Institut f. Statistik u. Wahrscheinlichkeitstheorie

1040 Wien, Wiedner Hauptstr. 8-10/107
AUSTRIA
http://www.statistik.tuwien.ac.at

Visualization of missing values using the R-package VIM

M. Templ and P. Filzmoser

Forschungsbericht CS-2008-1
Mai 2008

Kontakt: P.Filzmoser@tuwien.ac.at

Visualization of Missing Values using the R-Package VIM

Matthias Templ[1,2] and Peter Filzmoser[1]

[1] Department of Statistics and Probability Theory, Vienna University of Technology, 1040 Vienna, Austria

[2] Statistics Austria, 1110 Vienna, Austria

Abstract: This paper introduces new tools for the visualization of missing values. The tools can be used for exploring the data and the structure of the missing values. Depending on this structure, the tools can be helpful for identifying the mechanism generating the missings. This knowledge is important for selecting an appropriate imputation method to reliably estimate the missing values.

The visualization tools are implemented in the R-package VIM (visualization and imputation of missing values). A graphical user interface allows an easy handling of the plot methods. The developed tool can be used for data from official statistics, but also for data from various other fields. Special attention is given to data with spatial coordinates, and in that case the missing data information can be displayed with maps.

1 Introduction

Data often contain missing values, and the reasons are manifold. Missing values occur when measurements fail, or in case of non-respondents in surveys, when analysis results get lost, or when measurements do not fulfill some prior knowledge (edits) of the data, i.e. 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 which must be re-analyzed but where the soil samples are exhausted. Examples for missing values in official statistics are respondents who deny information about their income, small companies which do not report their turnover, values which do not fulfill pre-defined editing rules, etc.

Missing values are often of great interest and they must be replaced by meaningful values. Moreover, most of the standard statistical methods can only be applied to complete data, and deleting whole columns or rows of data matrices where missing values appear would result in a loss of important available information. The estimation of missing values is known under the name *imputation* (Little and Rubin, 1987). In order to be able to choose a proper imputation method one must be aware of the missing data mechanism(s). The quality of the imputed values depends on the imputation itself and on the imputation method used.

Although there exists a comprehensive literature on the estimation of missing values (see, e.g., Little and Rubin, 1987; Schafer, 1997), the visualization of missings is treated only in a few papers (e.g. Eaton et al., 2005; Cook and Swayne, 2007). This is also reflected in the statistical software. Visualization tools for missing values are rarely or not implemented in SAS, SPSS, or STATA, and only few plots are available in R and GGOBI (see, e.g., Cook and Swayne, 2007).

In this paper we introduce several graphical presentations to visualize missing values. We will demonstrate the usefulness of the plots with real data (Section 2). With an appropriate visualization of the missings it can also be possible to detect the missing values mechanism (see e.g. in Figure 8) which is important for the selection of an appropriate imputation method.

All visualization tools shown in the following are implemented in the R-package VIM (**V**isualization and **I**mputation of **M**issing Values) which has been written by the first author of this paper. A screen-shot of the implementation and few comments about the usage of the package are given in Section 3.

1.1 Missing Values Mechanisms

There are three important cases to distinguish that are the responsible generating processes behind the missing values (see Rubin, 1976; Little and Rubin, 1987; Schafer, 1997). The missing values are **M**issing **C**ompletely **A**t **R**andom (MCAR) if the distribution of missingness neither depends on the observed part X_{obs} nor on the missing part X_{miss}. Thus the probability of missingness is given by

$$P(X_{miss}|X) = P(X_{miss})$$

with the complete data $X = (X_{obs}, X_{miss})$. In other words, the missing data mechanism does not depend on the variable of interest, nor on any other variable which is observed in the data set.

When the distribution of missingness depends on the observed part X_{obs}, the missing values are said to be **M**issing **A**t **R**andom (MAR), and the probability of missingness is

$$P(X_{miss}|X) = P(X_{miss}|X_{obs}). \tag{1}$$

Hence the distribution of missingness depends not on the missing part X_{miss}.

When 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 X_{miss}, the missing data are said to be **M**issing **N**ot **A**t **R**andom (MNAR). This relates to the equation

$$P(X_{miss}|X) = P(X_{miss}|(X_{obs}, X_{miss})).$$

Hence, the missings can not be fully explained by the observed part of the data.

A practical example for the missing values mechanism which is adequate for the data used in this paper is given by Little and Rubin (1987). Considering two variables `age` and `income` the data are MCAR if the probability of missing 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 respondents (e.g. more missings for higher age) but does not vary according to the income of respondents with the same age (e.g. for a considered age class the distribution of the missings show a random pattern), then the missings in variable income are MAR. However, if the probability that income is recorded varies according to income for those with the same age (e.g. more missings for high income than for low income in certain age classes), then the missings in variable `income` are MNAR (Little and Rubin, 1987). Naturally, MNAR is difficult to be detected (see Section 1.2).

