

# Visualisation of Categorical Data

Martin Theus

Department of Computational Statistics and Data Analysis

Institute of Mathematics

University of Augsburg

86135 Augsburg

GERMANY

## Summary:

Many statistical graphics for the exploration as well as for the modelling of data measured on a continuous scale have been developed. In contrast to that, graphs for the interpretation and the modelling of categorical data are rarely to be found.

Hartigan & Kleiner (1981) proposed Mosaic Plots. Although this recursive visualising technique is very powerful, it has not proved popular. This is mainly due to the fact, that the visual impact of a mosaic plot depends considerably on the order of the variables. Static implementations (e.g. in SAS or S-Plus) are available, but cannot bypass this disadvantage.

An interactive environment like MANET (Unwin et al. 1996), offers very flexible means of rearranging the order of the variables manually and automatically. The paradigm of linked-highlighting can visualise categorical response-models easily by using mosaic plots as well as bar-charts. In addition to exploratory uses, superimposing residual information in the mosaic plots, makes possible a graphical stepwise modelling of categorical data, which reaches far beyond traditional methods.

Again, interactivity seems to be a key feature for achieving more powerful results.

# 1 Classical Parametric Approaches

Although this paper describes visualisation techniques for categorical data, I shall give a brief summary of the most common parametric competitors and their weaknesses.

## 1.1 Correspondence Analysis

Correspondence analysis is more a mathematical than a statistical technique, since no assumptions about the distribution of the investigated variables are made. The results of a correspondence analysis are often used to visualise categorical data. Given a  $c \times r$  contingency table, one calculates the singular value decomposition

$$X = U\Lambda V$$

where  $U$  are the eigenvectors of  $XX'$  and  $V$  the eigenvectors of  $X'X$ .  $X$  is the matrix of the standardized residuals of Pearson's  $\chi^2$ -statistic.

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad \text{and} \quad x_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

The  $o_{ij}$  denote the observed values, whereas the  $e_{ij}$  denote the expected values under the assumption of mutual independence. This kind of decomposition is a categorical equivalent to the principal components for continuous data and hence is very popular.

To bypass the limitation to 2-dimensional contingency tables, the  $n \times k$  data-matrix of the  $n$  observations on  $k$  variables is used, called a multivariate correspondence analysis (cf Nagel et al. (1996)).

Since the interpretation of eigenvalues and eigenvectors is hard, analysts tend to plot a scatterplot of  $v_1$  vs.  $v_2$  or  $u_1$  vs.  $u_2$  to obtain the so called row-profiles resp. column-profiles of the data-table. The  $v_i$  and  $u_j$  are the columns and rows of  $U$  resp.  $V$ . Figure 1 shows an example of a multivariate correspondence analysis for Bertin's accident dataset, including the variables Age and Vehicle (cf Bertin (1983) p.31). Although the distance between a row point and a column point has no meaning, the directions of the points from the origin have, which should serve for interpretation purposes.

For instance in figure 1 the directions of motorcyclists and the ages of 0–10 and 20–30 from the origin, suggest, that children are not involved in accidents with motorcycles, whereas young adults very frequently are.

