Interactive Data Visualization using Mondrian

Martin Theus
University of Augsburg
Department of Computeroriented Statistics and Data Analysis
Universitätsstr. 14, 86135 Augsburg, Germany
martin.theus@math.uni-augsburg.de

Abstract

This paper presents the Mondrian data visualization software. In addition to standard plots like histograms, barcharts, scatterplots or maps, Mondrian offers advanced plots for high dimensional categorical (mosaic plots) and continuous data (parallel coordinates). All plots are linked and offer various interaction techniques. A special focus is on the seamless integration of categorical data. Unique is Mondrian's special selection technique, which allows advanced selections in complex data sets.

Besides loading data from local (ASCII) files it can connect to databases, avoiding a local copy of the data on the client machine.
Mondrian is written in 100% pure JAVA.

1. Introduction

Data visualization has been acknowledged as an important tool in decision support. But usually visualizations are static and just used for presentation rather than exploration. Interactive statistical data visualization is a powerful tool which reaches beyond the limits of static graphs.

Although there was a big research effort in the mid 80s in interactive graphical statistics, this data analysis tool has not become widely used. One reason might be that the systems designed by researchers 15 years ago (cf. [1]) needed extremely expensive hardware and a big effort in software development. Certainly times have changed since then, and any desktop computer is capable of graphics today. Furthermore compatibility issues have become less important. Open source projects as well as platform independent programming languages like JAVA have made software more widely accessible.

Only few software packages are not only tailored towards one specific visualization task like e.g. network visualization or 3d-imaging, but offer a variety of plots for a general analysis of data.

This paper lists Mondrian's special features and novel implementations. Concepts of how to utilize the interactive visualization tools for an advanced data analysis are presented as well.

Mondrian is freely available and as a JAVA application runs on almost any platform.

2. Smart Selections

The main task in interactive data visualization is the identification of patterns and subgroups. Thus selecting and identifying data is of major importance. This section introduces the special selection technique implemented in Mondrian.

2.1. The Progress in Selection Techniques

The way how data are selected in interactive visualization software shows the steady advance of research results.

1. The *standard* way of selecting data is to select data and by doing so replace any other selection that might have been present. There is no way of refining a selection or selecting over different plots and/or variables. This standard selection technique is implemented e.g. in GGobi [7].

2. A more *advanced* way to handle selections is to allow to combine the current selection with a new selection with boolean functions like *and, or, Xor, not*. This allows the analyst to refine a selection step by step to drill down to a very specific subset of the data. DataDesk [11] implements this selection technique.

3. When dealing with a whole *sequence* of selections, it is often desirable to change a selection at an earlier stage, without having to redefine all preceding and successive selection steps. By storing the sequence of selections it is possible to make changes to any step in the

sequence. Selection Sequences have been first implemented in MANET [9].

4. Although a selection is always performed on the computer screen in the first place, i.e. in terms of screen coordinates, the data selection must be stored in terms of data coordinates. The approach used by Mondrian keeps a list of any selection associated with a data set. For each entry in the list the

   - selection area in screen coordinates and data coordinates,
   - selection step,
   - corresponding plot window and
   - selection mode (e.g. and, or, not)

   is stored. The currently selected subset of the data can then be determined by processing all elements of the list, no matter which kind of modification to the list was the reason for an update of the selection subset.

2.2. Selection Rectangles

Allowing multiple selections in a single window as well as across different windows makes a visual guide to the selections performed indispensable.

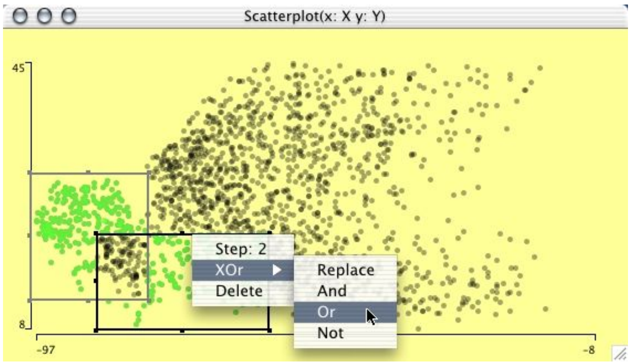


Figure 1. *Selection Rectangles* **in Mondrian.**

Mondrian introduces *Selection Rectangles*. Figure 1 gives an example of a scatterplot containing two selection rectangles. Selection rectangles indicate the area which was selected. An existing selection rectangle can be used as a brush by simply dragging the selection rectangle. The eight handles on the rectangle permit a flexible resizing of the rectangles. This enables various slicing techniques.

The selection mode can be changed via a pop-up menu. The deletion of a selection can be performed via this pop-up, too. An active (i.e. selected) selection can be deleted by simply pressing the backspace key. Only the active selection is plotted in black. All other selections are plotted in a lighter gray to make them less dominant in the plot.

