An application of **VIM**, the R package for visualization of missing values, to EU-SILC data

Matthias Templ and Andreas Alfons

August, 2009

Abstract

Package **VIM** allows to explore and to analyze the structure of missing values in data, as well as to produce high-quality graphics for publications. This paper illustrates an application of **VIM** to a highly complex data set – the European Statistics on Income and Living Conditions (EU-SILC).

1 The graphical user interface of VIM

The graphical user interface (GUI) has been developed using the R package **tcltk** [R Development Core Team, 2009] and allows easy handling of the functions included in package **VIM**. Figure 1 shows the GUI, which pops up automatically after loading the package.

```
> library(VIM)
```

If the GUI has been closed, it can be reopened with the following command. All selections and settings from the last session are thereby recovered.

```
> vmGUImenu()
```

For visualization, the most important menus are the *Data*, the *Visualization* and the *Options* menus.

1.1 Handling data

The *Data* menu allows to select a data frame from the R workspace (see Figure 2). In addition, a data set in `.RData` format can be imported from the file system into the R workspace, which is then loaded into the GUI directly.

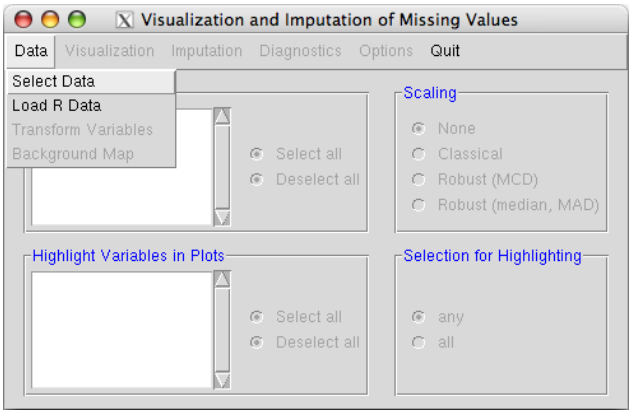

Figure 1: The **VIM** GUI and the *Data* menu.

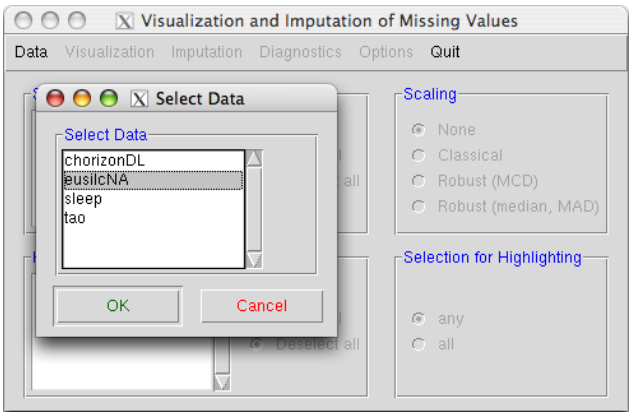

Figure 2: The dialog for data selection.

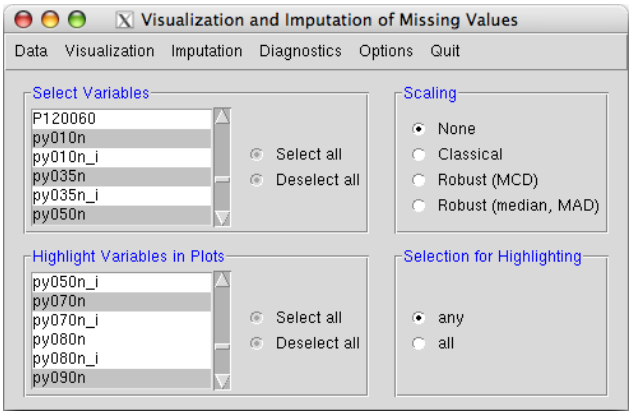

Figure 3: Variable selection with the **VIM** GUI.

Transformations of variables are available via `Data` → `Transform Variables`. The transformed variables are thereby appended to the data set in use. Commonly used transformations in official statistics are available, e.g., the Box-Cox transformation [Box and Cox, 1964] and the log-transformation as an important special case of the Box-Cox transformation. In addition, several other transformations that are frequently used for compositional data [Aitchison, 1986] are implemented. Background maps and coordinates for spatial data can be selected in the *Data* menu as well.

Functionality to scale variables, on the other hand, is offered in the upper right frame of the GUI. Note that scaling is performed on-the-fly, i.e., the scaled variables are simply passed to the underlying plot functions, they are not permanently stored.

