

## VISUALIZING CATEGORICAL DATA IN VISTA

**Forrest W. Young**

*The L.L. Thurstone Psychometric Laboratory.  
University of North Carolina at Chapel Hill. CB  
3270 DA, Chapel Hill NC., USA 27599-3270.  
e-mail: [forrest@unc.edu](mailto:forrest@unc.edu).  
web page: <http://forrest.psych.unc.edu/>*

**Pedro M. Valero Mora**

*Departamento de Metodología de las Ciencias del  
Comportamiento. Facultad de Psicología. Av.  
Blasco Ibáñez, 21. Valencia. CP: 46010. Spain.  
e-mail: [valerop@uv.es](mailto:valerop@uv.es).  
web page: <http://www.uv.es/~valerop/>*

**Rubén Daniel Ledesma Mouriño**

*Area de Investigación. Facultad de Psicología.  
Universidad Nacional de Mar del Plata.  
C/Funes, 3350. CP: 7600. Mar del Plata, Bs.  
As., Argentina.  
e-mail: [rledesma@mdp.edu.ar](mailto:rledesma@mdp.edu.ar)*

**Abstract.** This paper presents the modules in the statistical package ViSta<sup>14</sup> more related with categorical data analysis. These modules are: visualization of frequency data with mosaic and bar plots, correspondence analysis and multiple correspondence analysis. All these techniques are implemented in ViSta with a big emphasis on plots and graphical representations of data, as well as interactivity of the user with the system. This shapes a system that has shown to be easy, fun and useful both for novice and experienced users.

**Keywords:** Categorical data, visualization, statistical packages, programming languages.

## 1 INTRODUCTION

Young<sup>15,14</sup> has developed ViSta, a statistical package based on Lisp-Stat. ViSta incorporates the object-oriented approach as part of its internal and external functioning. In particular, it extends Lisp-Stat with additional graphical, statistical and data objects; it provides objects for mapping the process of data analyses and it has objects that guide novices through their early attempts to carry out analyses. All these characteristics shape a system that has been shown to be appropriate for students and teachers of statistics as well as for researchers and developers in computational and graphical statistics.

ViSta is focused on techniques for visualizing data. So, traditional statistical methods are considered from a graphical, dynamic, linked and interactive way by sets of plots denominated spreadplots. Spreadplots are one of the most innovative characteristics in ViSta<sup>16</sup>. At the time this is being written, ViSta integrates about 20 different kind of spreadplots. These include spreadplots for exploring raw data, there being a spreadplot specifically constructed for each of several kinds of data, including: Univariate, Bivariate, Multivariate (numeric and Guided Tour), Category, Classification, Frequency Classification (one-way and n-way), Frequency Table, Crosstabulation and Data Simulation. ViSta also has spreadplots for data transformations such as the Box-Cox (Figure 1), Folded Power<sup>13</sup> and Missing Data Imputation<sup>11</sup>. Finally, there are spreadplots for visualizing statistical models, including Analysis of Variance, Correspondence Analysis<sup>2</sup>, Multiple Correspondence Analysis<sup>8</sup>, Multidimensional Scaling, Multivariate Regression, Principal Components<sup>12</sup>, Regression Analysis<sup>1</sup>, Univariate Analysis, Cluster Analysis and Frequency Analysis.

This paper will focus on spreadplots for categorical data. As remarked by Friendly<sup>4,5</sup>, there is an almost paradoxical disparity between availability and use of methods for visualization of quantitative data and categorical data. So, while it is habitual that an analyst carrying out analysis of regression uses plots for exploration and model fitting on a routine basis, it is certainly unusual to see the same practice with categorical data. Also, software that implements these methods is still scarce, and, that includes features as found in ViSta (interactivity, dynamism, linking, etc.) almost not existent.

The plan of this paper will be the following: First, we will describe the different types or representations of categorical data that are built into ViSta. This information is important, because ViSta decides which analysis are acceptable as a function of the type of data selected so the analyst will have to know how to carry out the proper transformation to the right type when necessary. Second, we will describe the Visualization for Raw Frequency data available in ViSta. This visualization provides a preliminary examination of the data that can help to decide the type of analysis to be carried out. Third, we will describe Frequency Analysis. Finally, we will present Correspondence Analysis and Homogeneity Analysis (also known as Multiple Correspondence Analysis). A discussion with some of the planned additions to the modules in ViSta will close this paper.

