

MANET

Extensions to Interactive Statistical Graphics for Missing Values

Martin THEUS, Heike HOFMANN, Bernd SIEGL & Antony UNWIN

Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany

email: Martin.Theus@Math.Uni-Augsburg.DE

Abstract. Dealing with missing values is inconvenient in many ways. Although every statistician has to admit that the structure of missing values can be very meaningful for the interpretation of data, neither statistical software nor statistical theory can deal with missing values in a convincing way.

Interactive Statistical Graphics offer a possibility of getting more insight into the structure of missing values. This is done by a natural integration of missing values with well known statistical graphics in the package MANET.

MANET also offers some new interactive graphs including mosaic-plots, weighted histograms and weighted barcharts, as well as generalised brushing, together with extensive interactive features.

Easy to use, object-oriented applications overtax the user with management of all the windows. The context sensitive window-management in MANET enables the user to handle many more plots, and thus many more views of the data, in a smooth way.

1. Historic View

Reviewing the collection of articles in Cleveland & McGill [2] one could get the impression, that nearly ten years later, the progress of Interactive Statistical Graphics has stalled. Many of the techniques described there are not available in statistical software yet, and do not seem to have had much influence on further development. This impression is also supported by Cleveland's latest book [3], which covers dynamic graphics in just one and a half pages. The only exceptions seem to be DataDesk being developed by Paul Velleman [12] and SAS's JMP [7]. Thus the question arises, whether the power of interactive statistical techniques is too weak or the effort for developing this kind of software is too big.

The MANET project shows, that neither statement holds true, as will be shown in this paper.

2. Dealing with Missing Values

Handling missing values seems to be an unsolvable problem – how to handle information which was not collected? But often it is of great interest to know how many missing values were recorded in one variable, or in a multi-

variate context, for which values of a certain variable do we find missings in another variable. This leads to the following procedures:

2.1 The Missing Value Chart

The Missing value chart draws a horizontal bar for each variable. The left part of the bar represents the proportion of data that is not missing, whereas the right part represents the proportion of missing values inside the selected variable. Using highlighting it is now very easy to access the different kinds of data. Figure 1 shows an example of this missing value chart for the Crash dataset, see Velleman [12]. This is a version of the missing value chart originally implemented in REGARD, see Unwin [10].

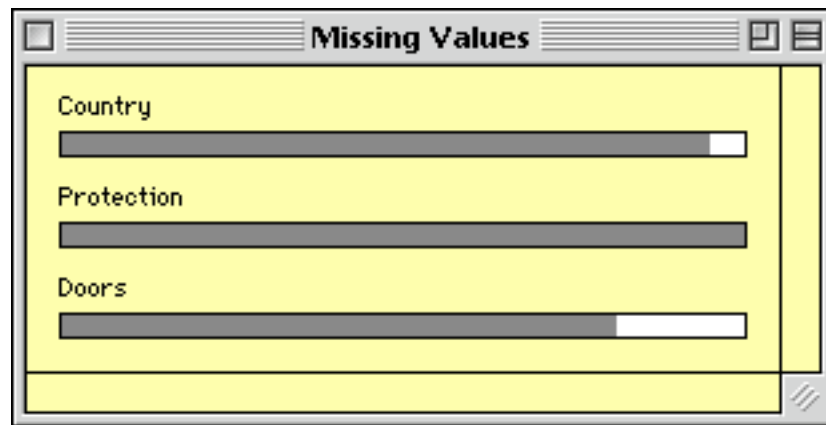


Figure 1: A Missing Value Chart for the Crash data

2.2 Univariate Plots

Missings can be incorporated in a very natural way for *histograms* and *bar-charts* by adding a bar corresponding to the amount of missing data. Figure 2 shows two examples, where we find the number of missing values represented by an additional white bar at the right (barchart) or left (histogram) of the plot separated by a small gap.