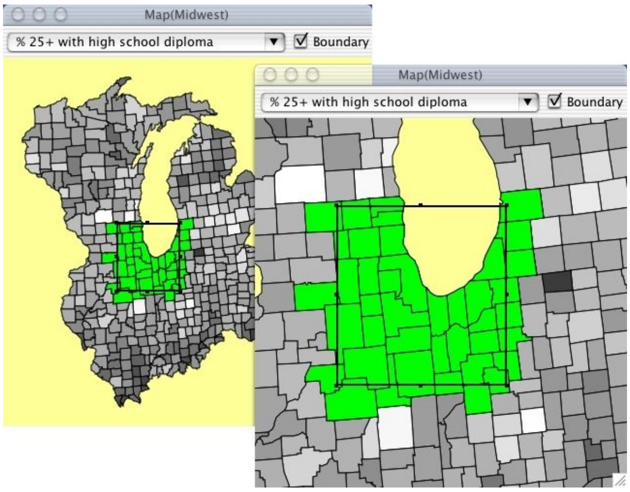

Map(Midwest)
% 25+ with high school diploma
☑ Boundary
Map(Midwest)
% 25+ with high school diploma
☑ Boundary

Figure 2. Zooming in a map: the Selection Rectangle changes accordingly.

Since selections are stored in terms of the data coordinates they are invariant to any alterations of a plot. Typical scenarios are things like interactive reordering of the axes in a parallel coordinate plot, flipping the axes in a scatterplot or zooming. These operations automatically update the selection rectangles. The new screen coordinates of the selection rectangles are calculated from the data coordinates. Figure 2 shows how a selection rectangle reacts on a zoom inside a map.

The ability to handle more than one selection in one window is indispensable when dealing with parallel coordinates.

The way Mondrian handles selections is particularly useful when working with databases, since the selection translate easily into SQL code. At this point it is important to be sure about the precedence of boolean operators. Mondrian always performs selections sequentially, which is in most cases the way the user thinks. Thus an example selection `S1 OR S2 AND S3` reads as `(S1 OR S2) AND S3`, ignoring the usual precedence of boolean operators, where `AND` has a higher precedence as `OR`. The `WHERE`-clause in an SQL-query thus is explicitly bracketed to ensure the sequential order of the operators.

The use of JAVA 2D would make the implementation of arbitrary shapes of a selection area relatively simple. Whereas this would allow very flexible selections, it is not obvious how a resizing of such a selection would look like. A more structured approach to more general selection areas could be to allow a rhombus shape of the selection area, which can be resized at the four corner handles. The remaining four handles would then be used to enlarge or reduce the size of the rhombus as needed.

3. Conventions

One of the keys to the success of a graphical user interface are conventions. Conventions enable the user to perform tasks without learning new, specific interactions. E.g. most graphical user interfaces allow for a change of the window size by dragging the lower right corner of the window. Once the user knows of this behavior he/she is able to resize a window no matter what application or operating system he/she uses. A brilliant collection of good and many bad examples of user interface design is given at `http://www.iarchitects.com`. A broader discussion of user interface design for interactive visualization software can be found in [8].

High interaction graphics with direct manipulation interfaces offer a lot of interactions. To ease the use of high interaction graphics it is necessary to gather the various interactions into different groups like queries, zooming, selection, reordering. Once these groups are identified, we can assign the various user interface interactions to them; e.g. shift-mouse-click, pop-up-trigger etc.

Inside Mondrian the following groups of interactions have been identified to be crucial to perform steps in an interactive graphical data analysis.

- **Selections**
  - *Creating a selection rectangle*
    Click and drag.
  - *Brushing*
    Click inside a selection rectangle and drag.
  - *Resize a selection (Slice)*
    Click-drag a handle of a selection rectangle
  - *Change the selection mode*
    Shift-click inside the selection rectangle
- **Queries**
  - *Popup trigger on an object*
    (i.e. right mouse button on most systems).
- **Alterations**
  - *Zoom-out (-in)*
    Meta-click (and drag).
  - *Change the plot settings*
    Popup trigger on the plot background.
  - *Reorder objects in the plot*
    Alternate-click on the object and drag to new position.

Obviously some of these interactions are identical in all plots (e.g. interactions with selections), and some depend on the plot-type. Whereas there is nothing to reorder in a map, you can reorder the categories in a barchart or the axes in a parallel coordinate plot. Given these conventions it is very easy to get used to all the different functions inside Mondrian.

4. Special Plots for High Dimensional Data

Although linking and highlighting across different plots can already increase the number of dimensions to look at simultaneously, it is very desirable to find visualizations which include many variables at a time. Mosaic plots for categorical data and parallel coordinate plots for continuous data are ideal for gaining insight into high dimensional data.

