

Visualizing Loglinear Models

Martin THEUS and Stephan R. W. LAUER

We consider visual methods based on mosaic plots for interpreting and modeling categorical data. Categorical data are most often modeled using loglinear models. For certain loglinear models, mosaic plots have unique shapes that do not depend on the actual data being modeled. These shapes reflect the structure of a model, defined by the presence and absence of particular model coefficients. Displaying the expected values of a loglinear model allows one to incorporate the residuals of the model graphically and to visually judge the adequacy of the loglinear fit. This procedure leads to stepwise interactive graphical modeling of loglinear models. We show that it often results in a deeper understanding of the structure of the data. Linking mosaic plots to other interactive displays offers additional power that allows the investigation of more complex dependence models than provided by static displays.

Key Words: Interactive stepwise graphical modeling; Loglinear models; Mosaic plots; Response models.

1. MOSAIC PLOTS

Mosaic plots (Hartigan and Kleiner 1981; Friendly 1994, 1995; Theus 1996, 1997) are defined as a recursive generalization of barcharts. In a simple barchart, bars represent categories. The heights of the bars are proportional to the number of observations falling into this particular category. Since the width of each bar is identical, the area is proportional to the number of observations. Starting with a barchart, Figure 1 shows the systematic construction of a mosaic plot for the Titanic data, given in Table 1.

- The barchart of the variable *Class* of the Titanic dataset is shown in the upper left plot in Figure 1. In every bar of the barchart the proportion of survivors is highlighted. It is not easy to compare the proportions in each class, since the bars have different heights.
- To solve this problem, we modify the barchart to have identical height instead of width, and proportional width instead of height. In this plot, called *spineplot* (see Hummel 1996), the proportions of survivors, still drawn from bottom to top, can

Martin Theus is Senior Technical Staff Member, Statistics Research, AT&T Labs, 180 Park Avenue, Florham Park, NJ 07932 (Email: theus@research.att.com). Stephan R. W. Lauer is Research Assistant, Department of Computer Oriented Statistics and Data Analysis, University of Augsburg, 86135 Augsburg, Germany (Email: lauer@math.uni-augsburg.de).

©1999 American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America
Journal of Computational and Graphical Statistics, Volume 8, Number 3, Pages 396–412

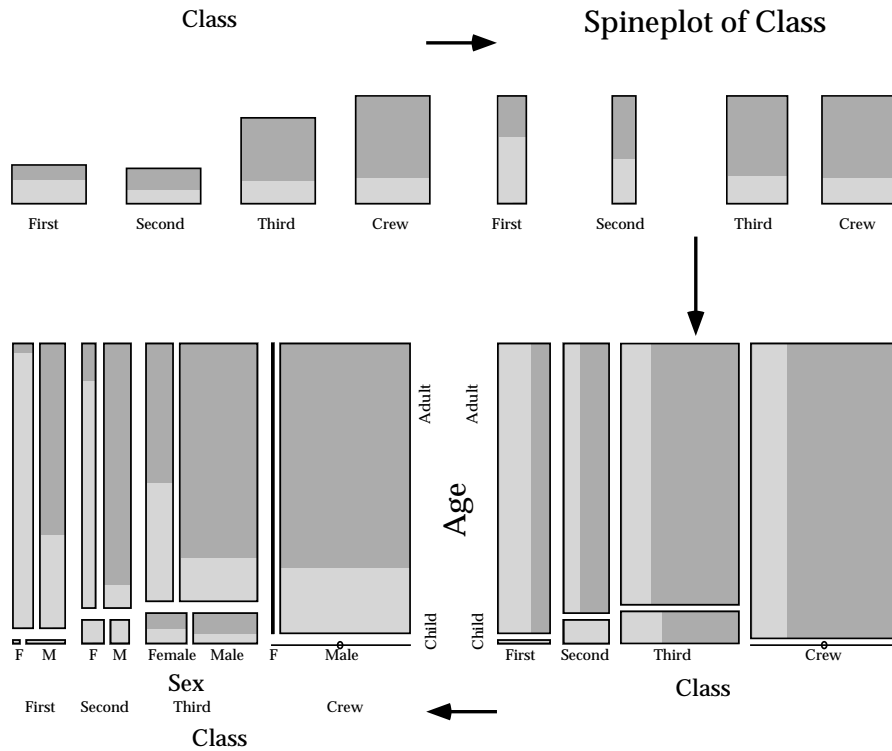


Figure 1. Systematic development of a mosaic plot for the Titanic data. Starting with a one-dimensional bar chart for *Class* the mosaic plot for the variables *Class*, *Age* and *Sex* evolves as we follow the plots clockwise. Passengers who survived are highlighted, adding another binary variable to the plot.

be compared easily. A spineplot for the variable *Class* is shown in Figure 1 top right.

- A spineplot can already be regarded as a one-dimensional mosaic plot. In the lower right plot in Figure 1 an example of a *two-dimensional mosaic plot* is shown. Whereas we divided the first variable along the *x*-axis, we divide for the second variable along the *y*-axis. This division is done conditional upon the different levels of variable 1. In the Titanic example the different classes have been divided according to the conditional proportions of adults and children inside the classes. The areas are thus proportional to the absolute numbers inside the particular class. Since there were no children in the crew, this category is empty, indicated by an area of size 0—that is, a line. To distinguish between very small nonempty groups, which may be rendered by just a single line, and really empty groups, an additional zero (“0”) is drawn centered in this line. This feature extends the definition of mosaic plots, and can be found implemented in the MANET software (Theus, Hofmann, Siegl, and Unwin 1997). The highlighting of the survivor proportions, formerly drawn from bottom to top, now is drawn from left to right. This is sensible, since further divisions in a mosaic plot will alternate between *x*- and *y*-direction. If the last division in a mosaic plot was

Table 1. Tabulation of the 2201 Titanic passengers

Titanic			Class			
Survived	Age	Gender	First	Second	Third	Crew
No	Child	Male	0	0	35	0
		Female	0	0	17	0
	Adult	Male	118	154	387	670
		Female	4	13	89	3
Yes	Child	Male	5	11	13	0
		Female	1	13	14	0
	Adult	Male	57	14	75	192
		Female	140	80	76	20

made along x , a potential highlighting is drawn along y , and vice versa.

- Incorporating the variable *Gender* we get a *three-dimensional mosaic plot*, which is shown in the lower left plot of Figure 1. Following the above algorithm this subdivision is done along x , conditional upon each combination of the variables *Class* and *Age*.

