CREATING AND MANIPULATING N-DIMENSIONAL BRUSHES

Matthew O. Ward, Worcester Polytechnic Institute
Computer Science Department, 100 Institute Rd., Worcester, MA 01609

Key Words: linked brushing, multivariate data visualization, N-dimensional brushes

Abstract

Techniques for interactive brushing, as found in multivariate data visualization systems, can be categorized as either screen-space or data-space. Screen-space techniques consider a brush to be a 2-dimensional shape (usually a rectangle) which can be used to select points which map to a particular region of the display. Data-space techniques assume the brush has as many dimensions as the data set, i.e. the brush specifies an N-dimensional subspace of the entire data space. In this paper, I describe methods for specifying and manipulating N-dimensional brushes and show their implementation in the public-domain visualization package **XmdvTool**. Among the topics covered are user-driven versus data-driven brushing, direct and indirect brush specification and modification, management of multiple simultaneous brushes, composite brush creation, and using brushes with ramped boundaries.

1 Introduction

The term *brushing* in the context of data visualization refers to the process of interactively selecting a subset of the data [3]. The purpose of the selection may be to highlight the data, delete it, mask it (removing all data not covered by the brush), perform statistical analysis on the subset, or any other user-specified operation. In this same context, *linking* refers to the communication of information between multiple views of a data set. This may consist of viewing parameters, analysis results, or any of a wide range of useful information. *Linked brushing* is therefore the communication of brush parameters, such as location and extents, from one view of the data to another. This is a very powerful technique, as it permits users to view a particular subset of data from several distinct perspectives.

XmdvTool [11]is a public-domain software package which integrates several techniques for multivariate data visualization, including scatterplot matrices [3, 4], parallel coordinates [6], star glyphs [10], and dimensional stacking [7]. By providing multiple display methods the system allows users to obtain a richer understanding of their data sets, drawing on the strengths of each of the supported techniques. Within **XmdvTool** is implemented the concept of an *n-dimensional brush* [8], which is a technique for interactively specifying and manipulating a hyperbox in n-dimensional space. In this paper, I describe the concepts and implementation of n-dimensional brushes and show how they can be used to perform data exploration tasks.

2 Background

The idea of brushing for data visualization has been around for more than 20 years. Fisherkeller, Friedman, and Tukey used the idea of interactively selecting a region in their PRIM-9 system [4], although it was not termed “brushing.” Becker and Cleveland were the first to formalize the concepts of data brushing [3], where the characteristics of brushes were derived and a system that implemented brushing with masking and highlighting was developed. Martin and Ward [8], extended brushes to be defined in either data-space or display-space, and provided techniques for manipulating data-space (N-dimensional) brushes in the **XmdvTool** package (see below). Recent work by Wills [12] has produced a taxonomy of data selection operations, categorizing methods based on whether they do or do not “remember” past selections, the type of area differentiation used, whether they are data dependent or independent, and the resulting change to the visualization considering the previously and newly selected points (replace, add, toggle, subtract, intersect).

The philosophy behind **XmdvTool** is that by integrating multiple methods of displaying high dimensional data one can take advantage of the strengths and overcome the weaknesses of each individual method. The four display techniques implemented in **XmdvTool** are:

- Scatterplots – Scatterplots are perhaps one of the oldest and most commonly used high dimensional visualization methods [3, 4]. Each of $N * N$ pairwise parallel projections are generated and arranged in a grid structure (Fig. 1a).

- Glyphs – A *glyph* is a generic term describing a graphical entity that is generated by mapping data values to graphical attributes. **XmdvTool** uses star glyphs [10], where the data value from each dimension maps to the length of a line. Each of these lines has a common center and radiates at uniformly spaced angles outwards. The outer endpoints are connected to form a polygon (Fig. 1b).

- Parallel Coordinates – This method was developed by Inselberg and Dimsdale [6], and consists of N vertically-aligned axes with each data point represented as a polyline through the N axes (Fig. 1c).

- Dimensional Stacking – The dimensional stacking technique is a recursive projection method developed by LeBlanc et. al. [7]. Two dimensions are used to define a discrete horizontal and vertical axis, creating a grid on the display. Within each box of this grid this process is applied again with the next two dimensions, and this process continues until all dimensions are assigned (Fig. 1d.).

