Visual Data Analysis using Tracked Statistical Measures within Parallel Coordinate Representations

Daniel Ericson, Jimmy Johansson and Matthew Cooper
NVIS - Norrköping Visualization and Interaction Studio
Linköping University
S-58183 Linköping, Sweden
daner292@student.liu.se, jimjo@itn.liu.se, matco@itn.liu.se

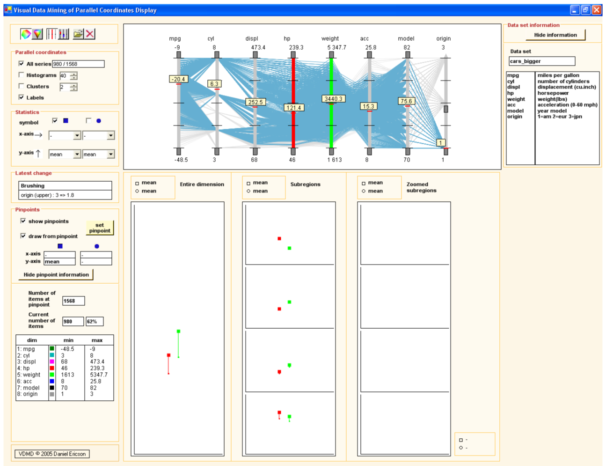
Abstract

With our increasing ability to capture or produce and to store large multivariate data, these data sets are increasing both in size and complexity. Many conventional techniques for visualizing multivariate data suffer from problems like cluttered displays since they are not designed to handle these amounts of entries. We present a novel method to overcome this problem by interactively selecting and displaying statistics derived from the data in a separate view. Changes in the display are visually tracked by animation and vector plotting for easy comparison of various measures applied to different subsets of the data.

1 Introduction

The parallel coordinates technique [7] is a powerful standard method for visualizing and analyzing multivariate data sets in a two-dimensional representation. With sophisticated modern acquisition and storage methods, data sets are rapidly increasing in size and complexity. Visualizing these large data sets with traditional methods like parallel coordinates, or any other method that displays a single data item per point, we frequently encounter over-plotting problems like display cluttering. Single data observations become impossible to distinguish and trends are no longer discernible (see figure 2).

In this paper, we propose a method called Visual Data Mining Display (VDMD) to display statistical measures of a data set, where its representation in parallel coordinates is too densely plotted for us to recognize significant changes in the distribution of points on each axis. The parallel coordinates technique was chosen to build upon because of its rich possibilities for interaction with multivariate data. The statistics in the VDMD are plotted as static or animated glyphs in a separate coordinate system in conjunction with the parallel coordinates. Figure 1 shows a screen shot of the graphical user interface of the application. Interaction with the parallel coordinates are reflected as changes in the statistics display. Any change in the statistics can be tracked by drawing a vector from the previous value, or from another point of interest, to the new statistical value. With this approach we aim to explore the data and find structures that are hidden in the cluttered parallel coordinates view.


Visual Data Mining of Parallel Coordinates Display

Parallel coordinates
☑ All series 980 / 1568
☐ Histograms 40
☐ Clusters 2
☑ Labels

Statistics
symbol ☑ ■ ☐ ●
x-axis → - -
y-axis ↑ mean mean

Latest change
Brushing
origin (upper) : 3 => 1.8

Pinpoints
☑ show pinpoints
☑ draw from pinpoint
set pinpoint
x-axis ■ - ● -
y-axis ■ mean ● -
Hide pinpoint information

Number of items at pinpoint 1568
Current number of items 980 62%

dim		min	max
1: mpg	🟩	-48.5	-9
2: cyl	🟦	3	8
3: displ	🟪	68	473.4
4: hp	🟥	46	239.3
5: weight	🟩	1613	5347.7
6: acc	🟦	8	25.8
7: model	⬛	70	82
8: origin	⬜	1	3

VDMD © 2005 Daniel Ericson

Data set information
Hide information
Data set
cars_bigger

mpg	miles per gallon
cyl	number of cylinders
displ	displacement (cu.inch)
hp	horsepower
weight	weight(lbs)
acc	acceleration (0-60 mph)
model	year model
origin	1=am 2=eur 3=jpn

□ mean ○ mean Entire dimension
□ mean ○ mean Subregions
□ mean ○ mean Zoomed subregions

Figure 1. Graphical user interface of the application. Three axes are selected in the parallel coordinates and its corresponding statistical values are presented in the displays below.