4.1. Parallel Coordinates/Boxplots

Parallel Coordinates are a powerful tool to analyze a high dimensional data sets graphically. Since static representations of parallel coordinates are usually not very revealing several interactive implementations arose very early. These implementations are restricted to very special computing environments and thus not easily accessible for most people. The probably most advanced implementations are the ones of Inselberg [6] and Wegman [13].

Figure 3 shows a parallel coordinate plot of the Midwest data including not less than 14 variables, of which 13 are continuous and one categorical. In addition to the standard selection, highlighting and interrogation methods parallel coordinates in Mondrian support the following features:

Coordinates can be rearranged manually to look at the most interesting adjacencies. Usually only a few adjacencies are of interest. [1]

Zooming is implemented for each axis individually. Since parallel coordinates are cluttered very much with an increasing number of observations displayed, zooming can focus on a more detailed view of the variable. E.g. for the variable *'% American-Indian-Eskimo-Aleut'* it would be desirable to simply zoom in, in order to get rid of the outliers and see the shape of the distribution, i.e. the box of the boxplot.

Mondrian offers a special feature to plot categorical variables in parallel box/coordinate plots. Whereas most implementations only use the number coding of a categorical variable, Mondrian plots a stacked barchart, with left-to-right highlighting for each categorical variable. This display is consistent with all other plots representing counts. Additionally lines can be displayed for the highlighted points in boxplot mode.

In Figure 4 the same data as shown in figure 3 is displayed. Whereas in Figure 3 one cannot really find out

[1] Only $\frac{k+1}{2}$ permutations of the k variable-axis are needed to display all possible adjacencies, cf. [13]

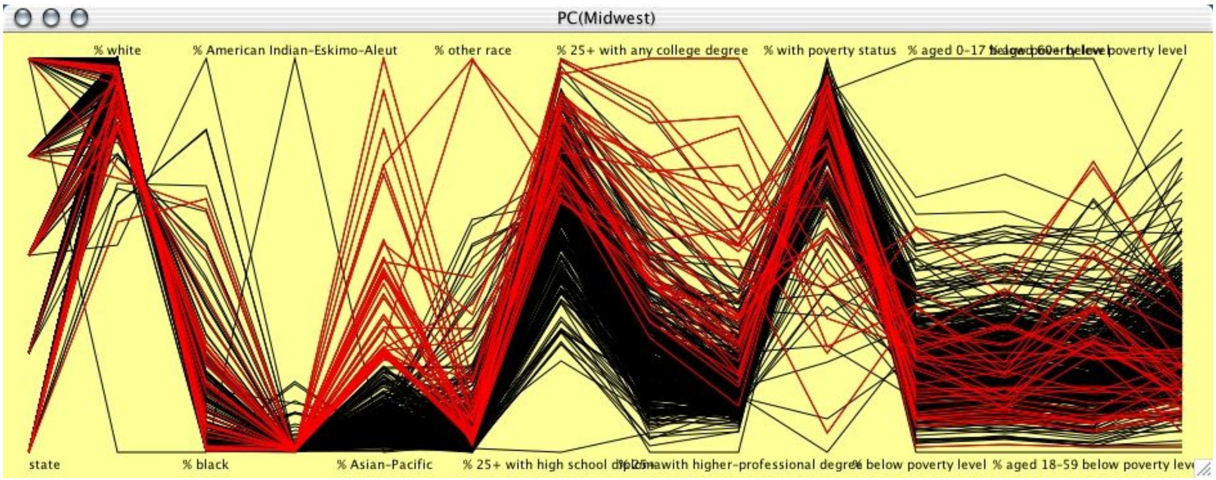
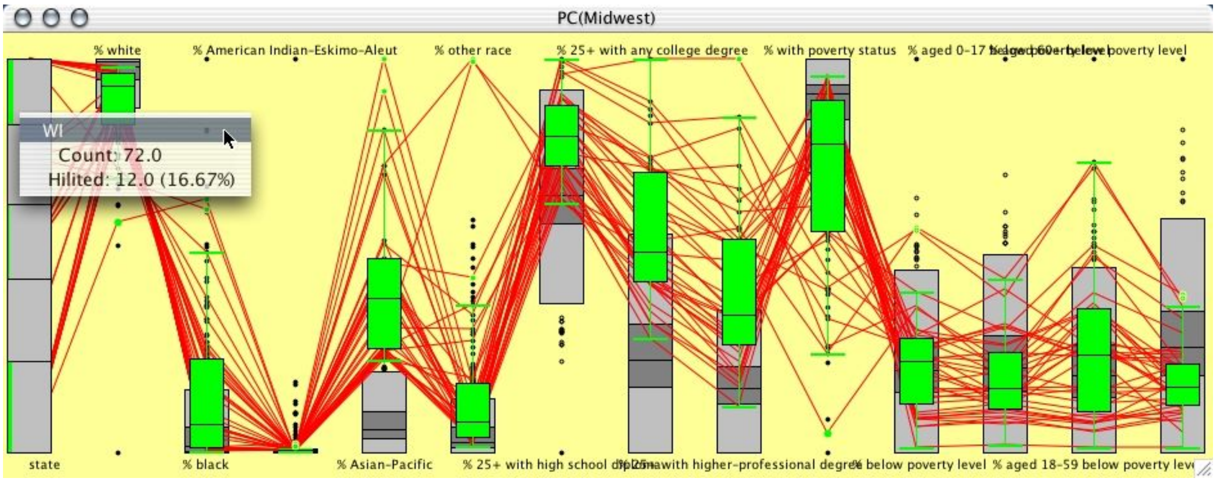