## 2 DATA REPRESENTATIONS OF CATEGORICAL DATA IN VISTA.

Categorical data can be represented in different ways in ViSta. Many of them are equivalent, and ViSta provides transformations that change data from a representation to other. Statistical analysis methods will sometimes expect data with a particular shape so it is important to know the different representations and the possible actions that can be pursued with each one.

Data objects in ViSta are represented with icons. The table 1 shows all the different icons used in ViSta to represent categorical data with the type of analysis that can be carried out with them.




| ANALYSIS  | ICON  | DATA SHAPE   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
|---|---|--|-----------|-----------|-----------|-----------|----------|----------|----------|----------|------|---------|------|------|---------|------|------|---------|------|------|---------|------|------|---------|------|---------|---------|------|--------|---------|------|------|---------|---------|------|---------|------|------|---------|------|------|
| <p><b>Table data</b></p> <p>Frequency analysis,</p> <p>Correspondence Analysis</p>  |    | <table><tr><th>2 Vars</th><th>drug</th><th>placebo</th></tr><tr><th>3 Obs</th><th>Numeric</th><th>Numeric</th></tr><tr><td>none</td><td>13.</td><td>29.</td></tr><tr><td>some</td><td>7.</td><td>7.</td></tr><tr><td>marked</td><td>21.</td><td>7.</td></tr></table>   | 2 Vars    | drug      | placebo   | 3 Obs     | Numeric  | Numeric  | none     | 13.      | 29.  | some    | 7.   | 7.   | marked  | 21.  | 7.   |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| 2 Vars  | drug  | placebo  |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| 3 Obs   | Numeric   | Numeric  |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| none  | 13.   | 29.  |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| some  | 7.  | 7.   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| marked  | 21.   | 7.   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| <p><b>Frequency classification data</b></p> <p>Frequency analysis</p>   |  | <table><tr><th>3 Vars</th><th>Frequency</th><th>Improve</th><th>Treatment</th></tr><tr><th>6 Obs</th><th>Numeric</th><th>Category</th><th>Category</th></tr><tr><td>Obs1</td><td>13.</td><td>none</td><td>drug</td></tr><tr><td>Obs2</td><td>29.</td><td>none</td><td>placebo</td></tr><tr><td>Obs3</td><td>7.</td><td>some</td><td>drug</td></tr><tr><td>Obs4</td><td>7.</td><td>some</td><td>placebo</td></tr><tr><td>Obs5</td><td>21.</td><td>marked</td><td>drug</td></tr><tr><td>Obs6</td><td>7.</td><td>marked</td><td>placebo</td></tr></table>   | 3 Vars    | Frequency | Improve   | Treatment | 6 Obs    | Numeric  | Category | Category | Obs1 | 13.     | none | drug | Obs2    | 29.  | none | placebo | Obs3 | 7.   | some    | drug | Obs4 | 7.      | some | placebo | Obs5    | 21.  | marked | drug    | Obs6 | 7.   | marked  | placebo |      |         |      |      |         |      |      |
| 3 Vars  | Frequency   | Improve  | Treatment |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| 6 Obs   | Numeric   | Category   | Category  |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| Obs1  | 13.   | none   | drug      |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| Obs2  | 29.   | none   | placebo   |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| Obs3  | 7.  | some   | drug      |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| Obs4  | 7.  | some   | placebo   |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| Obs5  | 21.   | marked   | drug      |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| Obs6  | 7.  | marked   | placebo   |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| <p><b>Category data</b></p> <p>Frequency analysis.</p> <p>Homogeneity Analysis<br/>(Multiple Correspondence Analysis)</p> |  | <table><tr><th>2 Vars</th><th>Improve</th><th>Treatment</th></tr><tr><th>84 Obs</th><th>Category</th><th>Category</th></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr><tr><td>non dru</td><td>none</td><td>drug</td></tr></table> | 2 Vars    | Improve   | Treatment | 84 Obs    | Category | Category | non dru  | none     | drug | non dru | none | drug | non dru | none | drug | non dru | none | drug | non dru | none | drug | non dru | none | drug    | non dru | none | drug   | non dru | none | drug | non dru | none    | drug | non dru | none | drug | non dru | none | drug |
| 2 Vars  | Improve   | Treatment  |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| 84 Obs  | Category  | Category   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |
| non dru   | none  | drug   |           |           |           |           |          |          |          |          |      |         |      |      |         |      |      |         |      |      |         |      |      |         |      |         |         |      |        |         |      |      |         |         |      |         |      |      |         |      |      |

Table 1: Data representations for category data

Category data is the most general categorical data type in ViSta. It only differs of the most general data in ViSta, multivariate data, in that only includes categorical variables while multivariate data can include a mixture of numerical and categorical variables.