Since the passengers who survived the sinking of the Titanic are highlighted in all four plots of Figure 1, the reader may check whether the often-claimed policy “Women and children first!” holds true or not.

2. LOGLINEAR MODELS

2.1 TWO DIMENSIONS

In two dimensions only two distinct models can occur, in general. Either the two variables are independent or they are associated. These two cases are depicted in Figure 2 for two binary variables. In this section we prefer to use modeled (and thus artificial) data, to meet the graphical representation exactly. Thus, we leave out all labels, in order to focus the reader’s attention on the principal shape of the plots.

The case of independence is shown in the left plot of Figure 2. The separating gaps between all levels and variables line up like a city-block layout. This property, indicating the presence of mutual independence, generalizes to more than two levels—that is, $r \times s$ tables, but is visually more demanding. The case of interaction is depicted in the right plot. The separating gaps between the levels of the second variable, which is displayed conditional upon the levels of the first variable, do not line up. This kind of departure from equal conditional probabilities can be detected easily for more than two levels.

The question whether this interaction is statistically significant or not remains unrevealed, as long as the number of underlying observations is unknown.

2.2 THREE DIMENSIONS

Given a three-dimensional contingency table the model assumes a sample of size n distributed over $IJK = N$ cells. Under multinomial sampling (no fixed margins), the probability that an observation will fall into cell ijk is then π_{ijk} for all $i = 1, \dots, I, j =$

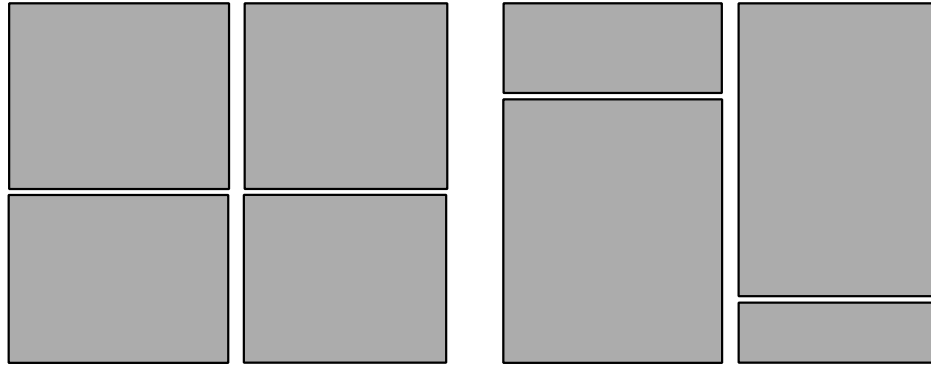


Figure 2. Independence (left) and interaction (right) in a 2x2 mosaic plot.

$1, \dots, J, k = 1, \dots, K$. The expected value m_{ijk} thus is $n\pi_{ijk}$.

Mutual independence of the three variables is now equivalent to $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ as well as $m_{ijk} = n\pi_{i++}\pi_{+j+}\pi_{++k}$, where the index “+” means to sum up the values over this index. Switching to a logarithmic scale we obtain

$$\log(m_{ijk}) = \log(n) + \log(\pi_{i++}) + \log(\pi_{+j+}) + \log(\pi_{++k}), \tag{2.1}$$

which is equivalent to

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z, \tag{2.2}$$

with

$$\begin{aligned} \lambda_i^X &= \log(\pi_{i++}) - \sum_{\nu} \log(\pi_{\nu++})/I \\ \lambda_j^Y &= \log(\pi_{+j+}) - \sum_{\nu} \log(\pi_{+\nu+})/J \\ \lambda_k^Z &= \log(\pi_{++k}) - \sum_{\nu} \log(\pi_{++\nu})/K \\ \mu &= \log(n) + \sum_{\nu} \log(\pi_{\nu++})/I + \sum_{\nu} \log(\pi_{+\nu+})/J + \sum_{\nu} \log(\pi_{++\nu})/K. \end{aligned}$$

The parameters $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$ and $\{\lambda_k^Z\}$ satisfy

$$\sum \lambda_i^X = \sum \lambda_j^Y = \sum \lambda_k^Z = 0. \tag{2.3}$$

Model (2.2) is the model of *mutual independence* for a three-dimensional contingency table. Without the constraint (2.3) it would not be possible to identify the parameters uniquely.

Analogous to the classical ANOVA modeling, interactions between two or all three variables can be modeled. Introducing the additional terms $\{\lambda_{ij}^{XY}\}$, $\{\lambda_{ik}^{XZ}\}$, $\{\lambda_{jk}^{YZ}\}$, and $\{\lambda_{ijk}^{XYZ}\}$ we obtain

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \tag{2.4}$$

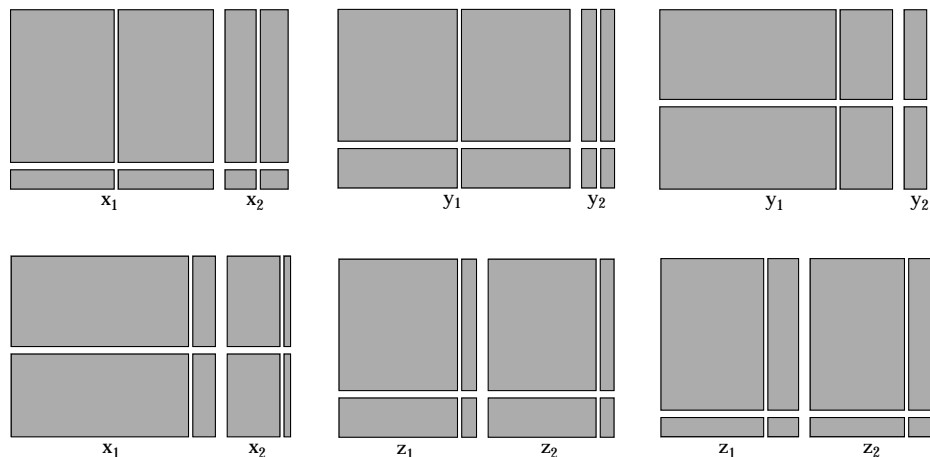


Figure 3. Mutual independence. Top: xyz , yxz , yzx . Bottom: xzy , zxy , zyx . (left to right)

with zero sums over the parameters—that is,

$$\begin{aligned} \sum_i \lambda_i^X &= \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} \\ &= \sum_j \lambda_{ij}^{XY} = \dots = \sum_k \lambda_{ijk}^{XYZ} = 0. \end{aligned} \quad (2.5)$$

Using this model, we are now able to formulate the following interaction structures:

1. *Mutual independence*—(presence of λ^X , λ^Y , and λ^Z).
2. *Partial independence*—(additional presence of one λ^{AB} , $A, B \in \{X, Y, Z\}$, $A \neq B$).
3. *Conditional independence*—(all parameters except λ^{XYZ} and one λ^{AB} , $A, B \in \{X, Y, Z\}$, $A \neq B$).
4. *No three-way interaction*—(all parameters except λ^{XYZ}).
5. *Three-way interaction*—(includes all parameters; the saturated model).