1.2 Selecting variables

After a data set has been chosen, variables can be selected in the main dialog (see Figure 3). An important feature is that the variables will be used in the same order as they were selected, which is especially useful for parallel coordinate plots.

Variables for highlighting are distinguished from the plot variables and can be selected separately (see the lower left frame in Figure 3). If more than one variable chosen for highlighting, it is possible to select whether observations

with missing values in any or in all of these variables should be highlighted (see in the lower right frame in Figure 3).

1.3 Selecting plots

A plot method can be selected from the *Visualization* menu. Note that plots that are not applicable with the selected variables are disabled, e.g., if only one plot variable is selected, multivariate plots are not available.

2 An application to EU-SILC data

In this section, some of the visualization tools are illustrated on the public use sample of the Austrian EU-SILC data from 2004 [Statistics Austria, 2007], which can be obtained from Statistics Austria (see Table 1 for an explanation of the variables used here). This well-known and complex data set is mainly used for measuring risk-of-poverty and social cohesion in Europe, and for monitoring the Lisbon 2010 strategy of the European Union. The raw data set contains a high amount of missing values, which are imputed with model-based and donor-based imputation methods before public release [Statistics Austria, 2006]. Since a high amount of missing values are not MCAR, the variables to be included for imputation need to be selected carefully. This problem can be solved with our proposed visualization tools.

Table 1: Explanation of the used variables from the EU-SILC data set.

name	meaning
age	Age
R007000	Occupation
P033000	Years of employment
py010n	Employee cash or near cash income
py035n	Contributions to individual private pension plans
py050n	Cash benefits or losses from self-employment
py070n	Values of goods produced by own-consumption
py080n	Pension from individual private plans
py090n	Unemployment benefits
py100n	Old-age benefits
py110n	Survivors' benefits
py120n	Sickness benefits
py130n	Disability benefits
py140n	Education-related allowances

```
> incvars <- c(paste("py", c("010", "035", "050", "070", "080",
+     "090", "100", "110", "120", "130", "140"), "n", sep=""))
> eusilcNA[, incvars] <- log10(eusilcNA[, incvars] + 1)
```

First of all, it may be of interest how many missing values are contained in each variable. Even more interesting, missing values may frequently occur in certain combinations of variables. This can easily investigated by selecting variables of interest (see Figure 3) and by clicking on `Visualization` → `Aggregate Missings`. If one prefers the command line language of R, the the plot in Figure **??** can be created by invoking:

```
> aggr(eusilcNA[, incvars], numbers=TRUE, prop = c(TRUE, FALSE))
```

Here `eusilcNA` denotes the data frame in use (see also Figure 2). The barplot on the left hand side shows the proportion of missing values in each

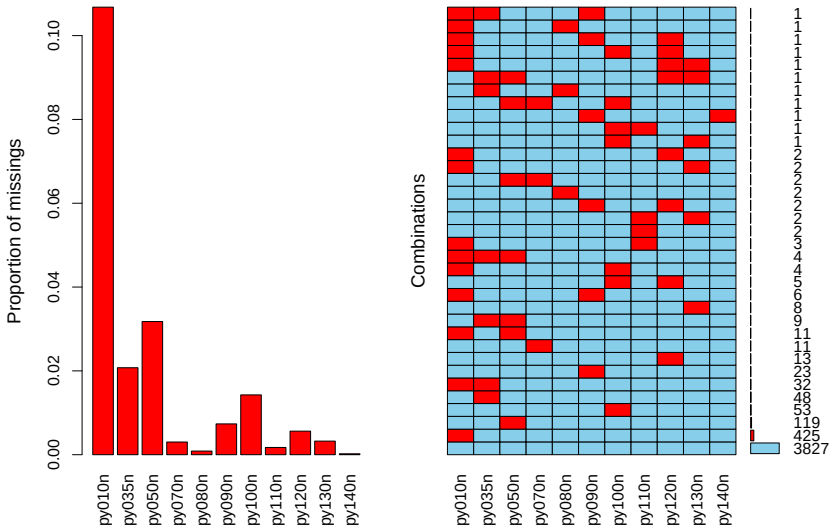

Figure 4: Aggregation plot of the income components in the public use sample of the Austrian EU-SILC data from 2004. *Left*: barplot of the proportions of missing values in each of the income components. *Right*: all existing combinations of missing (red) and non-missing (blue) values in the observations. The frequencies of the combinations are visualized by small horizontal bars.

of the selected variables. On the right hand side, all existing combinations of missing and non-missing values in the observations are visualized. A red rectangle indicates missingness in the corresponding variable, a blue rectangle represents available data. In addition, the frequencies of the different combinations are represented by a small bar plot and by numbers. Variables may be sorted by the number of missing values and combinations by the frequency of occurrence to give more power to finding the structure of missing values. For example, the top row in Figure 4 (right) represents the combination with missing values in variables *py010n* (employee cash or near cash income), *py035n* (contributions to individual private pension plans) and *py090n* (unemployment benefits), and observed values in the remaining variables, which appears only once in the data.