PC(Midwest)
state
% white
% black
% American Indian-Eskimo-Aleut
% Asian-Pacific
% other race
% 25+ with high school diploma
% 25+ with any college degree
% 25+ with higher-professional degree
% with poverty status
% below poverty level
% aged 0-17 below poverty level
% aged 18-59 below poverty level
% aged 60+ below poverty level

Figure 3. Parallel Coordinates for the Midwest data. Counties with high proportion of Asian-Pacifics are selected.


PC(Midwest)
state
% white
% black
% American Indian-Eskimo-Aleut
% Asian-Pacific
% other race
% 25+ with high school diploma
% 25+ with any college degree
% 25+ with higher-professional degree
% with poverty status
% below poverty level
% aged 0-17 below poverty level
% aged 18-59 below poverty level
% aged 60+ below poverty level

WI
Count: 72.0
Hilited: 12.0 (16.67%)

Figure 4. Parallel boxplots for the same data as in figure 3. The categorical variable in the plot is shown as spineplot.

about how many counties are selected in each state, interrogating Figure 4 shows, that the selected counties are mainly in Wisconsin, Illinois and Michigan.

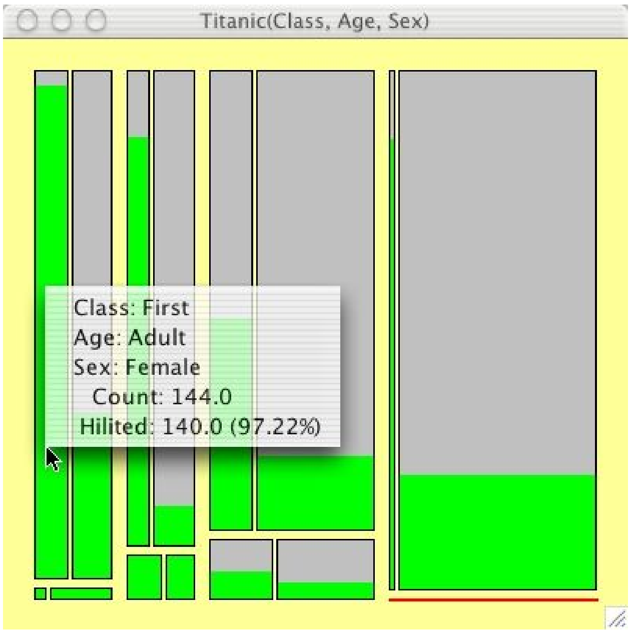
Wills [14] gives an alternate method of incorporating categorical variables into parallel coordinates based on circles-sizes, which is not compatible to the way counts are displayed in barcharts.

4.2. Mosaic Plots

Mosaic plots are a relatively new development. Recent implementations include a static version for S-Plus and R by Emmerson [3] and an interactive version by Hofmann [4] within the MANET software.

Within Mondrian to flexibly reorder the variables in the plot and to include and exclude variables the four arrow keys can be used. Empty cells which occur very often if the number of crossed categories is very high, are not subdivided on lower levels. In situations with many crossed variables this usually reduces the number of cells to draw drastically. To make empty cells visually more prominent, they are plotted in red.

Since Mondrian supports queries there are no labels printed around a mosaic plot. With only a few variables in a Mosaic plot labels would fit around the plot. But more complex plots with, e.g. 8 binary variables would need twice as



Class: First
Age: Adult
Sex: Female
Count: 144.0
Hilited: 140.0 (97.22%)

Figure 5. The Titanic Data in a Mosaic Plot.

much space for the labels as for the plot itself. Figure 5 gives an example of a mosaic plot with a query. Besides the query the name of the data set and the names of the variables in the plot are shown in the title bar of the plot window.

A special feature of the mosaic plots inside Mondrian is the interactive graphical modeling of loglinear models, based on mosaic plots (cf. [10]).

