# Exploring incomplete data using visualization techniques

**Matthias Templ · Andreas Alfons · Peter Filzmoser**

**Abstract** Visualization of incomplete data allows to simultaneously explore the data and the structure of missing values. This is helpful for learning about the distribution of the incomplete information in the data, and to identify possible structures of the missing values and their relation to the available information. The main goal of this contribution is to stress the importance of exploring missing values using visualization methods and to present a collection of such visualization techniques for incomplete data, all of which are implemented in the R package **VIM**. Providing such functionality for this widely used statistical environment, visualization of missing values, imputation and data analysis can all be done from within R without the need of additional software.

## 1 Introduction

Data often contain missing values, and the reasons are manifold. Missing values occur when measurements fail, in case of nonrespondents in surveys, when analysis results get lost, or when measurements are implausible. Examples for missing values in the natural sciences are broken measurement units for measurements of ground water quality or temperature, lost soil samples in geochemistry, or soil samples that would need to be re-analyzed but are exhausted. Examples for missing values in official statistics are respondents who deny information about their income or small companies that do not report their turnover.

M. Templ (corresponding author) · A. Alfons · P. Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstraße 7, 1040 Vienna, Austria
E-mail: templ@tuwien.ac.at

M. Templ
Methods Unit, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria

A. Alfons
ORSTAT Research Center, Faculty of Business and Economics, K.U.Leuven,
Naamsestraat 69, 3000 Leuven, Belgium

Subject matter specialists in statistical agencies work with only one (periodically sampled) survey and typically know the most relevant relationships in their data, but this knowledge is hard-earned through experience and many years of working on the same survey. Appropriate visualization tools may help to gain this knowledge in much less time. This information then helps to decide how the data will be prepared, either whether some respondents should be contacted once more, if some parts of the data should be calibrated for missing values, or if imputation (i.e., estimation and replacement of missing values; see, e.g., Little and Rubin, 2002) should be performed. Imputation specialists themselves frequently have only little knowledge about the underlying complex data and therefore require such tools to understand the dependencies of missing values in the data.

With appropriate visualization techniques, the structure of the missing and non-missing data parts, as well as their relations can be explored. Especially the latter is of great importance to subject matter specialists who need to understand the special characteristics of the nonresponses.

Although comprehensive literature on the estimation of missing values is available, including standards books (e.g., Schafer, 1997; Little and Rubin, 2002; Rubin, 2004) and active developments in many fields of research (e.g., Vanden Branden and Verboven, 2009; Hron et al, 2010; Templ et al, 2011b; Josse et al, 2011), visualization of data with missing values is treated in far less publications (e.g., Unwin et al, 1996; Eaton et al, 2005; Young et al, 2006; Cook and Swayne, 2007). This is also reflected in statistical software. Visualization tools for missing values are rarely or not at all implemented in SAS, SPSS, STATA or even R (R Development Core Team, 2011). Through interaction, observations with missing values can be highlighted in Mondrian (Theus, 2002) and GGobi (Swayne et al, 2003; Cook and Swayne, 2007). Users of legacy Mac OS operating systems may still be able to use MANET (Unwin et al, 1996; Theus et al, 1997). As for GGobi, the power of MANET lies in its interactive features. Furthermore, some visualization tools for missing values in data are implemented in ViSta (Young, 1996; Young et al, 2006) and REGARD (Unwin et al, 1990; Unwin, 1994). Nevertheless, it should be noted that MANET, ViSta and REGARD are not actively developed anymore.

The package **VIM** (**V**isualization and **I**mputation of **M**issing Values; Templ et al, 2011a) introduces visualization techniques for missing values to the R community. Data preparation and manipulation, exploration of missing values, as well as statistical estimation can therefore be done within the same software framework, which is not the case with the other mentioned pieces of software for exploring missing values. As mentioned above, it is possible in Mondrian and GGobi to highlight observations with missing values through linking. However, Mondrian and GGobi are general tools for data visualization with a focus on interactive data exploration with linked graphics. Even though interactive highlighting of other subsets of the data through linking is not as important for visualizing missing values, one drawback of the R environment is that the possibilities for interactive graphics are limited. The aim of the package **ix** (Urbanek, 2011), which is currently under development, is to bring extensible interactive graphics to R. At the time of writing this paper, a stable version of **ix** has not yet been released, but using **ix** to incorporate interactive features such as linked graphics into **VIM** is possible future work. For other applications of dynamically linked graphics, see for example Perrotta et al (2009). In any case, certain interactive features focused on exploring missing values are already implemented in **VIM**. Furthermore, **VIM** allows to create high-quality graphics for publications, including modifications and additional

information added by the user. With Mondrian, only screenshots can be taken, and modifying as well as adding to plots is limited. This is because the aim of Mondrian is to gain insight into the data and not to provide paper-quality graphics. Nevertheless, users often would not accept a piece of software for visualizing missing values without the possibility of producing high-quality graphics.

Most commonly used imputation procedures (e.g., Schafer, 1997; Raghunathan et al, 2001) use a sequence of regression models, imputing one variable at a time. Accordingly, various plots in **VIM** may be used to analyze missing values in one variable in relation to other variables (e.g., parallel coordinate plots, see Section 3.7). Other plots, however, focus on missing values in the multivariate structure of the data (e.g., matrix plots, see Section 3.8). Furthermore, if geographical coordinates are available, maps can be used to analyze whether missing data corresponds to spatial patterns.

The rest of the paper is organized as follows. Section 2 discusses mechanisms generating missing values and limitations for their detection. In Section 3, various visualization methods for missing values are presented using real data. A few notes on the R package **VIM** are given in Section 4, before Section 5 summarizes.