The plot reveals an exceptionally high number of missing values in variable *py010n*. The combination with variable *py035n* still contains 32 missing values. Note that it is possible to display proportions of missing values and combinations rather than absolute numbers.

2.1 Univariate plots

When only one variable is selected, only plots emphasized in Figure 5 can be applied. Standard univariate plots, such as barplots and spine plots for categorical variables and histograms, spinograms and different types of boxplots for continuous variables, have been adapted to display information about missing values.

For example, it may be of interest to display the distribution of years of employment, with missing values in *py010n* (employee cash or near cash income) highlighted. A spinogram [Hofmann and Theus, 2005] can easily be generated by clicking `Visualization` → `Spinogram with Missings`. Alternatively, the output shown in Figure 6 can be produced with the following command:

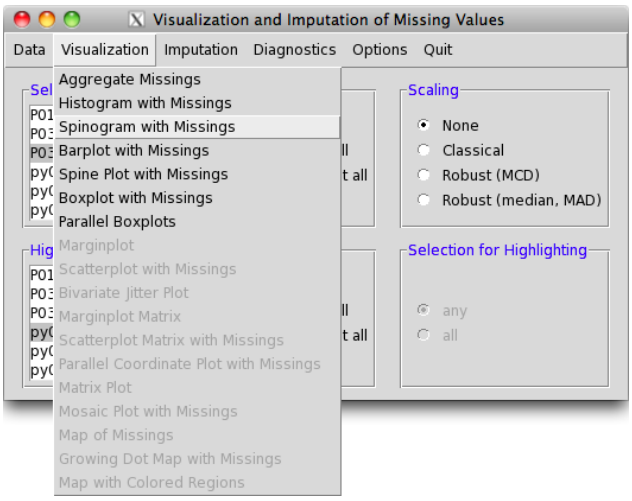

Figure 5: Univariate Plots supported by the **VIM** GUI.

```
> spineMiss(eusilcNA[, c("P033000", "py010n")])
```

Figure 6 indicates that the probability of missingness in *py010n* depends on the years of employment.

2.2 Bivariate plots

For bivariate data, different kinds of scatterplots are implemented. Figure 7 lists the plots applicable when two variables are selected. Multivariate plots are also highlighted because they can be used in the bivariate case, too.

Figure 8 shows a scatterplot with information about the univariate distributions and missingness of the variables in the plot margins (`Visualization` → `Marginplot`). The boxplots in red indicate observations with missing values in the other variable. It is clearly visible that the amount of missingness in *py010n* (employee cash or near cash income) is less for older people. Note that semi-transparent colors are used to prevent overplotting. The figure can also be produced with the command line interface of R, using the following command:

```
> marginplot(eusilcNA[, c("age", "py010n")], alpha = 0.6)
```

2.3 Multivariate plots

Parallel coordinate plots [Wegman, 1990] are very powerful for displaying multivariate relationships in data. A natural way of displaying information about missing data is to highlight observations according to missingness in a certain variable or a combination of variables. However, plotting variables with missing values results in disconnected lines, making it impossible to trace the respective observations across the graph. As a remedy, missing values may be represented by a point above the corresponding coordinate axis, which is separated from the main plot by a small gap and a horizontal line (see Figure 9). Connected lines can then be drawn for all observations.