Weighted Plots

Many data sets and most database queries present data in an already summarized form, i.e. a table. In Mondrian Mosaic plots as well as barcharts can handle data which is summarized, specifying attribute variables and a count variable. Obviously any non-negative numeric variable can be used as a weight variable, which allows for very flexible plots, which might be hard to interpret. A simple look up of values can be performed with barcharts by weighting case names by their values, as shown in figure 7.

5. Working with Categorical Data

Mondrian can handle categorical variables in both ways, as non-informative number coding, or full text labels. It implements interactive barcharts and mosaic plots for analyzing categorical data. Neither plot is very revealing in a static setting, but are very insightful in an interactive environment providing linked highlighting and interactive reordering of variables and categories.

The *Housing Factors* Example

The *Housing Factors* example will underline why interactivity is a key-feature for a graphical exploration of categorical data. The data are taken from Cox & Snell [2] resp. Venables [12] investigations (cf. pp 155 resp. pp 226).

Data on the housing situation of 1681 tenants in Copenhagen has been classified according to:

- **Housing Type**
  Apartments, Atrium House, Terraced House, Tower Block
- **Influence on the housing situation**
  low, medium high
- **Contact to other tenants**
  low, high
- **Satisfaction with the housing situation**
  low, medium, high

The data are distributed over all 72 cells, i.e. there are no empty cells. Table 1 lists the complete data set.

Figure 6 shows the default barcharts and mosaic plot for the four variables. The cases with high satisfaction are selected, to mark the most interesting response. Obviously the ordering of at least two of the variables makes no sense, and the mosaic plot does not reveal any systematic pattern, worth fitting a model for. The necessary steps to make the plots more insightful comprise:

Housing Factors			Housing Type			
Sat.	Infl.	Cont.	App.	Atr.	Terr.	Tower
low	low	low	61	13	18	21
		high	78	20	57	14
	med	low	43	8	15	34
		high	48	10	31	17
	high	low	26	6	7	10
		high	15	7	5	3
med	low	low	23	9	6	21
		high	46	23	23	19
	med	low	35	8	13	22
		high	45	22	21	23
	high	low	18	7	5	11
		high	25	10	6	5
high	low	low	17	10	7	28
		high	43	20	13	37
	med	low	40	12	13	36
		high	86	24	13	40
	high	low	54	9	11	36
		high	62	21	13	23

Table 1. Cross-classification of 1681 tenants

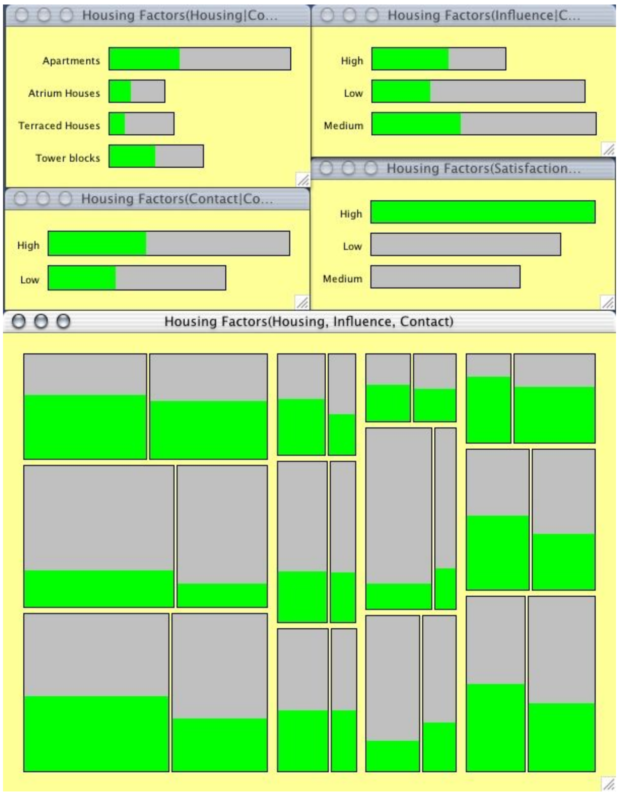
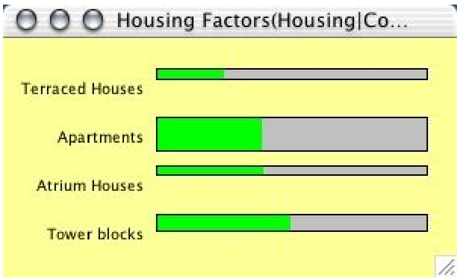
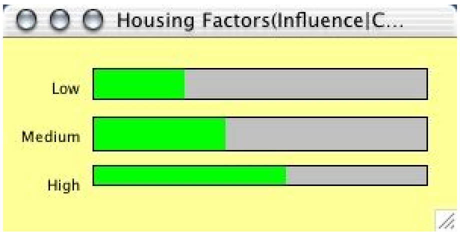

Housing Factors(Housing|Co...
Apartments
Atrium Houses
Terraced Houses
Tower blocks
Housing Factors(Influence|C...
High
Low
Medium
Housing Factors(Satisfaction...
High
Low
Medium
Housing Factors(Contact|Co...
High
Low
Housing Factors(Housing, Influence, Contact)

Figure 6. The Housing Factors Data in default view.

