Exploring the multivariate structure of missing values using the R package VIM

Matthias Templ[1,2], Andreas Alfons[1], Peter Filzmoser[1]

[1] Department of Statistics and Probability Theory, Vienna University of Technology
[2] Department of Methodology, Statistics Austria

Rennes, July 8, 2009

Content

Missing values

- Real data sets often contain missing values:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & \dots & x_{1p} \\ \vdots & \text{NA} & & \vdots \\ & & & \text{NA} \\ \vdots & & \text{NA} & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{pmatrix},$$

  with n observations, p variables, and some missing values. (NA)
- Examples: nonresponse in surveys, element concentration below detection limit in chemical analyses.

Comments on missing values

- Most statistical methods can **only** be applied to complete data.
- In order to select an appropriate imputation method (especially for model-based imputation), it is necessary to know the multivariate structure of the missing values beforehand.
- Visualizing missing values may not only help to detect the missing value mechanisms, but also to gain insight into the quality and various other aspects of the data.

Missing value mechanisms

Three important cases (e.g., Little and Rubin 2002):

- **MCAR** (**M**issing **C**ompletely **A**t **R**andom):

$$P(\mathbf{X}_{miss}|\mathbf{X}) = P(\mathbf{X}_{miss})$$

- **MAR** (**M**issing **A**t **R**andom):

$$P(\mathbf{X}_{miss}|\mathbf{X}) = P(\mathbf{X}_{miss}|\mathbf{X}_{obs})$$

- **MNAR** (**M**issing **N**ot **A**t **R**andom):

$$P(\mathbf{X}_{miss}|\mathbf{X}) = P(\mathbf{X}_{miss}|\mathbf{X}_{obs}, \mathbf{X}_{miss})$$

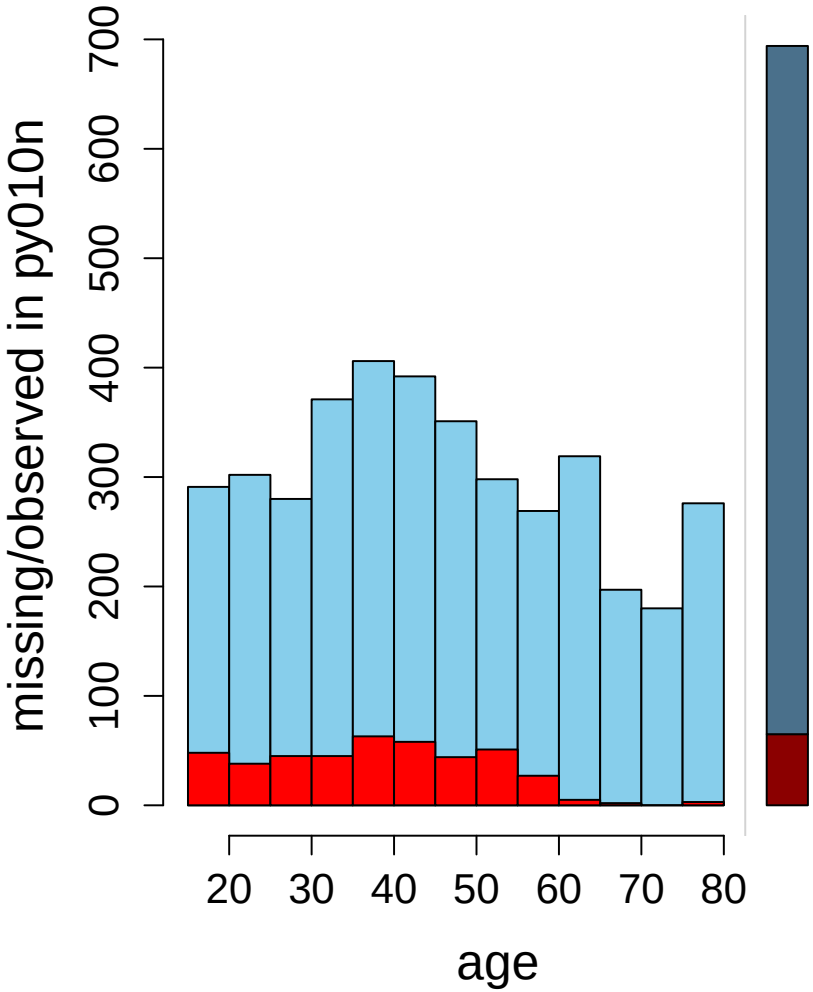
where $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$ denotes the complete data, and $\mathbf{X}_{obs}$ and $\mathbf{X}_{miss}$ are the observed and missing parts, respectively.