Such parallel coordinate plots can be genereated by clicking `Visualization` → `Parallel Coordinate Plot with Missings` in the GUI or by using the function `parcoordMiss()` on the command line. The example in Figure 9 can be produced with:

```
> parcoordMiss(eusilcNA[, c("age", "P033000", "py010n", "py035n",
+     "py050n")], plotvars = 1:4, highlight = 5, alpha = 0.2)
```

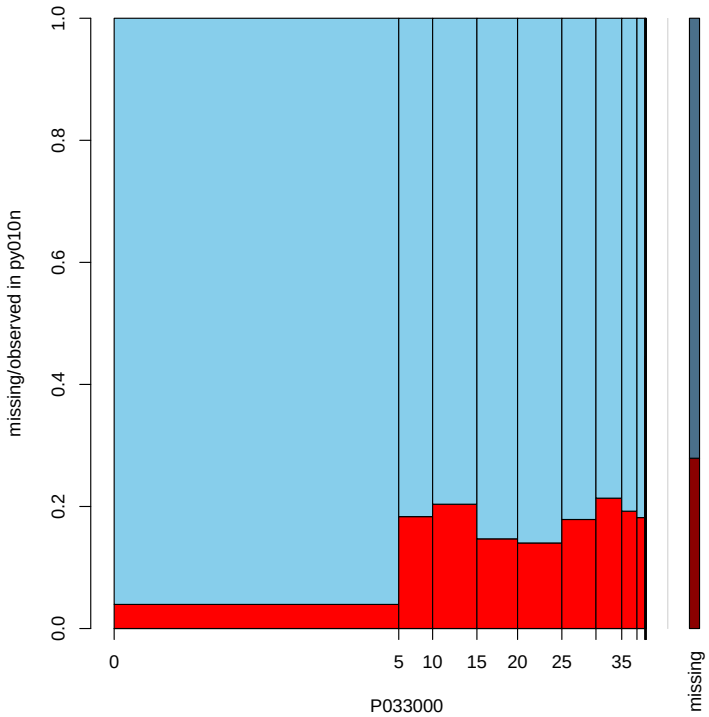

Figure 6: Spinogram of *P033000* (years of employment) with color coding for missing (red) and available (blue) data in *py010n* (employee cash or near cash income).

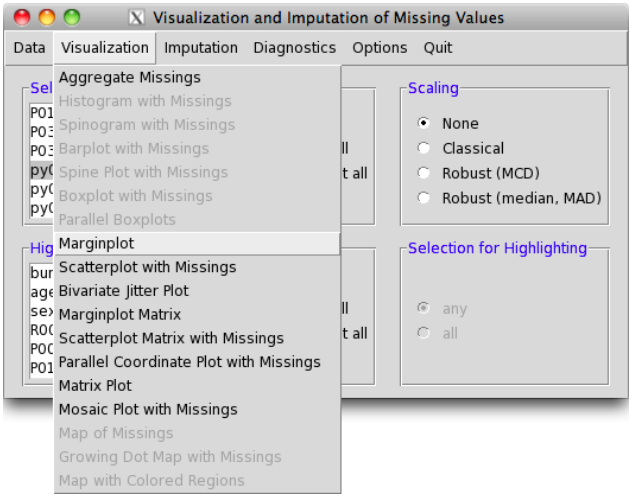

Figure 7: Bivariate plots (as well as multivariate plots) available in the **VIM** GUI.

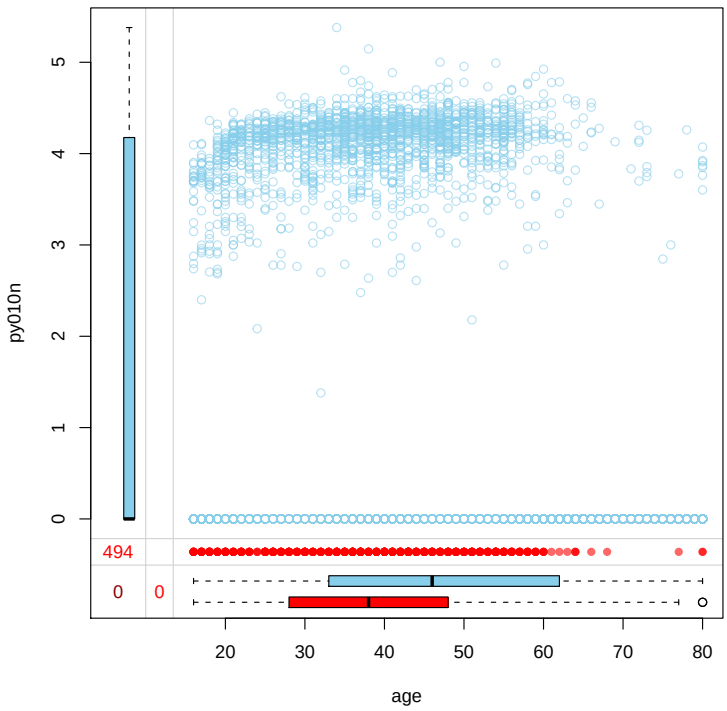

Figure 8: Scatterplot of *age* and transformed *py010n* (employee cash or near cash income) with information about missing values in the plot margins.