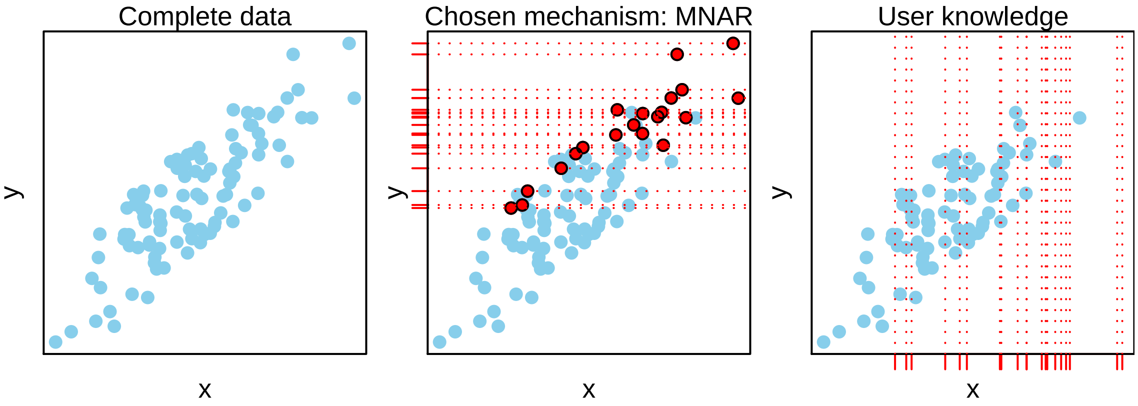


Figure 1: Simulated bivariate data set. LEFT: Complete data. MIDDLE: Red points are chosen as missings in y depending on the value of y (MNAR). RIGHT: Information is only available for x-values in practice.

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

1.2 Limitations for the Detection of the Missing Values Mechanism

In practice it is often difficult to detect the missing values mechanism exactly, because this would require the knowledge of the missing values themselves (Little and Rubin, 1987). In the following we want to give a simple example in order to show the limitations for the detection of the missing values mechanism.

Figure 1 shows a highly correlated bivariate data set. From the complete data (graphic on the left side of the figure) some observations are marked as missings in y (graphic in the middle of the figure) depending on the value of y (the higher y the higher the probability of missingness). So, the missing values mechanism is constructed as MNAR. In practice, however, we only know the x-part of the observations with missing values in y (graphic on the right of the figure) and can only observe that for increasing x-values the amount of missingness also increases. Therefore, we will assume a MAR situation knowing that this could also be a MNAR situation, i.e. we cannot distinguish between MAR and MNAR. We know from our construction that after taking the relationship between x and the missing data pattern into account, the probability of missingness in y still depends on y (see the graphic in the middle of Figure 1 where for approximately equal x-values only high y-values were set to be missing). However, in real world situations we will determine a MAR situation for this example because of the high correlation between x and y (see, e.g., Little and Rubin, 1987). This results in good estimation by using well-established imputation methods for MAR situations (see, e.g., Dempster et al., 1977).

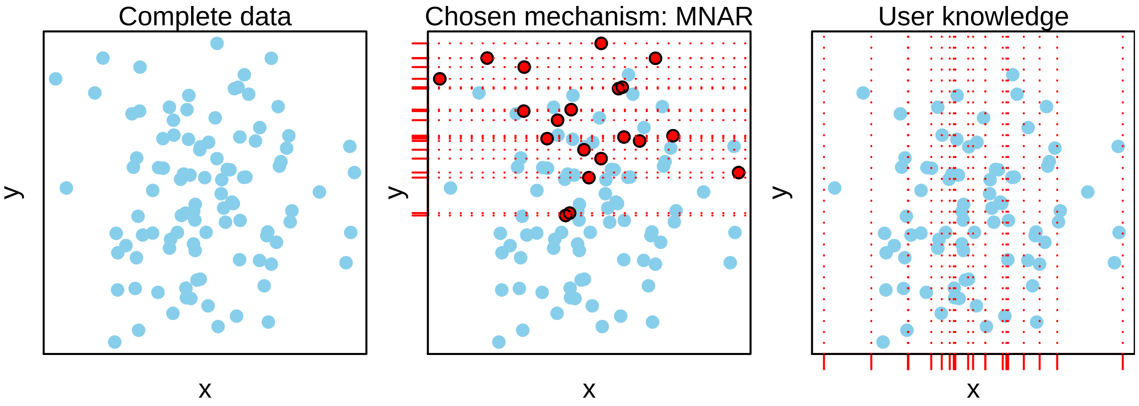


Figure 2: Simulated bivariate data set. LEFT: Complete data. MIDDLE: Red points are chosen as missings in y depending on the value of y (MNAR). RIGHT: Information is only available for x-values in practice.