The rest of this paper is organized as follows. Section 2 discusses previous work that addresses the problem of overplotted visual displays and related multivariate visualization techniques. Section 3 gives a thorough explanation of the features implemented in our system. There is also a description of the software used in the implementation as well as performance results. Section 4 evaluates the method and

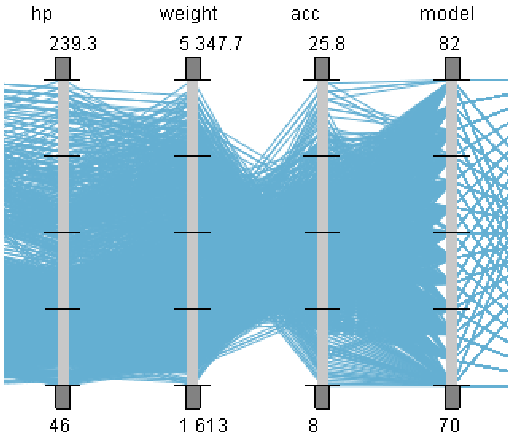


Figure 2. When too many lines are plotted in parallel coordinates, the view will become cluttered.

compares it with current existing extensions to parallel coordinates. The evaluation is performed with a set of case studies using data sets from different potential application areas. In section 5, we give conclusions about the method and suggest some possible future improvements.

2 Related work

There have been many previous efforts to overcome the problem of overplotted visualization displays. While some focus on reducing the number of items to display, other approaches aim to summarize the whole data set and display important characteristics. The main categories of previous work are described in the following sections.

2.1 Brushing methods

Brushing is the operation of interactively selecting subsets of the data to be highlighted, masked or deleted. Brushing multivariate data requires a concept of how to map all dimensions onto the two dimensions of the screen. Martin and Ward [11] describe the design of several brushes that are defined in data space rather than screen space. Many different types of brushing methods have been developed. Hauser et al. [6] present a method called angular brushing to brush data with respect to the correlation between adjacent axes in parallel coordinates.

2.2 Overview and Detail methods

There are several well-known techniques that take interesting parts of the data into consideration while ignoring the rest. The advantage of this is the ability to keep the overview of the whole data set, while focusing on a subset of the data. Examples are distortion techniques like the fish-eye lens [5, 16] and the Perspective Wall [10].

2.3 Summarization methods

When the whole data set is of interest, displaying the distribution of values rather than focusing on each single item of the data, can be more beneficial. Histograms are intuitive ways of showing such distributions, and have been implemented with parallel coordinates in, for example, [15] and [6].

In the scatter plot matrix, all dimensions in a data set are compared pairwise and mapped onto a two-dimensional projection. The projections are arranged in a grid structure to give the user an impression of the overall relations between dimensions in the data.

Another way of summarizing data is to group it into clusters to see if there are any obvious groupings in the data, that can be further manipulated and analyzed. Extensions to traditional clustering techniques have been proposed in order to gain more information about the overall structure and information about single clusters. Fua, Ward and Rundensteiner [4] propose a multiresolution view of parallel coordinates to obtain a level of detail structure via hierarchical clustering.

Johansson, Cooper and Jern [8] extend standard parallel coordinates to 3-dimensional clustered multirelational parallel coordinates. This allows for a simultaneous one-to-one dimension analysis between a focused dimension and all other dimensions.

Siirtola [17] describes two novel techniques to manipulate parallel coordinates. The polyline averaging dynamically summarizes a set of polylines by displaying the average line in a selection. The other technique enhances the knowledge of correlation between variables by plotting a bar between the ranges. The bar points upwards or downwards, indicating if the correlation between adjacent axes is positive or negative.

An approach to interactive data summarization is proposed in [9]. The idea is to have the computer do the exhaustive search rather than the user, and inform the user in advance about which manipulations would change the summary the most.

The box plot was introduced in the 1970’s and is still a widely used technique to display variable distributions in many statistical software packages. It encodes minimum, maximum, mean, median and quartile information in a compact representation. The Mondrian tool [20] uses multiple boxplots on top of the axes in a parallel coordinate plot to display statistics of a given subset of the data. In [2], box plots are extended to “ellipse plots” to compare subsets of the data with the whole data set.