## 2 Missing value mechanisms

There are three important cases to distinguish for the responsible generating processes behind missing values (see Rubin, 1976; Schafer, 1997; Little and Rubin, 2002). Let $\boldsymbol{X} = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ denote the data, where $n$ is the number of observations and $p$ the number of observed variables (dimensions), and let $\boldsymbol{M} = (M_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ be an indicator whether an observation is missing ($M_{ij} = 1$) or not ($M_{ij} = 0$). The missing data mechanism is characterized by the conditional distribution of $\boldsymbol{M}$ given $\boldsymbol{X}$, denoted by $f(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ indicates unknown parameters. Then the missing values are **M**issing **A**t **R**andom (MAR) if it holds for the probability of missingness that

$$f(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{\phi}) = f(\boldsymbol{M}|\boldsymbol{X}_{obs}, \boldsymbol{\phi}), \tag{1}$$

where $\boldsymbol{X} = (\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss})$ denotes the complete data, and $\boldsymbol{X}_{obs}$ and $\boldsymbol{X}_{miss}$ are the observed and missing parts, respectively. Hence the distribution of missingness does not depend on the missing part $\boldsymbol{X}_{miss}$.

If in addition the distribution of missingness does not depend on the observed part $\boldsymbol{X}_{obs}$, the important special case of MAR called **M**issing **C**ompletely **A**t **R**andom (MCAR) is obtained, given by

$$f(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{\phi}) = f(\boldsymbol{M}|\boldsymbol{\phi}). \tag{2}$$

If Equation (1) is violated and the patterns of missingness are in some way related to the outcome variables, i.e., the probability of missingness depends on $\boldsymbol{X}_{miss}$, the missing values are said to be **M**issing **N**ot **A**t **R**andom (MNAR). This relates to the equation

$$f(\boldsymbol{M}|\boldsymbol{X}, \boldsymbol{\phi}) = f(\boldsymbol{M}|(\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}), \boldsymbol{\phi}). \tag{3}$$

Hence the missing values cannot be fully explained by the observed part of the data.

A practical example for the different missing value mechanisms, which is adequate for the data used in this paper, is given by Little and Rubin (2002). Considering two variables *age* and *income*, the data are MCAR if the probability of missingness is the

same for all individuals, regardless of their age or income. If the probability that income is missing varies according to the age of the respondent, but does not vary according to the income of respondents with the same age, then the missing values in variable income are MAR. On the other hand, if the probability that income is recorded varies according to income for those with the same age, then the missing values in variable income are MNAR. Naturally, MNAR could hardly be detected (see below).

Appropriate visualization tools for missing values should be helpful for distinguishing between the three missing value mechanisms. However, there are some limitations that will be described in the following.

*Limitations for the detection of the missing value mechanisms*

It is often difficult to detect the missing values mechanism in practice exactly, because this would require the knowledge of the missing values themselves (Little and Rubin, 2002). For example, construct a non-correlated bivariate data set with variables $x$ and $y$, where only large values of $y$ are set to be missing (MNAR situation). A data analyst then cannot distinguish between MCAR and MNAR. On the other hand, for the same situation with highly correlated variables, a MAR situation is observed because the data analyst can only observe that the amount of missingness increases for increasing $x$-values. Nevertheless, well-established imputation methods for MAR situations still yield good estimates in such cases (see, e.g., Dempster et al, 1977).

Multivariate data with missing values in several variables can make it even more complicated to distinguish between the missing value mechanisms. The situation can become even worse in case of outliers, inhomogeneous data or very skewed data distributions. Nevertheless, those are general limitations for detecting missing value mechanisms not only affecting visualization. Visualization of missing values provides a fast way to distinguish between MCAR and MAR situations, as well as to gain insight into the quality and various other aspects of the underlying data at the same time.

## 3 Visualization methods for missing values

The visualization tools proposed in this section do not rely on any statistical model assumptions. The aggregation plot described in Section 3.2 is useful to gain an overview of the amount of missing values and to detect monotone missing values patterns. The plots described in Sections 3.3–3.9 are useful to explore the data and to gain insight into the distribution and structure of missing values. They often allow to detect MAR situations. Highlighting missing values in maps (see Section 3.10) may help detect MAR situations with respect to geographical positions of samples.

All plots are available in the R package **VIM**, and a graphical user interface allows easy handling.

3.1 Data sets

Before various visualization methods for missing values are discussed, a brief introduction to the data sets used in the examples is given.

### 3.1.1 Austrian EU-SILC data

Most of the visualization tools are illustrated on data from the *European Union Statistics on Income and Living Conditions* (EU-SILC). This well known survey produces highly complex data sets, which are mainly used for measuring risk-of-poverty and social cohesion in Europe in order to monitor the Lisbon 2010 strategy and Europe 2020 goals of the European Union. In particular, the Austrian EU-SILC public use data set from 2004 (Statistics Austria, 2007) is used in this paper. The variable description and more information on this data set is provided by the manual of package **VIM** (Templ et al, 2011a). The raw data contain a large amount of missing values, which are imputed with model-based imputation methods before public release (Statistics Austria, 2006). Since a considerable amount of the missing values are not MCAR, the variables to be included for imputation need to be selected carefully. This problem can be solved with the proposed visualization tools.

### 3.1.2 Kola C-horizon data

The *Kola Ecogeochemistry Project* was a geochemical survey of the Barents region whose aim was to reveal the environmental conditions in the European arctic. Soil samples were taken at different levels and are linked to spatial coordinates. In this paper, the C-horizon data (Reimann et al, 2008) are used. The raw data set, in which missing values are the result of element concentrations below the detection limit, is included in **VIM**.

### 3.1.3 Mammal sleep data

This data set contains sleep data and other characteristics of mammals (Allison and Cichetti, 1976). It is used in Young et al (2006) to illustrate visualization techniques for missing values in ViSta. Furthermore, it is also available in GGobi.