Notice that ViSta Correspondence Analysis can only carried out with table data and that Homogeneity Analysis will only work with categorical data. Frequency analysis is

allowed with all kinds of data, though. Fortunately, ViSta can transform data from a representation to other. This is carried out through the menu item **Create-Convert Data** in the menu **Data**. This results in the dialog box in figure 1. Selecting an option will output the data in the corresponding shape.

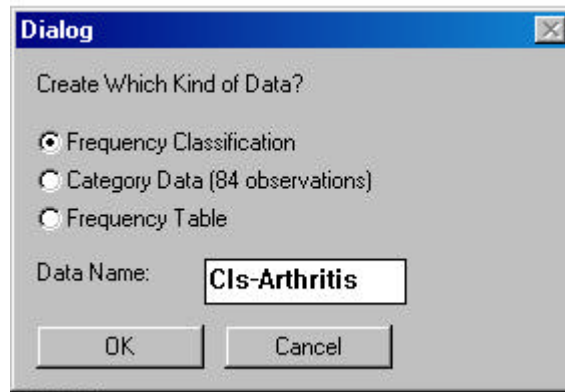


Figure 1: Transformations between types of data in ViSta

### 3 VISUALIZATION OF CATEGORICAL DATA

Visualization of raw frequency data is a topic that has recently received considerable attention in recent times. We illustrate ViSta's frequency data spreadplot with data from the sinking of the Titanic. In these data, the fate of each person on the ship (Lived, Died) is cross-classified with the person's Age (Child, Adult), Sex (Male, Female) and Class (First, Second, Third, Crew). Figure 2 shows the initial spreadplot for these data, a "two-way" spreadplot in which each cell presents information about the cross-classification of the first two ways of the data (which are survival and age). These data were obtained from Dawson<sup>3</sup>.

This spreadplot includes the following plots for the combinations of up to four classification variables (more variables may be used, but four is the maximum which can be combined).

A mosaic plot of frequencies. This plot visualizes an n-way contingency table by portraying the frequencies as "tiles" whose size (area) is proportional to the table's frequencies. The colors encode the Pearson residuals of the cell with respect to the model of mutual independence (the probabilities in a cell are products of one way marginal probabilities). Blue means a positive residual. Red indicates a negative residual. The names of the categories, the value of the observed frequency and the value of the residual of the associated cell in the table data are shown when the pointer of the mouse is moved across the mosaic plot.

A stacked bar graph of the frequencies. This plot shows bars that have a length proportional to the values of the cells. A variable can be crossed with these bars so they are split into stacks of rectangles that are proportional to the absolute value of the corresponding cell. Colors of the bars are the same as the colors of the Mosaic plot's tiles. Mouse movement has a similar effect on the bars as on the tiles.