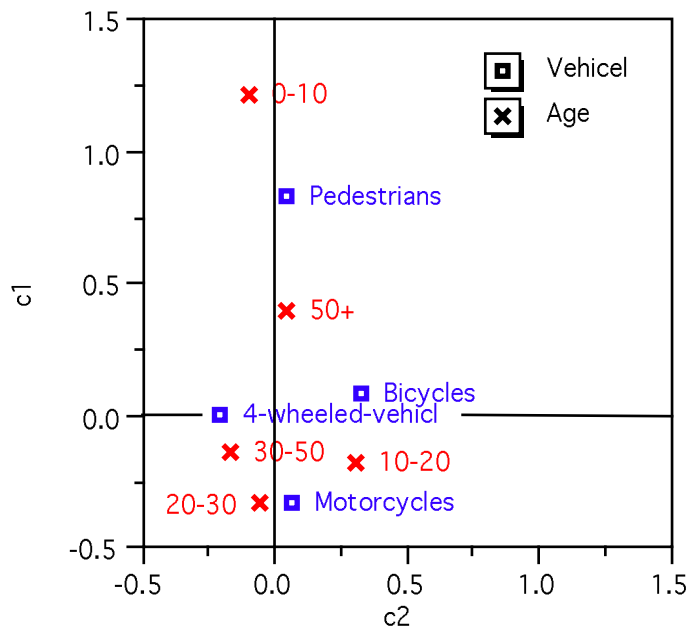


Figure 1: Correspondence analysis for Bertin's accident data

## 1.2 Loglinear Models

Loglinear models (cf Agresti (1990)) derive directly from linear models. Whereas correspondence analysis should visualise the interaction structure of the variables, loglinear models are defined by their interaction structure. A suitable model is usually judged by the corresponding  $\chi^2$ -statistic or  $G^2$ -statistic. Although the modelling of categorical data via loglinear models is elegant, there has been no proposal yet to visualise a model properly. A scatterplot of the observed vs. the expected values is often used for visualisation purposes, but incorporates neither the structure of the data, nor the structure of the model. This holds true for residual plots as well.

Often automatic selection procedures are used to suggest models, but they usually cannot reveal the really relevant information of a dataset.

## 2 Graphical Approaches

### 2.1 Playfair's Great Grandsons?

When William Playfair started to report trade figures in a graphical way, he obviously needed plots to visualise amounts, split up by different grouping variables. He designed various barchart and piechart like plots, whose range is hardly exceeded by modern serious statistical software packages. Reviewing the current literature on statistical graphics, which culminates in the book of William S. Cleveland (1993), does not reveal any graphical technique to cope with multivariate categorical data. The recently introduced Trellis Displays (cf Becker et al. (1994) and Theus (1995)) are based on categorical variables for conditioning plots of continuous variables, but can hardly visualise the multivariate structure of purely categorical data with more than three variables.

## 2.2 Why Bertin Failed!

Jaques Bertin (1983) made a great effort to analyse, i.e. decompose graphs, and tried to synthesize them to more general entities. But reviewing his work on categorical data shows certain limitations, which violate some essential demands on statistical graphs:

1. Generalizability

A design of a graph should be generalizable to more than just the number of variables, it was initially designed for. E.g. a scatterplot generalizes easily to a 3-d rotating plot, and a further generalization is possible, even beyond human 3-d perception (c.f. Cook et al. (1995)).

2. Consistency

Data measured on the same scale should be plotted by the same method, i.e. counts by areas, points on a continuous scale by dots, etc.

3. Extendability

The basic design of a plot should allow to extend the plot for different purposes, e.g. highlighting and colouring of subgroups, superposing residuals or other modelling information. A barchart can easily be used to highlight a subgroup, whereas a piechart can not.

4. Interactivity

The functionality of a plot produced by a modern statistic package should reach beyond a simple 'drawing'.

- (a) Plots should be linked and show highlighting of selected data.
- (b) The user should be able to interrogate for information.
- (c) The parameterization of a plot should be easy to change dynamically.

It is obvious, that Bertin's work was done before interactivity came into being, thus we can not blame him for not mentioning it — others would have been able to in the last ten years! Figure 2 shows Bertin's proposal for visualising multivariate categorical data, here set up for the Titanic data, cf figure 5, 3 and (Titanic 1990). The reader may check all the above demands for figure 2, and find, that this plot is neither generalizable nor consistent. But there exist some other plots, which fulfill the four points partially. E.g. the fourfold plot designed by Michael Friendly (1995) is well able to visualise the differences of  $2 \times 2 \times k$  tables from the model of mutual independence, but is limited to that single feature.