## 3.2 Aggregation plot

It is often of interest how many missing values are contained in each variable. Even more interesting, missing values may frequently occur in certain combinations of variables. In Figure 1, this information is displayed for the income components in the Austrian EU-SILC data. The barplot on the left hand side shows the proportion of missing values in each of the selected variables. Alternatively, the absolute frequencies can be shown instead of proportions. On the right hand side, all existing combinations of missing and non-missing values in the observations are visualized. A dark grey rectangle indicates missingness in the corresponding variable, a light grey rectangle represents available data. In addition, the frequencies of the different combinations are represented by a small bar plot. 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 bottom row in Figure 1 (right) represents observations without any missing values, in this case the large majority of the observations. Hence the small barplot for the different combinations would be dominated by the corresponding bar,
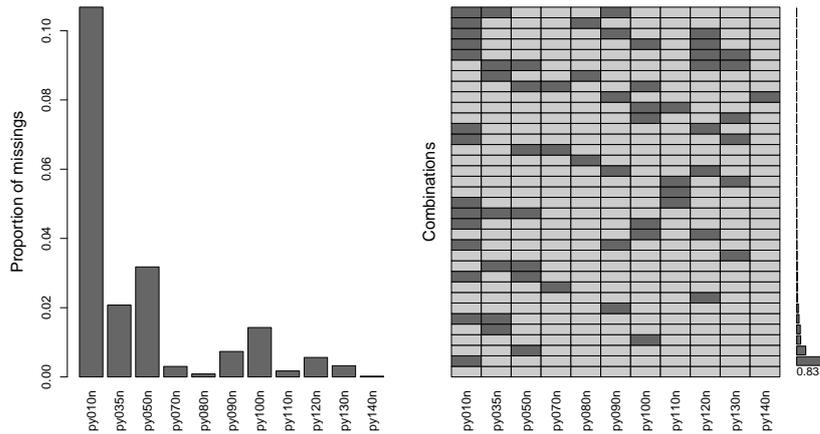
**Fig. 1** Aggregation plot of the income components in the public use sample of the Austrian EU-SILC data from 2004. Left: barplot of the proportions of missing values in each of the income components. Right: all existing combinations of missing (dark grey) and non-missing (light grey) values in the observations. The frequencies of the combinations are visualized by small horizontal bars.

leaving the bars for the combinations with missing values highly compressed. To increase the readability of the plot, **VIM** allows to represent the proportion or frequency of complete observations by a number and to rescale the bars for the remaining combinations. The least frequent combination is displayed in the top row: 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 income components. Furthermore, the plot reveals an exceptionally high number of missing values in variable *py010n* (second row from the bottom). Concerning combinations of variables, missing values in *py010n* and *py035n* are the most frequent (sixth row from the bottom). As mentioned in the introduction, such knowledge is useful when generating missing values in artificial data for simulation studies, and it is also useful for the data analyst during the data preparation process. In general, the plot showing the existing combinations of missing values (Figure 1, right) is helpful to detect monotone missing values patterns.

It should be noted that this plot is highly customizable in **VIM**. Instead of plotting a separate barplot of the amount of missing values in the variables on the left hand side, a smaller version of this barplot can be shown on top of the plot for the combinations. In addition, the frequencies of the combinations can also be visualized by adjusting the row heights instead of the small barplot on the right hand side. However, that type of plot has the disadvantage that it can easily become unreadable if there are many combinations with low frequencies of occurrence.

3.3 Histogram, barplot, spinogram and spine plot

When plotting a histogram of a continuous variable, the amount of missing values in this variable can be visualized by an extra bar. To emphasize that this bar does not
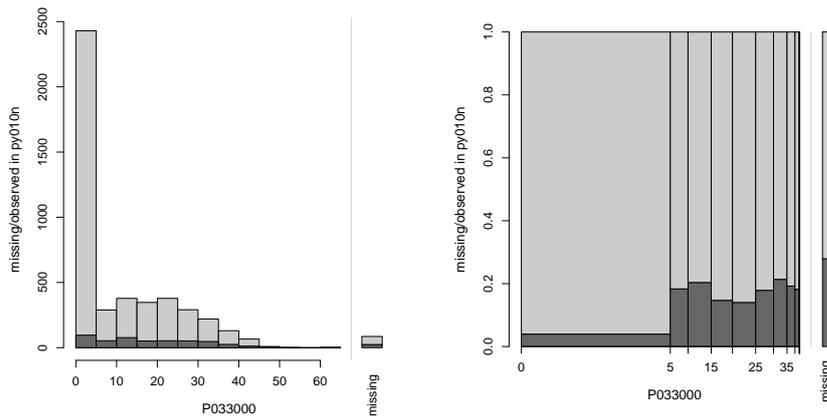
**Fig. 2** Histogram (left) and spinogram (right) of the variable *P033000* (years of employment) in the EU-SILC data with color coding for missing (dark grey) and available (light grey) data in variable *py010n* (employee cash or near cash income).

correspond to observed data, it can be separated from the rest of the plot by a small gap. In addition, the amount of missing values in other variables can be displayed by splitting each bin into two parts. Information about missingness can thereby be highlighted for more than one variable (see Section 4). Similar histograms are also implemented in MANET. Note that the same improvements for visualizing incomplete data can be made for barplots of categorical variables. As an example, Figure 2 (left) shows such a histogram of variable *P033000* (years of employment) of the EU-SILC data, where the bins are split according to the respective amount of missing (dark grey) and observed (light grey) values of variable *py010n* (employee cash or near cash income).

Alternative versions of histograms and barplots, which are not shown in this paper, are also available in **VIM**. Instead of plotting an extra bar for missing values in the variable of interest, a small additional barplot of observed and missing values can be displayed on the right hand side. Again, the two additional bars can be split according to information about missingness in other variables. This option has the advantage that it provides visual information of missingness in other variables for the total amount of observed values in the variable of interest. However, since the additional barplot is on a different scale, an extra $y$-axis is required.