Zhao et al. [21] introduce trend figures as an extension to parallel coordinates. The horizontal axis of the figure represents the sequence of the data record, and the vertical axis shows its value in each data record. Extending each axis of the parallel coordinates with these trend figures enables the user to quickly observe the variables that change in similar ways. Their work was specifically used to observe changes made in the design and test cycle of mobile phones.

In [1], axes in parallel coordinates are scaled according to statistics of the data set, which can be helpful in showing the distribution of data points on the axes.

2.4 Re-arrangement Methods

Axis re-arrangement is an old and now obvious extension to parallel coordinates. Comparison of two dimensions is best performed when the axes are placed next to each other. The XmdvTool is a public-domain software package for interactive multivariate data exploration where this feature is implemented. Deletion and addition of axes [6] can be useful to clear up the view, and to compare multiple variables with one specific dimension of interest.

Peng et al. [14] discuss more in detail the advantages of re-arranging dimensions in multi-dimensional visualizations, and propose algorithms to automatically find the optimal axis arrangement.

The Reorderable Matrix presents multivariate data graphically in a table that has objects in columns and properties of those objects as rows. Each crossing between rows and columns has a rectangle whose size is relative to the corresponding data value at that point.

Siirtola [18] examines the benefits of combining two conceptually different information visualization techniques. The parallel coordinates plot and the reorderable matrix were used to view the same data, with positive results. He concludes that linking different kinds of displays can enable users to see different things in their data, as well as reducing the cognitive load when they switch between the views.

2.5 Animated visualization methods

Animation has been widely used in visualizations to enhance the understanding of time-varying variables. Typically, animation has been used to demonstrate complex physical simulations, for example to show particle traces that change over time in a scientific visualization.

“The Animator” [12] was introduced by Barlow and Stuart with the purpose of illustrating how animation can be used in parallel coordinates. The line segments of the parallel coordinates are animated to enhance the understanding of how objects within the multidimensional space are changed over time.

Elmqvist and Tsigas [3] present a technique called “Growing Squares” where they use the metaphor of colour pools spreading over time on a piece of paper, to visualize causal relations in a system.

2.6 Summary of related work

The sections above describe previous work that is related to our ideas. These are extensions to parallel coordinates or other methods of displaying multivariate data. Solutions exist, that summarize the data displayed in a parallel coordinate plot using aggregated information, for example in [20, 2, 15]. However, most of them are focused on displaying statistics for a given state of the data, without letting the user follow changes in the statistics that may have occured as a consequence of interaction. The novelty of our method is the ability to follow and track changes in the data set statistics via animation and vector plotting.

3 Implementation

We have implemented the Visual Data Mining Display (VDMD) as a complement to parallel coordinates. The purpose of the VDMD is to display statistical analyses of the data in such a way that they enhance the user’s ability to understand correlations and structures in data sets where the number of items is too big for them to be visualized in a meaningful way by traditional multivariate techniques.

Several of the standard interaction and visualization extensions to the parallel coordinates technique described above are also implemented, such as axis re-arrangement, brushing and clustering. The purpose of this is twofold: in order to compare this new method with features already implemented in previous work, and to explore how it functions either as a replacement or as an extension to these.

The graphical user interface of the application (figure 1) consists of three parts. The upper central part is the parallel coordinate representation of the data set. When axes in the parallel coordinates are selected, their corresponding statistics are plotted in the central lower part of the VDMD interface. Menus for interaction with the VDMD and the parallel coordinates are located to the left of these displays.

3.1 Parallel coordinates

When the application starts, the data are visualized in parallel coordinates. Each axis is split into four parts that divide the range into four sub-ranges of equal size. Whole axes, as well as sub-regions of axes, can be selected, using simple mouse interaction, for further exploration in the VDMD. Figure 3 shows a zoom of three axes in the parallel coordinates, where one axis and one region have been

