

Interactive inspection of large data

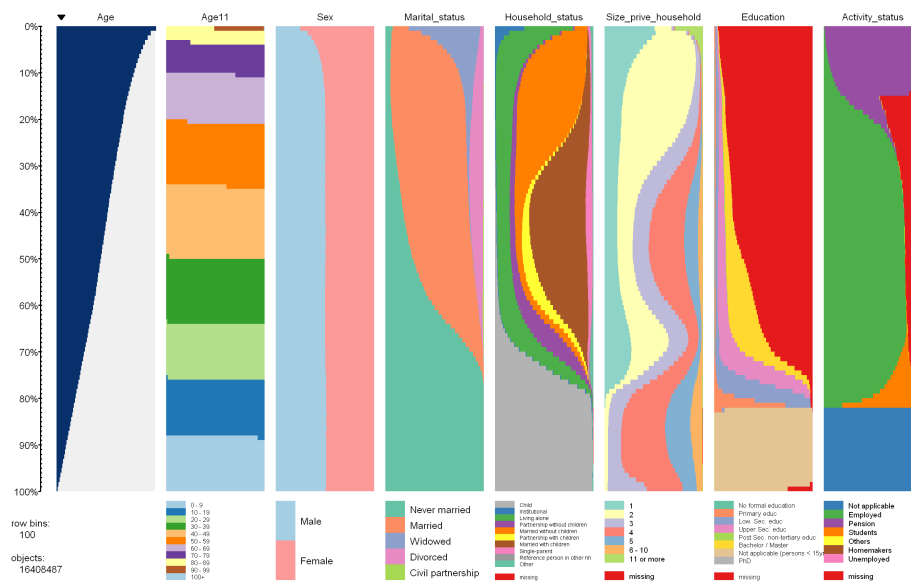
Edwin de Jonge^{1*}, Martijn Tennekes¹

1. Statistics Netherlands, The Hague/Heerlen

*Contact author: e.dejonge@cbs.nl

Keywords: multivariate analysis, visualization, large data, web application

We present **tabplotd3** an interactive tool in *R* for inspecting large multivariate data consisting of millions of observations. **tabplotd3** allows a user to visually check multivariate data for missing values, anomalies and patterns in bivariate relations, for categorical as well as numerical variables. The interface creates tableplots [1] from `data.frames` and allows to zoom in to individual observations. We applied the tool to the Dutch (virtual) census [2], creating a tableplot that depicts the 16.4 million inhabitants of the Netherlands.



The presentation will describe tableplots, their use and implementation in **tabplot**[3] and **tabplotd3**. The tool is a *R* package that spawns an interactive web application. To allow fast actions on very large `data.frames` we implemented some aggregations methods in *C*. The user interface is in *javascript* and *SVG*. The tool builds on several other packages including **Rook**, **RJSONIO**, **ff** and our non interactive **tabplot**.

References

- [1] Malik, W., A. Unwin, and A. Gribov (2010). An interactive graphical system for visualizing data quality - tableplot graphics. In H. Loracek-Junge and C. Weihs (Eds.), *Classification as a Tool for Research, Proceedings of the 11th IFCS Conference*, pp. 331–339. Berlin: Springer.
- [2] Nordholt, E. S. (2005). The dutch virtual census 2001: A new approach by combining different sources. *Statistical Journal of the United Nations Economic Commission for Europe* 22(1), 25–37.
- [3] Tennekes, M., E. de Jonge, and P. Daas (2011). Visual profiling of large statistical datasets. In *NTTS, New Techniques and Technologies for Statistics*.