection view visualized the cardinality with value of *category0* equal to **middle** and value of *category1* equal to **pretty good**. The numerical attributes are presented with a combination of box plots and scatterplots. The box plot shows the general distribution for values of numerical attributes and the scatterplot presents the individual points for the corresponding sets.

## 3 Perspective

In the future, our works will be focused on evaluating the prototype by conducting a user study to collect feedback from experts in different domains. Besides, we will also improve the prototype by adding more features such as ordering method, etc. Based on the pilot study, it shows that the circle combining number is not easy to be understood. Therefore, we will improve the matrix layout by rendering the circles with different sizes and colors. Finally, we will add a parallel view to present the distribution of multidimensional quantitative data.

## References


[1] N. Elmqvist, P. Dragicevic, and J. D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1141–1148, 2008. doi: 10.1109/TVCG.2008.153

[2] M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

[3] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, dec 2014. doi: 10.1109/TVCG.2014.2346248

[4] D. F. Williamson, R. A. Parker, and J. S. Kendrick. The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921, 1989.

[5] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton. Parallel bargrams for consumer-based information exploration and choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pp. 51–60, 2001.

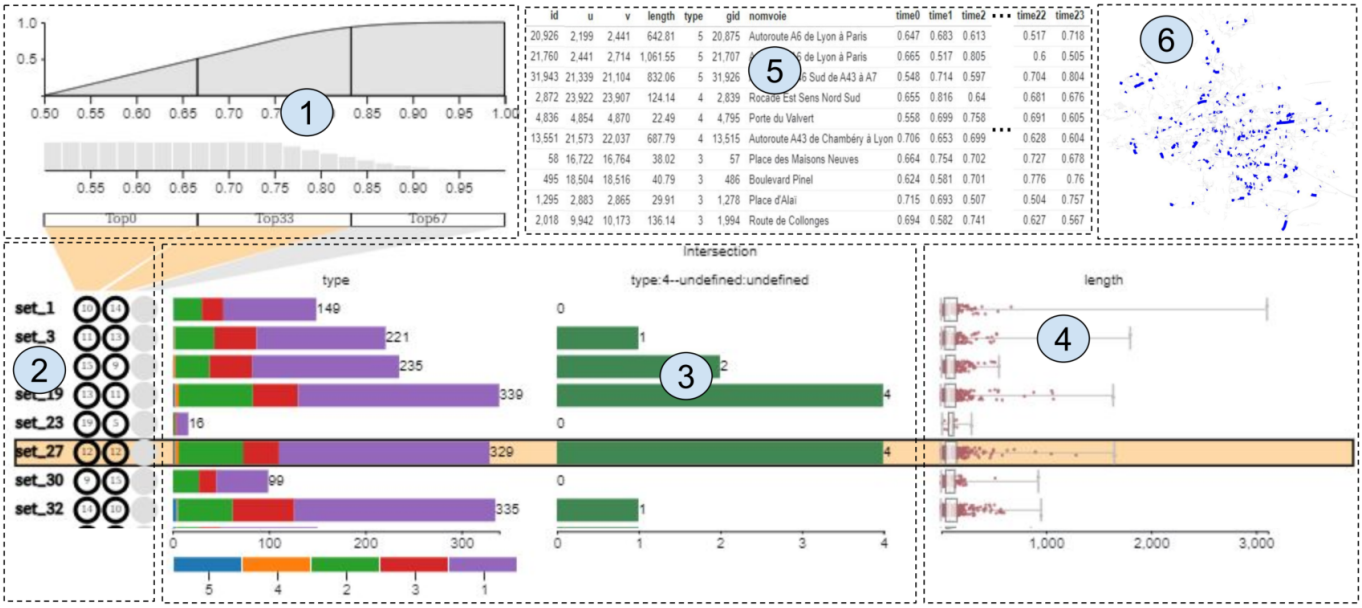


Figure 4: MULTISET first put together all the values of traffic density of different time, and then categorize the values into different groups in ① as well as a name is given for each category such as *Top33*. Road segments are gathered based on how many traffic density values being scaled in certain categories in ②. For example, in the row of highlighted **set_7**, the road segments include 24 traffic density values and these road segments have 12 traffic density values scaled into *Top0*, another 12 traffic density values scaled into *Top33* but no traffic density value scaled into *Top67*. The distribution of categorical attributes is illustrated as ③ such as attribute **type** and the specific values of categorical attributes are shown as *Intersection* bar chart. Except for categorical attributes, the numerical attributes are presented as box charts shown as ④. The elements of highlighted sets (**set_7**) are listed in ⑤ as a table and they are also visualized as a specific plot such as road map in ⑥.