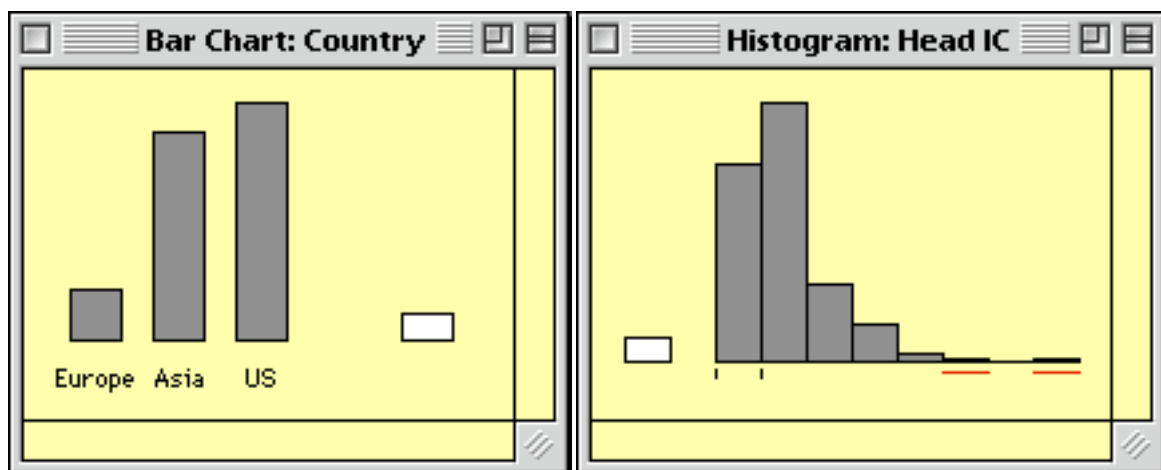


Figure 2: An example for missing values in barcharts and histograms

Boxplots and *Dotplots* can handle missing values only in a more rudimentary way. For each boxplot of a variable that includes missing values a missing values plot is drawn. Figure 3 shows boxplots with their corresponding missing values plots. Plotting several boxplots simultaneously yields one combined missing values window.

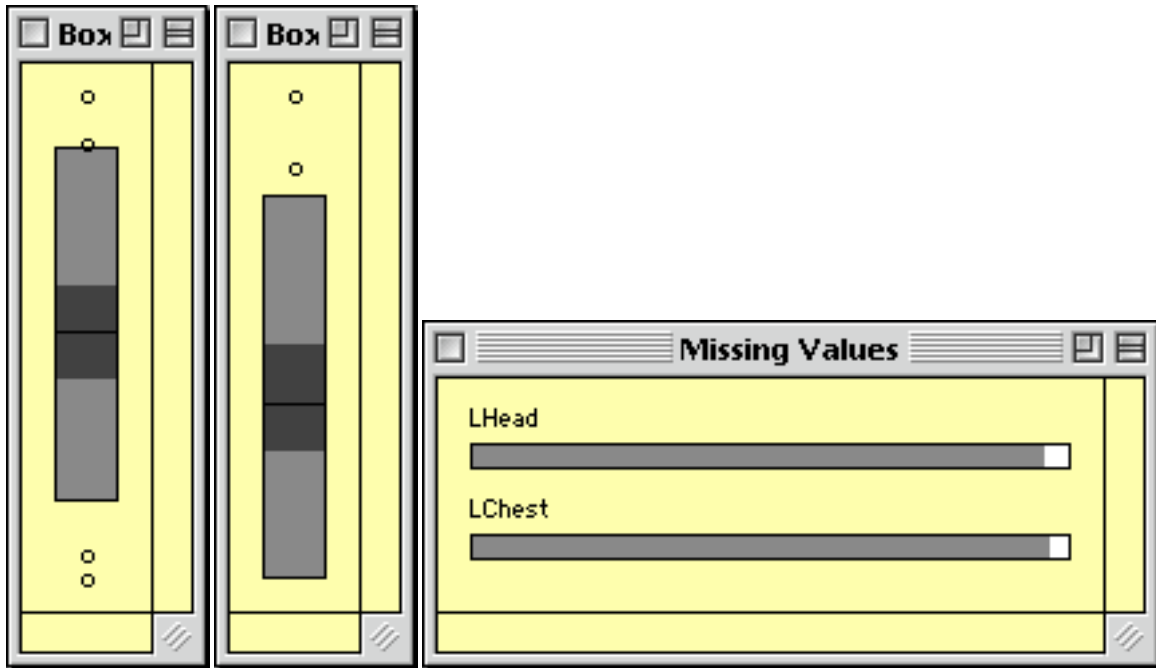


Figure 3: For each boxplot a missing value chart is plotted

2.3 Multivariate Plots

Multivariate plots can incorporate the information of missing values in a more powerful way. A *scatterplot* includes two variables. Thus each observation can belong to four different states:

1. Both values were recorded
2. The x-value was recorded, the y-value not.
3. The x-value was not recorded, but the y-value was.
4. Neither of the values was recorded.

Values belonging to the first case are plotted in the classical way. Values belonging to case two or three can be drawn as projections along the x- or y-axis. Only case four cannot be incorporated in the scatterplot. For this, each scatterplot has three additional boxes at the bottom. The leftmost box represents the proportion of missing x-values, the middle box the proportion of missing x- and y-values, and the rightmost box the proportion of missing y-values. Thus the user can easily select all cases belonging to case four from this middle box. Figure 4 shows a scatterplot inside MANET.

Note, that overlapping points add their brightness. This is a very good means of avoiding a loss of information due to an overplotting in areas of high density. The size of the points can be altered interactively, giving the user control over the amount of this smoothing.

A generalisation to a *3-D rotating plot* could be done in a related way, but has not yet been implemented in MANET.