In Figure 2 we see a similar picture as in Figure 1 but with uncorrelated variables. Some y-values are again marked as missing depending on the value of y (MNAR). In practice, however, we will detect a MCAR situation because the missingness seems to be completely independent from the data values. On the other hand, we know that this could also be a MNAR situation, and so we can not distinguish between MCAR and MNAR.

We can summarize that, for a bivariate data set where one variable contains missings which are generated with the MNAR mechanism, we can have the following situations:

(a) The variables are correlated: MNAR would be classified as MAR, and thus we only detect MCAR and MAR.

(b) The variables are uncorrelated: MNAR would be classified as MCAR, and thus we only detect MCAR and MAR.

Multivariate data with missing values in several variables can make it even more complicated to distinguish between the missing values mechanisms. The situation can become even worse in case of outliers, inhomogeneous data or very skewed data distributions.

2 Visualization Methods for Missing Values

The visualization tools proposed in this section do not rely on any statistical model assumptions. The tools are made available as R-package VIM, and through a graphical user interface they are easy to handle.

We illustrate the visualization tools on a subset of the European Survey of Income and Living Conditions (EU-SILC) from Statistics Austria from 2004, see Table 1. This very famous and complex data set is mainly used for measuring poverty and for monitoring

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

name	meaning
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	survivor benefits
py120n	sickness benefits
py130n	disability benefits
py140n	education-related allowances
pek_n	net income
pek_g	gross income
P001000	employment situation
r007000	occupation
P033000	years of employment
P029000	hours worked per week, miscellaneous employment
P014000	profession
bundesld	region
age	age
sex	sex

the Lisbon 2010 strategy of the European Union. This data set includes a high amount of missings which were imputed with model based imputation methods (Statistics Austria, 2006). Since a high amount of missings are not MCAR one has to think about which variables should be included for imputation. The proposed visualization tools are helpful for this decision.

2.1 Aggregation Plot

It is often of interest, how many missing values are contained in the single variables. Even more interesting, there can be certain combinations of variables with a high number of missing values. Figure 3 shows this information. The plot on the left hand side shows a bar for each considered variable, and the bar height corresponds to the number of missing values in the variable. The *aggregation plot* on the right hand side shows for the same variables (horizontal axis) all different combinations that are present in the observations with missing and non-missing values (vertical axis). The color red indicates missingness, the color blue represents available data. The numbers to the right are the frequencies of observations to the corresponding combinations. For example, the top row represents a combination where the first three variables of an observation have missing values and

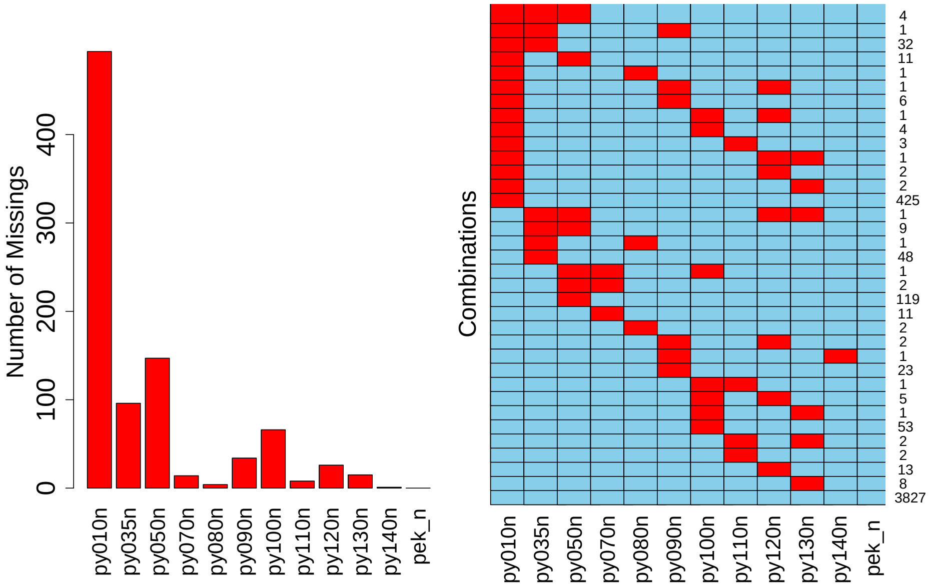


Figure 3: Number of missing values for a subsample of the EU-SILC data from Statistics Austria. LEFT: Barplots for the number of missing values in each variable. RIGHT: *Aggregation plot* showing all combinations of missing (red) and non-missing (blue) parts in the observations, and the corresponding frequencies.

the remaining variables have no missings. This combination appears 4 times in the data. Another example: The last row reflects the 3827 observations that have no missing values.

As a result we observe an exceptionally high number of missings for variable *py010n* (employee cash or near cash income). The combination with variable *py035n* (contributions to individual private pension plans) still has 32 missing values.