Visualization of missing values

- Famous books and almost all articles about missing values **do not** address vizualization.
- Visualization tools for missing values are rarely or not at all implemented in SAS, SPSS, STATA or even R.
- Through linking, missing values can be highlighted in GGobi (Cook and Swayne 2007) and Mondrian (Theus 2002).
- MANET (Unwin et al. 1996, Theus et al. 1997) is quite powerful, but only available for older Apple systems with PowerPC architecture and Mac OS.

Visualization tools for missing values **need to be available for the R community** so that visualization of missing valuess, imputation and analysis can all be done from within R, without the need of additional software.

Histogram and spinogram



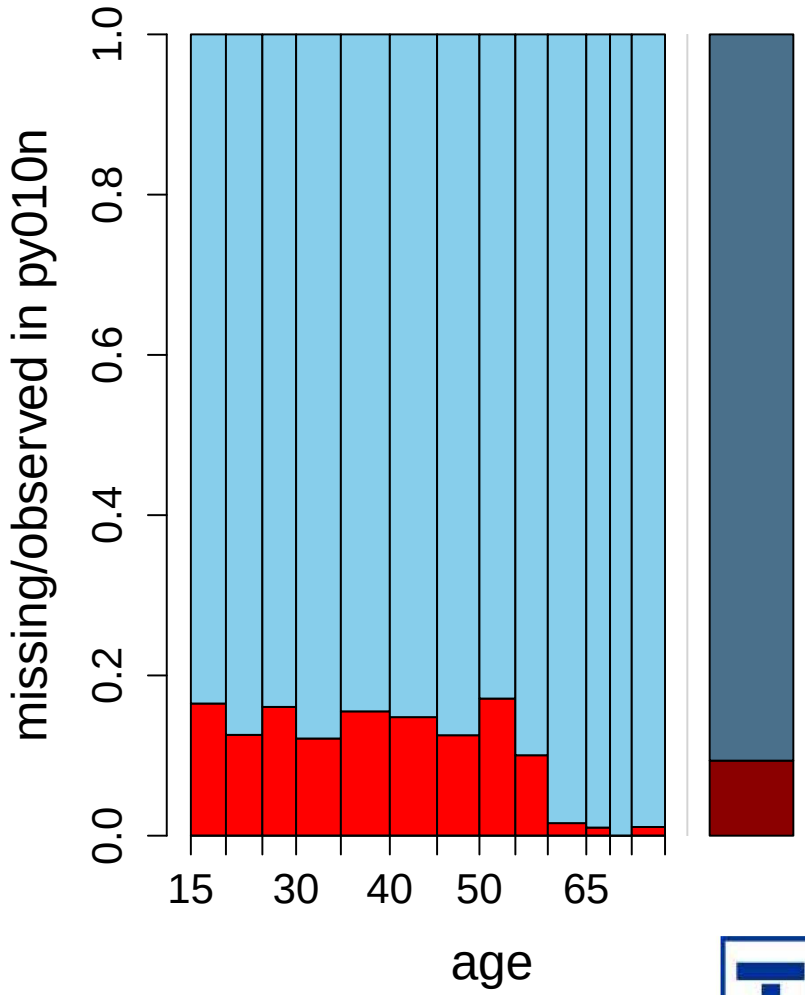


Figure: Austrian EU-SILC data from 2004 with missings generated in variable age.