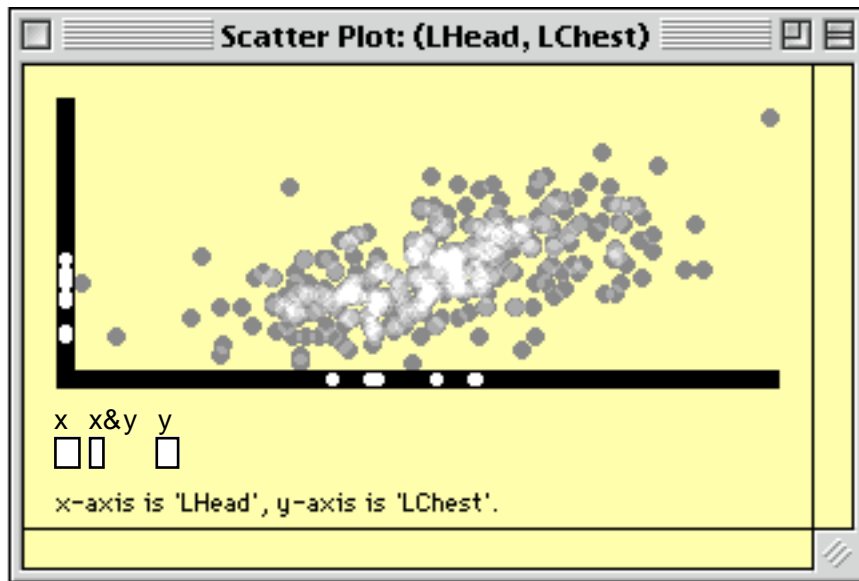


Figure 4: In scatterplots missing values are projected onto the axis

Mosaic plots, see Friendly (1994), can handle many categorical variables at once. MANET offers mosaic plots with up to eight variable or at most 256 categories. The incorporation of missing values is done in the same way as in a barchart by adding an additional category including all missing values of this variable. Figure 5 shows a mosaic plot for two variables, having 5 (year), 3+1 (country) categories, leading to $5 \times 4 = 20$ boxes inside the plot. The implementation of the mosaic plot is fully interactive, and an extension of the definition of Hartigan & Kleiner [5]. The highlighting information is added to the plot just like an extra binary variable, but without inserting a gap between the two categories (highlighted, not highlighted).

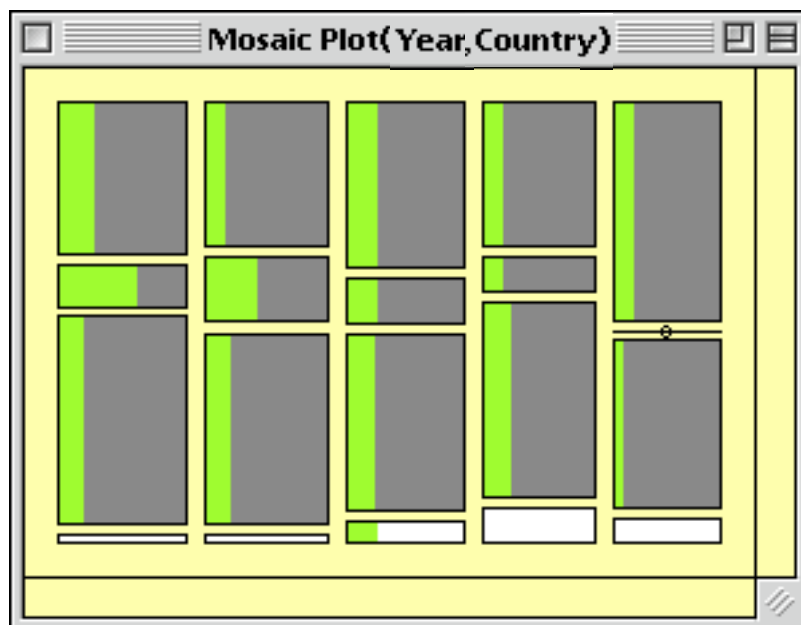


Figure 5: A mosaic plot with values of high head injuries highlighted

The practical use of mosaic plots with non trivial data, shows that often a lot of combinations of categories include no values, and thus are empty. To distinguish those boxes from boxes which include only very few values, and thus are only very small (e.g. one or two pixels of width or height), really empty

boxes are plotted with a „0“ in the middle, to indicate that they are of size zero. The visual perception depends not only on the sizes of the boxes, but also on the size of the gaps between the boxes. To achieve an alternate view of a mosaic plot, the user can choose the "mondrian"-display of a mosaic plot, where all gaps are left out. The name of this option can be understood by looking at paintings of Mondrian.

To find out more about the distribution of the different bin sizes inside a mosaic plot, an additional histogram of the bin sizes of the corresponding mosaic plot can be plotted. This histogram includes the option to show the real amount of values belonging to a bin of a certain size, or the number of bins having a certain size. This histogram is fully linked with all other plots, too.

3. New Interactive Plots

3.1 Mosaic Plots

In section 2 the *mosaic plot* has already been described. MANET is the first application which includes mosaic plots in an interactive context. Static representations have been developed for SAS as well as for S-Plus. But the most impressive advantages of mosaic plots in comparison to barcharts arise with interactivity. Highlighting a single box in a mosaic plot corresponds to highlighting the intersections of different categories in various barcharts – which is far more complicated. Thus the user can make a selection of subgroups by single mouseclicks, which would be very difficult with barcharts.