In addition to these data displays, **XmdvTool** provides an extended version of the Tukey Box Plot (Fig. 2), which displays the mean, standard deviation, median, quartiles, and outliers for the data point currently selected. Finally, the user is provided a separate window to view the text version of selected points.

XmdvTool is not a unique visualization tool. A growing number of other software packages are using brushing to link distinct methods of visualizing multivariate data. Some of these include EDV [12], VisuLab [9], and SPSS Diamond ^{TM}. The major differences between these techniques and **XmdvTool** are the specific display techniques supported, the degree to which they are integrated with statistical operations, and the variety of interactive brushing methods supported. The next section provides an overview of brushing in **XmdvTool**.

3 Basic Brush Creation, Manipulation, and Display

The basic brush in **XmdvTool** consists of two components: the specification of subranges of each of the N data dimensions (which we define as the *coverage*) and the operation to be performed. Each can be changed dynamically by the user. The currently supported operations consist of

1. Highlighting: set the color of each covered point to a distinct value (default).

2. Delete: remove the covered points from the display.

3. Mask: remove all but the covered points from the display.

4. Average: add a new point to the display based on the average values of the covered points.

5. Quantify: display the text representation of covered points in a separate window.

The default coverage for the brush is a hyperbox centered on each dimension with an extent equal to one half of the range for that dimension. There are numerous ways of manipulating this coverage:

Direct, data-driven: the user may create a brush by *painting* over displayed points while holding the shift key and left mouse button down. This will result in a brush whose extents are set to be the minimum bounding hyperbox containing the selected points. In scatterplot and parallel coordinate mode, this painting is continuous, i.e. the mouse button is held down throughout, while with glyphs and dimensional stacking, the user can click the mouse button several times on discrete data points (see Fig. 3a).

Direct, user-specified: in both the scatterplot and parallel coordinate displays, the extents of the brush can be displayed as a shaded region. The user can modify the location of one of the boundaries of this region by clicking the left mouse button on or near the boundary and dragging in the desired direction. In addition, the user can move the position of the brush by clicking the middle button and dragging. Note that in parallel coordinates, only one dimension may be modified at a time, while in scatterplots one or two dimensions may be changed simultaneously (see Fig. 3b).

Indirect, user-specified, local: two indirect tools are also provided in **XmdvTool**. The first is a series of sliders oriented horizontally and vertically to correspond with the configuration of dimensions in dimensional stacking. Left and middle mouse buttons control the coverage as in the direct user-specified mode. Another tool is shaped like a large glyph, with each ray acting as a slider. The minimum value for each dimension is offset from the center of the glyph to avoid ambiguity in selecting the dimension to be manipulated (see Fig. 4).

Indirect, user-specified, global: a useful function provided within **XmdvTool** is to increase or decrease the extents of all dimensions simultaneously. This permits the exploration of neighborhoods of selected data points as well as the distinction between points which are central or peripheral within the brush extents (see Fig. 5). This global brush manipulation is currently supported via buttons for increasing and decreasing all dimensions by ten percent of the entire range.

The final aspect of basic brushing in **XmdvTool** is how the brush is displayed. The option for displaying the brush exists for all projection techniques except for glyphs, although it is available on both indirect tools for brush manipulation. Each option has its advantages and disadvantages in terms of ease of distinguishing brush extents and amount of congestion on the screen.

Shaded region: the default display mode for brushes is to shade the subspace covered by the brush in a mute tone which conveys the brush extents without interfering significantly with the display of the data. The color used needs to be carefully selected to ensure sufficient contrast with both the original and highlighted data colors.

Brush outline: the original version of **XmdvTool** displayed the brush extents as an unfilled polygon. This was only applicable to scatterplots, parallel coordinates, and the indirect brush manipulation tools. However, because of the difficulty in discerning the boundaries in the parallel coordinates view when dense data sets were viewed, this option was dropped from later versions of the software.

Invisible extents: the user has the option of not drawing the brush extents. This helps remove clutter from the display, but makes direct brush manipulation more difficult. Thus it is common to toggle between this and the shaded region mode.