Marginplot

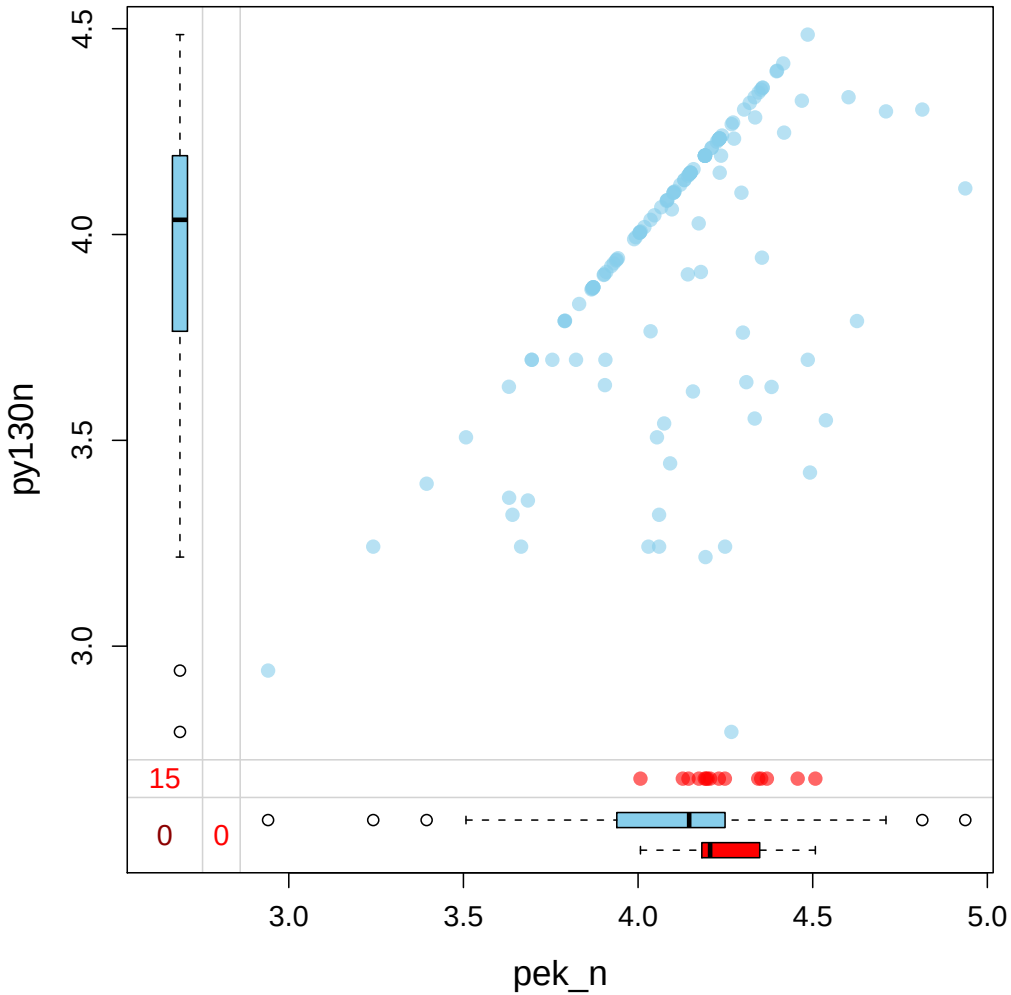


Figure: Austrian EU-SILC data from 2004.