Each of these models can be visualized using mosaic plots. The topological shape of the plot of each model-type is in principle uniquely defined, but the exact shape varies with the quantity of the interaction. Since the shape of mosaic plots is not invariant against a permutation of the variables, we like to show all possible permutations. A mosaic plot including three variables can be set up in $3! = 6$ different orderings of the included variables. The next sections will show the five different interaction-types for $2 \times 2 \times 2$ tables in detail.

2.2.1 Mutual Independence

The model of mutual independence is given in Equation (2.2).

All variables are pairwise independent. Thus, only the *main effects* $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$, and $\{\lambda_k^Z\}$ appear in the model. Figure 3 shows the city-block layout (see Section 2.1) for all six orderings, which depicts mutual independence. Note that the partitioning of

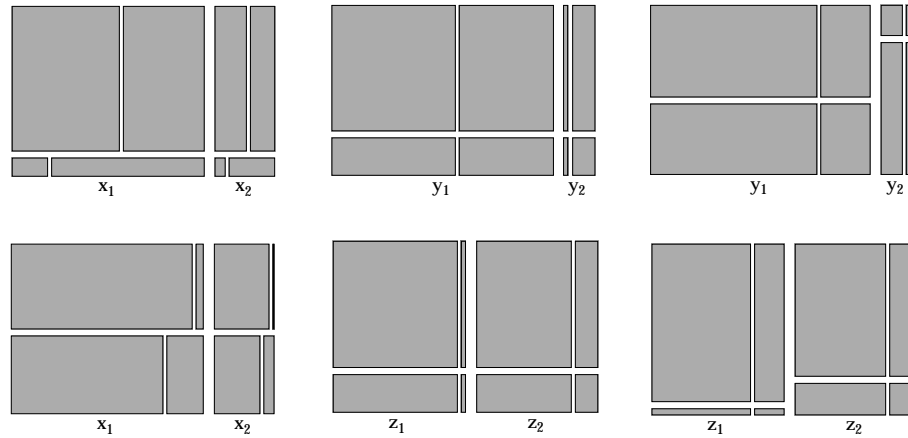


Figure 4. Partial independence of X versus YZ . Top: xyz , yxz , yzx . Bottom: xzy , zxy , zyx . (l. t. r.)

all subblocks follows the same proportions.

2.2.2 Partial Independence

The variable X is partially independent of Y and Z , if

$$\pi_{ijk} = \pi_{i++} \pi_{+jk} \quad \forall i, j, k \tag{2.6}$$

holds. Thus, the composite variable YZ , which has JK different levels combinations of Y and Z is mutually independent of X . The corresponding log-linear model is

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} \tag{2.7}$$

Looking at Figure 4 we find the interaction between the last two variables in the two leftmost plots, in which Y and Z are the last variables entered in the mosaic. The two plots in the middle show independence between the first two and the last two variables. The interaction between the first and the last variable (Y and Z) is reflected by the different partitioning inside the two subblocks of the first variable.

The two rightmost plots show the interaction between the two first variables. The presence of partial independence in each column in Figure 4 is shown in the same way, since the jointly dependent variables Y and Z are exchanged from the top display to the bottom display. This results in transposed plots, which represents only a quantitative, but not a qualitative change.

2.2.3 Conditional Independence

Conditional independence means

$$\pi_{ij|k} = \pi_{i+|k} \pi_{+j|k} \quad \forall i, j, k, \tag{2.8}$$

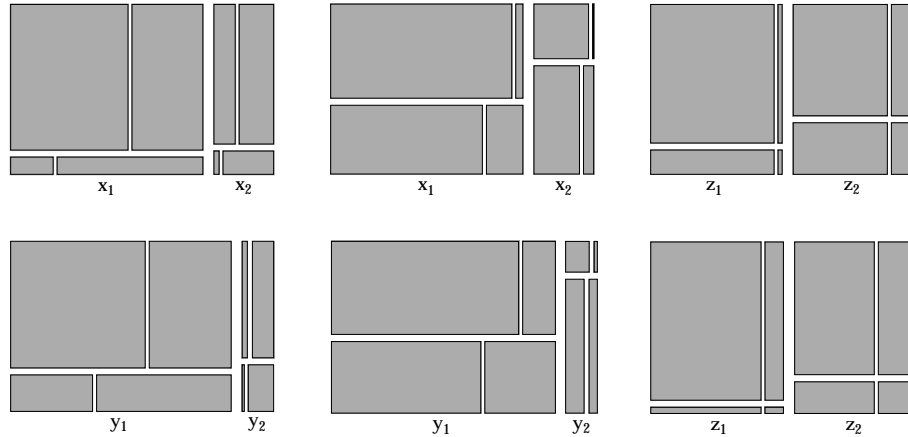


Figure 5. Conditional independence of X and Y , given Z . Top: xyz , xzy , zxy . Bottom: yxz , yzx , zyx .

with

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{ij+}}, \quad \pi_{i+|k} = \frac{\pi_{i+k}}{\pi_{i++}}, \quad \text{and} \quad \pi_{+j|k} = \frac{\pi_{+jk}}{\pi_{+j+}}.$$

The corresponding loglinear model thus is

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{2.9}$$

The graphical representation of Equation (2.9) is fairly obvious. If we look at the two rightmost plots, we find the typical independence structure of X and Y given the two levels (left and right subplot) of Z . Note that the partitioning inside the two levels of Z does *not* have the same proportions. This is the distinction between the conditional independence in Figure 5 and the similar shapes in Figure 4!

Looking at the other four plots in Figure 5 we find *no* further pairwise independence structures. Again two triples of variable orders (here: zxy and zyx) depict the model best.

2.2.4 No Three-Way Interaction

This model is closest to the saturated model. Although every variable interacts with each other variable, there is no interaction between all three variables. Although this might be hard to interpret, the mathematical representation is easy to obtain; see Equation (2.10).