4 Advanced Techniques

The basic brushing capabilities of the previous section provide a rich variety of methods to explore multivariate data sets. In this section we describe some more advanced brushing capabilities found in **XmdvTool**, along with justifications for their use.

Multiple brushes: a single brush allows users to isolate one hyperbox within the data space. This supports sequential, independent classification of features within the data. A useful facility within **XmdvTool** is to permit users to define and display up to four brushes simultaneously (see Fig. 6). The coverage of each brush and points contained within the brush are indicated in different colors (user-definable), and regions or points covered by multiple brushes are colored according to a user-specified ordering of the brushes. The user can rapidly move brushes to the top or bottom of the ordering to distinguish points with multiple coverages.

Logical operators: an alternative to dealing with multiply-covered points is to extend the selection operation to create new brushes which are logical combinations of other brushes. The default selection operation is simply the points which fall within the currently active brush. However, the user can specify selection to be include unary (NOT) and binary (AND, OR) combinations of some or all of the brushes (see Fig. 7). The logical expressions possible are not exhaustive (we do not parse an arbitrary expression), but cover a wide range of possibilities. Some examples can be seen in Fig. 8.

Ramped boundaries: another avenue we have been exploring is the notion of partial brush coverage. Rather than making a binary decision about whether a data point falls within the extents of a brush, we allow users to specify a ramped brush boundary. Thus points can either fall completely within the brush (all dimensions are within brush extents) or be partially covered (one or more dimensions falls between the original brush boundary and an outer boundary corresponding to the bottom of the ramp). By adjusting the shade of the highlight color proportional to the degree of coverage (the lighter shades indicate less coverage), we can now visually distinguish between central and peripheral points (see Fig. 9). And unlike the global brush sizing controls, we can vary the width of the ramp for each dimension individually.

5 Summary and Conclusions

This paper has presented an overview of N-dimensional brushing and a description of its implementation within **XmdvTool**. Techniques are provided for creating and manipulating brushes with various characteristics and capabilities to support the exploration of multivariate data sets. Our current research on brushing includes investigating non-linear brush ramps, set-based rather than space-based brush membership, brush annotation, and automated, data-driven brush formation.

XmdvTool is freely available via anonymous ftp at ftp.wpi.edu in the directory contrib/Xstuff. The filename is XmdvTool3_1.tar.gz. It runs on any UNIX or Linux platform running X11R4 or higher, and uses Athena Widgets and the Widget Creation Library.

References

[1] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[2] R. A. Becker, W. S. Cleveland, and A. R. Wilks. Dynamic graphics for data analysis. In W. S. Cleveland and M. E. McGill, editors, *Dynamic Graphics for Statistics*, pages 1–50. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988.

[3] R. A. Becker, W. S. Cleveland, and A. R. Wilks. The use of brushing and rotation for data analysis. In *Dynamic Graphics for Statistics*, pages 247–275. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1988.

[4] M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey. Prim-9, an interactive multidimensional data display and analysis system. In *Dynamic Graphics for Statistics*, pages 91–109. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1975.

[5] J. Haslett, R. Bradley, P. Craig, A. Unwin, and G Wills. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *Statistical Computing*, 45(3):234–242, 1991.

[6] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multidimensional geometry. In *Proceedings of Visualization '90*, pages 361–378, 1990.

[7] J. LeBlanc, M. O. Ward, and N. Wittels. Exploring n-dimensional databases. In *Proceedings of Visualization '90*, pages 230–237, 1990.

[8] A. R. Martin and M. O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of Visualization '95*, pages 271–278, 1995.

[9] C. Schmid and H. Hinterberger. Comparative multivariate visualization across conceptually different graphics displays. In *Proceedings of SSDBM '94*, pages 42–51, 1994.

[10] J. H. Siegel, E. J. Farrell, R. M. Goldwyn, and H. P. Friedman. The surgical implication of physiologic patterns in myocardial infarction shock. *Surgery*, 72:126–141, 1972.

[11] M. O. Ward. Xmdvtool: integrating multiple methods for visualizing multivariate data. In *Proceedings of Visualization '94*, pages 326–333, 1994.

[12] G. J. Wills. Selection: 524,288 wayts to say "this is interesting". In *Proceedings of Information Visualization '96*, pages 54–60, 1996.