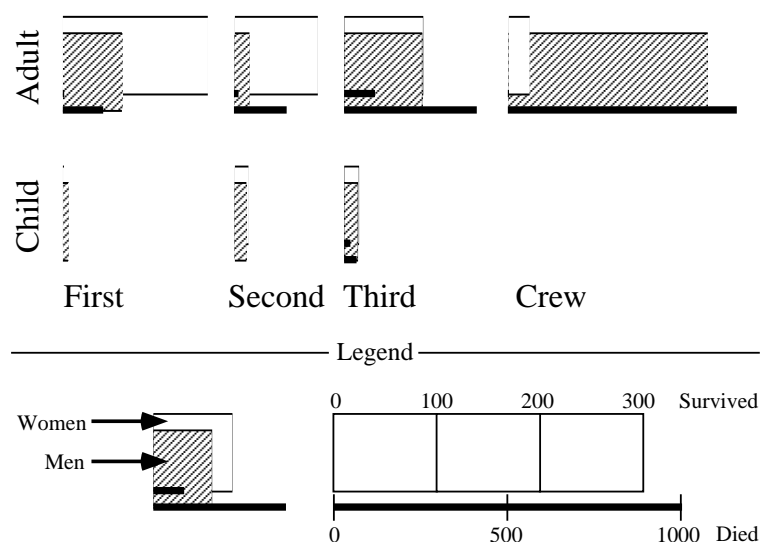


Figure 2: Bertin's proposal to visualise multivariate categorical data

The work of Riedwyl & Schuepbach (1994) is close to what we demand, but lacks the generalizability and interactivity. We will see in the next section, that the interactive implementation of Mosaic Plots is able to meet all requirements.

### 2.3 Escaping the Univariate — Linked Barcharts

Working with simple barcharts can not visualise the multiple structure of the data. By linked highlighting this can be bypassed partially. Hummel (1996) shows examples of how to explore the five independence structures of three categorical variables by using linked highlighting.

Figure 3 shows an example of the four linked barcharts of the Titanic data. All charts except the one for **Sex** have been scaled the same, thus facilitating a direct comparison of the amounts. Note that the lower left barchart has been modified. In this barchart the height of each bar is no longer proportional to the amount of data in this category, but the width. This enables the user to compare the highlighted proportions directly. The modified barchart is called a *spineplot*.

Linking of barcharts and spineplots is a good support for investigating two or three categorical variables. But keeping track of more than three variables is nearly impossible. Looking at particular subsets, i.e. intersections of selections, is too complicated even for experienced users.

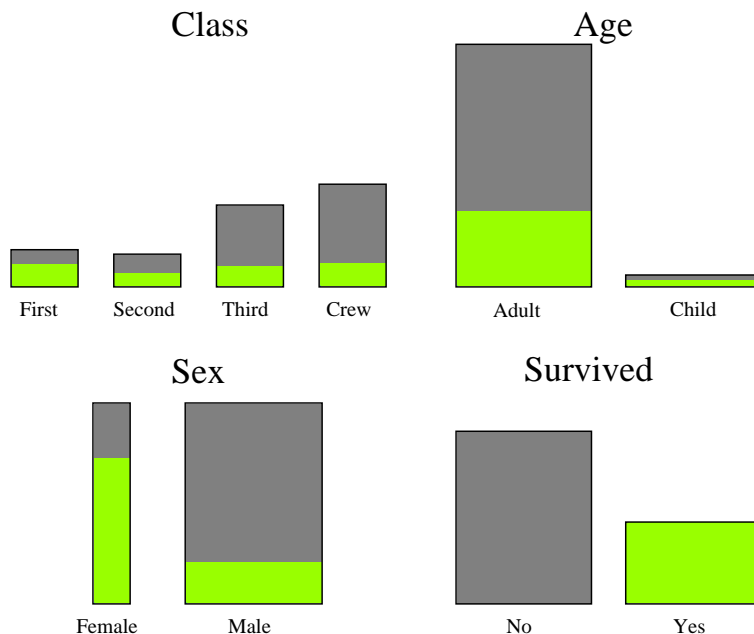


Figure 3: Four linked barcharts — Spineplots included

### 3 Interactive Mosaic Plots

Hartigan & Kleiner (1981) proposed mosaic plots. Figure 4 shows a Mosaic Plot for Bertin’s accident data. Whereas the variable **Age** has a given order, and the order of the binary variable **Sex** does not influence the shape significantly, the levels of **Vehicle** have been sorted to achieve a monotone decrease of the proportion of males (from top to bottom).

Although this recursive visualising technique of Mosaic Plots is very powerful, it has not proved popular yet. This is mainly due to the fact, that the visual impact of a mosaic plot depends considerably on the order of the variables. Static implementations (e.g. in SAS or S-Plus) are available, but cannot bypass this disadvantage. Mosaic Plots meet all the above mentioned demands on a statistical graph.