The loglinear model of no three-way interaction can be written as

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{2.10}$$

Comparing Figure 6 with Figure 7 we hardly find any differences in the plots. (Figure 7 shows the raw data which has been used to set up all the different models depicted in Figures 3 to 7.)

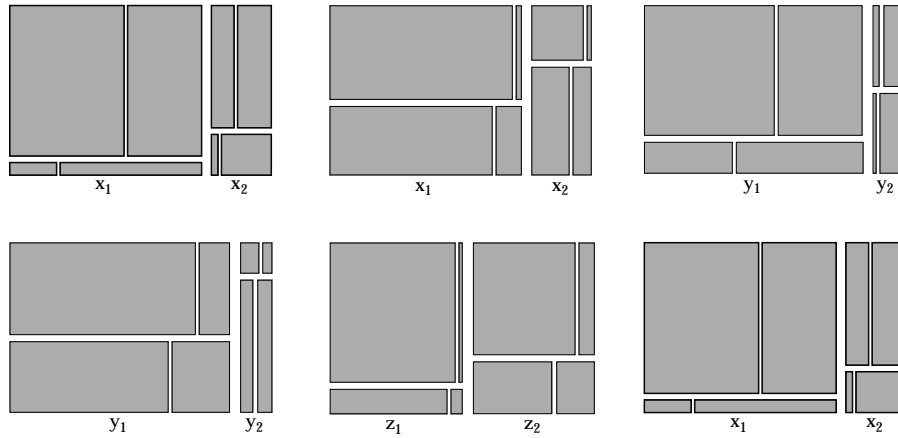


Figure 6. No three-way interaction. Top: xyz , xzy , yxz . Bottom: yzx , zxy , zyx . (left to right.)

2.2.5 Three-Way Interaction (Saturated Model)

This is the model where every possible interaction is included. Obviously the expected values e_{ijk} match the observed values o_{ijk} exactly, resulting in a zero χ^2 and G^2 statistic. The only interpretation of this model is the fact that apparently all other models failed to represent the data in a suitable way.

2.3 STEPWISE GRAPHICAL MODELING

The most common problem in loglinear modeling is to find the most suitable model. Suitable in this contexts means that the model includes as few interaction terms as possible and declares as much of the deviation from mutual independence as possible.

Since this problem is very similar to the stepwise regression process in classical

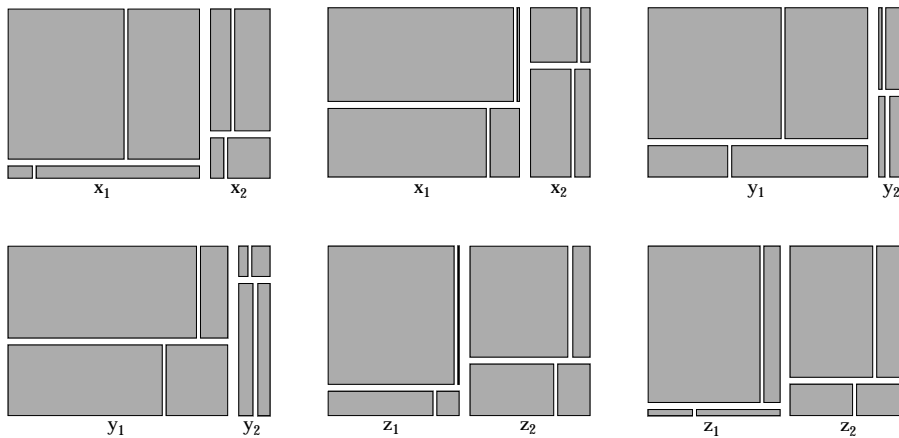


Figure 7. Three-way interaction. Top: xyz , xzy , yxz . Bottom: yzx , zxy , zyx . (left to right.)

Table 2. Cross-classification of 1008 Detergent Consumers

M-User?	Detergent usage		Watersoftness		
	Temperature	Preference	Hard	Medium	Soft
no	low	X	68	66	63
		M	42	50	53
	high	X	42	33	29
		M	30	23	27
yes	low	X	37	47	57
		M	52	55	49
	high	X	24	23	19
		M	43	47	29

multiple linear regression, stepwise procedures are also popular in loglinear modeling. These stepwise procedure for loglinear models (see Christensen 1997) usually use χ^2 and G^2 statistics to judge the adequacy of a model.

Using mosaic plots, a graphical stepwise selection can be applied to loglinear models. This allows a visual insight into the structure of a model. Analogous to the statistics-based selection procedures a *graphical backward selection* as well as a *graphical forward selection* can be applied. To illustrate both techniques, we use the *Detergent Usage* dataset (see Table 2), already analyzed in, for example, Venables and Ripley (1994, pp. 196–198); Cox and Snell (1989, pp. 86–90); and Fienberg (1985, pp. 71–80).

The mosaic plot of all four variables of the dataset is shown in Figure 8. The order of the variables is chosen as follows. Since *Watersoftness* and *Temperature* are given variables we put them at position one and two. The highest interaction is to be expected between *Preference* and *M-User?*. Putting these variables at position three and four—that is, the two last positions—enables us to investigate this interaction conditioned on the six different level-combinations of *Watersoftness* and *Temperature*.

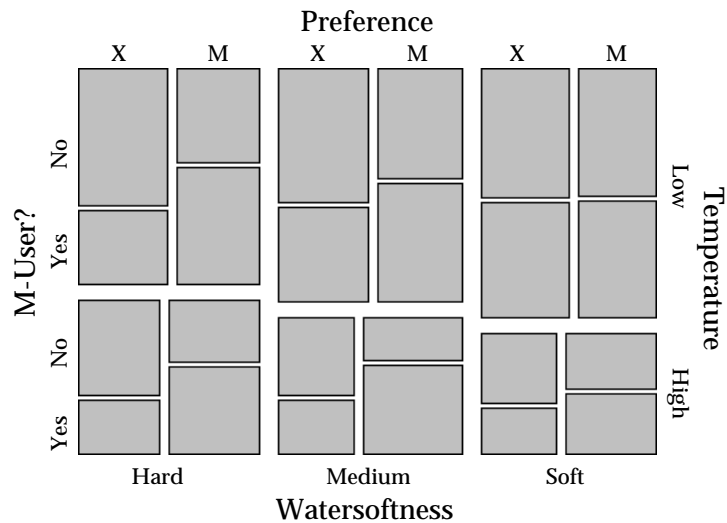


Figure 8. Mosaic plot for all four variables of the Detergent dataset. Raw data—that is, the saturated model shown.

