Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/
K. Hornik & F. Leisch (eds.) *ISSN 1609-395X*

Visualizing the independence problem using extended association and mosaic plots

David Meyer [*] **Achim Zeileis** [†] **Kurt Hornik** [‡]

1 Introduction

Statistical models built for the analysis of multivariate data, quickly become complex with increasing dimensionality. One idea of visualization techniques is to use the human visual system to detect structures in the data that possibly are not obvious from solely numeric output (e.g., test statistics). The R package 'vcd'—based on the Book 'Visualizing Categorical Data' (Friendly, 2000)—includes methods for the (mostly graphical) exploration of categorical data, such as:

- fitting and graphing of discrete data,
- plots for the independence and symmetry problems, and
- visualization techniques for log-linear models.

In this talk, we focus on the visualization of the independence problem, typically analyzed using a table of relative frequencies $\pi_{ij\dots}$ with two or more dimensions. In the setting of bivariate problems, the null hypotheses is $\pi_{ij} = \pi_{i+}\pi_{+j}$. With additional variables (often used to stratify the data), more independence models can be envisaged, including the null hypotheses of:

- total independence: $\pi_{ijk\dots} = \pi_{i++\dots}\pi_{+j+\dots}\pi_{++k\dots}\cdots$

[*]Institute of Statistics and Probability Theory, Vienna University of Technology
[†]Institute of Statistics and Probability Theory, Vienna University of Technology
[‡]Institute of Statistics, Vienna University of Economics and Business Administration

- conditional independence: $\pi_{ijk\ldots} = \pi_{i|k\ldots}\pi_{j|k\ldots}$
- joint independence: $\pi_{ijk\ldots} = \pi_{ij+\ldots}\pi_{++k\ldots}$

Classical non-graphical methods for these problems include the Chi-square test, Fisher's exact test (for fourfold tables), the Cochran-Mantel-Haenzel test (for $2 \times 2 \times n$-tables), and the analysis of log-linear models for more complex settings.

2 Association Plots

Association plots (Cohen, 1980) are inspired by the classical Chi-squared test of independence for two categorical variables, which—based on a $m \times n$-table of observed frequencies o_{ij}—is performed using the test statistic

$$\chi^2 = \sum_i^m \sum_j^n r_{ij}^2$$

where the *Pearson-residuals* r_{ij} are the standardized deviations of the observed from the expected frequencies:

$$r_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

Under the null of independence, $e_{ij} = o_{i+}o_{+j}/o_{++}$, and χ^2 follows a Chi-squared distribution with $(m-1)(n-1)$ degrees of freedom.

The association plot precisely visualizes the table of Pearson-residuals (see Figure 1): each cell is represented by a rectangle that has (signed) height proportional to the corresponding Pearson-residual r_{ij} and width proportional to the square root of the expected counts $\sqrt{e_{ij}}$. Thus the area is proportional to the raw residuals $o_{ij} - e_{ij}$. The sign of the residual is redundantly coded by orientation and color of the corresponding rectangle.

3 Mosaic Plots

Mosaic Plots can be seen as an extension of grouped bar charts, where width and heights of the bars show the relative frequencies of the two variables: a mosaic plot simply consists of a collection of tiles whose sizes are proportional to the observed cell frequencies (see Figure 2).

Sequential horizontal and vertical recursive splits are used to visualize the frequencies of more than 2 variables, each new variable conditional to the previously entered variables. A first extension by (Friendly, 1994) uses a color coding of the tiles to visualize deviations (residuals) from a given log-linear model fitted to the table, that is, from the expected frequencies under arbitrary (independence) hypotheses—the color of each tile depending on the sign, and the brightness being proportional to the corresponding residual.