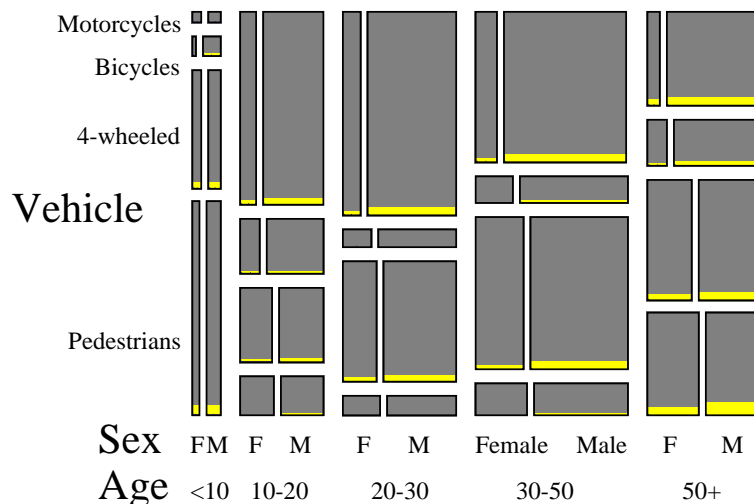


Figure 4: A Mosaic Plot for Bertin’s accident dataset

1. The recursive definition ensures a generalization to as many variables as sensible.

2. As the counts in a Mosaic Plots are all represented by the areas of the tiles, the plot is consistent.
3. Highlighting of subgroups can easily be done, since a subgroup can be added in the same way as a new binary variable would be added. This also allows the plotting of model residuals.
4. Interactivity can be achieved with the same tools used for barcharts. Optionally tools for a flexible reordering of the variables and the categories should be provided.

### 3.1 Linking Information

In many situations there exists one response variable, and several explanatory variables. In these cases it is desirable, to study the influence of particular explanatory variables resp. the influence of combinations of explanatory variables. A really excellent method to do this, is to put all the explanatory variables in one Mosaic Plot and display the response variable in a barchart. This technique can be applied to all

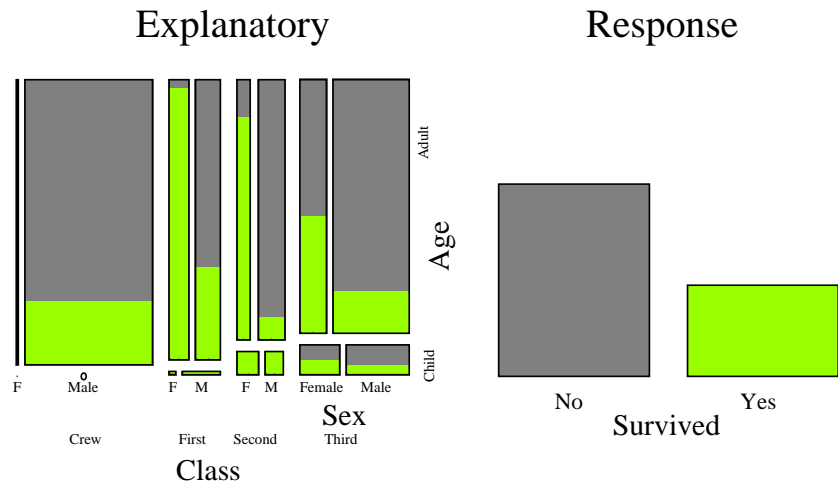


Figure 5: Linking the response variable to the Mosaic Plot

datasets of this structure. Figure 5 shows this situation for the Titanic data. All passengers who survived are highlighted. It is very easy to see, under which circumstances who was likely to survive or not. The policy 'Women and children first' holds true, but is affected by class.

### 3.2 Rotating Categorical Variables

The 3-d rotating plot was one of the first dynamic and interactive graphics implemented. It enabled the user to investigate 3 dimensional continuous data from different angles. The Grand Tour (Cook et al. 1995) is a generalization of this technique to more than three variables.

With categorical data there exists a similar point. Since the order of the variables inside a Mosaic Plot determines, which variables are plotted conditioned by other variables, changing the order of the variables is a similar procedure to rotating in a 3-dimensional space of continuous data. Similar to rotation in 3-d space, the 'rotation' of the variable order needs a fast and flexible control. The implementation in MANET (cf Unwin et al. (1996) and Theus (1996a)) allows a change of the order by drag & drop inside a list or by keyboard control. Besides manual rotation, there exist two automatic versions of finding meaningful variable orders.

