Bringing the power of complex data analysis methods into *R*

Teh Amouh[1,⋆], **Benot Macq**[2], **Monique Noirhomme-Fraiture**[1]

1. School of Computer Science, University of Namur, Belgium
2. School of Engineering, University of Louvain-la-neuve, Belgium
⋆Contact author: tam@info.fundp.ac.be

Keywords: Set-valued data, Modal data, Histogram data, Multidimensional data

In classical statistical data analysis, individuals are described by single-valued scalar type variables. A rectangular data matrix defines the relation between a set of individuals and a series of variables such that each cell of the data matrix contains one single scalar value (quantitative or categorical).

Sometimes, however, individuals require to be described by set-valued type variables. For example, in a time budget study, a variable that records the daily time spent watching television would not allow single-valued answer (such as 3 hours), because this value normally varies from day to day for each individual. An interval-valued answer like "between 2.5 and 3.5 hours", reflecting an internal variability, would be more appropriate. When considering a stock market, information on an unpredictable share price would include the degree of uncertainty. A possible statement would be: "the price of share S varies between 130 and 140, with probability 30%", leading to data expressed as an histogram for variable *price* and individual *S*. The classical single-valued cell data table can hardly cope with such complex descriptions. There is a need on a data table model in which each cell could contain a (weighted) listing of values taken from a predefined set, and not just a single quantitative or categorical value. The `data.frame` model provided by *R* does not apply.

A research field named *symbolic data analysis* (Bock and Diday (2000)) and defined as the extension of standard data analysis to complex data, proposes a great deal of methods for set-valued data analysis (Diday and Noirhomme-Fraiture (2008)). These data are called *symbolic data* and encompass interval type data (which means subsets of the set of real values), multi-valued categorical type data (which means listings of ordinal or nominal categories) and multi-valued quantitative type data (which means listings of numerical values). A listing of categories can be summarized as a weighted set of distinct categories or as a discrete probability distribution. In either case we talk about modal data. A listing of numerical values can be summarized as an interval with the lower and upper bounds being respectively the minimum and maximum values in the listing. A listing of numerical values can also be summarized or as a histogram (if the range of values in the listing is segmented into intervals) or as a cumulative density function. A *symbolic data table* is a rectangular data matrix which allows such complex data values in each of its cells. Powerful data analysis methods are available for these types of data.

In order to bring the power of set-valued data analysis methods into *R*, we develop appropriate data structures using both *S3* and *S4* object approaches available in *R* (Chambers (2008)). Our data structures include table objects that extend the `data.frame` object and allow complex data values in each cell. This talk is about the design and implementation of these data structures. An *R* package containing these data structures will be available as a basic building bloc for the *R* implementation of complex data analysis methods.

References

Bock, H.-H. and E. Diday (Eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin: Springer-Verlag.

Chambers, J. (2008). *Software for Data Analysis: Programming with R*. Springer.

Diday, E. and M. Noirhomme-Fraiture (Eds.) (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.