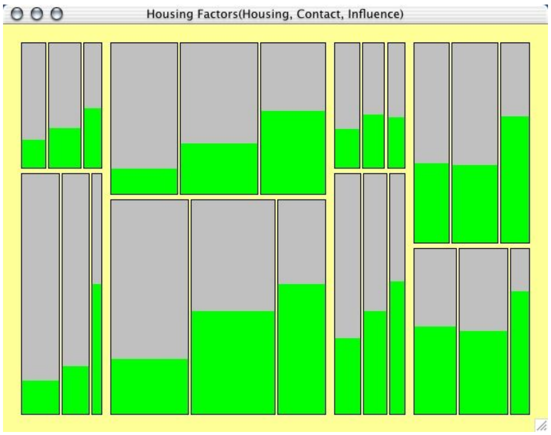
- Sort the categories of Housing Type according to the relative amount of high satisfaction cases (via the plot-option pop-up). The plot has been switched to the Spineplot view, to make the sorting more obvious.



- Sort Influence and Satisfaction to: *low, medium, high* (via alt-click and drag):



- Reorder the variables in the mosaic plot such that the plot is conditioned upon the *Housing Type* and put *Influence* - as a variable with many categories - at the deepest stage. The order is then: Housing Type, Contact, Influence. The reordering is done with the four arrow keys.



Certainly it is still hard to read the plots without the interactive queries. But in contrast to the default views, the reordered plots now reveal a clear pattern along with some deviations, which can now be investigated more closely using statistical models as well as other relevant information.

6. Special Features in Standard Plots

6.1. Barcharts

In Mondrian the layout of the bars in barcharts is chosen to be horizontal rather than vertical. This allows full-length printing of category names. The usual barchart view can be switched to a spine-plot view (cf. Hummel [5]), so that the height, not the width, is proportional to the number of cases in a category. If the highlighting is still done from left to right, the highlighted proportions can then be compared directly.

When working with large data sets (10,000 to 50,000 cases are usually already sufficient) the number of categories will grow as well. No matter how big the screen/window is, we will encounter situations where we can not see all bars/categories at the same time. Making barcharts scrollable allows the investigation of variables with dozens of categories.

Obviously the ordering of the categories then becomes very important. Mondrian offers four ways to order categories in barcharts.

1. **Lexicographic Order**
   This is the default order, which is presented after the plot is constructed and displayed. This ordering is best for looking up categories.

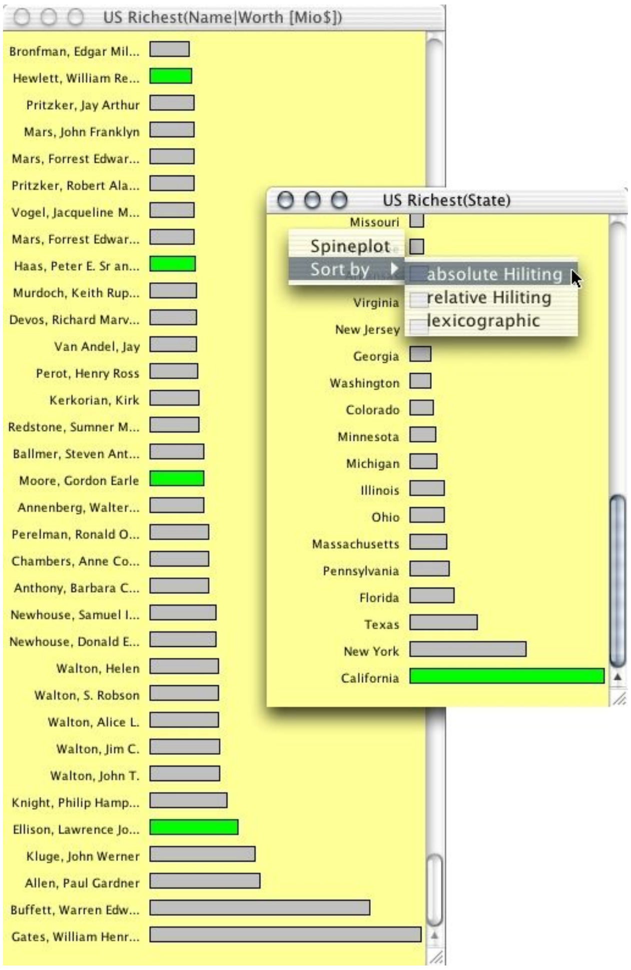


Figure 7. An example of two linked barcharts.