For continuous variables, *spinograms* (Hofmann and Theus, 2005) are closely related to histograms. The horizontal axis is scaled according to relative frequencies of the bins, i.e., the widths of the bars reflect the frequencies rather than their height. For categorical variables, the same kind of plot can be produced, which in this case is referred to as *spine plot*. As for histograms and barplots, the proportion of missing values in the variable of interest can be represented by an extra bar. On the vertical axis, the proportion of missing and observed values in other variables can be displayed. Since the height of each cell corresponds to the proportion of missing/observed values in those other variables, it is now possible to compare the proportions of missing values across the different bins. Significant differences in these proportions indicate a MAR situation, which should be considered, e.g., when generating close-to-reality scenarios for missing data in simulation studies. Continuing the example from the histogram
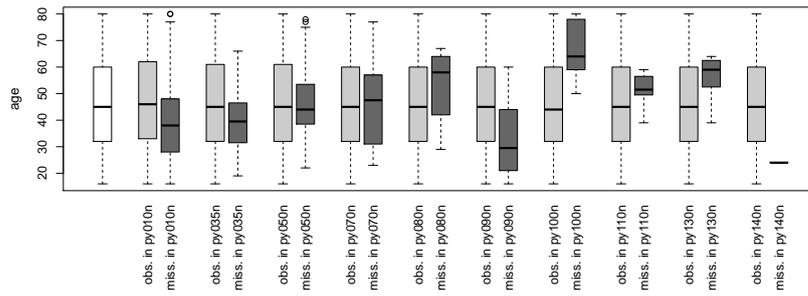
**Fig. 3** Values of the variable *age* in the EU-SILC data are grouped according to missingness in the different income components and presented in parallel boxplots.

in Figure 2 (left), Figure 2 (right) contains a spinogram of variable *P033000* (years of employment) with bins split according to the respective amount of missing (dark grey) and observed (light grey) values of variable *py010n* (employee cash or near cash income). While the histogram is clearly dominated by inactive (e.g., unemployed or retired) persons, for whom *P033000* is set to 0, the spinogram in Figure 2 shows a clear picture of the structure of the missing values. It should also be noted that **VIM** offers an option to plot a small additional spine plot on the right hand side for observed and missing data in the variable of interest instead of the extra bar for missing values.

Furthermore, it is possible to interactively switch between the variables, e.g., for histograms, clicking in the right plot margin corresponds with creating a histogram or barplot of the next variable, while clicking on the left variable switches to the previous variable.

3.4 Parallel boxplots

For a continuous variable, the conditional distributions according to a set of variables with values recoded as missing or non-missing can be compared by multiple parallel boxplots. This plot is therefore especially useful to explore whether one continuous variable explains the distribution of missing values in any another variable. Figure 3 shows an example of *age* in the EU-SILC data. In addition to a standard boxplot (left), boxplots grouped by observed (light grey) and missing (dark grey) values in the different income components are drawn. **VIM** thereby allows to draw the box widths either proportional to the group size or of equal size. Unfortunately, the first option is not possible in this example because the proportion of missing values is close to 0 for some of the income components (see Figure 1).

For many components, the presence of missing values clearly depends on the magnitude of the values of *age*, e.g., missing values in variable *py090n* (unemployment benefits) occur predominantly for individuals with lower age. This indicates MAR situations for missing values in these variables, which is a useful information for subject matter specialists.

As for the other univariate plots in **VIM**, the variable of interest can be switched interactively, i.e., clicking in the right margin creates parallel boxplots for the next variable, while clicking in the left margin switches to the previous variable. This allows

to view all possible $p(p-1)$ combinations with $p-1$ clicks, where $p$ is the number of variables.

## 3.5 Scatterplots

In addition to a standard scatterplot of two numeric variables, information about missing values can be displayed. A straightforward approach is to show observations with missing values in only one of the variables as univariate dot plots (sometimes also referred to as stripcharts or one-dimensional scatterplots) along the $x$- or $y$-axis, similar to implementations in MANET and GGobi. However, the implementation in **VIM** also includes boxplots for available and missing data in the plot margins. The frequencies of missing values in one or both variables are represented by numbers in the lower left corner. This plot will henceforth be referred to as *marginplot*. Figure 4 (left) shows an example using the variables *age* and *py090n* (unemployment benefits) in the EU-SILC data. Note that *alpha blending* is used to prevent overplotting, i.e., the plot colors are converted to translucent colors. The degree of transparancy is thereby controlled by the *alpha channel*, which can be varied easily with **VIM**. Furthermore, the variable *py090n* was log-transformed (with base 10) after the constant 1 had been added. The reason for this choice of transformation is twofold. First, income distributions are typically heavily right-skewed. A log-transformation results in a more symmetric distribution and prevents the data points from being plotted in a tight cluster of points in the corner of the graph. Second, *py090n* is semi-continuous, i.e., contains a large amount of zeros. Adding a positive constant is thus necessary to take the logarithms, and the constant 1 has the advantage that the zeros are preserved in the transformed variable. However, a transformation might fundamentally change the nature of the variable, making the interpretation of the plots more complex (for a detailed discussion of transformations, see, e.g., Osborne, 1999). Along the horizontal axis, the light grey box corresponds to observed data and the dark grey box to missing data in *py090n*, which indicate a MAR situation (cf. Figure 3). Furthermore, the plot reveals some outliers in *py090n*.

Another type of scatterplot is shown in Figure 4 (right). For observations with missing values in only one variable, a rug representation (i.e., small tickmarks on the corresponding plot axis) and dashed lines are drawn to indicate the observed part. Whether these are drawn for the $x$- or $y$-variable can be selected interactively. Optionally, tolerance ellipses, which are defined as the set of $p$-dimensional points whose Mahalanobis distance from the center equals the square root of a certain quantile of the $\chi^2$ distribution with 2 degrees of freedom, can be displayed to indicate the bivariate structure of the data. The dashed lines can then only be drawn within the largest ellipse, which may improve the readability of the plot. The second advantage of the tolerance ellipses is an improved visualization of the outliers. Points falling outside a large tolerance ellipse are easily identified as outliers. For this purpose, a robust estimator of location and scatter needs to be computed. **VIM** therefore uses the MCD estimator (Rousseeuw and Van Driessen, 1999). In this example, the largest tolerance ellipse is given by the 0.975 quantile of the $\chi^2$ distribution with 2 degrees of freedom.