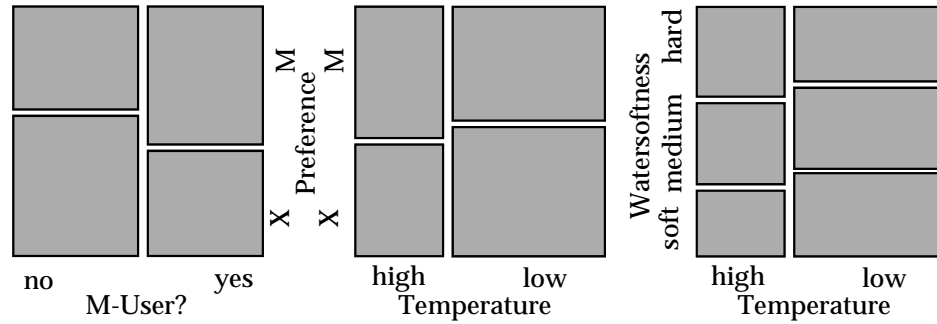


Figure 9. Two-way-interactions in the Detergent dataset.

2.3.1 Graphical Backward Selection

As known from the standard stepwise procedures, a backward selection starts with the saturated model, which includes all interaction terms. Here we start with three-way and two-way interactions, and remove all independent tuples, step by step, assuming that real datasets do not include high-order terms without the corresponding significant low-order terms. Whenever we remove a term, all interactions of higher order which include this term are also excluded, since we are only interested in hierarchical models. Investigating all two-way interactions (there are $6 = \binom{4}{2}$ for a dataset with 4 variables), we find each three interactions *M-User?:Preference*, *Temperature:Preference*, *Temperature:Watersoftness* depicted in Figure 9, and three nonsignificant pairs *Temperature:M-User?*, *Watersoftness:M-User?*, *Watersoftness:Preference* depicted in Figure 10. In the example of the *Detergent Usage* dataset, we remove the three insignificant interaction-terms shown in the corresponding mosaic plots in Figure 10. Although we now know the different shapes of mosaic plots of interactions for three variables (see Section 2.2), there is no need to look for the presence of a three-way interaction, since there are no three pairwise independent variables. Thus, we have to remove all higher interaction terms of order three and four, resulting in the model which includes the interaction shown in Figure 9.

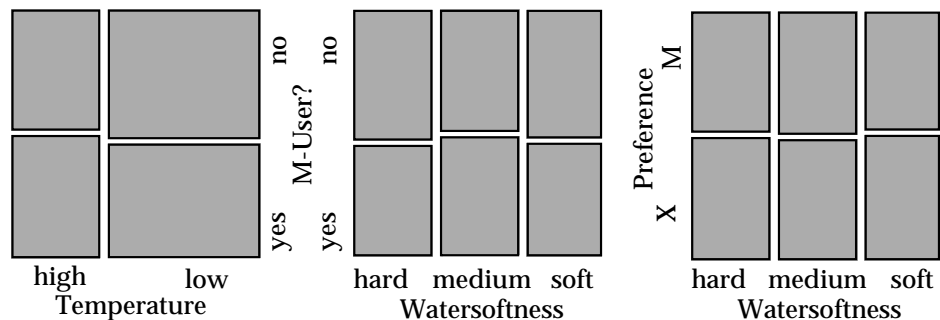


Figure 10. Independent pairs in the Detergent dataset.

2.3.2 Graphical Forward Selection

Graphical forward selection starts with plotting the mosaic plot for the model of mutual independence. To visualize this loglinear model we have basically three options:

- 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. As shown in Section 2, standard dependency structures for two and three variables have certain shapes, which do not depend on the underlying data.

The second approach was used by Friendly (1994). The sign 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 certain 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 and Schuepbach (1994), who used cross-hatching to reflect cell counts. In both cases the use of hatching seems to be a concession to the underlying low-level plotting routines and journal publication practices.

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, what cannot occur in a plot of the expected values of a model.

This leads to the third approach and the following procedure:

1. Display the residual information in the mosaic plot of the corresponding model.
2. Use color (red or blue) for the direction (negative or positive) of the standardized error: $r_i := (o_i - e_i) / \sqrt{e_i}$.
3. Use areal highlighting for the quantity of the error. Since the absolute difference between observed and expected values can be larger than the expected value, the highlighting can only be done by using relative residuals
4. Scale all residuals according to the absolute sum of all residuals—that is, calculate $r_i^* := \frac{|r_i|}{\sum (|r_i|)} \cdot 4$. It is easy to prove that all r_i^* lie in the interval between 0 and 1, because no single residual can contribute more than 25% to the absolute total sum of the residuals.
5. Scale the brightness of the red and blue error highlighting (i.e., the corresponding value of the RGB channel) with the α -quantile of the G^2 -statistics, such that errors of statistically insignificant models can hardly be seen.

Basically the incorporation of residuals into the plot should enable three things:

- The ability to judge the structure of the residuals; that is, to find potential not yet included interaction terms (pt. 2). In Figure 11 there is obviously a strong interaction between *Preference* and *M-User?*, revealed by the regular overcrossing of blue and red highlighting in all six subgroups of *Watersoftness* and *Temperature*.
- When modeling high-dimensional categorical data, we are interested in the low-dimensional interactions that contribute most to the sum of errors. Including in-

