The Dataset Project: Handling survey data in R

E. Rousseaux[a] and **G. Ritschard**[a]

[a] NCCR LIVES
Institute for Demographic and Life Course Studies
University of Geneva, Switzerland
emmanuel.rousseaux@unige.ch

Mots clefs : Survey, Data Management, Data processing, Data analysis, Panel Data.

Population studies strongly rely on survey data. To meet the needs of recent research questions in social sciences, data collected have becomed in the past decades more and more complex, such as longitudinal data, network data and spatial data. These high volumes of structured data complicate the task of both documenting data and manipulating data, as for example when we want to prepare data for a specific study. There is a need for specific tools to assist the user in handling these complex data. The Dataset software is an effort in this direction. It aims at providing a framework for handling survey data in R, especially network and biographical data. More precisely, the software aims at facilitating the management of survey data by providing researchers in social sciences with high-level tools for storing, documenting, sharing, exploring and recoding survey data in a secure and efficient way. This initiative, conducted within the NCCR LIVES project, targets mainly life course data and especially data types collected and used within the NCCR LIVES project. Thus, the current roadmap includes the development of the framework to support (1) cross-sectional data, (2) network data with a specific handling of demographic data from people cited in the network of each respondent, and (3) panel data organized in successive waves. The software comes in the form of an R package which is currently available on the R-Forge platform. R is a powerful statistical tool, freely available and multi-platform which is nowadays more and more often used in the social sciences as an alternative to classical commercial software (SPSS, SAS, Stata). As R is open-source, a lot of researchers in methods appreciate to be able to share their work through this software. As a consequence, most of the recent state-of-the-art methods are available in R and many of them in R only. This is especially the case for the newest tools for life course analysis (e.g. the `TraMineR` package for life course sequence analysis and the `ltm` package for latent class modeling). Moreover, working in R allows benefiting from the numerous statistical procedures already optimized in R and taking advantages of the R powerful graphical capability.
From a general point of view, the Dataset software follows three goals:

- *Providing an efficient framework for storing and documenting complex survey data.* As a key point, the software aims at storing data together with the design of the survey within which data were collected. Thus, the data and the user manual describing the data are merged together. Among the different features provided to describe data we can mention the possibility of assigning short and long labels to variables and variable values, to refer each variable to its question number within the survey, to declare user-defined types of missing values, and to account for cross-sectional and longitudinal weights. Many important metadata can also be stored such as the population concerned by the survey, the used sampling method, the organization releasing the data, the user license type, etc. As all information is stored within the data object, we provide a method for such objects for generating a summary of the whole data base. The summary gives for each variable

its long label, the percent of valid cases and basic descriptive statistics. This summary can be directly exported as a PDF file and serves as a basic user manual of the data base that proves particularly useful for detecting errors and for sharing data with others.

- *Saving the scientist's time spent on data processing in favor of time devoted to the research question.* Preparing data for a study is often a very burden task. The Dataset software is intended to help the analyst in this task, allowing him to focus more quickly on the analysis. As data bases are generally large, the package provides a search function allowing to explore the whole data base and retrieve relevant variables for the study. It provides efficient tools for recoding categorical and quantitative variables. The Dataset software also provides support for handling missing values and allows to easily turn a missing value into a valid case and vice-versa. Furthermore, the software provides, for some classical statistical methods, front-ends especially designed for scientists in social sciences. These front-ends facilitate the scientist daily work within the R environment. As a key point, the software performs systematic data consistency checks to ensure that data were not altered during data preprocessing operations. When filtering out cases and using weights when available, the software also processes automatic checks to prevent the loss of representativeness with respect to control variables defined by the user.

- *Facilitating reproducible research.* Demographic and sociologic questions are generally complex and require a lot of work to be understood. Reproducible research, meaning attaching sufficient information about the performed data analysis to allow anyone to retrieve the same results, is a helpful methodology when studying social dynamics. Having the possibility to rerun an experiment made by other researchers, or by oneself several months ago, gives the possibility to verify, better understand, and pursue an already done work. The Dataset software works in this direction by tracing operations made on data, so that the user can find back previously performed operations. Furthermore, for each statistical method provided by the package, results can be printed in a PDF file which also provides all settings used for calibrating the method. Outputs are displayed with a "ready-to-publish" formatting, allowing to quickly focusing on result interpretation.

In addition of these tools for cross-sectional data, the proposed software solution provides efficient methods for handling panel data organized in successive waves such as in the Swiss Household Panel. The user can directly extract whole trajectories from the panel data without having to bother with extracting the same variable independently from each yearly wave. The software automatically checks for each variable that it shares the same missing values and valid cases across years. By specifying ".." in place of the two year digits in the variable names, the user can extract a whole sequence in a single step. Likewise, the user can recode or merge values, or turn a missing value into a valid case directly for all waves where the variable exists. There also is a method for exporting a trajectory as a sequence object ready to be analyzed with the `TraMineR` package.

References

[1] Gabadinho, A., Ritschard, G., Müller, N. S., Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software, **40**(4), 1-37.
[2] Dimitris Rizopoulos (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, Journal of Statistical Software, **17**(5), 1-25.