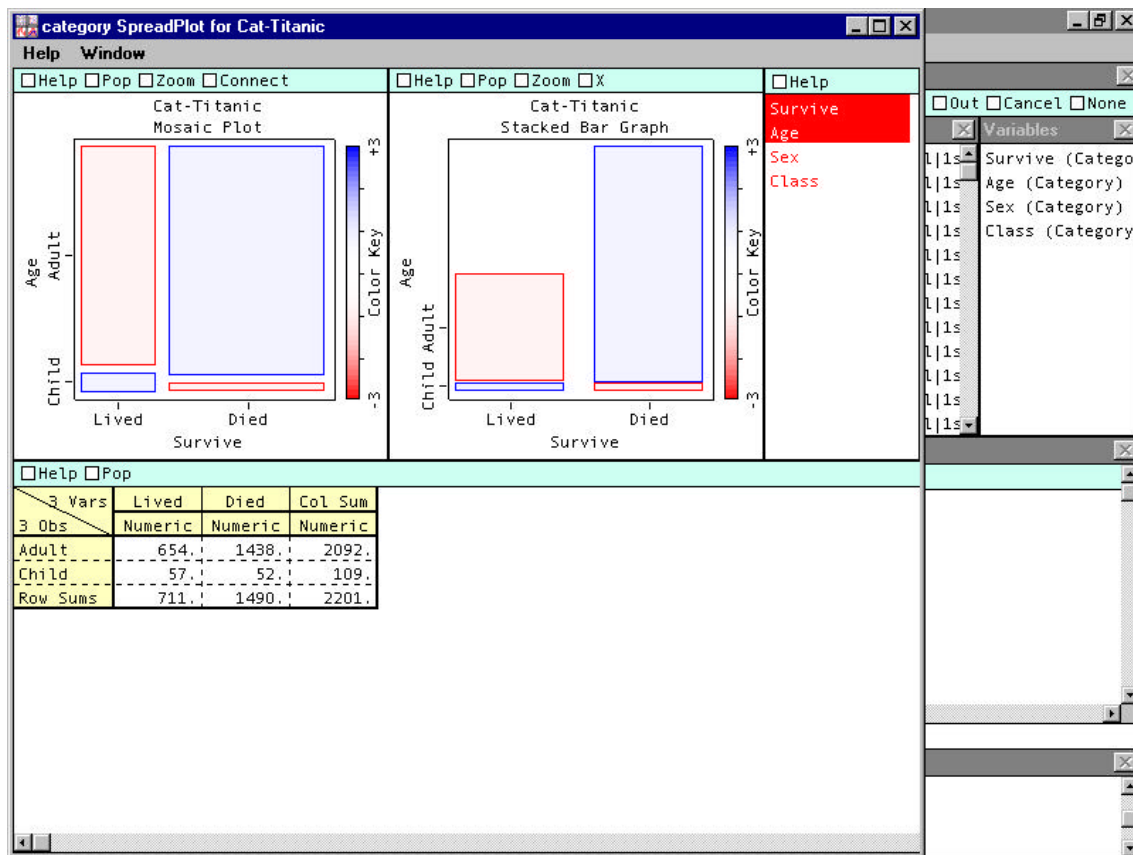


Figure 2: Spreadplot for visualization of frequency data in ViSta

A table of frequencies for the cross-classification of the categories of the variables included in the visualization.

A list of ways of the table (classification variable names) which permits us to determine which variables are visualized in the other cells of the spreadplot.

The spreadplot for frequency data is designed to explore the effect of interactively adding new variables to it. This is controlled by the “Change Plots” window in the upper right corner, a window which displays the names of the ways of the table. Clicking on a name produces a “one-way” spreadplot for the variable selected. A ctrl-click on another name adds that name to those already selected and changes the spreadplot to a “two-way” spreadplot. Three-way and four-way spreadplots can be created by additional ctrl-clicks. Similar effects can be had by dragging the mouse across adjacent names.

This spreadplot provides a first evaluation of the data that an analyst can use to decide whether a more sophisticated exploration or analysis is necessary. In this case, an examination of the different combinations of variables lead us to the conclusion that a more sophisticated approach would be necessary.

## 4 FREQUENCY ANALYSIS

Frequency analysis in ViSta at this moment has the same visualization as previously discussed. However, work for including loglinear analysis into it is under way.

ViSta offers the possibility of obtaining crosstabulations of two variables like shown in table 1. Output for the association of the variables is also shown. This crosstabulations can be splitted by control variables. In case that the matrix is 2x2, it prints a test for the homogeneity of association through the strata. An example of the printed report is in figure 3.

```
Expected Values:
      AGE
SURVIVE Adult  Child  Row Sums
Lived    675.79  35.21   711.00
Died     1416.21  73.79  1490.00
Column Sums 2092.00  109.00  2201.00

Row Percentages:
      AGE
SURVIVE Adult  Child  Row Sums
Lived    91.98   8.02  100.00
Died     96.51   3.49  100.00
Column Sums 188.49  11.51  200.00

Column Percentages:
      AGE
SURVIVE Adult  Child  Row Sums
Lived    31.26  52.29   83.56
Died     68.74  47.71  116.44
Column Sums 100.00  100.00  200.00

Statistics:
Chi Square Coefficient = 20.9555  DF=1  P= 0.0000
Cochran-Mantel-Haenszel = 20.9460  DF=1  P= 0.0000
Phi Coefficient = 0.0976
Contingency Coefficient = 0.0971
```

Figure 3: An example of the text report for frequency analysis in ViSta

## 5 CORRESPONDENCE ANALYSIS

Correspondence Analysis is a exploratory technique that displays the rows and columns of a contingency table as points in a graph. This technique is regarded as a dimensionality reduction method because tries to fit the data in usually two or three dimensions accounting with the maximum of variance<sup>7</sup>.