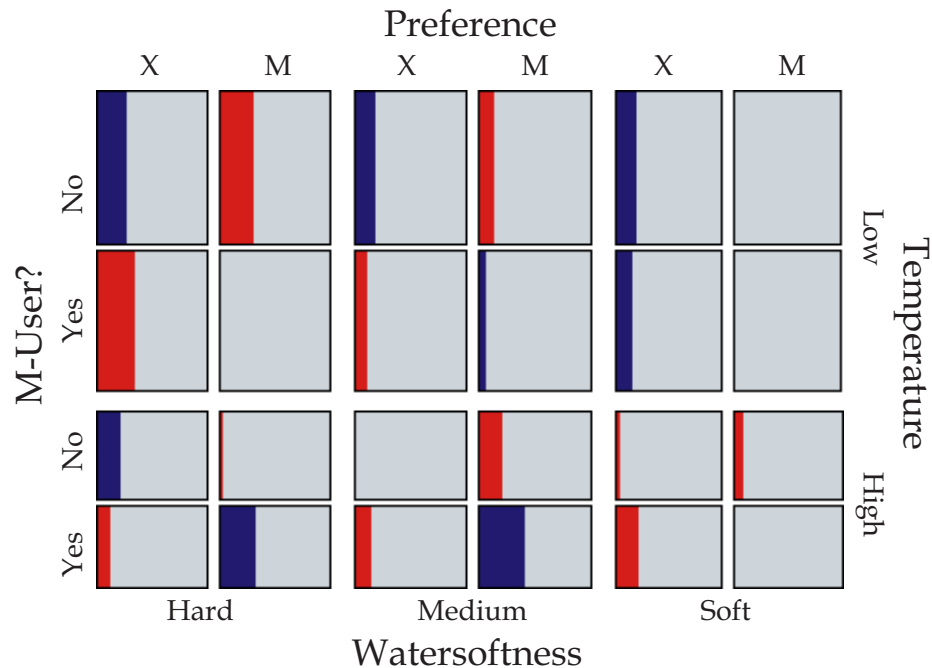


Figure 11. Mosaic plot for the model of mutual independence. The residuals are incorporated according to the previously mentioned scheme. The variable order is the same as in Figure 8.

teractions which show a great amount of error highlighting in the corresponding mosaic plot will thus decrease the G^2 -statistics most (Point 4). The top left plot in Figure 12 shows the interaction between *M-User?* and *Preference*. This interaction contributes by far the biggest amount of residuals to the current model, indicated by the big amount of error highlighting in the four cells. Adding this interaction will reduce the G^2 -statistics to almost 50%.

- Looking for further interactions may be senseless if the model is already statistically insignificant. Point 5 thus implements an optical stopping criterion for the graphical forward selection. Figure 11 shows bright colors in the error highlighting. This is due to the corresponding p value being less than .001. After we added the first interaction the p value of the resulting model increased to .17, which is no longer significant. Thus, the brightness is increased and the colors start to fade; see Figure 12 top right. Including more and more interactions—see Figure 12—increases the brightness more and more, indicating the statistical irrelevance of the last interactions added.

In Figure 12 the entire stepwise modeling process is depicted. The resulting model includes all significant two-way interactions. Starting from the regular shape in Figure 11, every interaction adds more deviation from the regular grid. The inclusion of the interaction of *M-User?* and *Preference* is shown in the mosaic plot in Figure 12 top

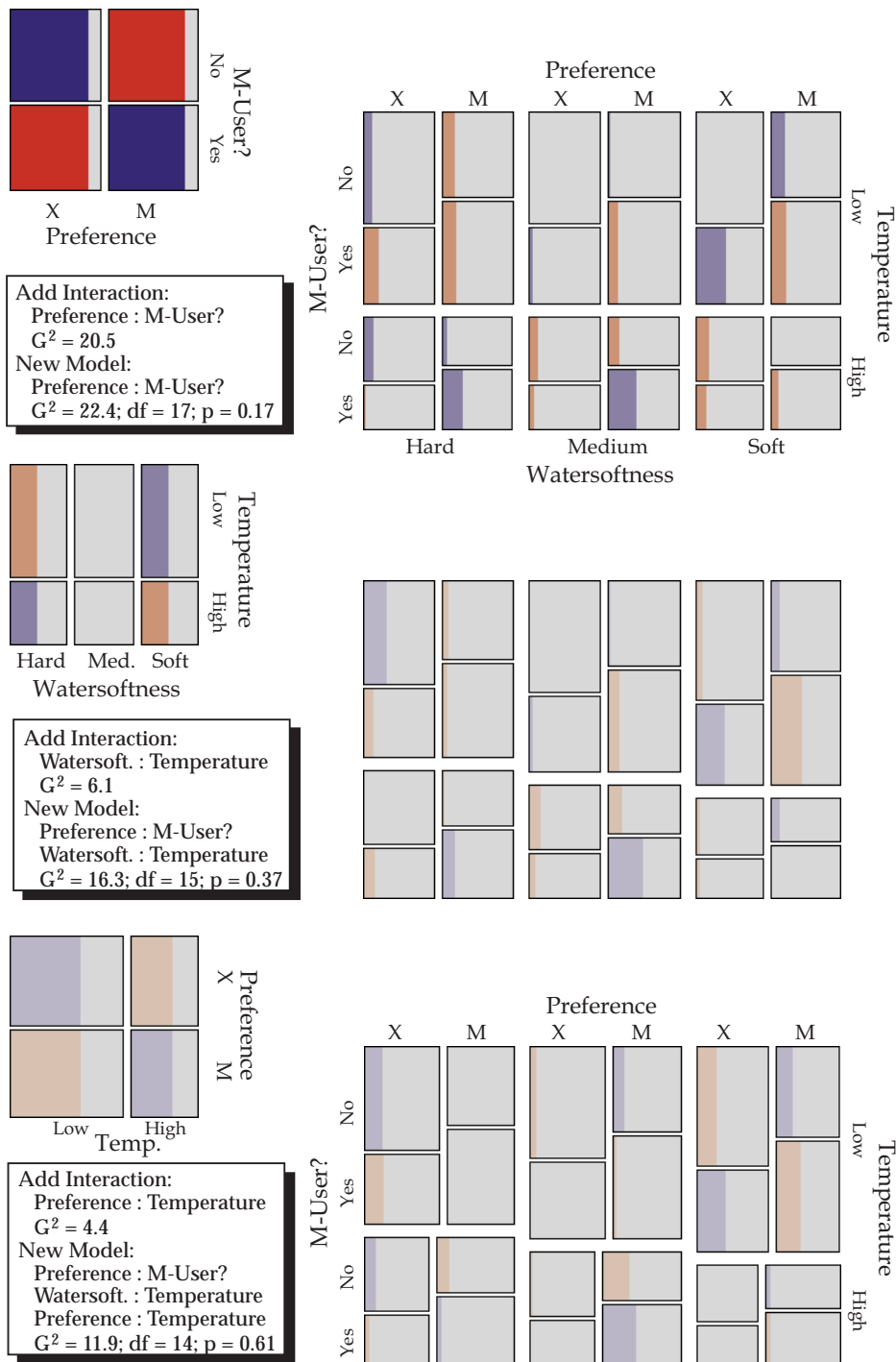


Figure 12. Evolving modeling process for the Detergent data. The left-hand plots show the interaction which is added. The right-hand plots show the resulting, updated full mosaic. Each step has its corresponding statistics attached.