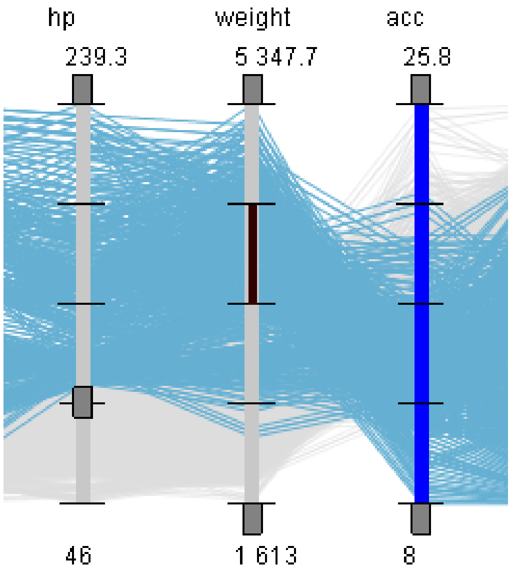


Figure 3. Three axes from the parallel coordinates. The rightmost axis and the third region on the middle axis are selected.

selected. Brushing is performed by moving 'handles', attached to each axis, to crop the data in this dimension. In figure 3, objects from the lower part of the leftmost axis have been brushed with the handle. The cropping is propagated to all axes and complete polylines will be removed. Objects that are removed from the selection are shaded in grey but still visible in order to retain the visual information about the complete data set. To enable comparison of arbitrary dimensions side by side, axes can be re-arranged by simple drag-and-drop mouse interaction.

Classifications of the currently selected data, as clusters and histograms, can be turned on and off as layers on top of the parallel coordinates. The addition of histograms (figure 4) on each axis helps the user to get a quick overview of the distribution of data points in an initial stage as well as to observe changes of the distribution after brushing or interaction. The drawbacks of this representation are its static properties and the inability to mediate changes of multiple dimensions simultaneously. Even though changes may occur in many of the histograms, it can be hard for the user to focus on more than one or two dimensions simultaneously.

The clusters are calculated using the K-means technique, where the user specifies the number of clusters to be identified in the data. Figure 5 shows a number of clusters displayed on top of the parallel coordinates. The cluster centroids are re-calculated after every change in the data selection to update the groupings of the data in a given state.

The reason we have chosen to build upon the parallel coordinates technique is its advantage of easy and intuitive selection and brushing of the data. Our approach might, however, be usefully applied in conjunction with other multivariate visualization techniques.

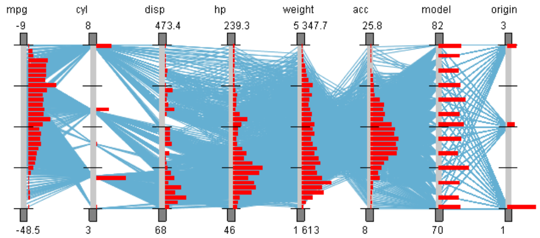


Figure 4. Histograms are placed on top of the parallel coordinates to show the distribution of objects along each axis.

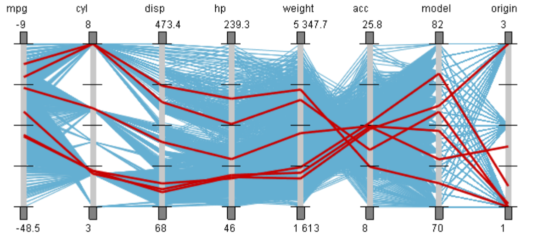


Figure 5. Additional lines, representing cluster centroids of the data set, are placed on top of the parallel coordinates.

3.2 Visual Data Mining Display (VDMD)

The VDMD consists of several two-dimensional coordinate systems, that are linked with the parallel coordinate view to display statistics from selected dimensions and subranges of the data. It is divided into three parts that represent different levels of detail within the selected axes. The leftmost part of figure 6 shows statistics for a whole axis, while the middle part is divided into four regions, each displaying statistics of the corresponding sub-region of the axis. Each subregion is considered independently, in order to gain more information, at a local level, about the axes. Observations of changes at this level can be valuable in understanding what causes the statistical value for a whole axis to change. The rightmost part of the VDMD further divides a single region of an axis, selected by clicking on that region in the parallel coordinates axis display, into four sub-regions of equal size, and statistics of these sub-regions are displayed in the same manner. This level-of-detail feature could be extended indefinitely by dividing sub-regions