This technique shows the deviation of the expected values for two variables with regard to the model of independence. Therefore, a table with a non-significant Chi Square Coefficient will result in an uninteresting output. On the other hand, correspondence analysis is usually more interesting when the variables have many categories and understanding of their interrelations is the main goal of the analysis.

However, in order not to produce an overwhelming example we will use a moderately complex contingency table. This example corresponds to the hair and eye color data used by Friendly in his papers on Mosaic Displays. Data reported by Snee<sup>9</sup>. They are shown in table 2.

| 4 Vars | Black   | Brown   | Red     | Blond   |
|--------|---------|---------|---------|---------|
| 4 Obs  | Numeric | Numeric | Numeric | Numeric |
| Brown  | 68.     | 119.    | 26.     | 7.      |
| Blue   | 20.     | 84.     | 17.     | 94.     |
| Hazel  | 15.     | 54.     | 14.     | 10.     |
| Green  | 5.      | 29.     | 14.     | 16.     |

Table 2: Crosstabulation of hair and eye color for a group of people.

The visualization for correspondence analysis for this table is shown in figure 4. This visualization includes five plots and a list. We will describe them from left to right and top to down.

- 1) List for categories of variables analyzed. It is placed on the left of the spreadplot. This list is linked to the spin plot, the scatterplot, the scatterplot-matrix and the boxplot of categories of variables. Colors are used to distinguish between the row variable (eye color in this case) and the column variable (hair color).
- 2) Scatterplot-matrix. This plot is linked with the spin-plot, the scatterplot and the boxplot. Using the hand tool, clicking on a plot cell selects the dimensions shown in other plots of the spreadplot.
- 3) Spin-plot of row and column-points. This plot portrays three dimensions of the correspondence analysis solution.
- 4) Scatterplot of row and column points. This plot shows only the first two dimensions of the correspondence analysis solution.
- 5) Boxplot of row and column points. This plot shows the coordinates of the correspondence analysis solution. It can show as many dimensions as computed in the analysis.
- 6) Mosaic plot of the data. The colors of the cells symbolize the standardized residuals from the model of independence as in the basic visualization of the data described previously. The rows and columns are sorted according to the scores of the objects in the first dimension of the correspondence analysis. This makes that cells with positive residuals fall approximately on the diagonal running from left down to right top. Negative residuals will fall on the diagonal running from right top to left down.
- 7) Scree plot. Represents the inertia of the dimensions of the result. Brushing the spreadplot or selecting a point will show the value of the inertia and the percentage of total inertia of each dimension.

Interpretation of this example is quite straightforward. The first dimension accounts for almost a 90% of the total inertia so we may focus on it without much loss. Traditionally, interpretation of correspondence analysis has been done using scatterplots, but we will use the mosaic plot. Both the rows and columns are sorted from light to dark colors, and cells corresponding reveal positive residuals. For example, hair blond and blue eyes or black hair and brown eyes.