2.2 Matrix Plot

The *matrix plot* visualizes all cells of the data matrix by horizontal lines. Red lines are drawn for missing values, and a grey scale is used for the available data values. For the grey scale the data are first scaled to the interval [0,1] for each variable by subtracting the mean and dividing by the range. Then small values are assigned a light grey, high values a dark grey and values equal to 0 are colored in white. Additionally, the observations can be sorted by the magnitude of a selected variable.

Figure 4 shows a matrix plot of a subset of the EU-SILC data where the sorting was done for variable *pek_n* (net income). It can be seen that the higher the net income, the more missing values in the variables *py010y* (employee cash or near cash income), and *py050n* (employees income). Thus, the missing data mechanism was detected to be MAR for these two variables which should be considered when applying imputation methods on these variables.

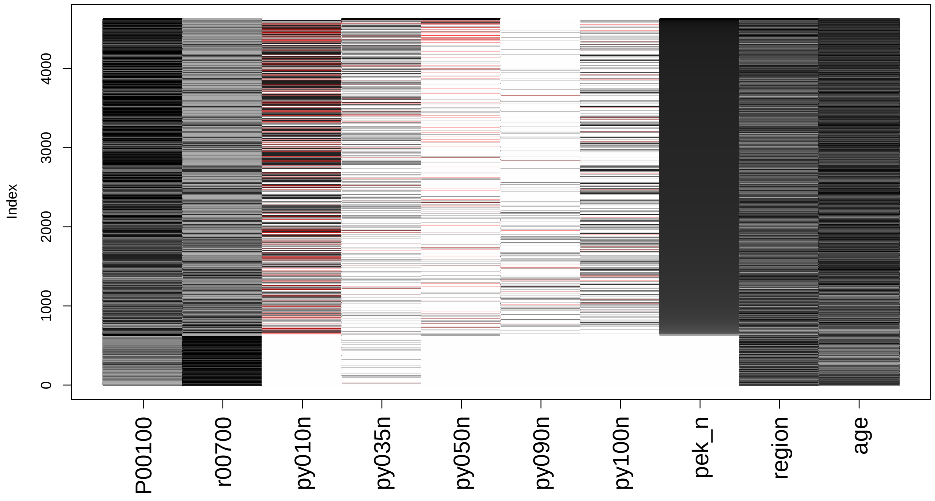


Figure 4: Matrix plot of a subset of the EU-SILC data which are sorted by variable *pek_n*.

2.3 Spineplot with Missings

When plotting a histogram of a variable one could show the amount of missings of a second variable by splitting each bin into two parts. This is shown in Figure 5 (left), where the histogram represents the frequencies of the variable *age* and the bins (age classes) are split according to the numbers of missing (red) and observed (blue) values of variable it py010n (employee cash or near cash income) for the corresponding age class.

Another way of presenting this information can be done by the *spineplot* (see, e.g., Hummel, 1996). Figure 5 (right) shows such a spineplot for the same two variables as used in Figure 5 (left). The horizontal axis is scaled according to relative frequencies of the age classes. The vertical axis shows the proportion of missing (red) and observed (blue) values of variable *py010n* (employee cash or near cash income) for each age class, and thus they sum up to 1. Since the area of the bin for each age interval reflects the number of values, it is now possible to compare the areas for the missing proportions among the different age classes. For example, strictly decreasing red areas for increasing age would indicate a MAR situation for variable *py050n* because the proportion of missing values in *py050n* would then also strictly decrease. Another typical MAR situation could be found when using variable *py100n* (old-age benefits) and *age*. The higher the age the higher the probability of missing values in *py100n*. In this spineplot, however, one can only see that for age classes > 60 years the proportion of missing values of the employees' income becomes very small, which is logical.

2.4 Scatterplots with Missing Data Information in the Margins

Additionally to a standard scatterplot that shows the data values for a pair of variables, information about missing values can be shown. Figure 6 uses the variables *pek_n* (net income) and *py130n* (unemployability) of the EU-SILC data and shows the scatterplot. In addition, boxplots for missing (red) and available (blue) data are shown in the outer