1. Finding the largest cells

Given a combination of levels of variables, the order of the variables is chosen as follows. The first variable is the variable, which includes most observations in its particular level. The second variable is the variable, where the intersection of the first and the second variable on their particular levels is maximized, etc.

## 2. Minimize the number of empty cells

Looking at many variables, implies a more or less 'empty' sample space. This 'curse of dimensionality' as Huber called it, will guarantee, that many of the level-combinations will be empty, resulting in empty cells in the Mosaic Plot. A first step to decrease the number of plotted empty cells is not to split up empty cells, if a cell is already empty on a higher level. E.g. there are no children in the Crew of the Titanic, so we do not gain more information by splitting up this empty group by sex.

The Quine/McCluskey algorithm, formerly devised for the minimization of boolean polynomes can be generalized to find out an optimal ordering of the variables, to minimize the number of plotted empty cells (cf Theus & Wilhelm (1996b)).

Another important point is the order of the categories inside the Mosaic Plot. Often categorical variables are also ordered variables. In these cases it is desirable to sort the categories. If the category-names are simple integer numbers, it is easy to build in an automatic sorting of the categories. But often the categories are labeled by strings. MANET offers a simple drag & drop interface to reorder the categories, in cases, where the alphabetic order does not suit. Once the order of the categories is defined, it applies to all current and further produced plots.

### 3.3 Viewing Loglinear Models

In principle there are three ways of viewing a loglinear model:

- Plot the expected values of the model in a Mosaic Plot
- Superimpose the residual information of the model onto the plot of the raw data.
- Superimpose the residual information of the model onto the plot of the model data.

The first approach is useful for understanding the graphical representation of loglinear models in Mosaic Plots in general. Standard dependency structures for two and three variables have certain shapes, which can be understood very easily, and do not depend on the underlying data.

The second approach was used by Friendly (1994). The direction of the residuals is coded by the direction of the hatching of a cell and the amount is coded by the hatching-density. This representation has some disadvantages. The perception of the density of hatching does not project linearly to the amount of the underlying residuals. A similar approach was used by Riedwyl & Schuepbach (1994). In both cases the use of hatching seems to be a concession to the underlying lowlevel plotting routines of a pen-plotter, which was the standard output device for statistical graphs until 1985. Plotting the residuals onto the plot of the raw data can be misleading, since empty or very small cells are not visible in the plot, which does not occur in a plot of the model data.

This leads to the third approach (cf Theus & Wilhelm (1996a)), and the following procedure:

1. Display the residual information in the Mosaic Plot of the corresponding model.
2. Use colour (red or blue) for the direction (negative or positive) of the error:  $r_i := \frac{o_i - e_i}{\sqrt{e_i}}$ .
3. Scale all residuals according to the largest cell, i.e. calculate  $r_i^* := \frac{|r_i|}{\max(|r_i|)}$
4. Scale all residuals with the  $\alpha$ -quantile of the  $\chi^2$ -statistic of the model.

This scaling incorporates the corresponding  $p$ -value of the model into the plot, which can be read from the cell with the largest residual. A model is only significant to the level  $p$ , if at least one cell is highlighted by a residual to more than  $(100 - p)\%$ .

5. Use areal highlighting for the quantity of the error.
6. Optionally raise all residuals to the power of  $\beta$ .
  - $\beta \geq 2$ , only bad fits are visually significant
  - $\beta \leq 2$ , structure of the residuals can be seen clearly