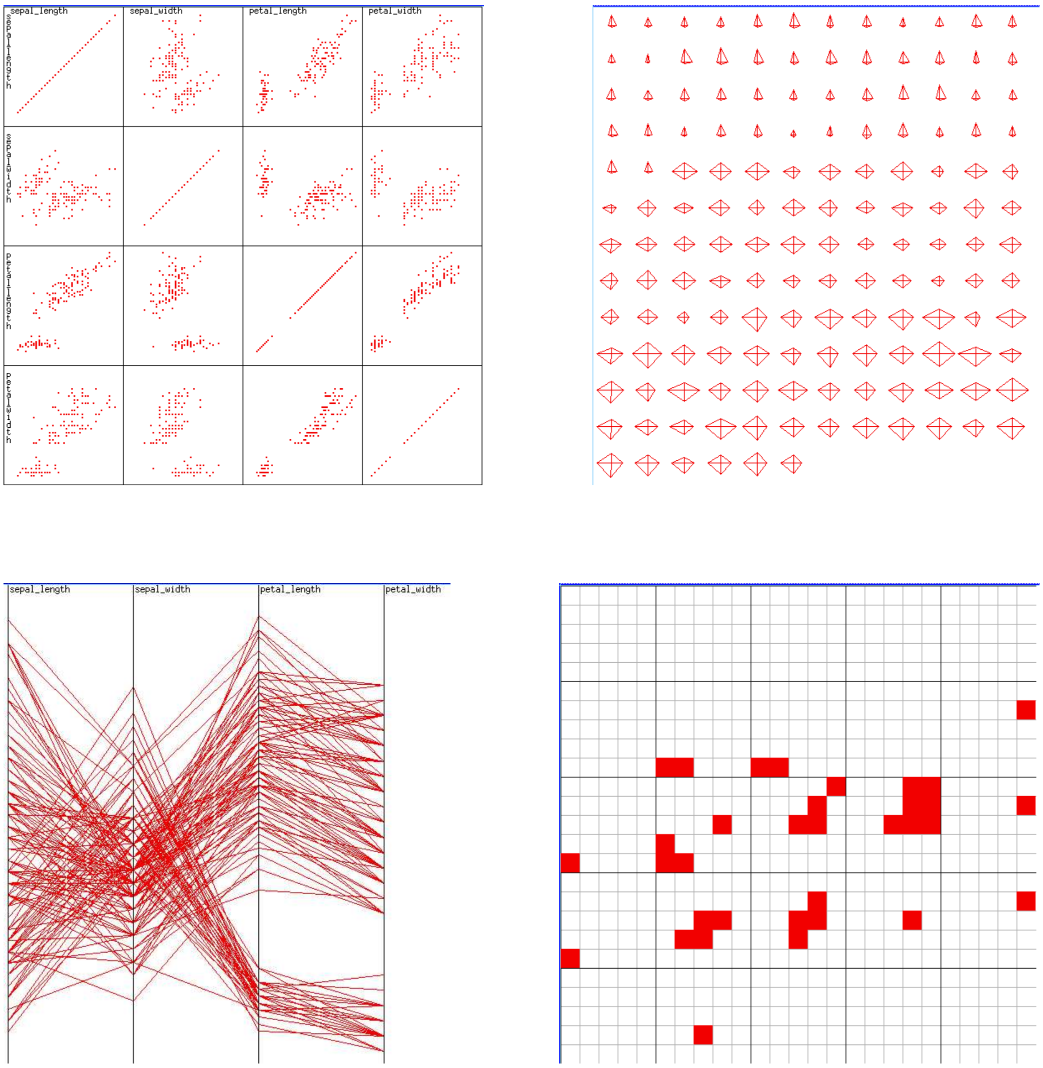

Figure 1: Four views of the Iris data set using XmdvTool: Scatterplot matrix, star glyphs, parallel coordinates, and dimensional stacking (5 discrete levels for each dimension).