Both scatterplots in **VIM** allow to specify whether the displayed variables are semi-continuous. In this case, only the non-zero observations are used for drawing the boxplots and computing the tolerance ellipses, respectively, as the large point mass of the marginal distributions at 0 would distort them otherwise. In addition, both scatterplots provide useful information on the structure of the missing values in the data.
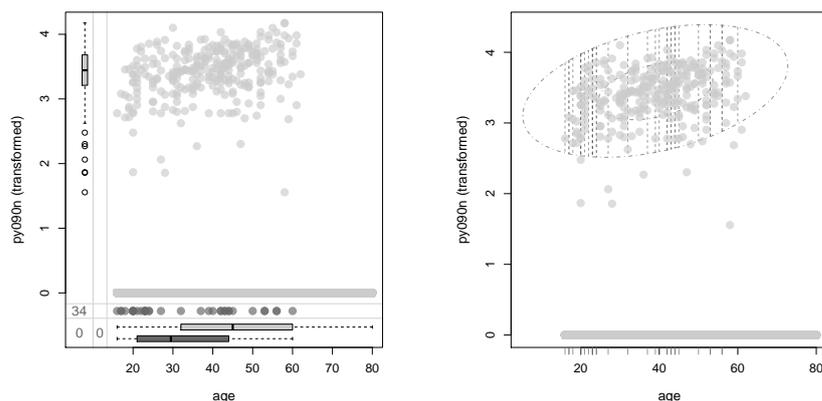
**Fig. 4** Scatterplots of the variables *age* and transformed *py090n* (unemployment benefits) in the EU-SILC data with information about missing values in the margins (left) and displayed as rug representation and lines limited by tolerance ellipses (right).

This knowledge may be used for generating close-to-reality data sets for simulation (for an application, see Todorov et al, 2011), and it allows subject matter specialists to learn about the characteristics of their data.

3.6 Scatterplot matrices

Scatterplot matrices are a straightforward generalization of scatterplots to the multivariate case. If the structure of missing values between each pair of variables is of interest, a marginplot matrix can be produced. Another type of scatterplot matrix is also available in **VIM**, in which observations with missing values in a certain variable or combination of variables are highlighted in the pairwise scatterplots, thus allowing for more than two-dimensional relations. Rug representations are drawn for observations with missing values in one variable. In the diagonal panels, density plots of highlighted and non-highlighted observations may be shown for comparison of the univariate distributions. By clicking in these diagonal panels, variables to be used for highlighting can be selected or deselected interactively. Information about the current selection is then printed on the R console.

Figure 5 displays an example using the mammal sleep data (see Section 3.1.3). Log-transformed *BodyWgt* (body weight), log-transformed *BrainWgt* (brain weight), *Dream* (amount of sleep with rapid eye movement) and *Sleep* (total amount of sleep) are plotted and observations with missing values in variable *Dream* are highlighted in all bivariate plots that do not include that variable. Log-transformation of the first two variables is thereby necessary to obtain more symmetric distributions. The plot shows that missing values in *Dream* do not occur for low values of body and brain weight. Furthermore, it is clearly visible that body and brain weight are highly correlated, and that the data contain some outliers with unusually high values in *Dream*.

In addition to the two scatterplot matrices described above, the underlying workhorse function in **VIM** can be used to create custom-made scatterplot matrices.
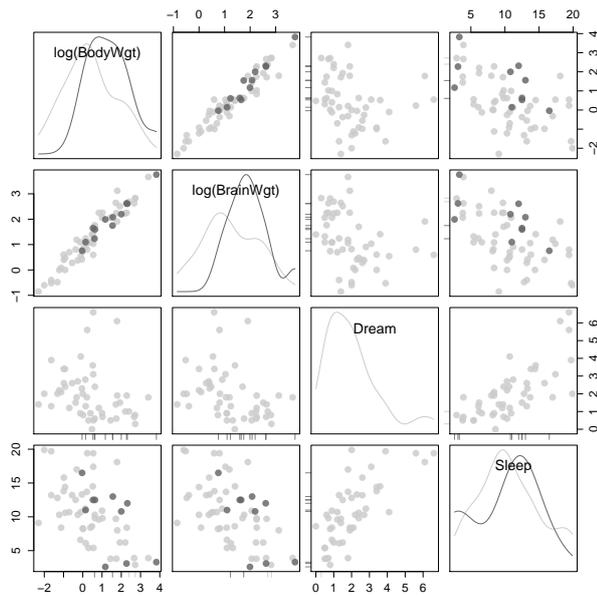
**Fig. 5** Scatterplot matrix of log-transformed *BodyWgt* (body weight), log-transformed *Brain-Wgt* (brain weight), *Dream* (amount of sleep with rapid eye movement) and *Sleep* (total amount of sleep) in the mammal sleep data. Observations with missing values in *Dream* are highlighted in all bivariate plots that do not include that variable.

### 3.7 Parallel coordinate plot

In parallel coordinate plots, each variable is first transformed to the same scale (usually the interval [0,1]). Each observation is then shown as a line, while the variables are represented by parallel axes (Wegman, 1990). Note that this plot is not limited to numeric variables, categorical variables can be used as well. In the latter case, the scale of the coordinate axis is broken down into $m$ equidistant points, where $m$ denotes the number of categories. Even though there is no specific order of the categories for nominal variables, observations with the same outcome are grouped at the same point of the corresponding coordinate axis. It is thus still possible to detect patterns in the data, which is the main aim of a parallel coordinate plot. 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. Alpha blending significantly increases the readability of parallel coordinate plots for large data sets. Similar plots are available in **Mondrian** and **GGobi** through linking. However, plotting observations with variables having missing values results in disconnected lines, making it impossible to trace the respective observations across the graph. As a remedy, lines may be connected outside the corresponding coordinate axes, e.g., above a horizontal line in the upper part of the plot that separates this representation of the missing values from the observed data (see Figure 6). Nevertheless, a caveat of this display is that it may draw attention away from the main relationships between the variables. **VIM** allows to interactively switch between this display and the standard display without the separate level for missing values by clicking in the top margin of the plot. Another interactive feature is
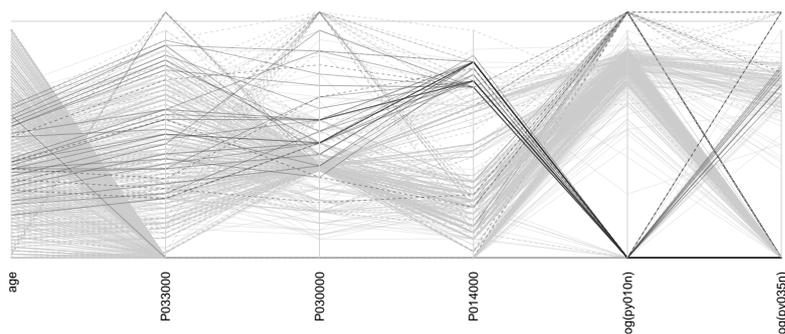
**Fig. 6** Parallel coordinate plot of selected variables of the subset of the EU-SILC data for the federal state Lower Austria. Dark grey lines indicate missing values in variable *py050n* (cash benefits or losses from self-employment). For missing values in the displayed variables, lines are connected above the corresponding coordinate axes, separated from the observed data by a horizontal line.

that the variables to be used for highlighting can be selected or deselected interactively by clicking near the corresponding coordinate axis. The current selection is thereby printed on the R console.