Variables in the plot: var1, var2, var3, var4

(...) Variables included

[...] Variables excluded

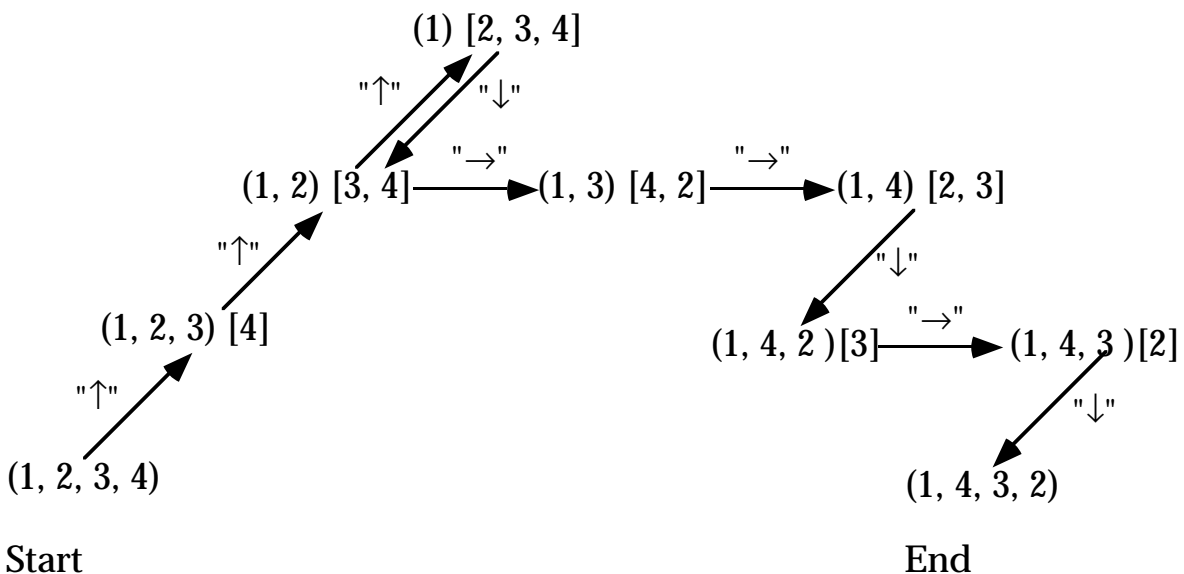


Figure 6: A sample path through the permutation tree of a mosaic plot

Mosaic plots which display many variables vary much with the change of the order of the variables. The order of the variables of a plot can be easily changed in two ways:

1. Use the parameters box of a mosaic plot to drag & drop the variable to its

new position in the list.

2. Use the four arrow keys to add or drop variables at the end of the variable list, or to change the last variable in the list cyclical with all variables currently not in the plot. Figure 6 shows a schematic view of changing the order and inclusion of variables in a sample mosaic plot using the arrow keys

The mosaic plot updates after every keystroke, enabling the user to get a succession of alternative views very quickly.

3.2 Weighted Histograms and Barcharts

In analysing surveys it is often necessary to weight the values of one variable by another. Weighting in MANET is done as follows. Histograms as well as barcharts are based on bins, which represent the number of observations inside the bin or interval. The weighting variable is standardised to a range of 0 to c and the area which each observation contributes to a bin (which is one in the standard unweighted case) is multiplied by the standardised weight. Missing values in the weighting variable have the weight zero, so for each class, which includes missing values in the weighting variable, an extra bin is plotted below the bar, to enable the user to judge the influence of the missing weights. The layout of weighted barcharts is analogous to that for weighted histograms. Figure 7 shows an example of a weighted histogram inside MANET.

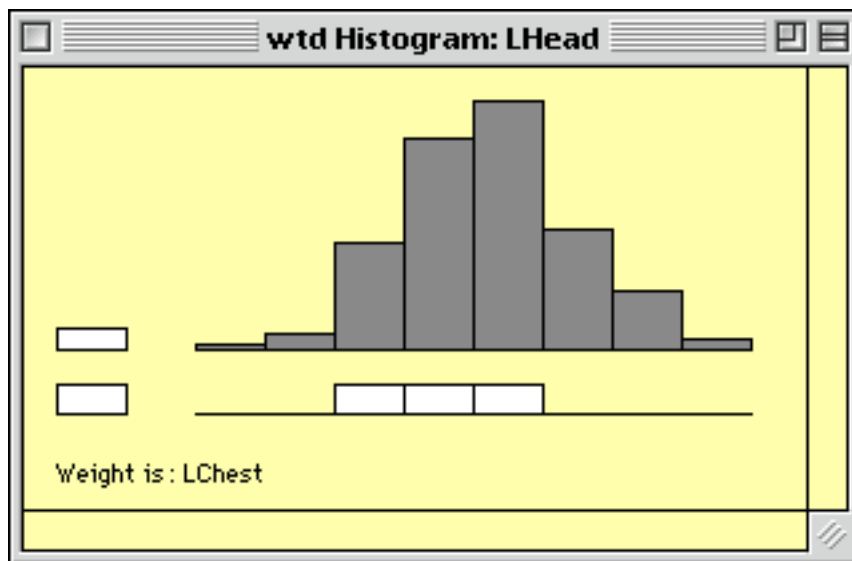


Figure 7: Histogram of head injuries weighted by chest injuries

4. New Interactive Features

4.1 Interrogating Plots

The possibility of interrogating plots is usually limited to identifying points. This is implemented in DataDesk in a very attractive manner. But interrogation of plots can be generalised. Based on the Macintosh philosophy, that standard functions are the same in any application MANET offers the possibility of interrogating a plot by option-clicking any part of the plot. The result of the interrogation depends on the part of the plot, and the plot itself. For the

different plots several interrogations are defined. (In all the following cases, clicking means option-clicking).