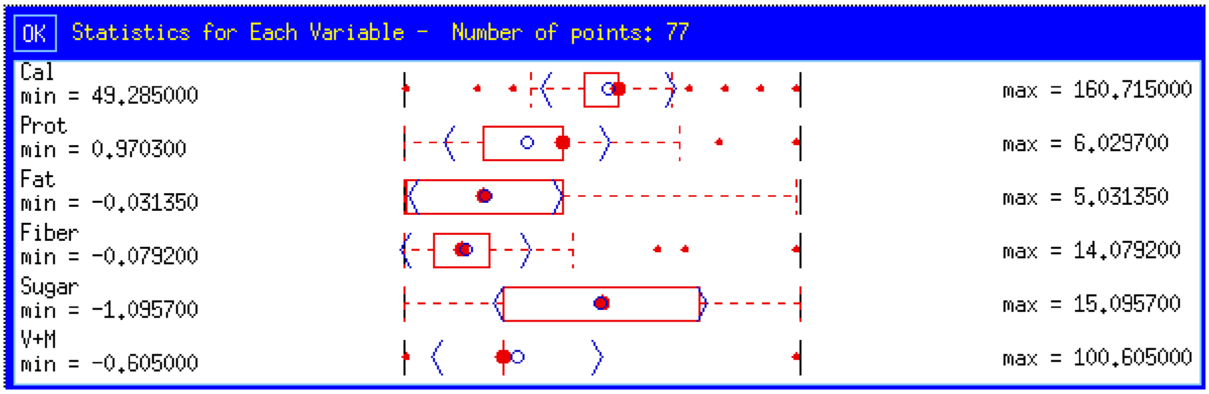

Figure 2: Extended Tukey Box Plots of data covered by brush (6 dimensions of breakfast cereal data). Filled circle is median value, box represents extents of middle two quartiles, dashed line represents median plus/minus 1.5 times size of middle two quartiles, and dots are outliers. Hollow circle is the mean value, and angled brackets represent standard deviation.

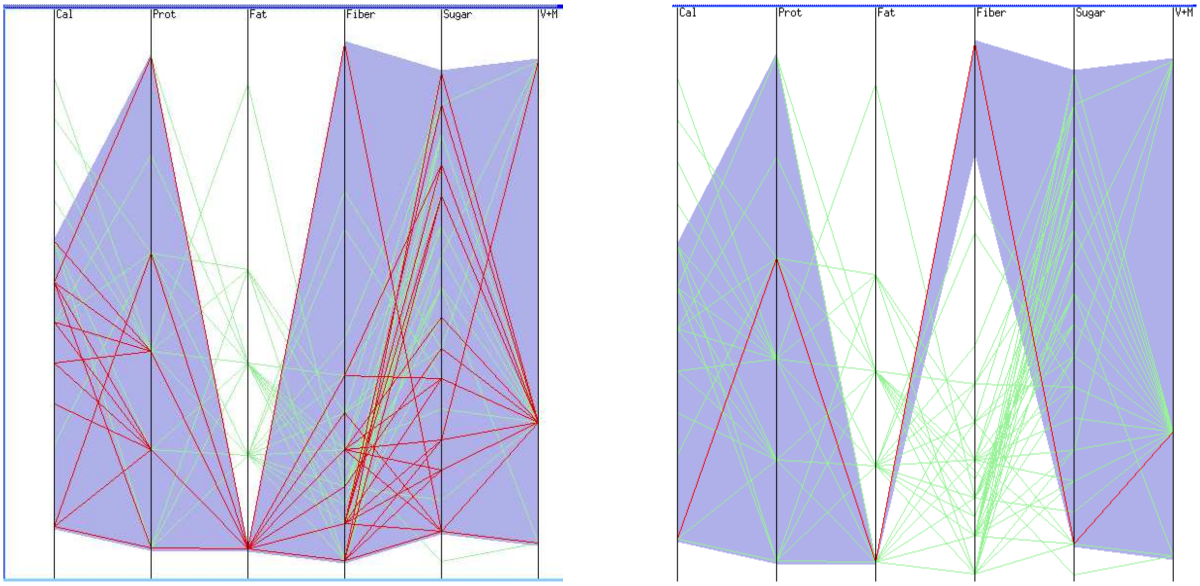

Figure 3: Parallel coordinates view of breakfast cereal data (6 dimensions selected), first with a data-driven brush focussed on low-fat data points, and then a user-controlled adjustment to highlight the high fiber points in the subset.