Scatterplot matrix

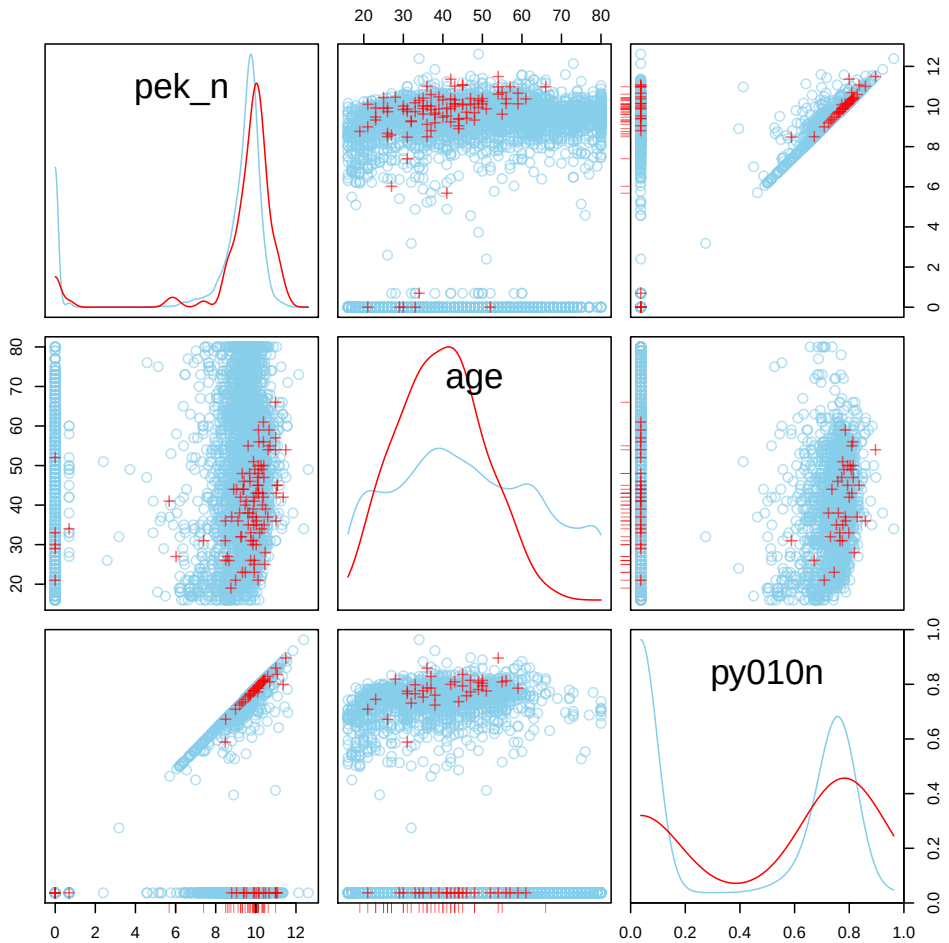


Figure: Austrian EU-SILC data from 2004.