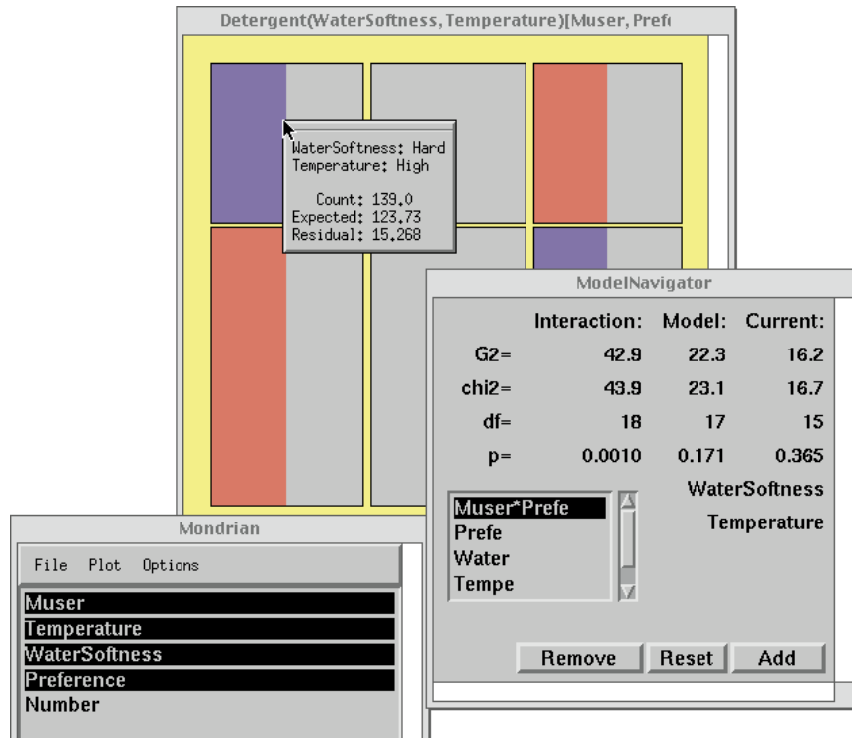


Figure 13. Sample screen-shot of a Mondrian session.

right. It is easy to see that the interaction is the same for all combinations of *WaterSoftness* and *Temperature*, which is the next interaction to be added. Adding this interaction (see Figure 12, middle right) adjusts the sizes of the different groups of *WaterSoftness* and *Temperature*. The last interaction cannot be seen as easily as the first two, since it is not between the first two or last two variables. The interaction between *Temperature* and *Preference* (variable two and three in the chosen order) can be seen in the splitting between *Preference X* and *Preference M*, which is different for the two levels of *Temperature* (see Figure 12 lower right).

2.3.3 Software Support

Whereas the graphical backward selection could probably be done by using static print-outs of the different interactions, the graphical forward selection definitely needs support by an interactive computing environment.

The previously described method for incorporating the residuals of a model into a mosaic plot is implemented in the *Mondrian* (Theus in press) data-visualization tool. A sample screen-shot of a *Mondrian* session is shown in Figure 13. The lower left plot in Figure 13 shows the start-up window of *Mondrian*. The upper plot shows the same mosaic plot as in Figure 12 middle left; that is, the interaction between *WaterSoftness* and *Temperature*, after including the interaction between *M-User?* and *Preference*. The right-hand window shows the *ModelNavigator*. Basically the *ModelNavigator* consists of

Table 3. Data on 251 Cesarean Birth

<i>Cesarean birth</i>			<i>Type of infection</i>		
<i>Planned</i>	<i>Antibiotics</i>	<i>Risk factor</i>	<i>T1</i>	<i>T2</i>	<i>None</i>
yes	yes	yes	0	1	17
		no	0	0	2
	no	yes	11	17	30
		no	4	4	32
no	yes	yes	4	7	87
		no	0	0	0
	no	yes	10	13	3
		no	0	0	9

a spreadsheet-like arrangement of the following model information: G^2, χ^2, df, p . These parameters are provided for

- The model set up so far (Model column).
- The model which results if the interaction selected in the list would be removed from the current model; that is, a potential backward step (Interaction column and list below).
- The model that results if the currently in the mosaic plot displayed interaction would be included; that is, a potential forward step (Current column and variable names below).

The reader may compare the model information displayed in the *ModelNavigator* with the information provided in Figure 12.

3. RESPONSE MODELS

In Section 1 we saw an example of a typical response model with the Titanic dataset in Figure 1. In this situation—several influencing variables and few depending variables—we are usually not interested in interactions inside the groups, but in interactions across the two groups.

To achieve this, we put all influence variables into a mosaic plot, and all dependent variables in another mosaic plot. If there is only one dependent variable, we can use a simple barchart. Linking and highlighting the two plots can show the influence of the different variables easily. Table 3 shows a dataset on 251 cesarean births. For each case it was recorded whether the cesarean was *planned* or not; whether there was a certain *risk factor* present or not; and whether *antibiotics* have been given or not. Depending on these influence variables, it was recorded whether an *infection* of type I, type II, or no infection occurred. A more detailed description of both the dataset and its source is given in Fahrmeir and Tutz (1994).

The main question when investigating this dataset is the influence of antibiotics on the infection risk, given the two other variables *risk factor* and *planned*. In Figure 14 three mosaic plots of the variables *risk factor*, *antibiotics*, and *planned* are plotted, together with a barchart of the variable *infection*. The two infection types have been selected and are reflected by the corresponding highlighting in the mosaic plots.