Missing Value Chart:

- Clicking a horizontal bar of the chart shows the number of recorded and missing values in the bar and the numbers highlighted.

Barchart:

- Clicking a bar of a barchart shows the number of values in the bar and the numbers highlighted

Histogram:

- Clicking a bar in the histogram shows the number of values in the bar and the numbers highlighted and the limits of the interval of that bar

Boxplot:

- Clicking any part of the box shows the values of the median, upper and lower hinges as well as inner and outer fences.

Scatterplot:

- Clicking an axis shows a horizontal respectively a vertical line together with its position in the scatterplot.
- Clicking any horizontal bar at the bottom of the scatterplot shows the number of recorded and missing values in this bar and the numbers highlighted.

Mosaic Plot:

- Clicking any bin in the plot shows a list of the variables and their values corresponding to that bin as well as the number of values falling in this bin and the numbers highlighted.

Weighted Histogram & Barchart:

- The interrogation results include the percentage of the weightings in the class and the percentage highlighted.

4.2 The Plot Parameters

Each plot offer various parameters to change the look of a plot. This section summarises the most common parameters in MANET:

- Display missing values
- Show scale
- Set the scale of a plot

More parameters can be accessed via the plot parameter box depending on the specific plot type.

4.3 Cues

Cues in MANET are related to hyperviews in Data Desk which were introduced by Velleman several years ago. Inside Data Desk they represent special pop-up menus, which can be used to introduce further steps of an analysis. They are used to guide the user through an analysis in a context sensitive manner.

Cues in MANET are used to change the settings of a plot. Whenever the cursor reaches a certain position in a plot the shape of the cursor changes. A subsequent mouse click initiates the change of the plot. Several cues have been implemented in MANET. We summarise some of them here.

Inside barcharts the shape of the cursor changes whenever it moves to the lower part of the plot. If the barchart is plotted in the classical manner the pointer changes to a double-headed vertical arrow. Clicking the

mouse now changes the barchart to an alternate display, the so called spine-plot, see Hummel [6]. In a spine plot the height of all bars is the same, but the width is proportional to the number of cases falling into this category. This alternate view enables the user to compare the proportions of highlighted observations between the different categories very easily. Note that a spine plot is just a univariate mosaic plot!

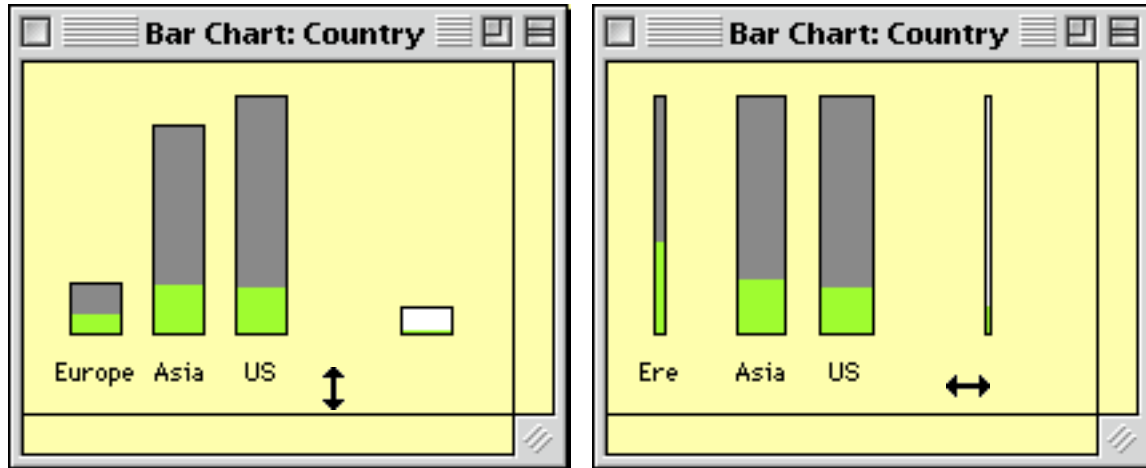


Figure 8: Cue to flip the barchart

In scatterplots we find another cue. At the lower right corner of a scatterplot the shape of the cursor changes to a curved arrow pointing to two axes. Clicking the mousebutton whenever this cue appears, switches the x and the y-axis of a scatterplot.

Figure 9 shows another example of cues inside MANET. Mosaic plots can be switched between the display of the raw data and the display of the expected values for a certain loglinear model.