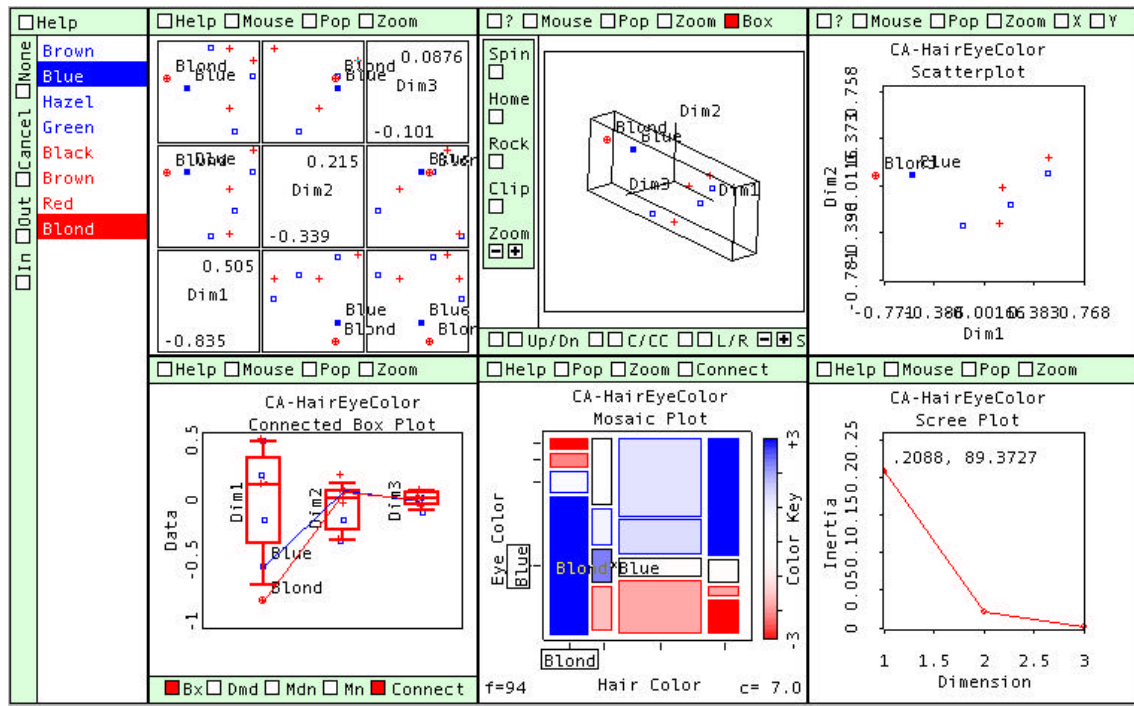


Figure 4: Visualization for correspondence analysis.

## HOMOGENEITY ANALYSIS-MULTIPLE CORRESPONDENCE ANALYSIS

ViSta performs Multiple Correspondence Analysis (MCA) using the Homogeneity Analysis by Alternating Least Squares (HOMALS) method<sup>6</sup>. MCA provides a graphic portrayal of the bivariate relationships between categorical variables, and can be useful to understand large, multivariate categorical datasets. MCA is a technique that has received many different names and that has been derived in different ways in different disciplines and contexts<sup>10</sup>. We will analyze again the data about the titanic sinking. As indicated in table 1, this method needs a categorical data matrix for computing the analysis. The analysis starts with the following dialog box for options for the analysis. Using the default options, the computer will print the history for iterations shown in the figure 6.



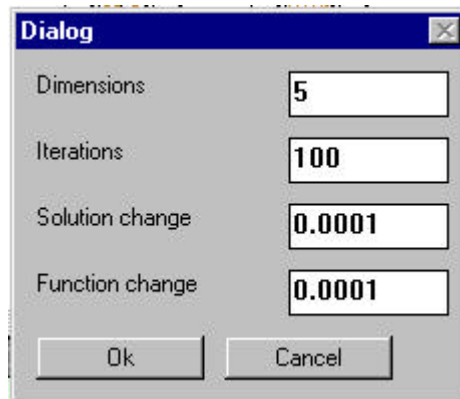


Figure 5: Dialog box for Homogeneity Analysis

```

Homals iterations.
Homals computation is an iterative process.
The iteration history appears below.

Iteration    0, Loss 10998.23847, Change 0.01662
Iteration    1, Loss 7994.19336, Change 0.98672
Iteration    2, Loss 7967.09261, Change 0.99013
Iteration    3, Loss 7961.58382, Change 0.99326
Iteration    4, Loss 7960.17633, Change 0.99577
Iteration    5, Loss 7959.74803, Change 0.99742
Iteration    6, Loss 7959.60264, Change 0.99840
Iteration    7, Loss 7959.54982, Change 0.99899
Iteration    8, Loss 7959.52977, Change 0.99934
Iteration    9, Loss 7959.52193, Change 0.99956
Iteration   10, Loss 7959.51880, Change 0.99970
Iteration   11, Loss 7959.51753, Change 0.99979
Iteration   12, Loss 7959.51701, Change 0.99985
Iteration   13, Loss 7959.51679, Change 0.99989
Converged....

```

Figure 6: History of iterations for Homogeneity Analysis for the Titanic data.

