

# Analysing the Structure of Categorical Data using Interactive Mosaic Plots and the Minimisation of Boolean Functions

Martin Theus, Adalbert Wilhelm

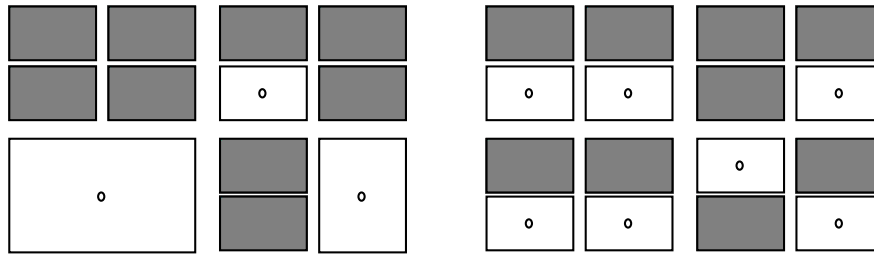
Universität Augsburg  
Institut für Mathematik  
D-86135 Augsburg, Germany

Traditional parametric statistics handles categorical data with  $\chi^2$ -tests in the case of two dimensional data and with log-linear models if more than two variables are involved. Whereas the choice of the model is trivial in the two dimensional case, the choice of a suitable log-linear model is often hard.

Graphical techniques based upon barcharts are not very helpful unless they are embedded in an interactive environment (Hummel, 1995) as e.g. Data-Desk (Velleman, 1995). But looking at particular subsets can only be done by intersecting different barcharts successively. The analyst may lose track of what particular group he is looking at.

Mosaic-plots (Hartigan & Kleiner 1981) seem to solve the problem of single barcharts for each variable. They are able to incorporate lots of variables without a theoretical limit. But the practical use of mosaic-plots (implementations for SAS (Friendly, 1992), S-Plus (Theus, 1995) and MANET (Unwin, 1995) are available) shows quickly, that the possibilities of an interpretation of the mosaic-plot shrinks with the number of variables in the plot. This is because the analyst often has no hint for a specific ordering of the variables. But the visual perception of the plot varies very much with the order of the variables. Few statistical packages offer mosaic-plots, often restricted to two variables as a visualisation of a corresponding  $r \times c$  contingency table, though JMP (JMP, 1995).

Friendly (1994) describes briefly how mosaic-plots can be used to judge log-linear models. But these examples also do not exceed four binary variables. When using even more variables loglinear models do not seem to be the right tool any longer. This is due to the "curse of dimensionality", which yields a lot of empty cells in the mosaic plot. A natural criteria for the optimisation of the ordering of the variables would be to unite as many empty cells as soon as possible (i.e. empty subgroups are not divided if another variable is added to the plot), which is equivalent to a minimisation of the number of cells in the plot. Figure 1 shows an example of a maximum and the minimum number of cells in a mosaic plot of artificial data. A solution for this optimisation problem can be obtained via the minimisation of Boolean functions. A standard technique is the Quine/McCluskey-algorithm (Quine, 1955). To apply the algorithm, which



**Fig. 1.** Best case (left) and worst case (right) empty cells: 'o'.

was devised for Boolean polynomials, to categorical data, each variable has to be taken as a Boolean variable and thus may only have two levels coded by "0" and "1". The function is now constructed via the conjunctive normal form (CNF), which is the product of the sums of the variables (Wegener, 1987), yielding a "1" if a combination was observed, a "0" otherwise. Although the algorithm was constructed for the disjunctive normal form (DNF) it can be applied to the CNF straightforwardly. Resulting prime implicants correspond directly with empty intersections. The restriction to binary variables can be bypassed by generalising the algorithm to multinomial variables. Ongoing research includes the interpretation of the prime implicants for observed cells.

## References

- 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
- 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
- Hummel, Jürgen (1996) Linked Bar Charts: Analysing Categorical Data Graphically *Computational Statistics* Vol. 11, Issue 1.96
- JMP (1995) *Statistical and Graphical Guide*, SAS Institute Inc., Cary, NC
- Quine, W.V. (1955) A way to simplify truth functions *American Mathematical Society* 62, pp627-631.
- Theus, Martin (1995) Implementation of mosaic-plots in S-Plus  
<http://www1.math.uni-augsburg.de/~theus/manet/mosaic.splus>
- Unwin, Antony R. (1995) Interactive Graphics for Data Sets with missing values MANET *Journal of Computational and Graphical Statistics* Vol. 5, No. 2
- Velleman, Paul F. (1995) *Data Desk 5.0* Data Description, Ithaca, New York.
- Wegener, Ingo (1987) *The complexity of Boolean functions* Wiley-Teubner, Chichester-Stuttgart.