Matrixplot

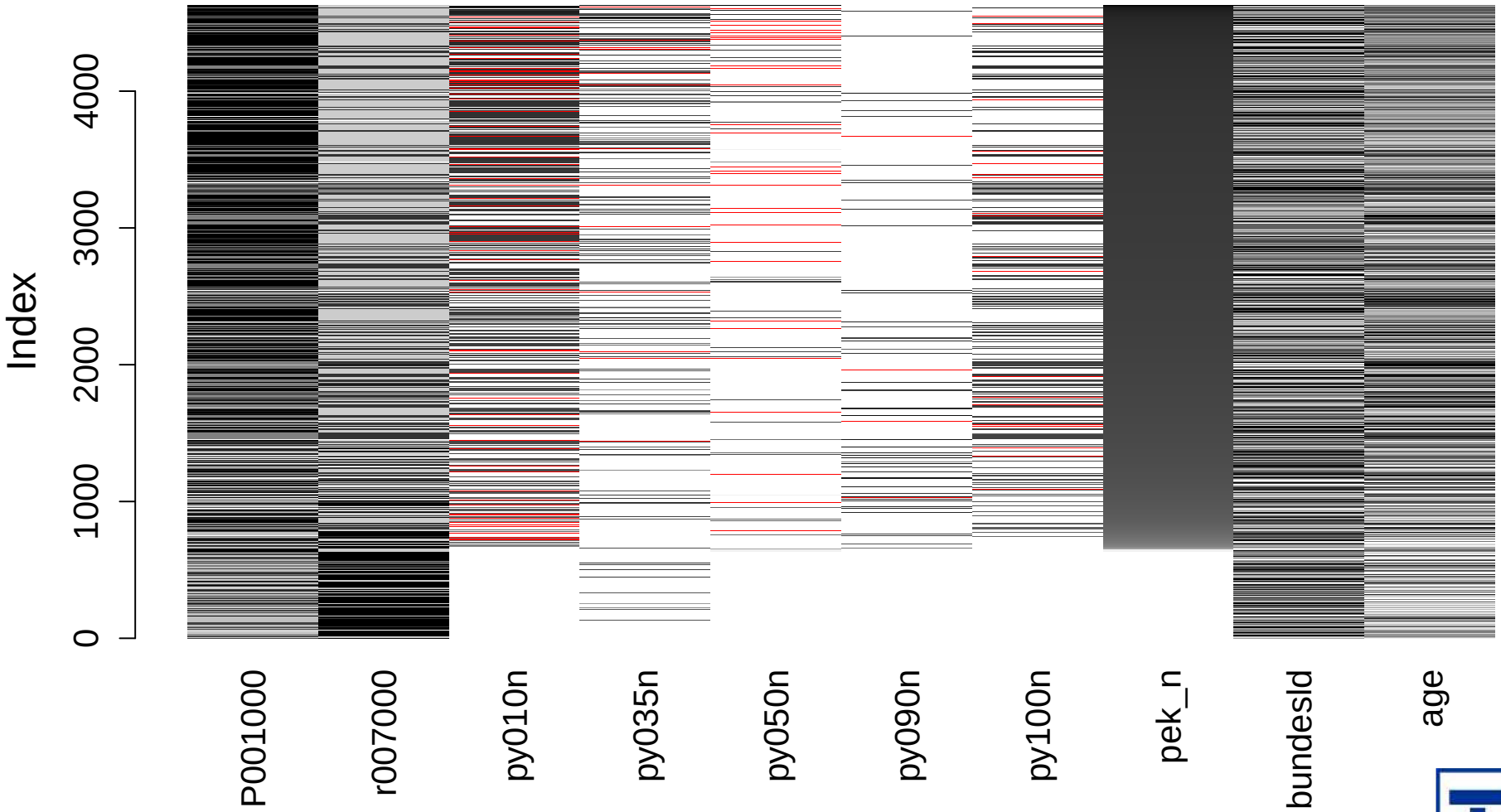


Figure: Austrian EU-SILC data from 2004.