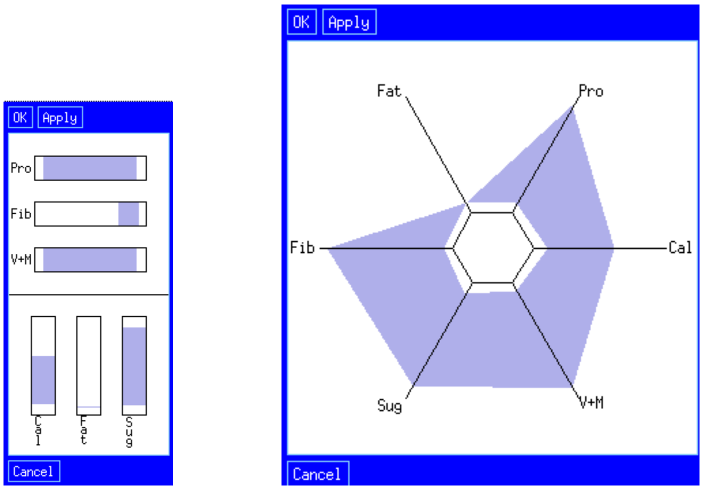

Figure 4: Dimensional Stacking and Glyph Brush Tools defining low fat subset of breakfast cereal data.

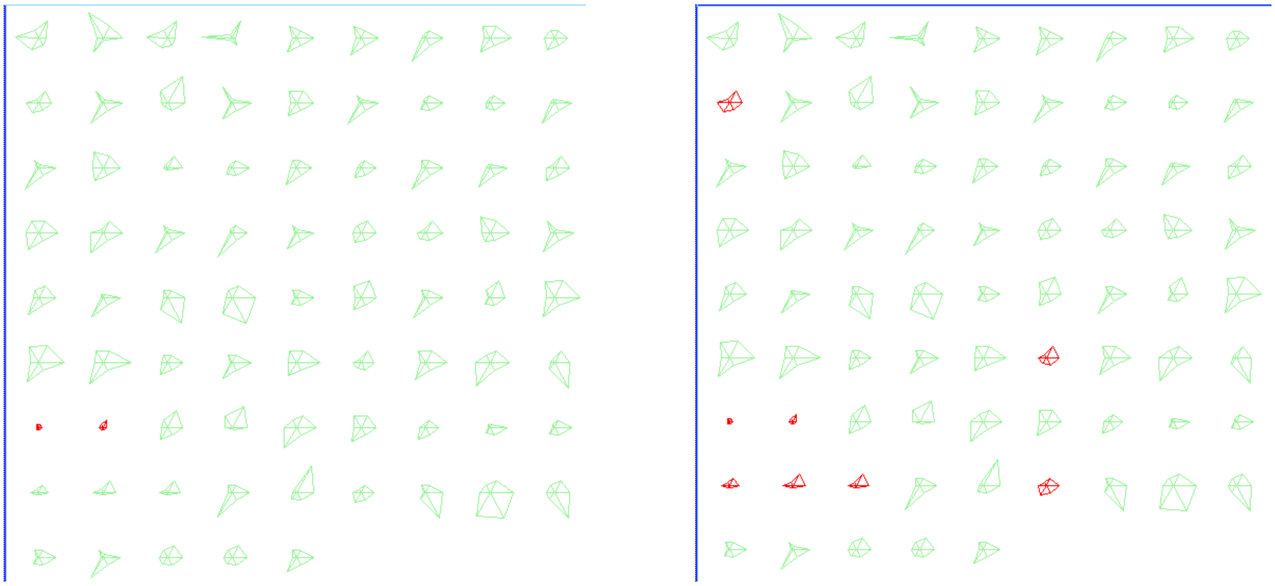

Figure 5: Data-driven brush choosing glyphs representing breakfast cereals with low values on all 6 selected dimensions, and then using the global size adjustments to expand the brush in all dimensions uniformly, thus showing points near the original selection.