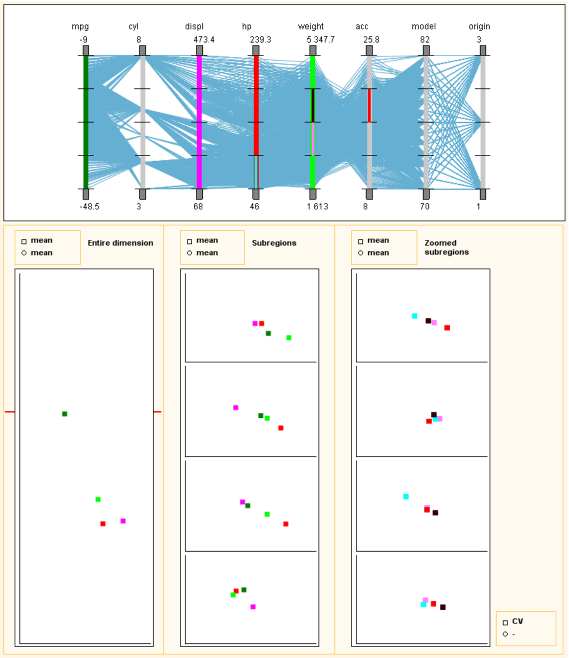


Figure 6. Plotting of statistical values from four dimensions and four regions of the parallel coordinates. The coefficient of variance is plotted on the x-axis and the mean value is plotted on the y-axis.

into even smaller regions but, at present, experiments have only been carried out with three levels since, in the data sets used so far, the number of items would then probably be too small to calculate any useful statistics from.

The plots in the VDMD represent statistical values of selected dimensions. They are, when changed, animated to their new positions, and leave a visible vector as a track of the last movement. Each of these features are described in detail in the following sections.

3.2.1 Plotting

Selected axes in the parallel coordinates display are highlighted with predefined colours that are associated with the specific dimensions even after re-ordering of the axes. We use the "colorcube" command in Matlab to achieve as many regularly spaced colours in RGB colour space as possible. It varies the hue component of the HSV colour model, but tries to provide more steps of grey, pure red, pure green, and pure blue.

For each axis selected, graphical symbols will be plotted in the VDMD to display statistics from the specific dimension. One symbol is plotted in the left display and the middle and right displays will be populated if the corresponding subregion of the axis contains any data item.

The graphics in the VDMD follows the same colour coding as the parallel coordinates and the glyphs can thus easily be connected with the corresponding axes. Each symbol in the VDMD is positioned to display two statistical values for the associated axis in a two-dimensional coordinate system. When axes in the parallel coordinates are brushed, so that the plotted statistical value changes, the symbol will be moved to represent the new value. The symbol will also be moved if it is set to represent a different aggregation function not equal to that of the previous one. The statistical aggregation functions that are mapped to the x- and y-axes respectively can be changed by the user in drop-down menus.Figure 7 demonstrates how the median and mean values of a selected axis change when the brushing of another axis is performed. To provide comparison between selected dimensions, multiple symbols can be plotted side by side in the same display. Since all statistical values are normalized to keep a value between 0 and 1, even dimensions with different ranges can be compared.

The default graphical glyph in the VDMD is a filled square. Additional symbols can be added to display other statistics from the same dimension.