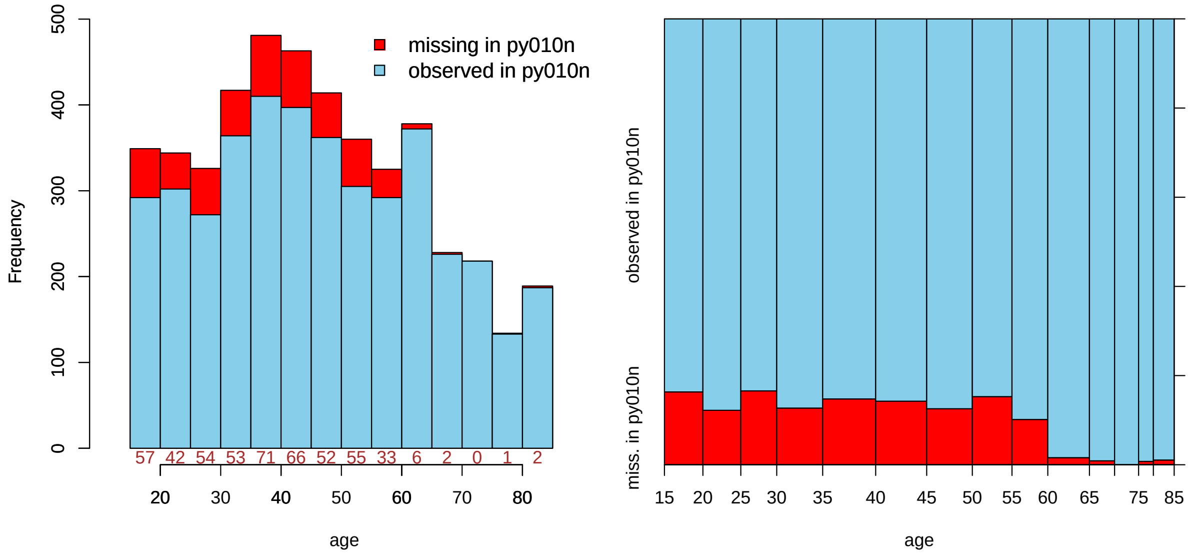


Figure 5: Histogram (left) and Spineplot (right) of *age* with color coding of the amount of missing and available data in variable *py010n* (employee cash or near cash income).

plot margins. For example, along the horizontal axis the two parallel boxplots both represent the variable net income, but the red boxplot is for those values of the net income, where no values for unemployability are available, and the blue boxplot for net income values, where also information for unemployability is available. A comparison of the two boxplots can indicate the missing data mechanism. In this example the net income for persons that provided no information for unemployability seems to be higher than that where information is available.

The plot additionally shows a univariate scatterplot of the values that are missing in the second variable (red points), and the frequency of these values by a number (lower left). The number in the lower left corner (here 0) is the number of observations that are missing in both variables.

This kind of bivariate scatterplot can easily be extended to a scatterplot matrix representing all combinations of a set of variables by such scatterplots. Another extension already included in the R-package VIM is that the information about missingness of more than one variable is represented, allowing for more than two-dimensional relations.

2.5 Parallel Coordinate Plots with Missings

Parallel coordinate plots show each observation of the scaled data (usually interval scaled in [0,1]) by a line, where the variables are presented by parallel axes (Wegman, 1990). Similar to previous plots, the information of missingness of another variable can be color coded. Figure 7 shows a parallel coordinate plot of a subset of the EU-SILC data where the lines color refers to observations which are missing (red) of available (blue) for variable *py050n* (employees income). Missingness in *py050n* is related with several of the

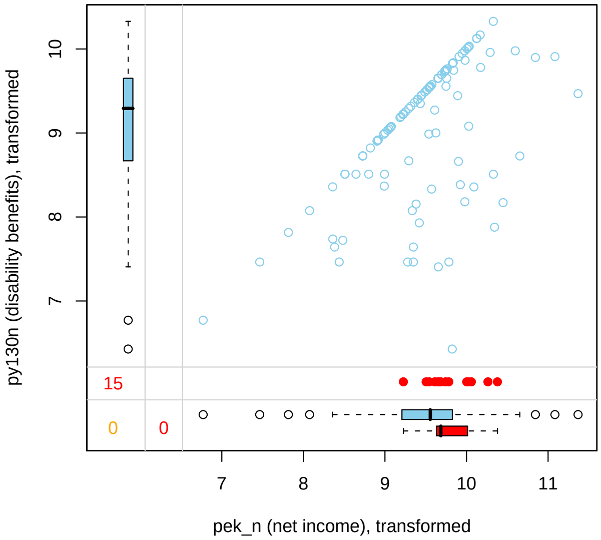


Figure 6: Scatterplot of net income (*pek_n*) and unemployability (*py130n*) with missings data information in the plot margins.

presented variables. In Figure 7 we can see that a high portion of small values in *P033000* (years of employment) implies missings in *py050n*. Additionally, missings in *py050n* occur only for employees which have more than one employment (in this case the values were set to 0 in variable *P029000*). Furthermore, Figure 7 shows that the amount of missings depends on the actual values for the variables *P001000* (different employment situation) and *bundesld* (region). Finally, for variable *pek_g* (gross income) missing values only occur in a certain range.