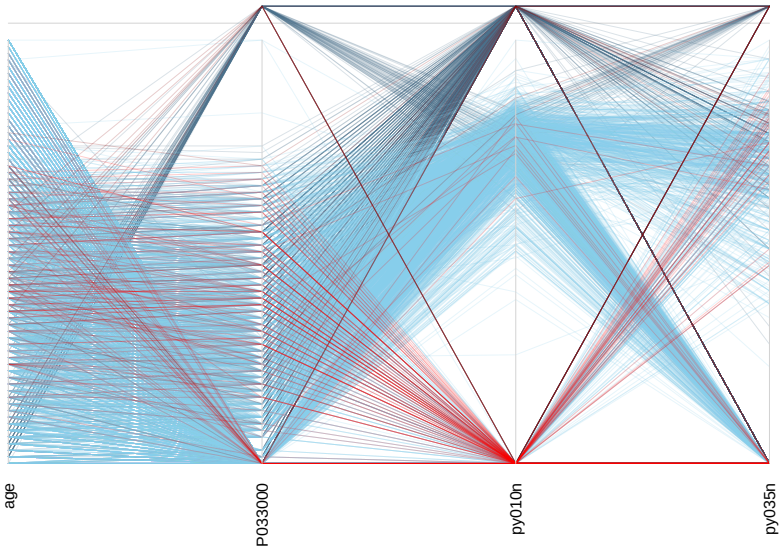

Figure 9: Parallel coordinate plot of *age*, *P033000* (years of employment), transformed *py010n* (employee cash or near cash income) and transformed *py035n* (contributions to individual private pension plans), with color coding for missing (red) and available (blue) data in variable *py050n* (cash benefits or losses from self-employment).

A data frame containing all variables of interest needs to supplied as the first argument, the variables to be plotted are given by argument `plotvars`, and the variables to be used for highlighting are specified by argument `highlight`.

Due to the large number of lines, a very low alpha value (i.e., very high transparancy) is used in Figure 9 to prevent overplotting. Missing values in *py050n* occur mainly for middle-aged people. Moreover, observations with missing values in *py050n* behave in an entire different way for the variables *py010n* (employee cash or near cash income) and *py035n* (contributions to individual private pension plans) than the main part of the data.

The *matrix plot* is an even more powerful multivariate plot. It visualizes all cells of the data matrix by (small) rectangles. In the example in Figure 10, red rectangles are drawn for missing values, and a greyscale is used for the available data. To determine the grey level, the variables are first scaled to the interval $[0, 1]$. Small values are assigned a light grey and high values a dark grey (0 corresponds to white, 1 to black). In addition, the observations can be sorted by the magnitude of a selected variable, which can also be done interactively by clicking in the corresponding column of the plot. Using the GUI, a matrix plot can be produced by clicking `Visualization` → `Matrix Plot`. The example in Figure 10 can also be created on the command line by invoking the following command:

```
> matrixplot(eusilcNA[, c("age", "R007000", incvars)],
+     sortby = "R007000")
```

Figure 10 shows a matrix plot of *age*, *R007000* (occupation) and the transformed income components, sorted by variable *R007000* (occupation). It is clearly visible that missing values in most income components depend on the occupation of the corresponding person. Thus the missing data mechanism was found to be MAR for these variables, which should be considered when applying imputation methods.

2.4 Other plots

Various other plots are availabe in the package and can also be created with the GUI (see Figures 5 and 7). For spatial data, mapping is supported if a background map is provided by the user, e.g., as a shape file, data frame or list of coordinates.

3 Fine tuning

In the *Preferences* dialog from the *Options* menu (click `Options` → `Preferences`), which is displayed in Figure 11, the colors and alpha channel to be used in the plots can be set. In addition, it contains an option to embed multivariate plots in Tcl/Tk windows. This is useful if the number of observations and/or variables is large, because scrollbars allow to move from one part of the plot to another.

4 Interactive features

Many interactive features are implemented in the plot functions in order to allow easy modification of the plots.