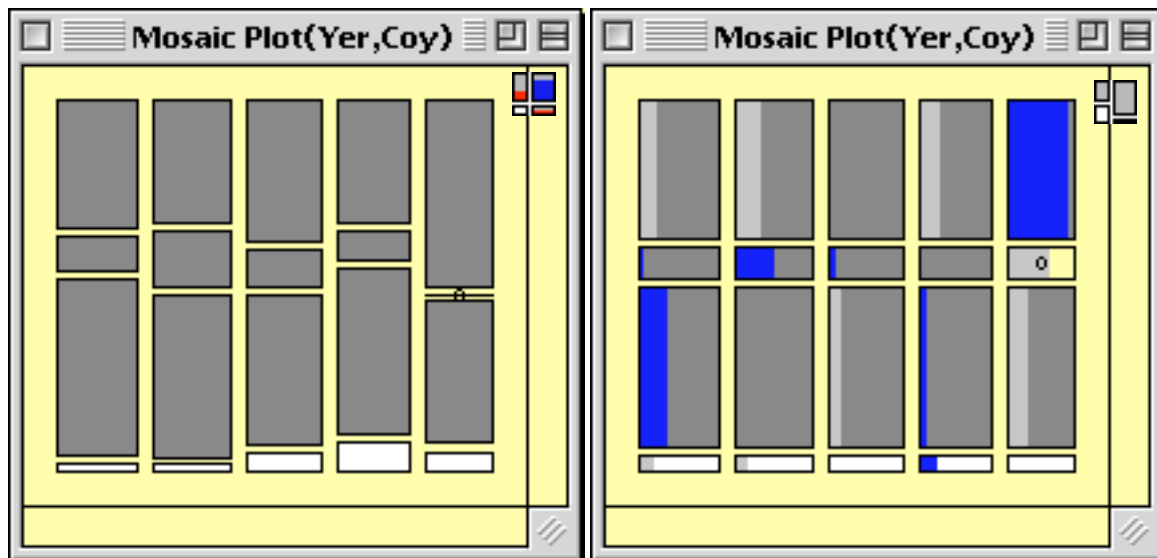


Figure 9: Cue to flip between raw values and expected values (total independence)

Positive (black) and negative (light grey) residual are shown as a highlight proportion in the bins. The highlighting is scaled by the G^2 - statistics of the corresponding model.

The shape of a histogram varies very much with the starting point and the bin width of the histogram. Thus a single plot of a histogram can mislead the

user. MANET offers two cues, appearing at the two handles of the leftmost bin of a histogram. The first cue, a small triangle, moves the starting point of the histogram up and down, when moving the mouse right or left. The second cue, a left-right arrow between two vertical lines, changes the bin-width continuously.

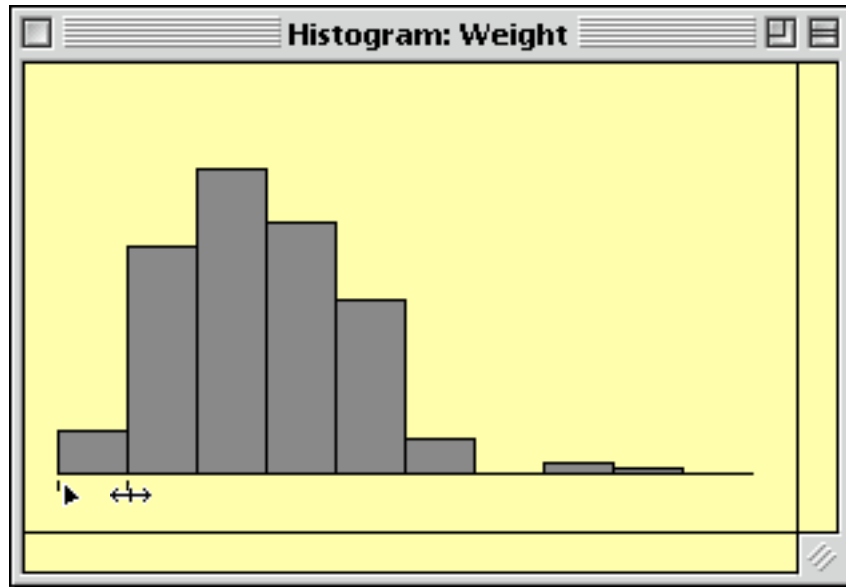


Figure 10: Cues to alter binsize and offset of a histogram

Figure 10 shows a histogram together with the two cues. Alternatively the changing can be done by using the arrow keys. The nature of an cue implies that there is at most one cue visible at a time, thus figure 10 has been set up manually to show the two cues at once.

4.4 Warnings

Warnings given by alert boxes can be very annoying in an analysis. A preferable way to warn a user against pitfalls is to add the warning graphically. MANET has three different situations where graphical warnings appear.