The visualization for the result is shown in figure 7. There are four elements in this visualization.

- 1) List of the objects or rows in the dataset and the categories of the variables or columns. This list is linked with both the spinplot and the scatterplot matrix. Colors are used to distinguish between variables, so, for example, the categories of the variable survive are coded in red, adult and child in green, sex in grey and class in purple (not seen in black and white versions of this paper).
- 2) Spinplot for the categories of the variables and the objects in the dataset. Notice that when there are several cases with the same combination of categories they will be represented with exactly the same score in the plot, so each visible point will correspond to many points. Notice that the picture has been zoomed out so the category child is out of it, on the top extreme of the figure. First dimensions runs left to right and second dimension top to down.

- 3) A scatterplot matrix of the first 5 dimensions of the result. This is similar to the spinplot but by each pair of variables.
- 4) A plot of the discrimination measures for each variable and, simultaneously, the inertia by each dimension. The discrimination measures indicate how well represented is a variable by each dimension. They are the sum of the squares of the distances of the categories of every variable respect to the origin. The inertia of each dimension is the average of the discrimination measures of all the variables for that dimension and is represented by the green line in the plot. The labels of the variables and the profiles along the dimensions can be examined by using the brush tool in the scatterplot.

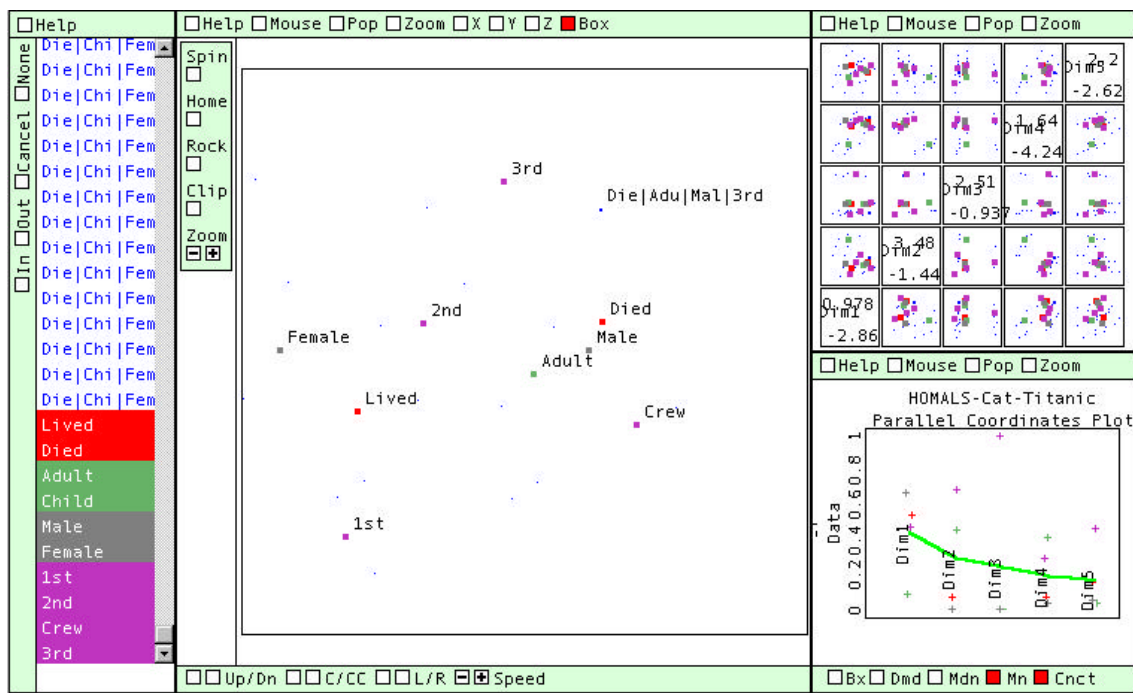


Figure 7: Visualization for homogeneity analysis

It can be seen in the plot of figure 7 that Survival in the Titanic was associated with being female, travelling in first or second class and being a child rather than an adult.

## DISCUSSION

ViSta can be regarded as an exceptional tool for data analysis because of the central role given to graphics in it. This has permitted investigating new ways of using and understanding data analysis techniques. In this sense, we are convinced that the use of this program for research and teaching is, by all means, highly recommendable. However, ViSta is an ongoing work so we expect to add more features and improve its capabilities in the close future. We will provide an overview of the possible extensions to be developed in the close future.