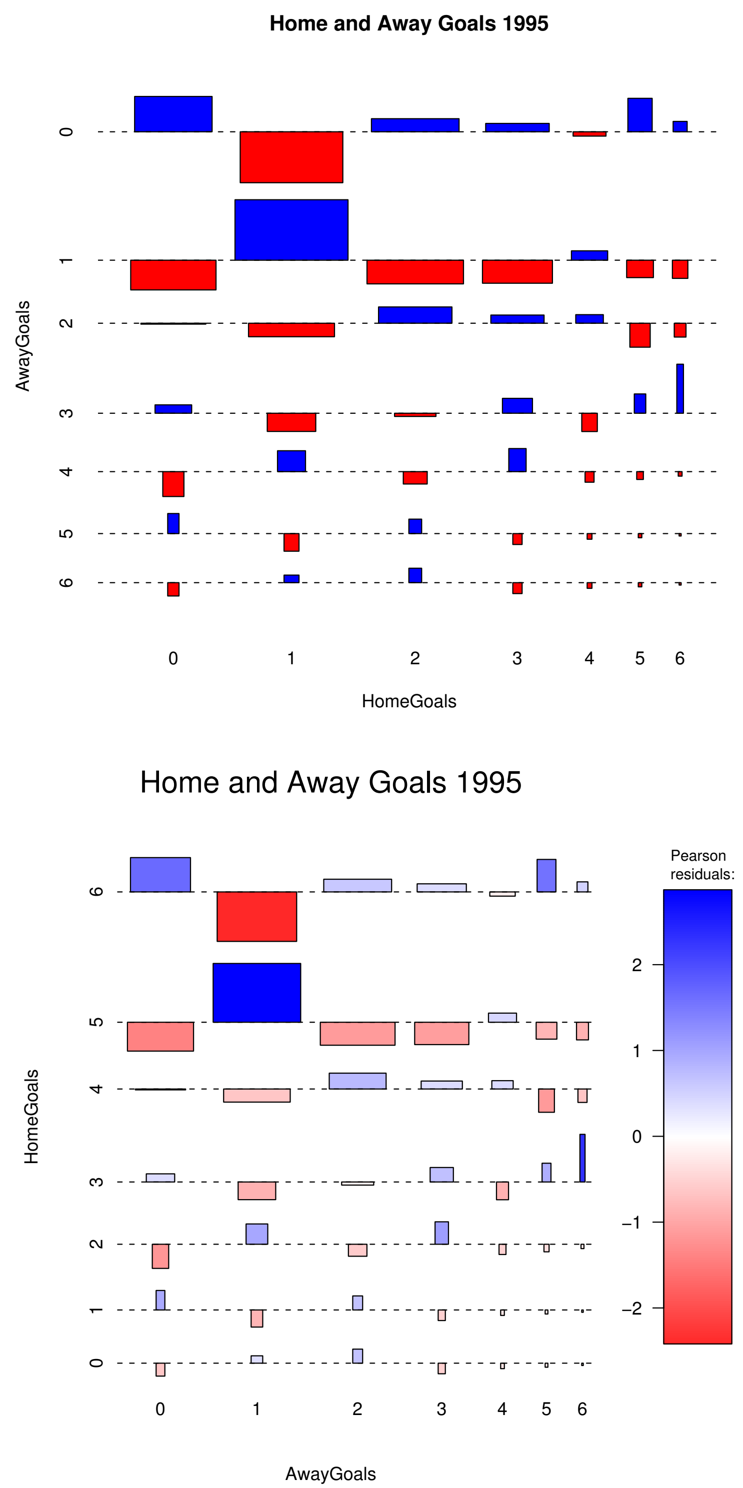

Figure 1: 'Classical' (top) and extended (bottom) association plot for the 1995 home and away goals of the Deutsche Bundesliga. The shading gives more information on the importance of the Pearson residuals.

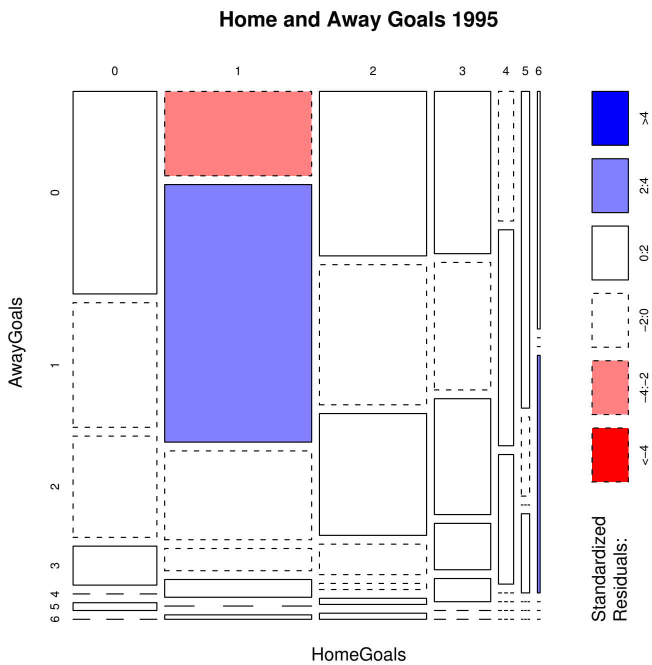

Figure 2: Mosaic plot (with Friendly extensions) for the 1995 home and away goals of the Deutsche Bundesliga. As in extended association plots, the shading visualizes the deviations (residuals) from the independence model.

4 Extensions

Our extensions to these two visualization methods cover three topics: improvement of color coding, visualization of the overall test of independence, and flexible specification of graphical parameters. The color schemes used by Michael Friendly in his implementations are rather poor, due to limitations of the SAS software used that only implements a discrete color palette in the 'RGB' color space that does not, e.g., allow for copier proofness and homogeneous saturations over different colors. Our implementation uses continuous color schemes in the device-independent 'HCL' (Hue-Chroma-Luminance) color space. The continuous scheme is more appropriate to visualize the importance of the Pearson residuals as any possible discretization of the scale.

In addition to the 'local' independence information (shading of the residuals), we also offer to simultaneously visualize the overall independence model by adding white stripes to the rectangles as long as a specified test does not reject the null—the colors, then, appear 'lighter' than without stripes. The user currently can choose the test—either the classical Chi-Square test, or the maximum-residual test. The latter uses a (sampled) conditional exact distribution of the maximal Pearson residual (compared to the sum of squared residuals used for the Chi-squared test). Using this maximum-test assures consistency between the 'local' and 'global' independence information: when the overall model is significant, the plot indeed visualizes the 'guilty' residuals.

Finally, one might wish to use graphical parameters varying from tile to tile to visualize additional information. Our implementation offers full control over each element by using a call-back function invoked each time a tile is plotted and that is given the position information (that is, the index in the visualized table).

5 Conclusion

Both methods described (association plot and mosaic plot) are used to visualize the independence problem. Our extensions include better color schemes, visualization of the overall independence model, and more flexible manipulation of the graphical parameters. Future Work will include extensions for more than two dimensions (such as conditional plots or trellis layouts), and improved labelling.

References

Cohen, A. (1980). On the graphical display of the significant componentsin a two-way contingency table. *Communications in Statistics-Theory and Methods*, A9:1025–1041.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200.

Friendly, M. (2000). *Visualizing Categorical Data*. SAS Insitute, Carey, NC.