Figure 6 shows a parallel coordinate plot of selected variables of the subset of the EU-SILC data for the federal state Lower Austria, in which light grey lines refer to observed values and dark grey lines to missing values in the variable *py050n* (cash benefits or losses from self-employment). The heavily skewed income variables *py010n* and *py035n* were thereby log-transformed after the constant 1 had been added to obtain more symmetric distributions. Observations with missing values in any of the displayed variables are represented by dashed lines. The plot clearly shows that the highlighted observations behave differently than the main part of the data. In particular for the last three displayed variables, missing values in *py050n* occur primarily in a certain range.

### 3.8 Matrix plot

The *matrix plot* visualizes all cells of the data matrix by rectangles, similar to heat maps. It is a much more powerful extension of the function `imagmiss()` in the R package **dprep** (Acuna et al, 2009). Available data are visualized by a continuous color scheme. To compute the colors via interpolation, the variables are first transformed to the interval $[0, 1]$. This transformation is done in the same manner as for the parallel coordinate plot. Hence the matrix plot can be applied to numeric and categorical data. Missing values can then be visualized by a clearly distinguishable color. The implementation in **VIM** allows to use colors in the *HCL* or *RGB* color space. Advantages of the HCL color space for statistical graphics are discussed in Zeileis et al (2009). A simple way of visualizing the magnitude of the transformed available data is to apply a greyscale, which has the advantage that missing values can easily be distinguished by using a color such as red (see Figure 7). Nevertheless, the amount of information in the plot is determined by the type of variable. For nominal variables, there is no meaningful order of the categories. While the different colors allow to see for which
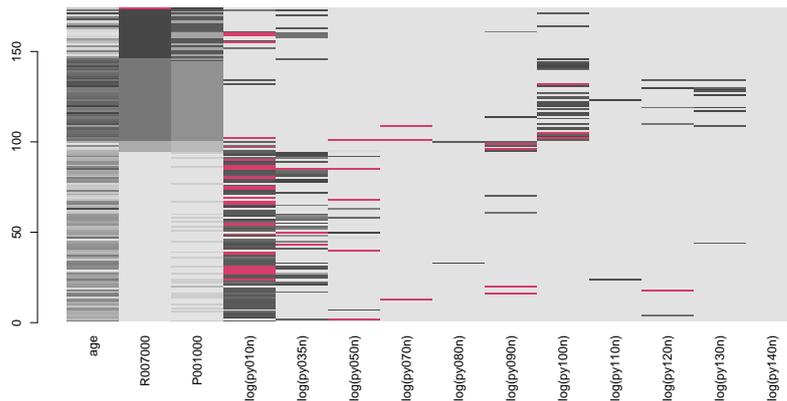
**Fig. 7** Matrix plot of selected variables of the subset of the EU-SILC data for the federal state Vorarlberg, sorted by variable *R007000* (occupation).

categories missing values in other variables occur predominantly, the ordering of the colors does not contain relevant information in this case. However, matrix plots are very powerful for finding the structure of missing values if the observations are sorted according to a selected variable. In **VIM**, this can be done interactively by clicking on the corresponding column of the plot.

Figure 7 presents a matrix plot of selected variables of the subset of the EU-SILC data for the federal state Vorarlberg, sorted by variable *R007000* (occupation). Note that more symmetric distributions of the heavily skewed income variables were obtained by log-transformation after the constant 1 had been added. Missing values in the income components clearly show a dependence, e.g., missing values in *py010n* (employee cash or near cash income) occur primarily for economically active persons (light grey cells).

3.9 Mosaic plot

Mosaic plots, introduced by Hartigan and Kleiner (1981, 1984), are graphical representations of multi-way contingency tables. The frequencies of the different cells of categorical variables are visualized by area-proportional rectangles (tiles). For constructing a mosaic plot, a rectangle is first split vertically at positions corresponding to the relative frequencies of the categories of a corresponding variable. Then the resulting smaller rectangles are again subdivided according to the conditional probabilities of a second variable. This can be continued for further variables accordingly. Hofmann (2003) provides an excellent description of the construction of mosaic plots and the underlying mathematical theory. Additional tiles can be used to display the frequencies of missing values. Furthermore, missing values in a certain variable or combination of variables can be highlighted in order to explore their structure. The implementation in **VIM** is based on the highly flexible `strucplot` framework of package **vcd** (Meyer et al, 2006, 2011). As an example, Figure 8 shows a mosaic plot of the variables *sex* and *R007000* (occupation) of the EU-SILC data, with missing values in *py010n* (em-

**Fig. 8** Mosaic plot of the variables *sex* and *R007000* (occupation) of the EU-SILC data, with missing values in *py010n* (employee cash or near cash income) highlighted in dark grey.

ployee cash or near cash income) highlighted in dark grey. This plot is thus useful for exploring the distribution of the missing values.

Similar implementations are available in Mondrian and MANET, although Mondrian does not show extra tiles for missing values.

3.10 Missing values in maps

If geographical coordinates are available for a data set, it can be of interest whether missingness in a variable corresponds to spatial patterns in a map. Values of one continuous or ordinal variable can be represented by growing dots in the map, reflecting their magnitude (Gustavsson et al, 1997). For the readability of such growing dot maps, it is very important that the area occupied by the symbols is reasonable with respect to the total area of the map. In the implementation in **VIM**, much attention has been directed towards finding sensible default values. More details on computing the dot sizes can be found in Gustavsson et al (1997).