2.6 Parallel Boxplots for Missing Values

Boxplots were already used in Figure 6 for comparing the values of a variable that are either available or missing in a second variable. This can be extended for a comparison of the information of missingness of several variables. Figure 8 shows the values of variable *pek_n* (net income) in form of a boxplot (left). The other boxplots shown in the figure also refer to the values of *pek_n*, but they are grouped according to missingness (red) or non-missingness (blue) of each observation in another variable. Some of the boxplots show a clear dependence between the magnitude of the values of *pek_n* and the presence of missing values. For example, missing values in variable *py080n* (annuity) appear especially for high values of the net income. This indicates a clear MAR situation for missings in *py080n*.

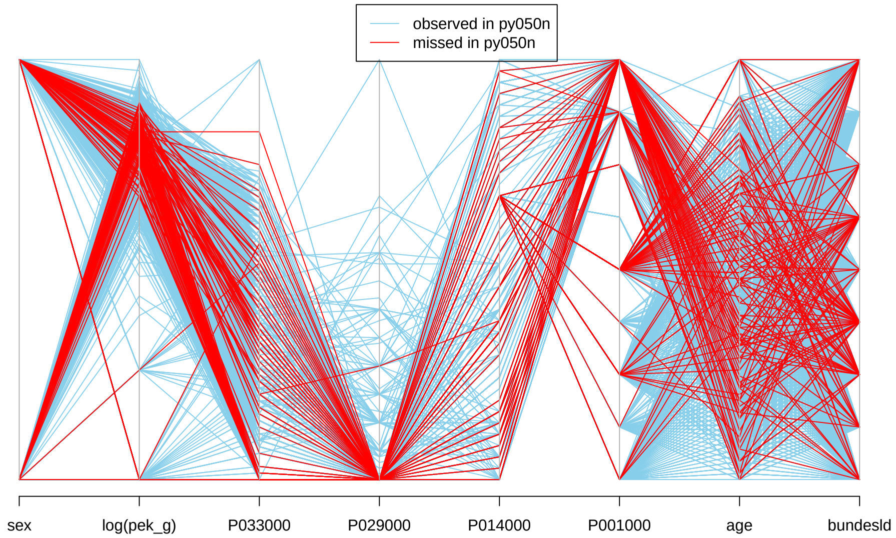


Figure 7: Parallel coordinate plot for a subset of the EU-SILC data. The color indicates missing values in variable *py050n* (employees income).

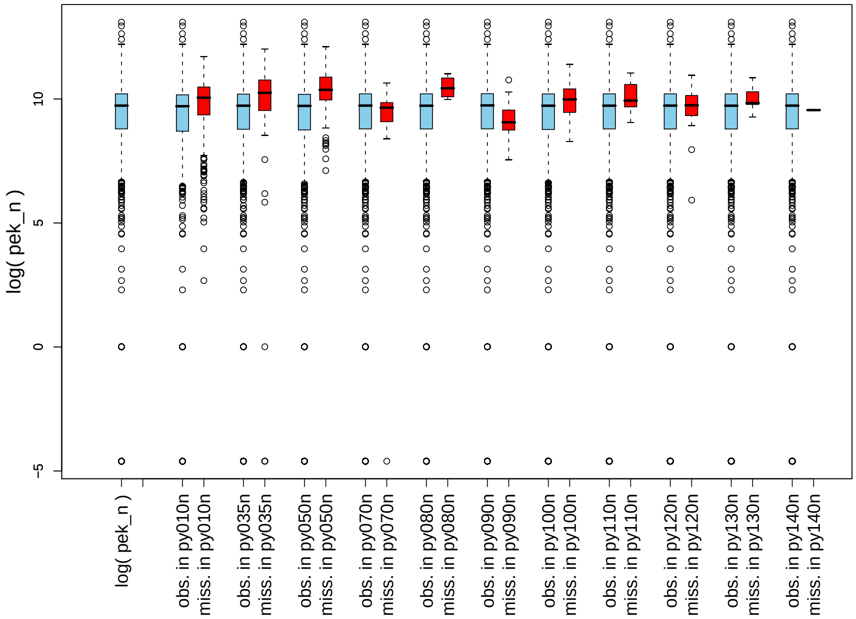


Figure 8: The values of a specific variable (here *pek_n* – net income) are grouped according to missingness in another variable and presented in parallel boxplots. This grouping is done for several variables of the EU-SILC data.

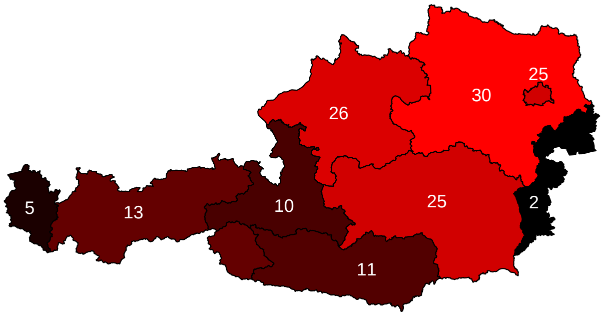


Figure 9: The number of missing values in variable *py050n* (employees income) are computed for the nine political regions of Austria. Regions with a higher amount of missing values (see the numbers in each region) receive a higher portion of red.