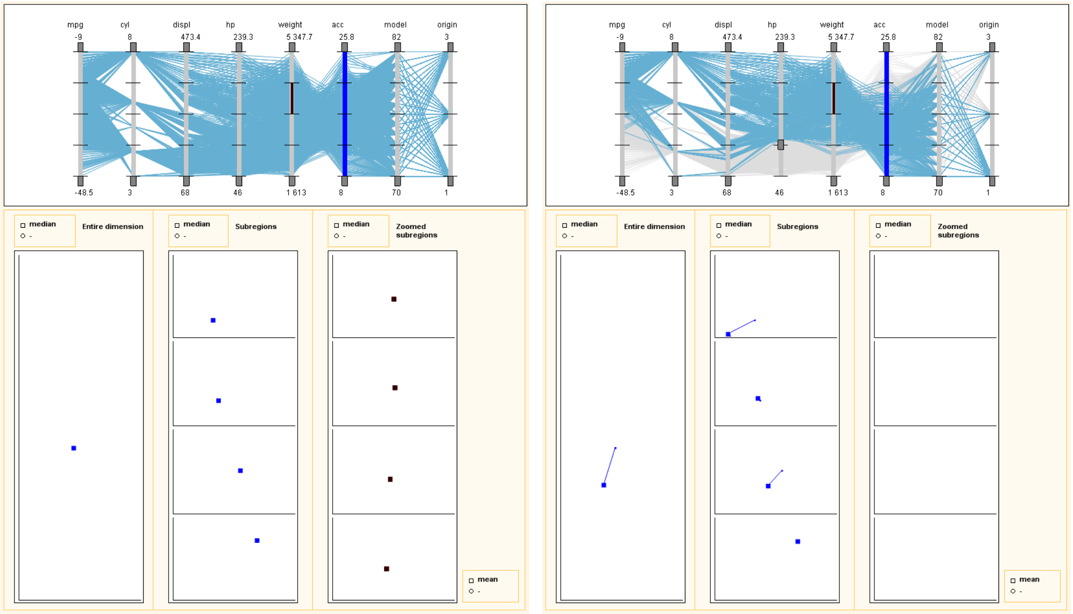
3.2.2 Animation

Once a change is made that affects the position of a symbol, it will be moved to its new position in the VDMD. To better be able to follow the changes, all movements of the symbols are animated along a vector to their new positions. This feature gives the user a clear indication of how different variables are affected by a particular interaction and hence which variables are correlated with each other. Since animation of multiple symbols starts and stops at the same times, the speed of the movement will indicate the size of the change of the statistical value. A faster movement means a bigger change and will tend to attract the user's attention more. Figure 8 demonstrates how animation is used to enhance the visual cues about changes in the display. The animation is run for 100 frames and takes two seconds.

3.2.3 Tracking

The user has the option to track the latest change in the display by having a vector drawn representing the movement of the glyph. The endpoint of the vector is always at the new position of the selected statistical value, that is, the position where the glyph is plotted as described in section 3.2.1.

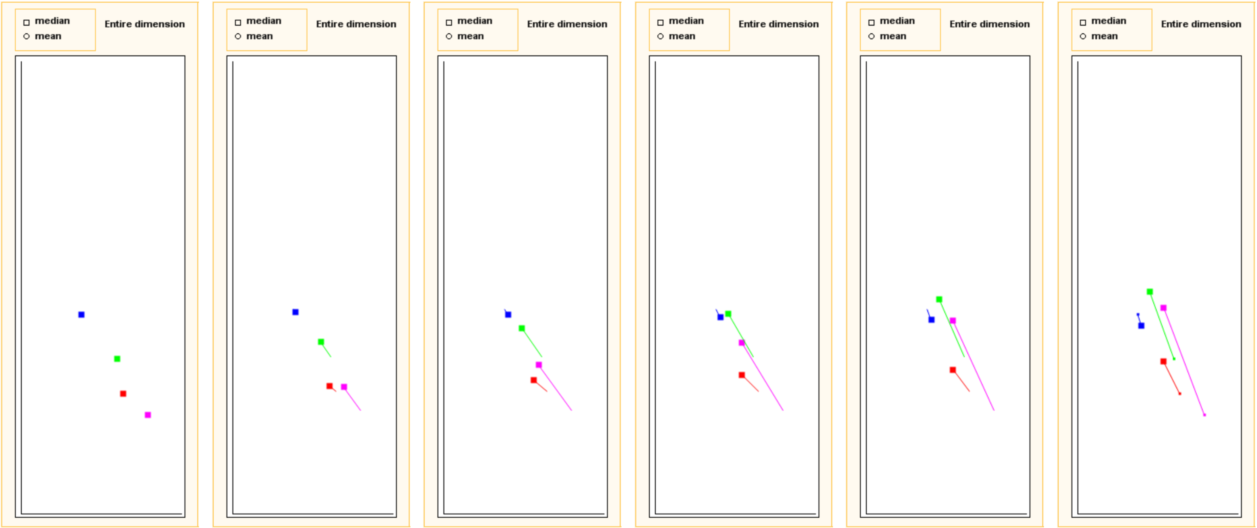
The tracking also follows the animation scheme described in section 3.2.2. By default, the origin of the vector is set to be the previous position of the glyph. The vector starts as a zero vector and is elongated until it reaches its endpoint. Consequently, the result is a vector representing the effect of the change that occurred in the last interaction,



(a) The VDMD before brushing.

(b) The VDMD after brushing.