When variables are selected for highlighting in univariate plots such as histograms, barplots, spine plots or spinograms, it is possible to switch between the variables. Clicking in the right plot margin of a histogram, for example, corresponds with creating a histogram (or barplot) for the next variable, and clicking in the left margin switches to the previous variable. This interactive feature is particularly usedful for parallel boxplots, as it allows to view all

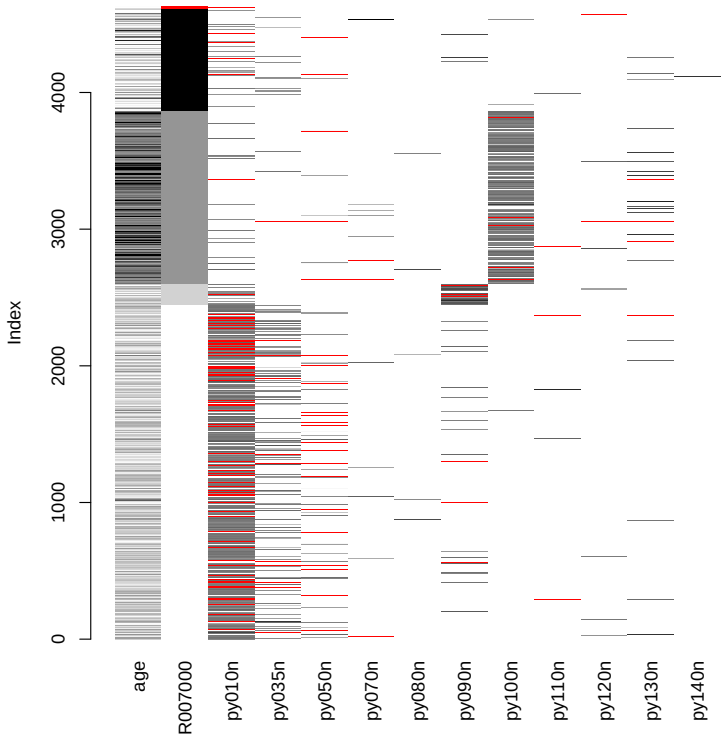

Figure 10: Matrixplot of *age*, *R007000* (occupation) and the transformed income components, sorted by variable *R007000* (occupation).

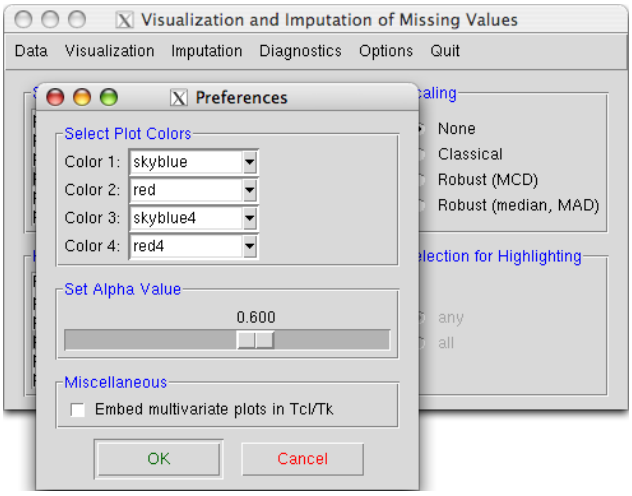

Figure 11: The *Preferences* dialog of the **VIM** GUI.

possible $p(p-1)$ combinations with $p-1$ clicks, where p denotes the number of variables.

For multivariate plots (scatterplot matrix and parallel coordinate plot), variables for highlighting can be selected and deselected interactively, by clicking in a diagonal panel of the scatterplot matrix or on a coordinate axis in the parallel coordinate plot. Information about the current selection is printed on the R-console.

The matrixplot is particulary powerful if the observations are sorted by a specific variable (see Figure 10). This can be done by clicking on the corresponding column.

5 Summary

We showed that the visualization of missing values is extremely simple with package **VIM**, either by using the GUI or by typing code on the R command line. With the visualization techniques in **VIM**, it is possible to gain insight into the data and to understand the structure of missing values. The latter is absolutely necessary when dealing with missing values, e.g., before imputation is performed.

6 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 on the project.

References

J. Aitchison. *The Statistical Analysis of Compositional Data*. Wiley, New York, 1986.

G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26:211–252, 1964.

H. Hofmann and M. Theus. Interactive graphics for visualizing conditional distributions. Unpublished manuscript, 2005.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

Statistics Austria. Einkommen, Armut und Lebensbedingungen 2004, Ergebnisse aus EU-SILC 2004, 2006. In German. ISBN 3-902479-59-0.

Statistics Austria. EU-SILC 2004. Erläuterungen: Mikrodaten-Subsample für externe Nutzer, 2007. In German.

E.J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.