Parallel coordinate plot

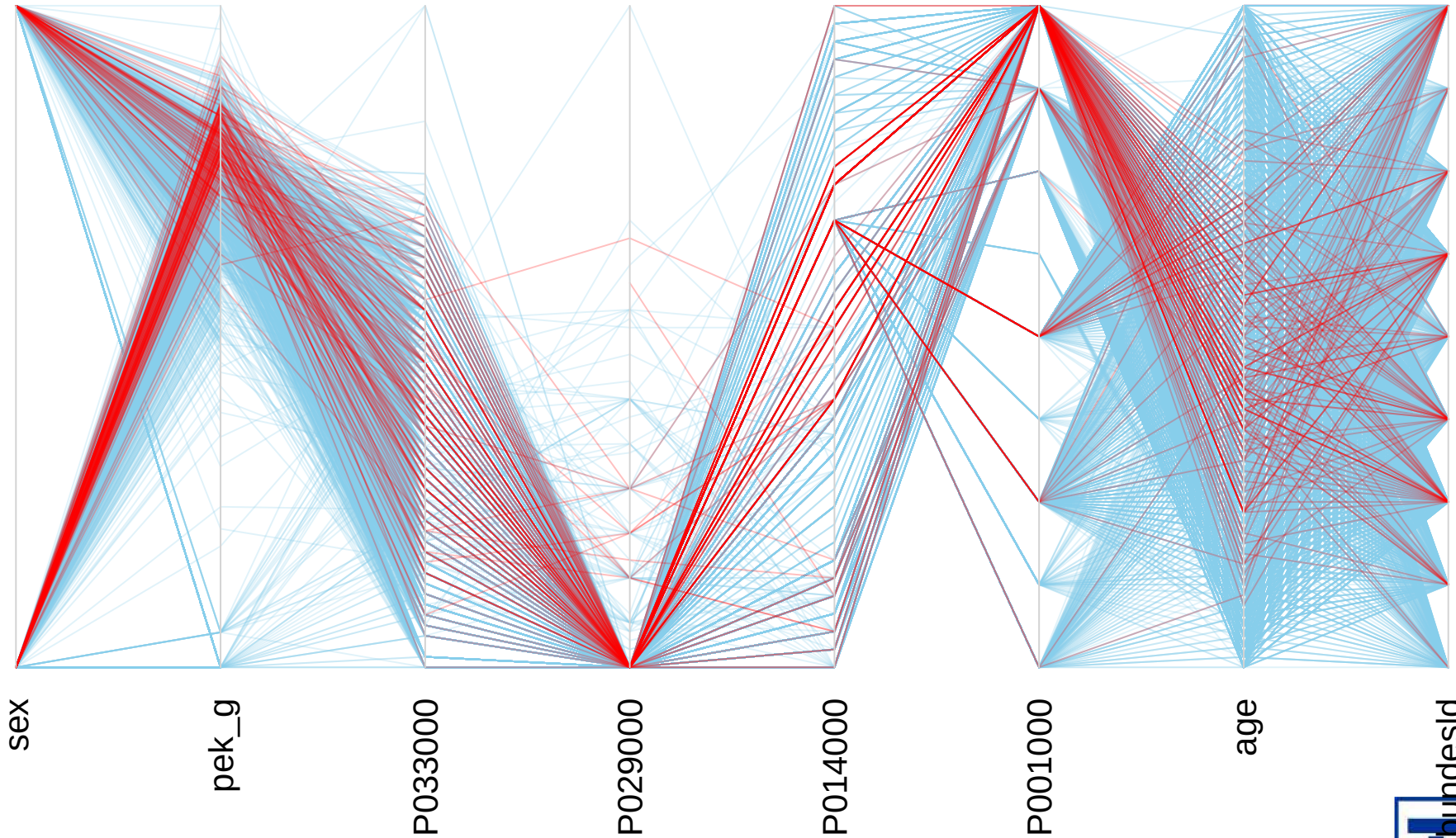


Figure: Austrian EU-SILC data from 2004

Parallel boxplots

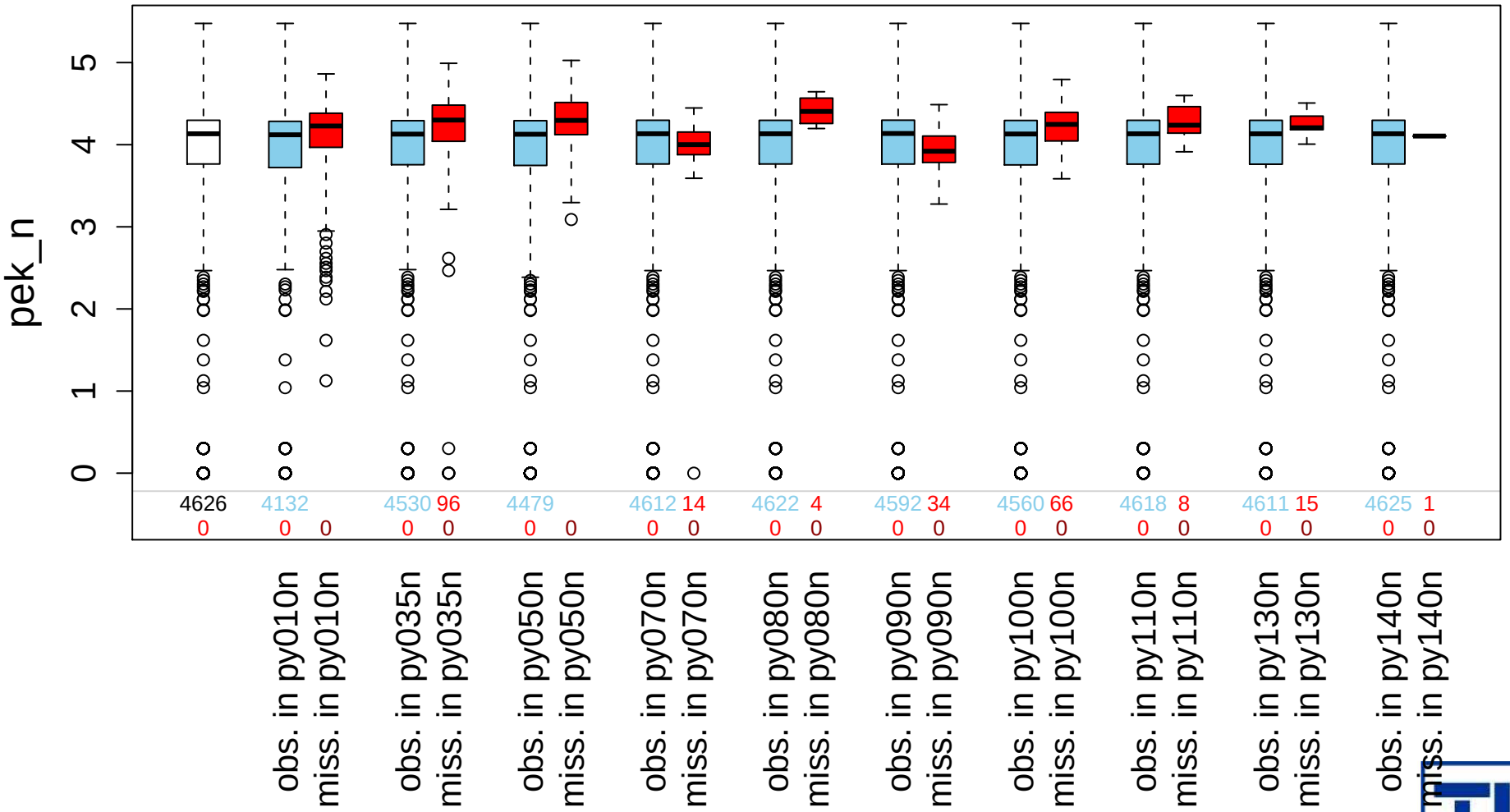


Figure: Austrian EU-SILC data from 2004.

General Statements

- The detection of missing value mechanisms is quite complex when using models or tests.
- Statistical methods frequently lead to only vague statements about the missing value mechanisms.
- Non-robust methods lead to erroneous statements about missing value mechanisms for data containing outliers.
- Visualization tools are easier to handle and more powerful, but flexible, easy-to-use visualization software is required.

The R package VIM

The R package VIM (Templ and Filzmoser 2008, Templ and Alfons 2009)

...

- has all previously shown plots implemented, along with some more.
- is a tool for explorative data analysis of data with missing values.
- makes it possible to analyze the multivariate structure of missing values.
- comes with a graphical user interface (GUI).
- contains interactive features.
- allows producing high-quality graphics for publications.
- is available on CRAN (http://cran.r-project.org/package=VIM).