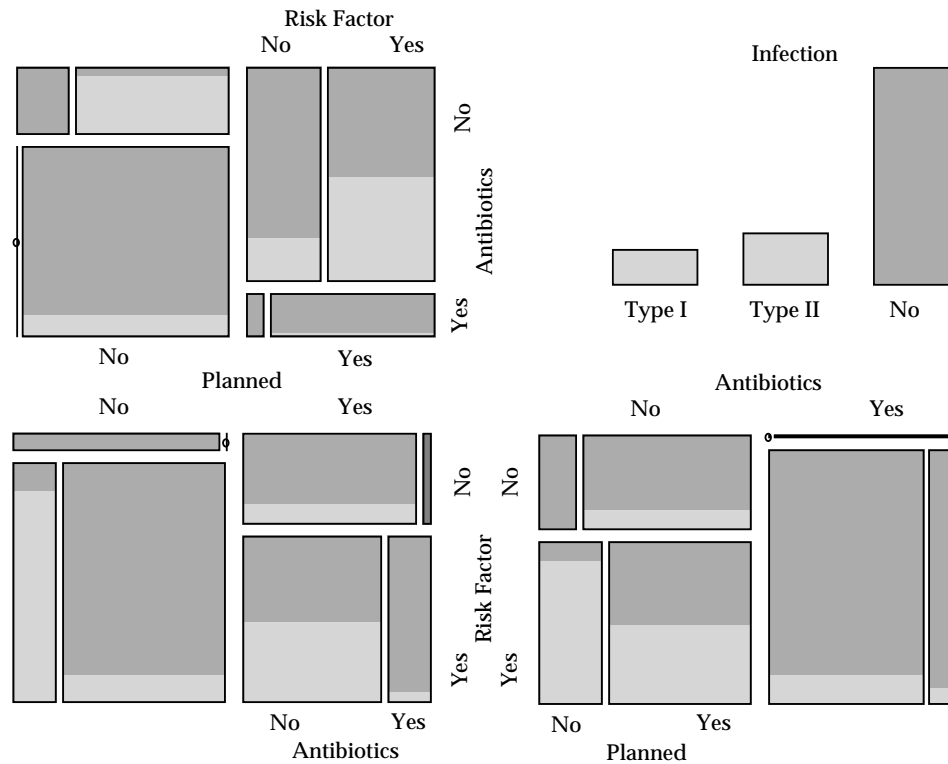


Figure 14. A setup of three different orders of the Cesarean dataset. Each variable is at position three in one of the mosaic plots.

To study the influence of administering antibiotics we look at the lower left mosaic plot in Figure 14. In this plot we can monitor the highlighted proportions of cases with infection according to whether antibiotics have been given or not in adjacent bins. For all four level-combinations of *planned* and *risk factor* we find a clear reduction of the infection risk. This reduction is about 8:1 for women with a *risk factor* and still about 4:1 for all women in the study. The apparently smaller impact of antibiotics for the complete sample is due to Simpson’s paradox; that is, the high interaction between *planned* and *antibiotics*, and the small absolute number of cases in cells with high infection risk.

As we take a closer look at the two remaining mosaic plots in Figure 14, we find one cell which is out of line. In both plots it is the cell for cases where the cesarean was not planned, no antibiotics have been given and no risk factor was found, which is the cell in the upper left corner in all three plots. Whereas in the upper left plot the absence of any case of infection could be explained by a very strong influence of *risk factor*, the absence of infection cases in this particular cell in the lower right plot is odd, since a planned cesarean should have a smaller infection risk than an unplanned one. But looking at the adjacent cell for unplanned cesareans we find 20% infection cases there.

Modeling the two two-way interactions *infection : risk factor* and *infection : planned* results in an expected value of 2.9 infection cases for the cell of “no antibiotics”, “no risk factor”, and “not planned”.

Obviously we have revealed an unexpected result, worthy of further study.

4. SOFTWARE

Currently three implementations of interactive mosaic plots exist. The most advanced and most general implementation can be found in the *MANET* data-visualization package running on Apple Macintosh. *MANET* offers a variety of different linked plots. For more information on *MANET* see the URL <http://www1.math.uni-augsburg.de/Manet>.

The implementation in *Mondrian* is at an earlier stage, but includes the more advanced modeling techniques described previously. *Mondrian* is implemented in JAVA and thus runs on any platform. For more information on *Mondrian* see <http://www1.math.uni-augsburg.de/Mondrian>. A third implementation can be found at <http://www.math.yorku.ca/SCS/Online/mosaics/>. This page, maintained by Michael Friendly, is a WWW front-end to an underlying SAS-function realized in Perl and JAVA-Script. A static implementation of mosaic plots for S-Plus (which is already part of the R-distribution) was presented by Emerson (1998).

[Received December 1997. Revised February 1998.]

REFERENCES

- Christensen, R. (1997), *Log-Linear Models and Logistic Regression*, New York: Springer.
- Cox, D. R., and Snell, E. J. (1981), *Applied Statistics — Principles and Examples*, London: Chapman and Hall.
- Emerson, J. (1998), "Mosaic Displays in S-Plus: A General Implementation and a Case Study," *Statistical Computing and Graphics Newsletter*, 9.
- Fahrmeir, L., and Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer.
- Fienberg, S. (1985), *The Analysis of Cross-Classified Categorical Data*, Cambridge and London: The MIT Press.
- Friendly, M. (1994), "Mosaic Displays for Multi-Way Contingency Tables," *Journal of the American Statistical Association*, 89, 190–200.
- (1995), "Conceptual and Visual Models for Categorical Data," *The American Statistician*, 49, 153–160.
- Hartigan, J. A., and Kleiner, B. (1981), "Mosaics for Contingency Tables," in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ed. W. F. Eddy, New York: Springer, pp. 268–273.
- Hummel, J. (1996), "Linked Bar Charts: Analysing Categorical Data Graphically," *Computational Statistics*, 11, 23–33.
- Riedwyl, H., and Schuepbach, M. (1994), "Parquet Diagram to Plot Contingency Tables," in *Softstat '93: Advances in Statistical Software 4*, ed. F. Faulbaum, New York: Gustav Fisher.
- Theus, M. (1996), *Theorie und Anwendung Interaktiver Statistischer Graphik*, Augsburg: Wißner.
- (1997), "Visualizing Categorical Data," in *Softstat '97: Advances in Statistical Software 6*, eds. W. Bandilla and F. Faulbaum, Stuttgart: Lucius and Lucius.
- (in press), "MONDRIAN—Interactive Statistical Graphics in JAVA," *Statistical Computing and Graphics Newsletter*, 10.
- Theus, M., and Wilhelm, A. (1996), "Modelling Categorical Data by Interactive Mosaic Plots and Tables," in *Statistical Modelling: Proceedings of the 11th International Workshop on Statistical Modelling*, eds. A. Forcina, G. M. Marchetti, R. Matzinger, and G. Galmacci, pp. 462–465.
- Theus, M., Hofmann, H., Siegl, B., and Unwin, A. (1997), "MANET—Interactive Graphics for Missing Values," in *New Techniques and Technologies for Statistics II*, Amsterdam: IOS Press.
- Venables, W. N., and Ripley, B. D. (1994), *Modern Applied Statistics in S-Plus*, New York: Springer.