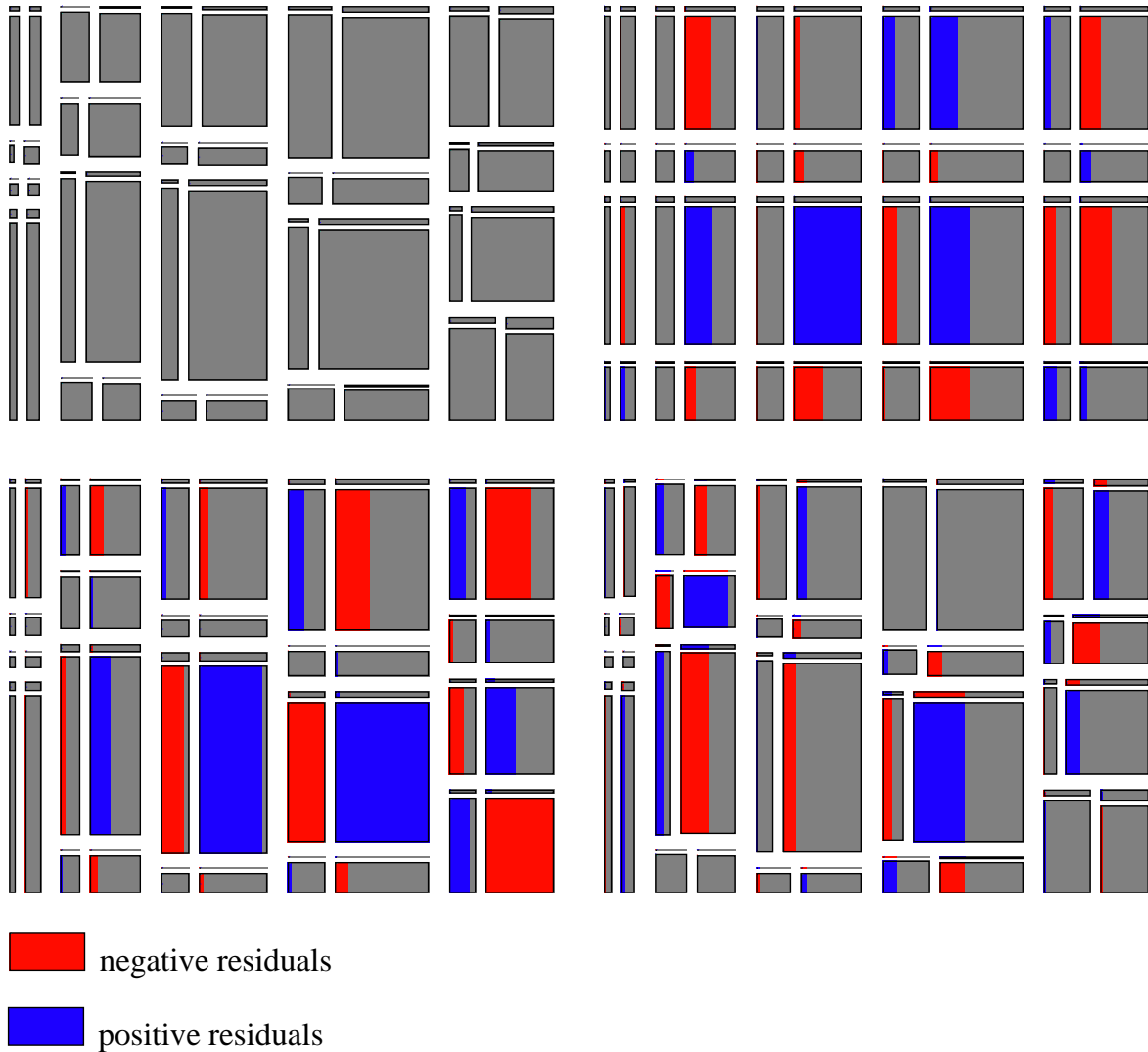


Figure 6: Modelling the Accident data:

- Upper left: raw data
- Upper right: mutual independence
- Lower left: Interaction: Age\*Vehicle included
- Lower right: All interaction except 4-way included

See the URL: <http://www1.math.uni-augsburg.de/~theus/Papers/SoftStat97.html> for a coloured version of this plot!

Figure 6 shows the graphical modelling process, starting with the raw data (upper left). The model of mutual independence is shown in the upper right panel. For this kind of model the Mosaic Plot has a totally regular shape, showing no interaction structure between any variable.

The lower left panel shows the model with the interaction of Age and Vehicle included. Since these variables are the first two in the plot, the shape of the Mosaic Plot for this model is very close to the

plot of the raw data. The interaction structure between **Age** and **Vehicle** supported by the residual highlighting disappeared, and a new interaction structure, now between **Sex** and **Vehicle** shows up.

The lower right panels shows the Mosaic Plot for the model with all interactions except the 4-way interaction included. There is no visual difference between the raw data and this complex model, which is the only model, which is not significant.

In this example the highlighting of residuals serves for judging the interaction structures, rather than the overall fitting of the model, since  $\beta$  was chosen to 1.