Figure 7. The appearance of the VDMD before and after the left axis in figure 3 has been brushed. The median value of the right axis of figure 3 is plotted on the y-axis. The x-axis represents the mean value.



(a) Frame 0

(b) Frame 20

(c) Frame 40.

(d) Frame 60

(e) Frame 80

(f) Frame 100

Figure 8. Sequence of six frames during the animation after one variable of the parallel coordinates has been brushed. The sequence demonstrates how the median and the coefficient of variance of four axes are changed when items on a fifth axis are brushed.

whether it occurred as a change of the selected data in the parallel coordinates or as change of aggregation function in the menu. As seen in figure 7(b), the vector is drawn from the previous value to the current.

In addition to this, the origin of the vector can be locked at the current position by clicking the button “Set Pinpoint” shown in figure 9. The current position will be set as a visible “pinpoint”, and future vectors can be chosen to have this point as its origin. If the checkbox “Draw from pinpoint” is checked and a change occurs in the VDMD, the vector will visualize the difference between the current value and the value represented by the pinpoint. In this case, the magnitude of the vector might increase or decrease depending on how the symbol is moved.

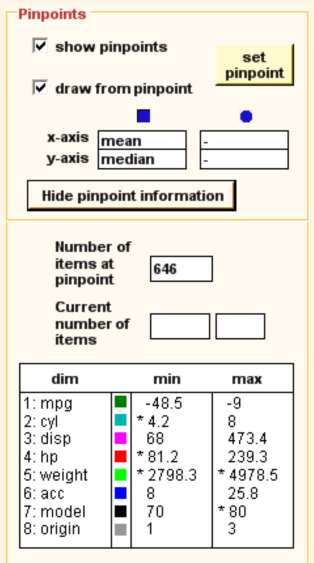
The purpose of the pinpoints is to help the user remember specific statistical values from specific subsets of the data. If multiple axes are selected in the parallel coordinates, vectors are drawn for each axis to provide comparison between variables. A great advantage of the vector plotting is the ability to track changes of arbitrary dimensions even if the corresponding axes were not selected at the time of interaction. When pinpoints are set they are calculated for all dimensions in the data set but are only visible for the ones selected in the parallel coordinates. If an axis or a region is selected after a change is made, the vector will still represent the effect of the last interaction, and can be compared with vectors already visible. This means that the user does not have to keep track of changes of all dimensions simultaneously to see how they are affected by an interaction. Figure 10 demonstrates a sequence where the median is set as a pinpoint, and later compared with the 25th and 75th percentile respectively.

Figure 9 shows the pin-point control interface, the bottom half of which shows the current pin-points that have been set. The max and min values are useful for determining how the selected data was restricted when the pinpoints were set. The values that are changed by the user to brush the data are marked with an asterisk. The colour legend to the right of the dimension names is an additional help in connecting the symbols with the correct axes. There are also notations about which statistical variables the pinpoints represent.

Information about the latest change in the VDMD is also given in text format (figure 11) as a reminder of what changes most recently affected the VDMD.

3.2.4 Linking with parallel coordinates

One of the main purposes of the VDMD is to see how statistical values are affected when the selected subset of the data is changed. Therefore, interactions in the parallel coordinates view are directly reflected in the VDMD. There is also a need for linking in the opposite direction to display the real statistics from the normalized values of the VDMD. The real values indicated by a symbol’s position can be obtained by clicking in the VDMD. Labels on each axis of the parallel coordinates will then be visible, showing the real value that corresponds to the y-position of the click. In figure 6, the lines on both sides of the leftmost display indicate which normalized value is being processed.



x-axis mean -
y-axis median -

Number of items at pinpoint 646

Current number of items

dim		min	max
1: mpg	🟩	-48.5	-9
2: cyl	🟦	* 4.2	8
3: disp	🟪	68	473.4
4: hp	🟥	* 81.2	239.3
5: weight	🟩	* 2798.3	* 4978.5
6: acc	🟦	8	25.8
7: model	⬛	70	* 80
8: origin	⬜	1	3

Figure 9. Information about the current pinpoints is given in a separate view.

The aggregation functions available in the menu are: mean, median, mode, 25th and 75th percentiles, geographical mean and coefficient of variance. All but the last of these are suitable to be mapped along an axis with real values. The coefficient of variance does not have a natural mapping onto the real values of the axes and, for that reason, is not suitable to be displayed on the labels on the axes.