### 1) Data manipulation

ViSta at this moment does not have a full language for crosstabulations of data like other statistical packages have. This can be a limitation for computing square tables from datasets with three or more variables. For example, given three categorical variables A, B, C, the crosstabulation A-B with C needs to write a small program in any of ViSta's programming languages. A concatenation function will solve this problem very easily.

## **2) Loglinear modeling**

Loglinear modeling is an approach for analysis of categorical data that provides a degree of accuracy for hypothesis testing that the methods currently implemented do not have. An implementation of these methods in ViSta is currently under development and it should be finished by the time that this work is presented.

## **3) Homogenization of principal component analysis correspondence analysis and multiple correspondence analysis.**

Methods for reduction of dimensionality are very related (see for example<sup>10</sup> for a reinterpretation of multiple correspondence analysis in terms of principal component analysis) but they have been often described separately without mention of the interrelationships they have. We have been working towards avoiding this disparity in ViSta by making the visualization for these three methods as similar as possible. The result should be a set of similar plots for the different methods that can be interpreted in similar ways. This would make teaching of the methods easier than now.

We would like to finish this work by mentioning one of the most important features we think that ViSta has: Fun. Our experience with users is that they often have fun doing their analysis when using ViSta. It is not uncommon to hear comments such as "Hey, this is fun!" or "It's like a video game", or "Lets play with our data"! The highly interactive, highly dynamic graphics that are supported by the ViSta architecture provide the foundation for building an environment in which data analysis is fun. And we believe that a data analyst who is having fun is one who is more likely to have insight.

## **REFERENCES**

- [1] C. M., Bann, *ViSta Regress: Univariate Regression with ViSta, the visual Statistics System*. L.L. Thurstone Psychometric Laboratory Research Memorandum (1996).
- [2] Lee Bee-Leng *Correspondence Analysis*. L.L. Thurstone Psychometric Laboratory Research Memorandum (1996).
- [3] R. J. M. Dawson, The "unusual episode" data revisited. *Journal of Statistics Education*, **3** (3) (1995).

- [4] M. Friendly, Mosaic displays for multi-way contingency tables. *Journal of the Marican Statistical Association*, 89: 190-200. (1994).
- [5] M. Friendly, Conceptual and Visual models for categorical data. *The American Statistician*, 49:153-160. (1995)
- [6] A. Gifi, Nonlinear multivariate Analysis. Chichester: Wiley. (1990).
- [7] M. J. Greenacre, *Theory and applications of correspondence analysis*. London, Academic Press. (1994).
- [8] R., Ledesma, P., Valero, F. W. Young, and J. de Leeuw, Multiple Correspondence Analysis. Comunicación presentada al I Congreso de Investigación mediante Encuestas. (2000).
- [9] R. Snee. Graphical Display of Two-Way Contingency Tables. *American Statistician*, 28, 9-12. (1974)
- [10] M., Tenenhaus & F. W. Young, An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 1, 91-119. (1985).
- [11] Valero, P. M. & Young, F. W. *Missing Data Analysis*. L.L. Thurstone Psychometric Laboratory Research Memorandum. (2000),
- [12] F. W. Young,. & P. M. Valero, (1999), *Principal Component Analysis*. L.L. Thurstone Psychometric Laboratory Research Memorandum. (2000)
- [13] F. W. Young, & P. M. Valero, *Transformations*. L.L. Thurstone Psychometric Laboratory Research Memorandum. (2000)
- [14] F. W. Young, and C. Bann, ViSta: A Visual Statistics System. In Stine, R. and Fox, J. (Eds.), *Statistical Computing Environments for Social Research*, 207-235. Thousand Oaks: Sage. (1997)
- [15] F., Young, R. A. Faldowski, & M. M. McFarlane, Multivariate Statistical Visualization. En C. R. Rao, (Ed). *Handbook of Statistics*, Vol 9. (959-998). Amsterdam: Elsevier Science. (1993).
- [16] F., Young, P., Valero, R. A., Faldowsky, & C. Bann, *SpreadPlots*. The Visual Statistic project, L. L. Thurstone Psychometric Lab, Univ. N. Carolina, Chapel Hill. Report Number 2000-4, May 2000. (2000).