## 4 Conclusions

There is a lack of modern and efficient graphics for categorical data. Parametric methods are useful for achieving quantitative results, but weak for gaining insight into the data. This restricts data analysis and model diagnostics extremely.

Interactive Mosaic Plot, as they are implemented in MANET, offer a flexible environment for analysing multidimensional categorical data. Both, the exploration of data as well as the modeling of data are supported strongly, and offer easy to understand results.

The paradigm of linked highlighting is a key feature, and offers the possibility of bringing continuous data into the analysis smoothly.

## 5 Acknowledgements

I would like to thank Heike Hofmann, who programmed the major part of the MANET software, and all of the Mosaic Plots. I am also grateful to all other contributors, especially to Antony Unwin, who initiated the MANET project in 1994. The design of MANET is mainly founded on his ideas, and he contributed a lot of valuable remarks to the implementation of Mosaic Plots.

Since the Department of Computational Statistics and Data Analysis was founded with the assistance of the Volkswagen Stiftung, I also like to thank them for their support.

## References

- Agresti, Alan (1990). *Categorical Data Analysis*. Wiley, New York.
- Becker, Richard A., Cleveland, William S., Shyu, Ming-Jen Kaluzny & Stephen P. (1994). *Trellis Display: User's Guide*. AT&T Bell Laboratories Statistics Research Report No. 10/94.
- Bertin, Jaques (1983). *Semiology of Graphics*. The University of Wisconsin Press, Madison Wisconsin.
- Cleveland, William S., & McGill, M.E. (Ed.) (1988). *Dynamic Graphics for Statistics*. Wadsworth & Brooks/Cole, Pacific Grove California.
- Cleveland, William S. (1993). *Visualizing Data*. Hobart Press, Summit New Jersey.
- Cleveland, William S. (1994). *The Elements of Graphing Data*. Hobart Press, Summit New Jersey.
- Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995). Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*. Vol. 4 No. 3, 155–171.

- 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, 190–200.
- Friendly, Michael (1995). Conceptual and Visual Models for Categorical Data. *The American Statistician*, Vol. 49, No. 2, May 1995, 153–160.
- Hartigan, J.A. & Kleiner, B. (1981). Mosaics for Contingency Tables. *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, 268–273, ed. W.F. Eddy, Springer, New York.
- Hummel, Jürgen (1996). Linked Bar Charts: Analysing Categorical Data Graphically. *Computational Statistics*, Vol. 11, Issue 1, 23–33.
- Nagel, Matthias, Benner, Axel, Ostermann, Rüdiger & Henschke, K. (1996). *Graphische Datenanalyse*. Fischer, Stuttgart.
- Report on the Loss of the 'Titanic' (S.S.)* (1990). British Board of Trade Inquiry Report (reprint), Allan Sutton Publishing, Gloucester, UK.
- Riedwyl, H. , Schuepbach, M (1994). Parquet diagram to plot contingency tables. In *Softstat 93: Advances in Statistical Software*, F. Faulbaum (Ed.). New York, Gustav Fisher.
- Theus, Martin (1995). Trellis Displays vs. Interactive Statistical Graphics. *Computational Statistics*, Vol. 10 Issue 2, 113–127
- Theus, Martin (1996a). MANET — Interactive Graphics for Missing Values. *Proc. of the NTTS'95*.
- Theus, Martin (1996b). *Theorie und Anwendung Interaktiver Statistischer Graphik*. Wißner, Augsburg.
- Theus, Martin & Wilhelm, Adalbert (1996a). Modelling Categorical Data by Interactive Mosaic Plots and Tables. *Proceedings 11th Workshop on Statistical Modelling*.
- Theus, Martin & Wilhelm, Adalbert (1996b). Analysing the Structure of Categorical Data using Interactive Mosaic Plots and the Minimisation of Boolean Functions. *Proceedings in Computational Statistics 1996 / Short Communications and Posters*.
- Unwin, Antony R., Hawkins, G., Hofmann H. & Siegl B. (1996). Interactive Graphics for Data Sets with Missing Values — MANET. *Journal of Computational and Graphical Statistics*, Vol. 4, No. 6
- Velleman, Paul F. (1995). *Data Desk 5.0, Data Description*. Ithaca, New York.