- The first one arises if observations fall out of the plot area, and thus are clipped. To indicate this MANET plots an extra red frame inside the plot window.
- A second warning can be found in the context of highlighting. There are two special cases, when highlighting can give trouble. The first case occurs, whenever a very small bar or bin is highlighted with very few values. If the unhighlighted bar or bin is only 2 to 5 pixels of height on the screen, a proportion of less than 10% would have less than 1 pixel of height, and thus not be plotted at all. Although the user could interrogate the bar or bin by a single optionclick, finding out the right values, it is important to prevent the user from overlooking the highlighted values. This is done by changing the colour of the frame of the bar or bin from black to red. The complementary case, where nearly all values are highlighted except for a few, not plotted, values, is handled in the same way.
- The third situation where warnings can occur, are plots where the bin-size differs dramatically from the number of values that bin should represent. This is due to limited resolution of the computer screen. Figure 11 shows two situation of this third warning: A histogram with

very small tails and a mosaic plot with very many, and thus small cells. In the first case the warning is done by underlining, in the latter by a dialog, giving the cumulative error of all cells in the mosaic plot.

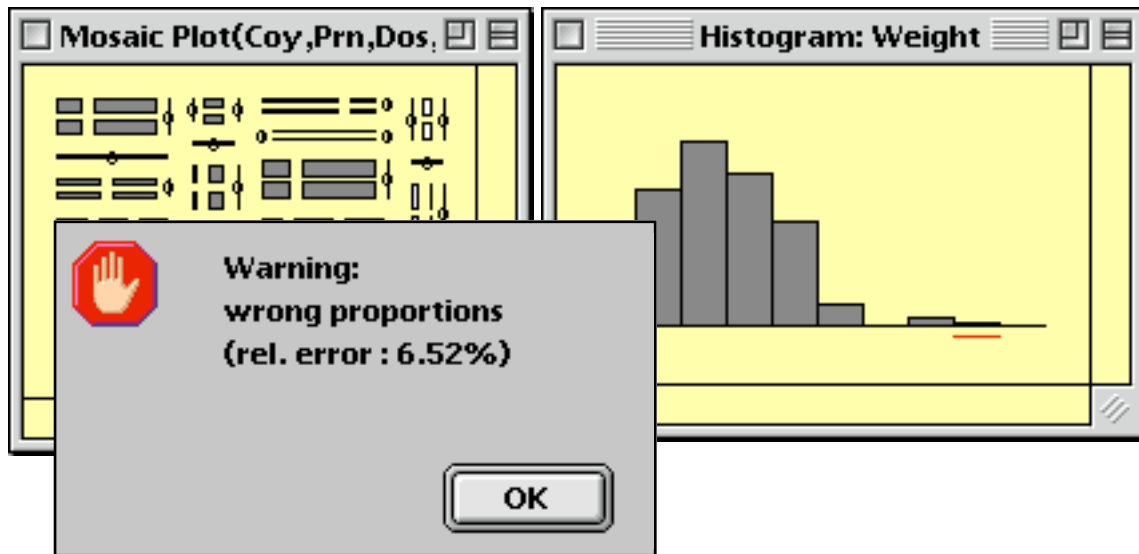


Figure 11: Two sample warnings inside MANET

Unfortunately, in black and white printing it does not make a difference to show black or red frames, so an example of the framing warnings is not possible here.

5. Managing Windows

5.1 Scaling and Sizing by Group

To achieve the right impact for a group of plots, it is often desirable to draw them to the same scale. This can be done in MANET by command-clicking a plot which brings up a list of plots that can be identically scaled.

5.2 Tiling and Stacking by Group

Menu based, mouse driven applications enable the user to bring up a lot of graphs and statistics in a few minutes. But often the screen is quickly jammed, and the user needs means of tidying up the screen. This can be done by tiling and stacking windows by group, i.e. by plot type.

5.3 Siblings

Data Desk has introduced the concept of siblings for closing whole sets of windows. In this concept siblings are all windows that were created at the same time, i.e. with the same command. MANET generalises this concept in two ways.

Siblings do not have to be necessarily created at the same time, but they are defined manually, which enables the user to set up different groups of plots that go together.

A second generalisation is, that the functionality of siblings is not restricted to closing windows. Siblings can optionally have the following properties:

- open/close all windows of a group by opening or closing a single window
- scale all plots identically
- resize all plots identically
- all plot parameters go together for all plots of the same kind.

5.4 Trellis Displays

Trellis Displays were introduced by Cleveland [1], and are only available inside S-Plus as yet. Although Trellis Displays are controversial, see Theus [8], they offer the possibility of achieving a systematic view of the data. Interactivity can be accomplished inside an interactive environment such as MANET offers.

Trellis Displays inside MANET are restricted to two categorical conditioning variables, forming the rows and columns of the trellis. Creating shingle variables has been left out because shingling seems to be a dangerous and misleading technique, although the shingling could be done in an interactive manner via sliders.