2. **Manual Order**
   Any current order can be changed by manually dragging a bar to its new position. This is useful if all automated sortings fail.

3. **Absolute Size of Highlighting**
   This option sorts the categories according to the absolute number of selected cases in a category. Selecting all data allows for a sorting according to the absolute size of the categories.

4. **Relative Size of Highlighting**
   This sorting option sorts corresponding to the relative amount of highlighting in the categories. In the spineplot view this option nicely shows the ordering of the selected proportions.

A change of the order of the categories of variables is automatically propagated to any other plot which holds information based on this variable and updated instantaneously.

This could be a mosaic plot, a parallel barchart or a choropleth map, which is shaded according to the levels of the categorical variable.

Figure 7 shows data of the fortune 400 private persons in the US taken from Forbes Magazine in 1996. The left barchart shows each individual weighted by its worth. The right barchart, showing the 50 US states, has been sorted according to the number of individuals in this state. California has been selected.

6.2. Histograms

The most crucial point in plotting a histograms is to choose the ”right” origin of the first bin and the ”right” number of bins. Since there exists a vast amount of rules and suggestions what ”right” means under different assumptions, the most important interactive manipulation inside histograms is changing the origin and the width of the bins.

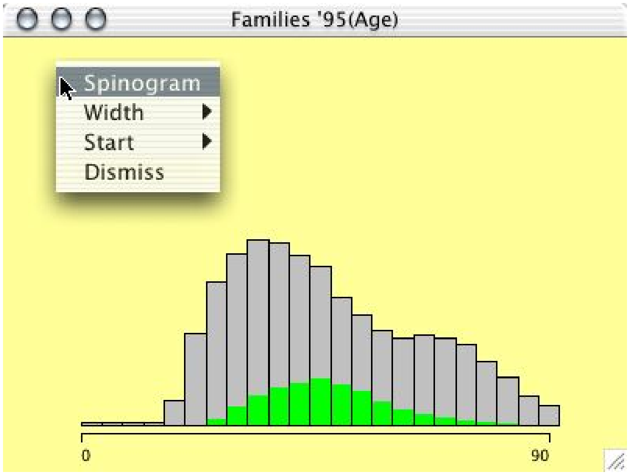


Figure 8. Histogram of the age distribution. Cases with more than 60K income are highlighted

These parameters can be altered by using the four arrow keys (left, right moves the origin; up, down changes the bin width). Additionally a popup-menu offers two sliders to set bin width and origin to whole numbers. This is especially useful when the user wants to set “pretty” ticks, i.e. multiples of 1, 2 or 5 to a power of 10.

In order to keep the visual distortion as small as possible, the scale of the histogram axis is not updated during the interactive reparametrization.

As barcharts can be switched to spineplots, histograms in Mondrian can be switched to the so called spinogram view. In a spinogram all bars of the histogram are scaled to be of same height and are plotted next to each other. Figures 8 and 9 show a corresponding pair of histogram and spinogram.

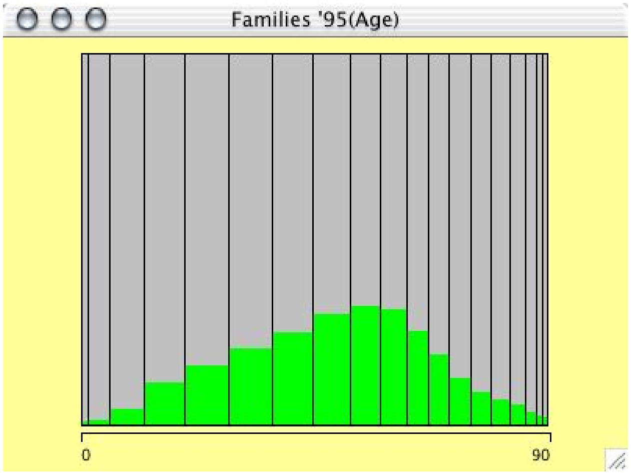
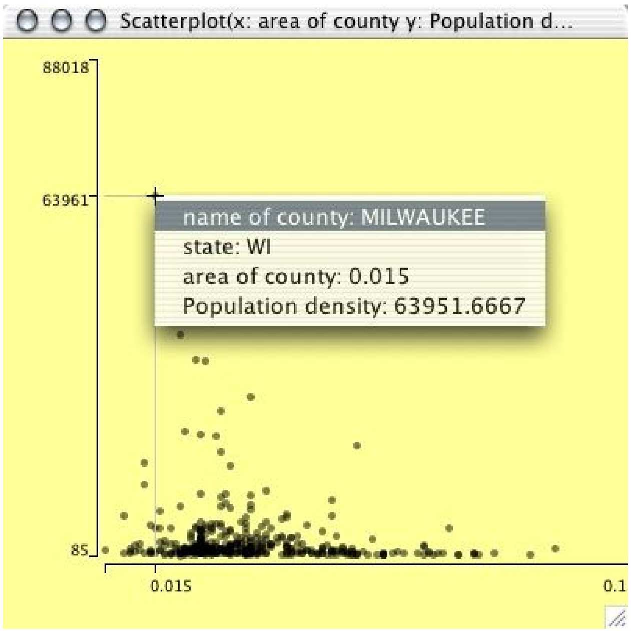


Figure 9. Spinogram of the age distribution.