Graphical user interface of the R Package VIM

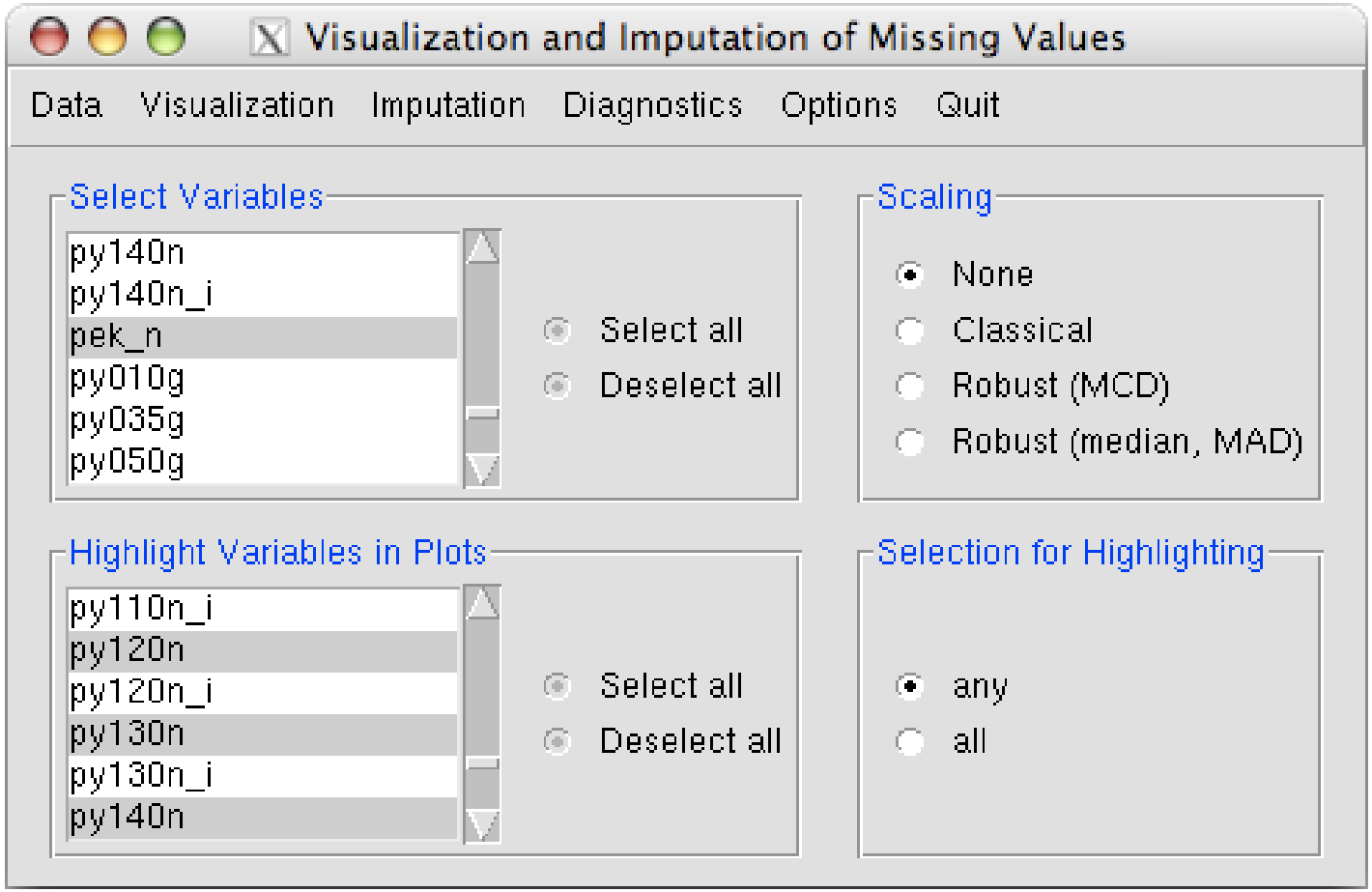


Figure: VIM GUI

Acknowledgments

This work was partly funded by the European Union (represented by the European Commission) within the 7^{th} framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). Visit http://ameli.surveystatistics.net for more information.

References I

D. Cook and D.F. Swayne. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi. Springer, New York, 2007. ISBN 978-0-387-71761-6.

R.J.A. Little and D.B. Rubin. Statistical Analysis with Missing Data. Wiley, New York, 2nd edition, 2002. ISBN 0-471-18386-5.

M. Templ and A. Alfons. VIM: Visualization and Imputation of Missing Values, 2009. URL http://cran.r-project.org/package=VIM. R package version 1.3.

M. Templ and P. Filzmoser. Visualization of missing values using the R-package VIM. Research Report CS-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology, 2008. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf.

References II

M. Theus. Interactive data visualization using mondrian. Journal of Statistical Software, 7(11): 1–9, 2002. URL http://www.jstatsoft.org/v07/i11.

M. Theus, H. Hofmann, B. Siegl, and A. Unwin. MANET - Extensions to interactive statistical graphics for missing values. In In New Techniques and Technologies for Statistics II, pages 247–259. IOS Press, 1997.

A. Unwin, G. Hawkins, H. Hofmann, and B. Siegl. Interactive graphics for data sets with missing values: MANET. Journal of Computational and Graphical Statistics, 5(2): 113–122, 1996.