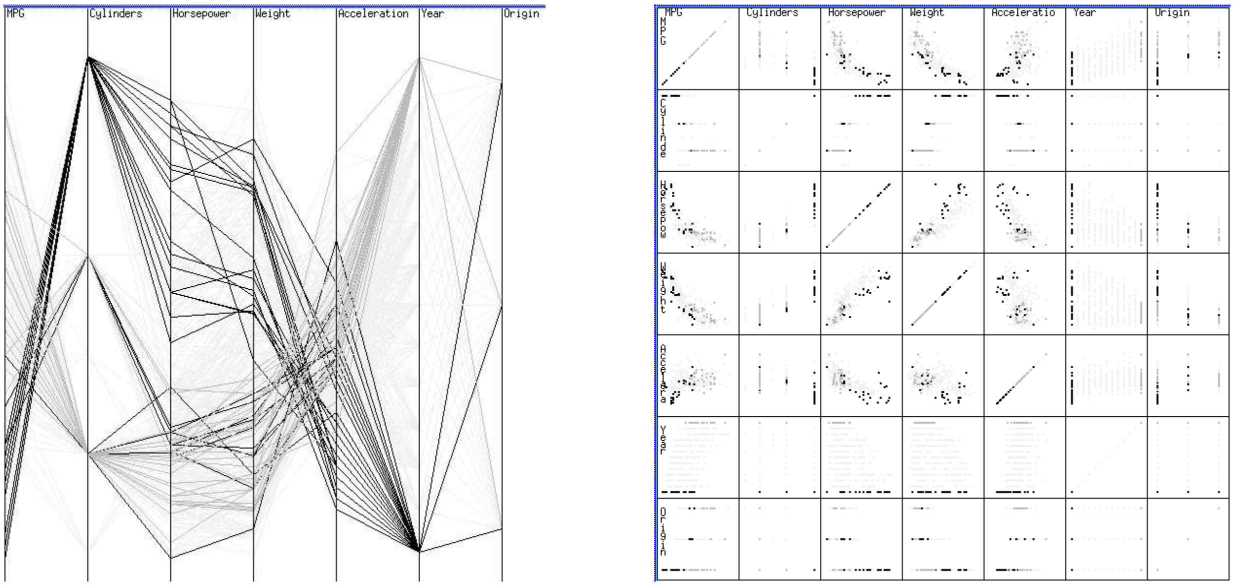

Figure 6: Multiple brushes, one representing all cars from 1970 and the other representing all cars from 1982.

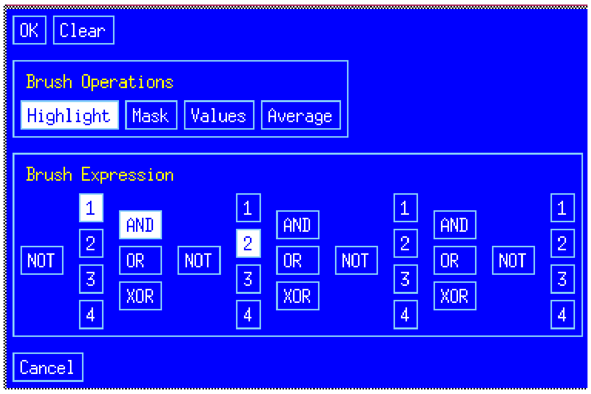

Figure 7: Control panel for combining brushes in logical expressions.