Scatterplots

In contrast to most other plots in Mondrian, scatterplots offer axes, showing the maximum and minimum as basic orientation. Interrogation methods inside scatterplots operate on two levels. The first level is a simple overview of the position of the cursor, which is displayed by projections onto the x- and y-axes. This interrogation is invoked by simply pressing the control key. A <ctrl-click> invokes the second level of interrogation, cf. Figure 10.



name of county: MILWAUKEE
state: WI
area of county: 0.015
Population density: 63951.6667

Figure 10. Both levels of interrogation in a scatterplot.

A pop-up is presented with the data of the x- and y-variables according to the closest point. By selecting variables in the main variables window, it is possible to specify the variables for which the pop-up will show the values. If more than one point is found at the same distance, a list of the cases is presented in the pop-up.

7. Conclusions

This paper shall encourage the reader to make use of interactive graphical software. Furthermore writing such software in JAVA is easier than ever. JAVA is capable of all graphical displays we can think of. Carefully designed JAVA applications run fast enough on today's hardware and can compete with classical implementations. The platform independence allows for a much wider distribution than we are used to from former development environments.

Although Mondrian was never designed to be a general purpose graphical data analysis package, it already offers most standard plots. Furthermore various features and ideas never implemented before are available.

Current development versions on Mondrian implement direct connections to databases. A general interface to databases via JDBC allows to work on huge data sets, reaching far behind current limits. Certainly display techniques must be adapted. Using $\alpha-$channel transparency allows for plotting vast amounts of data without cluttering the screen.

In order to allow simple extensions to plots like externally defined scatterplot smoothers, an interface to R is under development as well.

Download

Current versions of Mondrian can be downloaded at `http://stats.math.uni-augsburg.de/Mondrian` or `http://www.theusRus.de/Mondrian`. Versions for Windows[2] and Mac OS X — which can be started with no further installations — are provided. For all other platforms a JAR file is distributed.

The latest version covers $\alpha-$blending techniques, implemented in scatterplots and parallel coordinates, which are not mentioned in this paper, to cope with very large datasets. The current development version implements the seamless integration of database connections.

Acknowledgments

The development of Mondrian started in 1997 at AT&T Shannon Labs. Further development will be carried out in Augsburg, as well as at other sites. Those who want to join contributing to the development may contact the author.

[2] given an installation of SUN's JDK 1.3 or higher

References

[1] W. S. Cleveland and M. E. McGill. *Dynamic Graphics for Statistics*. Wadsworth & Brooks/Cole, Pacific Grove CA, 1988.

[2] D. R. Cox and E. J. Snell. *Applied Statistics — Principles and Examples*. Chapman & Hall, London, 1991.

[3] J. Emerson. Mosaic displays in s-plus: A general implementation and case study. *Statistical Computing & Statistical Graphics Newsletter*, 9(1):17–23, 1998.

[4] H. Hofmann. Simpson on board the titanic? interactive methods for dealing with multivariate categorical data. *Statistical Computing & Statistical Graphics Newsletter*, 9(2):16–19, 1998.

[5] J. Hummel. Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, 11(1):23–33, 1996.

[6] A. Inselberg. Visual data mining with parallel coordinates. *Computational Statistics*, 13(1):47–63, 1998.

[7] D. Swayne, D. Temple, A. Buja, and D. Cook. Ggobi: Xgobi redesigned and extended. In *Proceedings of the 33th Symposium on the Interface: Computing Science and Statistics*, 2001.

[8] M. Theus. User interfaces of interactive statistical graphics software. In *Proceedings of the 31th Symposium on the Interface: Computing Science and Statistics*, 1999.

[9] M. Theus, H. Hofmann, and W. A. Selection sequences — interactive analysis of massive data sets. In *Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics*, 1998.

[10] M. Theus and S. Lauer. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(3):396–412, 1999.

[11] P. F. Velleman. *DataDesk Version 6.0 — Statistics Guide*. Data Description Inc., Ithaka, NY, 1997.

[12] W. Venables and B. Ripley. *Modern Applied Statistics with S-PLUS, 3rd Ed.* Springer, New York, NY, 1999.

[13] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85:664–675, 1990.

[14] G. Wills. A good, simple axis. *Statistical Computing & Statistical Graphics Newsletter*, 11(1):20–25, 2000.