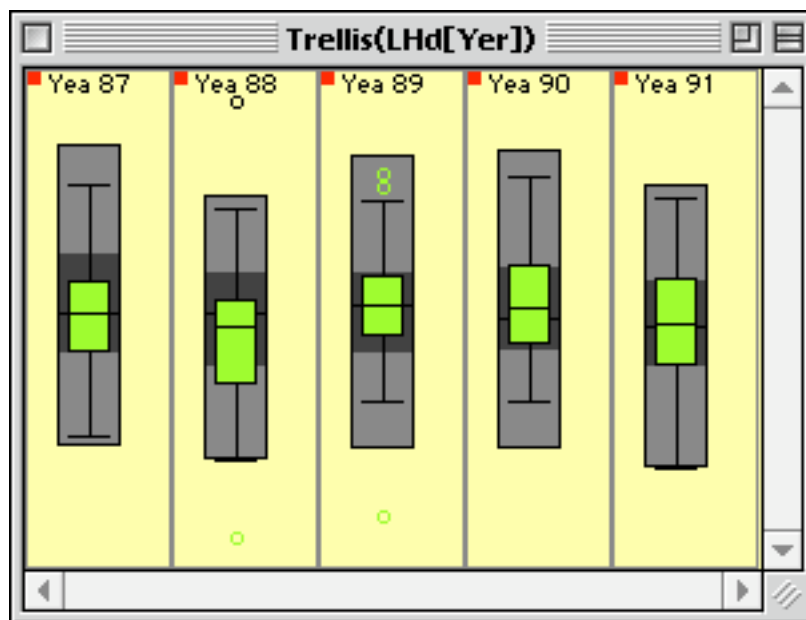


Figure 12: Head injuries split up by year – cars with 4 doors highlighted

5.5 The Index Window

Index windows have been introduced first in Diamond Fast, see Unwin [9]. They form a virtual display of the screen, showing all windows on the screen, except the Index Window itself. Every window is represented by a grey box, with an additional icon in the upper left corner, showing the type of plot inside the window. Every window inside the Index Window can be moved around. After dropping a window at a new place of the screen, this window will be popped up, to be the front most of all windows. This enables the user to tidy up and place windows easily. The Index Window is also used to define the sibling structure of groups.

In addition to the index window, the so called *catwalk* enables the user to place the graphics onto up to four virtual screens. This is very helpful when working with more than one dataset at a time. The analysis of different datasets or the of different aspects of one dataset can be carried out on distinct virtual screens more clearly. Related concepts can be found in SUN's `olvwmm`.

6. Conclusion

MANET enables the data analyst to handle missing values easily in an interactive graphical environment. The integration of missing values into standard statistical graphics is smooth and efficient. MANET covers the standard range of graphics, and even adds some more plots, which form a powerful addition to classical graphics in use.

Interactive techniques are further supported by adding new interactive features. Additionally cues have been introduced, which provide considerable analytic flexibility.

The extensions to the management of windows are valuable for improving efficiency of analyses.

The answer to the question, raised initially, whether the power of interactive statistical graphics is too weak, or the effort for implementing this kind of software is too big, is definitely: no!

References

- [1] BECKER, Richard. A., CLEVELAND, William. S., SHYU, Ming-Jen, KALUZNY, Stefan. P. (1994), *Trellis Display: User's Guide*. AT &T Bell Laboratories Statistics Research Report No. 10
- [2] CLEVELAND, William S. & McGill, Marylyn E. (1988), *Dynamic Graphics for Statistics*. Wadsworth Inc., Belmont California
- [3] CLEVELAND, William S. (1993), *Visualizing Data*. Hobart Press, Summit NJ.
- [4] FRIENDLY, Michael (1994), *Mosaic Displays for Multi-Way Contingency Tables*. Journal of the American Statistical Association, Vol. 89, No. 425, March 1994 Theory and Methods
- [5] HARTIGAN, J. A. & KLEINER, B. (1981), *Mosaics for Contingency Tables*. Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, ed. W. F. Eddy, Springer, New York, pp. 268–273
- [6] HUMMEL, Jürgen (1995), *Linked Bar Charts: Analysing Categorical Data Graphically*. Computational Statistics, Vol. 11 Issue 1
- [7] JMP (1995), *JMP® Statistical and Graphical Guide, Version 3.1*. SAS Institute Inc., Cary, NC
- [8] THEUS, Martin (1995), *Trellis Displays vs. Interactive Statistical Graphics*. Computational Statistics, Vol. 10 Issue 2, pp. 113–127
- [9] UNWIN, Antony R. and WILLS, Graham J. (1988), *Eyeballing Time Series*. Proceedings of the ASA Statistical Computing Section, pp. 263–268
- [10] UNWIN, Antony R. (1994), *REGARDing Geographic Data*. In P. Dirschedl and Ostermann, R. (Eds.), Computational Statistics, pp. 315–326. Heidelberg: Physica.
- [11] UNWIN, Antony R., HAWKINS, George, HOFMANN, Heike, SIEGL, Bernd (1995), *Interactive Graphics for Data Sets with missing values – MANET*. Journal of Computational and Graphical Statistics, Vol. 4, No. 6
- [12] VELLEMAN, Paul F. (1995), *Data Desk 5.0*. Data Description, Ithaca, New York.
- [13] WILHELM, Adalbert, UNWIN Antony R. & THEUS, Martin (1995), *Software for Interactive Statistical Graphics – A Review*. in: Advances in Statistical Software 5, Softstat '95, eds. Frank Faulbaum & Wolfgang Bandilla, Lucius & Lucius, Stuttgart