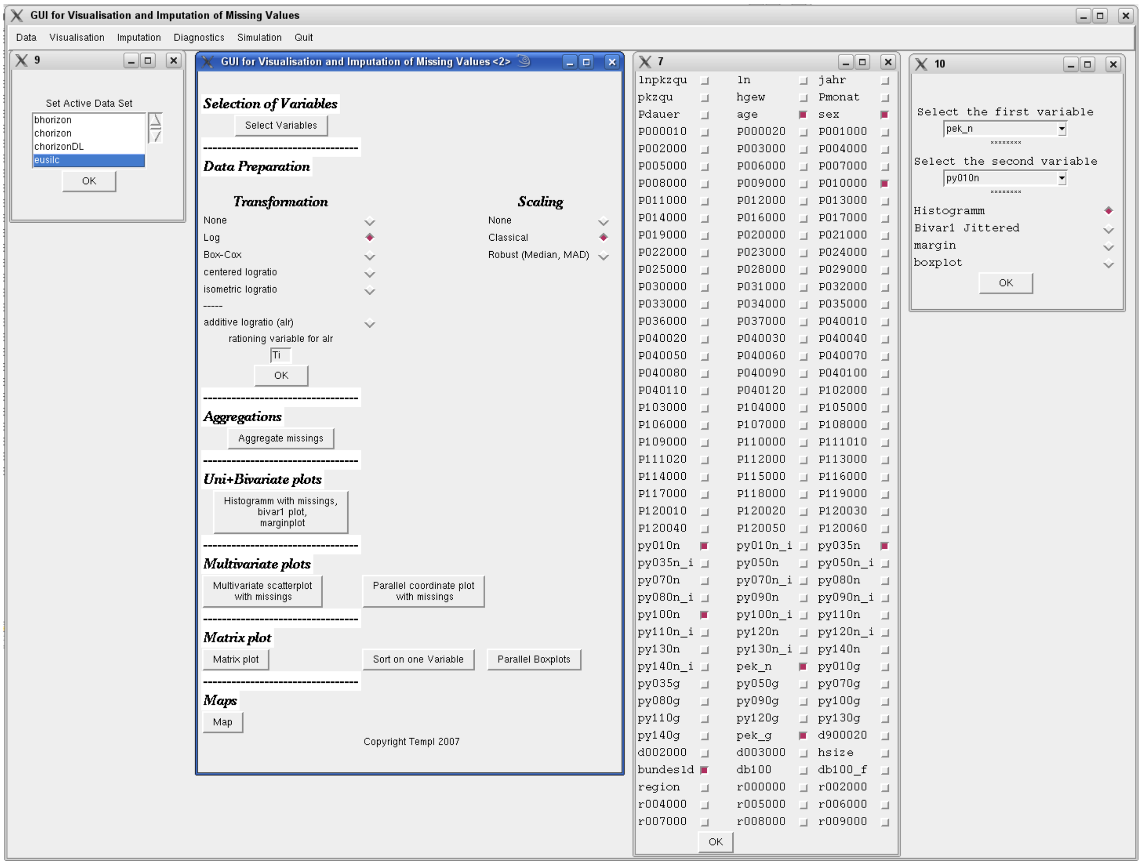
2.7 Plot Missings in Maps

If geographical coordinates are available for a data set, it can be of interest to check whether missingness of a variable corresponds to spatial patterns in a map. For example, the observations of a variable of interest could be drawn by growing dots in the map, reflecting the magnitude of the data values, and missingness in a second variable could be color coded. This allows conclusions about the relation of missingness to the values of the variable of interest, and to the spatial location.

For the EU-SILC data only the assignment of each observation to the nine political regions of Austria is available, and not the spatial coordinates. Therefore, we can only visualize the amount of missingness of a variable in the map. Figure 9 shows the map of Austria with the nine political regions. The number of missing values of variable *py050n* (employees income) is coded according to an *rgb* color scheme, resulting in a higher portion of red for regions with a higher number of missing values for the considered variable. Additionally, the numbers of missing values are shown in the regions.