Whenever an observation has missing values in another (highlighting) variable, the corresponding growing dots can be color coded (e.g., using a darker shade of grey as shown in the left plot of Figure 9). This allows conclusions about the relation of missingness to both the values of the variable of interest and their spatial location. Additionally, alpha blending can be used to prevent overplotting.

In the example in Figure 9 (left), the variable *Ca* of the Kola C-horizon data (see Section 3.1.2) is displayed with information about missing values in the chemical elements *As* or *Bi*. It shows that the missing values (dark grey points) have a regional dependency, i.e., they occur mainly in a certain part of the Kola project area. On the other hand, there does not seem to be any relation to the magnitude of the concentration of *Ca* (represented by the size of the dots).

The observations in the EU-SILC data are only assigned to one of the nine federal states of Austria, they do not have spatial coordinates. However, the proportion of
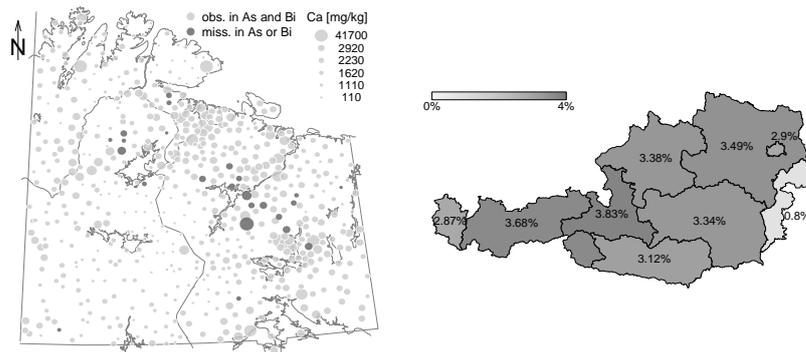
**Fig. 9** *Left*: Growing dot map of *Ca* in the Kola C-horizon data. Missing values in *As* or *Bi* show a spatial dependency. *Right*: Map of the nine federal states of Austria. Regions with a higher proportion of missing values (see the percentages in each region) in variable *py050n* (cash benefits or losses from self-employment) of the EU-SILC data receive a darker grey.

missing values or the absolute amount of missing values in a variable can still be visualized for the regions. In Figure 9 (right), the proportions of missing values in variable *py050n* (cash benefits or losses from self-employment) are coded according to a continuous color scheme, resulting in a darker grey for regions with a higher proportion of missing values. Since proportions of missing values are visualized, any type of variable can be used. Alternatively, equally spaced cut-off points may be used to discretize the color scheme. In package **VIM**, the sequential color palettes may thereby be computed in the *HCL* or the *RGB* color space. Further information on selecting colors in maps can be found in Harrower and Brewer (2003).

Both maps in **VIM** contain interactive features. By clicking on a data point in the growing dot map, detailed information about the corresponding observation is printed on the R console. Clicking inside a region in the regional map prints information about the included missing values and the corresponding sample size.

## 4 R package VIM

All visualization methods for missing values presented in Section 3 are implemented in the R package **VIM**. The figures in this paper were produced with **VIM** version 2.0.3 and R version 2.13.0. Since the development of this software is ongoing work, it is highly recommended to always use the latest version available from the comprehensive R archive network (CRAN, `http://cran.r-project.org`).

A graphical user interface (GUI), which has been developed using the R package **tcltk** (R Development Core Team, 2011), allows easy handling of the functions for quick data exploration. The full potential of **VIM** can be unleashed on the R command line. Figure 10 shows the **VIM** GUI. For visualization, only the *Data*, *Visualization* and *Options* menus are important. The *Data* menu allows to select a data set from the R workspace or load data into the workspace from `RData` files. Furthermore, it can be used to transform variables, which are then appended to the data set in use. Commonly
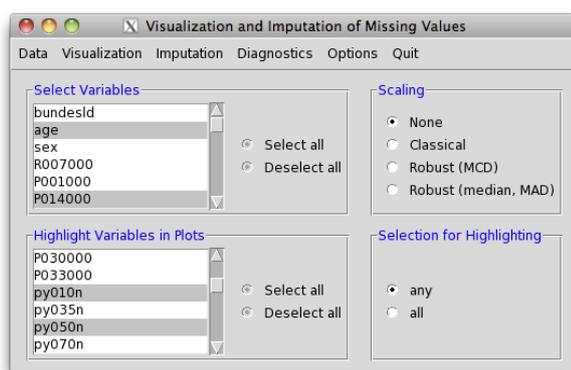
**Fig. 10** The **VIM** GUI. Here, the Austrian EU-SILC data set is already chosen and some variables are selected.

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. 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.

After a data set has been chosen, variables can be selected in the main window, along with a method for scaling. 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. For 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.

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

Last, but not least, the *Options* menu allows to set the colors and alpha channel to be used in the plots. In addition, it contains an option to embed multivariate plots in Tcl/Tk windows. This is useful if the number of observations or variables is large, because scrollbars allow to move from one part of the plot to another.

Interactive features are implemented in various plot methods. There are, however, limited possibilities for interactive graphics in standard R. Interactivity with respect to linked graphics is not in focus of this contribution and is possible future work once a stable version of package **ix** (Urbanek, 2011) for extensible interactive graphics in R is released.

## 5 Summary

The proposed visualization methods allow to combine information about the data with information about missingness in a certain variable or a certain combination of variables. All methods are implemented in the R package **VIM** and various plots thereby offer interactive features. The information resulting from the different graphics can be

used for detecting missing value mechanisms. The plots provide valuable information about the characteristics of missing values in the data set. This knowledge can then be used by subject matter specialists or statisticians in the data preparation procedure. The information on the structure of missing values can also be used to generate close-to-reality data as done in the AMELI project (`http://ameli.surveystatistics.net`). Realistic nonresponse mechanisms can then be simulated in order to evaluate imputation methods or to investigate the influence of missing values on point and variance estimates. **VIM** can easily be used within R without the need to install additional software. A simple graphical user interface allows an easy handling of the implemented plots. Moreover, users have the possibility to use the whole power of the statistical environment R at the same time. Even a re-implementation of some plot methods might be of high interest for the users. Using **VIM**, it is thus possible to explore and analyze the structure of missing values in data, as well as to produce high-quality graphics for publications.