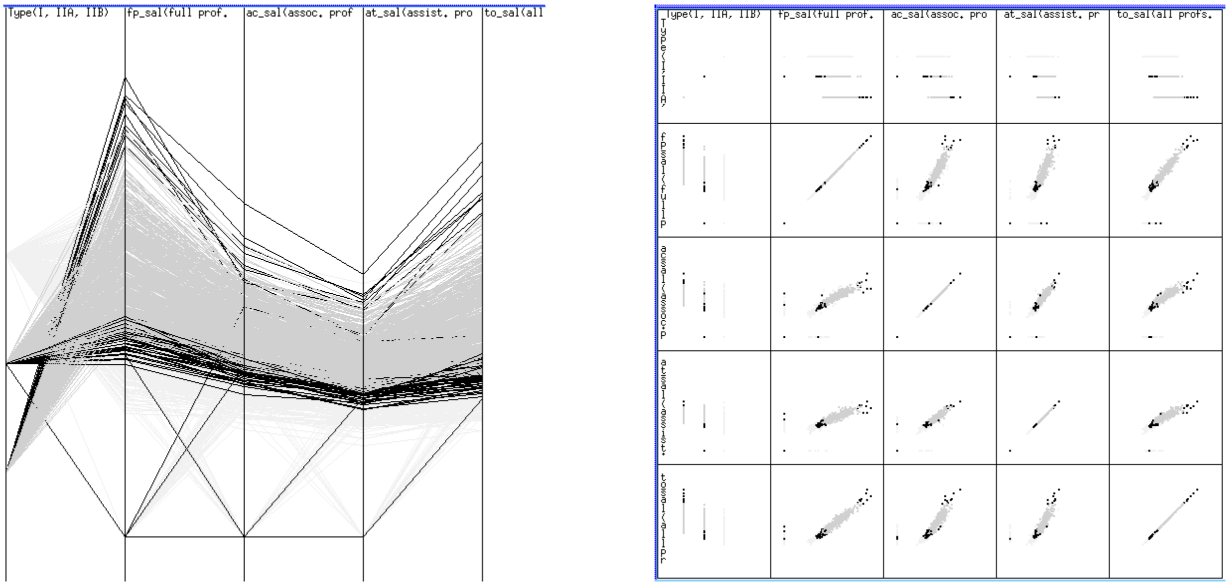

Figure 8: Two views of the exclusive OR of two brushes using the AAUP faculty salary data. The first is initially focussed on Type I universities, after which the Type extent is expanded to include Type IIA universities. The second is initially focussed on Type IIA universities after which the Type extent is expanded to include Type I universities. The exclusive OR shows the points belonging to one but not both brushes.

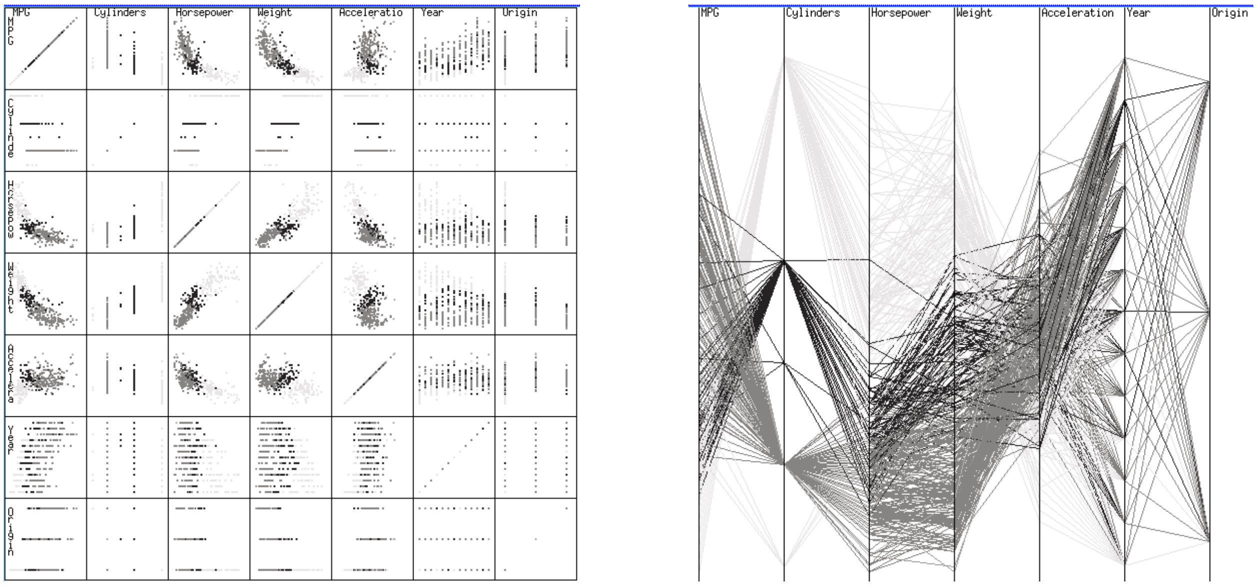

Figure 9: Two views of ramped brushes. Intensity indicates distance from a brush outlining 5 and 6 cylinder cars.