3 R-Package VIM

All tools shown in Section 2 for visualizing missing values have been implemented in the R-package VIM. A graphical user interface (GUI) which was developed with the help of the R-package `tcltk` (R Development Core Team, 2007) allows an easy handling of the functions. Figure 10 shows the main menu. For the visualization of the missing values only the *Data* and the *Visualization* menu are important. In the *Data* menu one can load data into the R workspace, select the data of interest, load a background map (optional), and specify coordinates (optional). After selecting a data set a new frame opens (the right frame of Figure 10) where variables can be selected. In the middle left area of Figure



Transformation: ○ None, ● Log, ○ Box-Cox, ○ centered logratio, ○ isometric logratio, -----, ○ additive logratio (alr); rationing variable for alr: Ti

Scaling: ○ None, ● Classical, ○ Robust (Median, MAD)

lnpkzqu	☐	ln	☐	jahr	☐
pkzqu	☐	hgew	☐	Pmonat	☐
Pdauer	☐	age	☒	sex	☒
P000010	☐	P000020	☐	P001000	☐
P002000	☐	P003000	☐	P004000	☐
P005000	☐	P006000	☐	P007000	☐
P008000	☐	P009000	☐	P010000	☒
P011000	☐	P012000	☐	P013000	☐
P014000	☐	P016000	☐	P017000	☐
P019000	☐	P020000	☐	P021000	☐
P022000	☐	P023000	☐	P024000	☐
P025000	☐	P028000	☐	P029000	☐
P030000	☐	P031000	☐	P032000	☐
P033000	☐	P034000	☐	P035000	☐
P036000	☐	P037000	☐	P040010	☐
P040020	☐	P040030	☐	P040040	☐
P040050	☐	P040060	☐	P040070	☐
P040080	☐	P040090	☐	P040100	☐
P040110	☐	P040120	☐	P102000	☐
P103000	☐	P104000	☐	P105000	☐
P106000	☐	P107000	☐	P108000	☐
P109000	☐	P110000	☐	P111010	☐
P111020	☐	P112000	☐	P113000	☐
P114000	☐	P115000	☐	P116000	☐
P117000	☐	P118000	☐	P119000	☐
P120010	☐	P120020	☐	P120030	☐
P120040	☐	P120050	☐	P120060	☐
py010n	☒	py010n_i	☐	py035n	☒
py035n_i	☐	py050n	☐	py050n_i	☐
py070n	☐	py070n_i	☐	py080n	☐
py080n_i	☐	py090n	☐	py090n_i	☐
py100n	☒	py100n_i	☐	py110n	☐
py110n_i	☐	py120n	☐	py120n_i	☐
py130n	☐	py130n_i	☐	py140n	☐
py140n_i	☐	pek_n	☒	py010g	☐
py035g	☐	py050g	☐	py070g	☐
py080g	☐	py090g	☐	py100g	☐
py110g	☐	py120g	☐	py130g	☐
py140g	☐	pek_g	☒	d900020	☐
d002000	☐	d003000	☐	hsize	☐
bundesld	☒	db100	☐	db100_f	☐
region	☐	r000000	☐	r002000	☐
r004000	☐	r005000	☐	r006000	☐
r007000	☐	r008000	☐	r009000	☐

Select the first variable: pek_n

Select the second variable: py010n

● Histogramm, ○ Bivar1 Jittered, ○ margin, ○ boxplot

Figure 10: The VIM package GUI menu together with the the active data selection frame, the visualization workout sheet, the variable selection frame and the combo box for the variable selection used for bivariate plots.

10 one can see the possible options and methods which can be chosen for visualizing the active data set.

Besides the selection of variables one can transform and scale the data with various techniques. The log-transformation or the box-cox transformation (Box and Cox, 1964) which are common transformations for data from official statistics can be used, but also various other transformations are implemented, especially for geochemical data. When selecting *bivariate plots* in the graphical user interface, a new selection of two variables can be done simply by clicking on two variables in a combo box.

4 Conclusions

The detection of missing values mechanisms is usually done by statistical tests or models. Visualization of the missing values can support the test decision, but also reveals more details about the data structure. Especially statistical requirements for a test can be checked graphically, and problems like outliers or skewed data distributions can be discovered.

A graphical user interface was developed and implemented in R to produce several possibilities for the visualization of data with missing values. The tools allow to combine the information of missingness in a variable with other variables. Some plots allow for an interactive handling, like the selection of the variable used for sorting in the matrix plot (see Figure 4). The information resulting from the different graphics can be used for detecting the missing values mechanism, and thus for selecting an appropriate method for data imputation.

The tool can be used for data from essentially any field. If data with spatial coordinates are available, the missing value information can be presented in maps. It is also possible to load a background map and show it on the plot. Spatial patterns of missingness can be very instructive for example in environmental sciences.

The package VIM is available on the comprehensive R archive network (CRAN).

References

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

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.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussions). *Journal of the Royal Statistical Society*, 39: 1–38, 1977.

C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: Graph interpretation user study. In *Human-Computer Interaction - INTERACT 2005, Lecture Notes in Computer Sciences, Springer*, pages 861–872, 2005. ISBN 978-3-540-28943-2.

J. Hummel. Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, 11:23–33, 1996.

R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*. Wiley, New York, 1987.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

J.L. Schafer. *Analysis of incomplete multivariate data*. Chapman & Hall, 1997.

Statistics Austria. Einkommen, armut und lebensbedingungen 2004, ergebnisse aus eu-silc 2004, 2006. Vienna, Austria. ISBN 3-902479-59-0.

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