# References

Acuna E, members of the CASTLE group at UPR-Mayaguez (2009) **dprep**: Data preprocessing and visualization functions for classification. URL `http://math.uprm.edu/ edgar/dprep.html`, R package version 2.1

Aitchison J (1986) The Statistical Analysis of Compositional Data. John Wiley & Sons, Hoboken, New Jersey

Allison T, Cichetti D (1976) Sleep in mammals: ecological and constitutional correlates. Science 194(4266):732–734

Box G, Cox D (1964) An analysis of transformations. J Royal Stat Soc B 26:211–252

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

Dempster A, Laird N, Rubin D (1977) Maximum likelihood for incomplete data via the EM algorithm (with discussions). J Royal Stat Soc B 39(1):1–38

Eaton C, Plaisant C, Drizd T (2005) Visualizing missing data: Graph interpretation user study. In: Costabile M, Paternò F (eds) Human-Computer Interaction - INTER-ACT 2005, Springer, Heidelberg, Lecture Notes in Computer Sciences, pp 861–872, ISBN 978-3-540-28943-2

Gustavsson N, Lampio E, Tarvainen T (1997) Visualization of geochemical data on maps at the Geological Survey of Finland. J Geochem Explor 59(3):197–2007

Harrower M, Brewer C (2003) ColorBrewer.org: An online tool for selecting colour schemes for maps. Cartogr J 40(1):27–37

Hartigan J, Kleiner B (1981) Mosaics for contingency tables. In: Eddy W (ed) Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, Springer, New York, pp 268–273

Hartigan J, Kleiner B (1984) A mosaic of television ratings. Am Stat 38(1):32–35

Hofmann H (2003) Constructing and reading mosaicplots. Comput Stat Data Anal 43(4):565–580

Hofmann H, Theus M (2005) Interactive graphics for visualizing conditional distributions, unpublished manuscript

Hron K, Templ M, Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. Comput Stat Data Anal 54(12):3095 – 3107

Josse J, Pagès J, Husson F (2011) Multiple imputation in principal component analysis. Adv Data Anal and Classif 5(3):231–246

Little R, Rubin D (2002) Statistical Analysis with Missing Data, 2nd edn. John Wiley & Sons, Hoboken, New Jersey, ISBN 0-471-18386-5

Meyer D, Zeileis A, Hornik K (2006) The `strucplot` framework: Visualizing multi-way contingency tables with **vcd**. J Stat Softw 17(3):1–48, URL `http://www.jstatsoft.org/v17/i03`

Meyer D, Zeileis A, Hornik K, Friendly M (2011) **vcd**: Visualizing Categorical Data. URL `http://CRAN.R-project.org/package=vcd`, R package version 1.2-11

Osborne J (1999) Notes on the use of data transformations. Pract Assess Res Eval 8(6):212–223, URL `http://pareonline.net/getvn.asp?v=8&n=6`

Perrotta D, Riani M, Torti F (2009) New robust dynamic plots for regression mixture detection. Adv Data Anal Classif 3:263–279

Raghunathan T, Lepkowski J, Van Hoewyk J, Solenberger P (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol 27(1):85–95

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0

Reimann C, Filzmoser P, Garrett R, Dutter R (2008) Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons, Hoboken, New Jersey

Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223

Rubin D (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin D (2004) Multiple Imputation for Nonresponse in Surveys, Wiley Classics Library edn. John Wiley & Sons, Hoboken, New Jersey, ISBN 0-471-65574-0

Schafer J (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London, ISBN 0-412-04061-1

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

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

Swayne D, Lang D, Buja A, Cook D (2003) GGobi: evolving from XGobi into an extensible framework for interactive data visualization. Comput Stat Data Anal 43(4):423–444

Templ M, Alfons A, Kowarik A (2011a) **VIM**: Visualization and Imputation of Missing Values. URL `http://CRAN.R-project.org/package=VIM`, R package version 2.0.4

Templ M, Kowarik A, Filzmoser P (2011b) Iterative stepwise regression imputation using standard and robust methods. Comput Stat Data Anal 55(10):2793–2806

Theus M (2002) Interactive data visualization using mondrian. J Stat Softw 7(11):1–9, URL `http://www.jstatsoft.org/v07/i11`

Theus M, Hofmann H, Siegl B, Unwin A (1997) MANET - extensions to interactive statistical graphics for missing values. In: New Techniques and Technologies for Statistics II, IOS Press, pp 247–259, ISBN 90 5119 326 9

Todorov V, Templ M, Filzmoser P (2011) Detection of multivariate outliers in business survey data with incomplete information. Adv Data Anal Classif 5(1):37–56

Unwin A (1994) Computational Statistics, Physica-Verlag, Heidelberg, chap REGARDing Geographic Data., pp 315–326

Unwin A, Wills G, Haslett J (1990) REGARD - graphical analysis of regional data. In: Proceedings of the Section on Statistical Graphics, American Statistical Association, pp 36–41

Unwin A, Hawkins G, Hofmann H, Siegl B (1996) Interactive graphics for data sets with missing values: MANET. J Comput Graph Stat 5(2):113–122

Urbanek S (2011) **Acinonyx**: iPlots Extreme. URL `http://www.RForge.net/Acinonyx/`, R package version 3.0-0

Vanden Branden K, Verboven S (2009) Robust data imputation. Comput Biol Chem 9(1):7–13

Wegman E (1990) Hyperdimensional data analysis using parallel coordinates. J Am Stat Assoc 85(411):664–675

Young F (1996) ViSta: The Visual Statistics System. UNC L.L. Thurstone Psychometric Laboratory Research Memorandum 94-1(c)

Young F, Valero-Mora P, Friendly M (2006) Visual Statistics. Seeing Data with Dynamic Interactive Graphics. John Wiley & Sons, Hoboken, New Jersey, ISBN 978-0-471-68160-1

Zeileis A, Hornik K, Murrell P (2009) Escaping RGBland: Selecting colors for statistical graphics. Comput Stat Data Anal 53(9):1259–1270