3.3 Software development

The application has been developed in Visual Basic.Net with the visualization library OpenViz [13] used to handle the interaction and graphical display components of the parallel coordinates. All calculations are performed in real-time using a Matlab engine which provides an efficient means to carry out this heavy computation. Figure 12 shows the data flow between the involved components and how interactions are propagated.

3.4 Performance

For our evaluations, we have used a standard desktop PC with a 2.08 GHz AMD processor and 512 MB of RAM. Us-

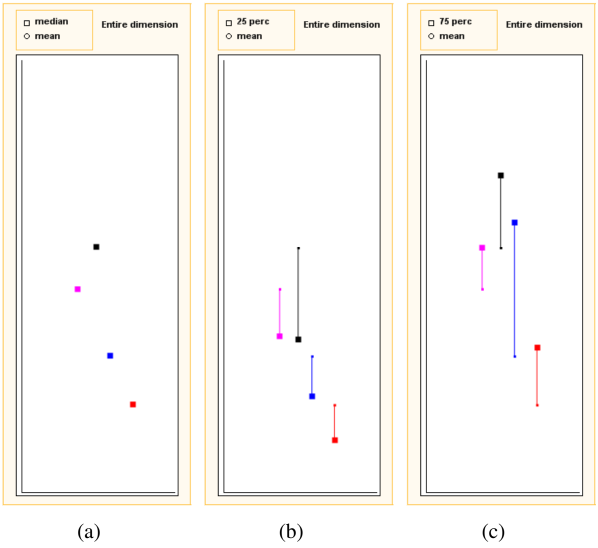
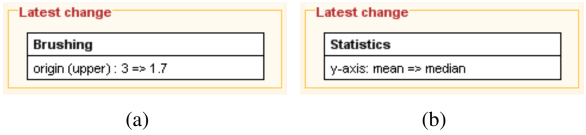


Figure 10. Vectors are plotted in the VDMD to show the difference between statistical measures. (a): Median is set as pinpoint for the y-axis. (b): The current aggregation function is changed to 25th percentile. (c): The aggregation function is changed to 75th percentile, while the pinpoint is still at median.

ing a data set with approximately 3000 objects, the calculation of statistics takes around 0.05 seconds per dimension. In order to provide faster response when selecting axes, statistics are always calculated for all dimensions, which results in calculation times of approximately 0.35 seconds if the data set has 7 dimensions. The more time-consuming part is the drawing of parallel coordinate lines and VDMD plots. The whole process of brushing half the 3000-item data set takes around 2 seconds, including calculation of new statistics. We have used the graphic card GeForce FX5200 from NVIDIA.



Brushing

origin (upper) : 3 => 1.7

Statistics

y-axis: mean => median

Figure 11. The latest interaction that affected the VDMD is displayed in the menu. (a): The parallel coordinates were brushed. (b): The aggregation function was changed.

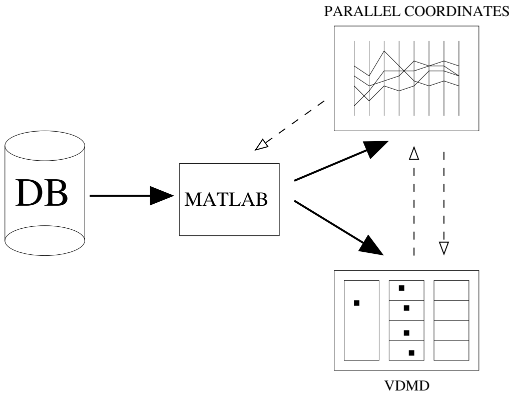
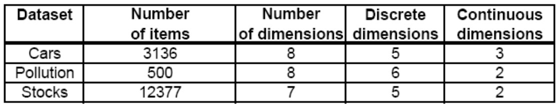


Figure 12. Data flow. Solid lines illustrate the data path. All data are processed and normalized in Matlab before being passed to the graphical components. Dashed lines illustrate interaction in the graphical components. Interaction causes the Matlab engine to recalculate statistics for the entire data set.



Dataset	Number of items	Number of dimensions	Discrete dimensions	Continuous dimensions
Cars	3136	8	5	3
Pollution	500	8	6	2
Stocks	12377	7	5	2

Figure 13. Table describing the data sets used for the evaluation.

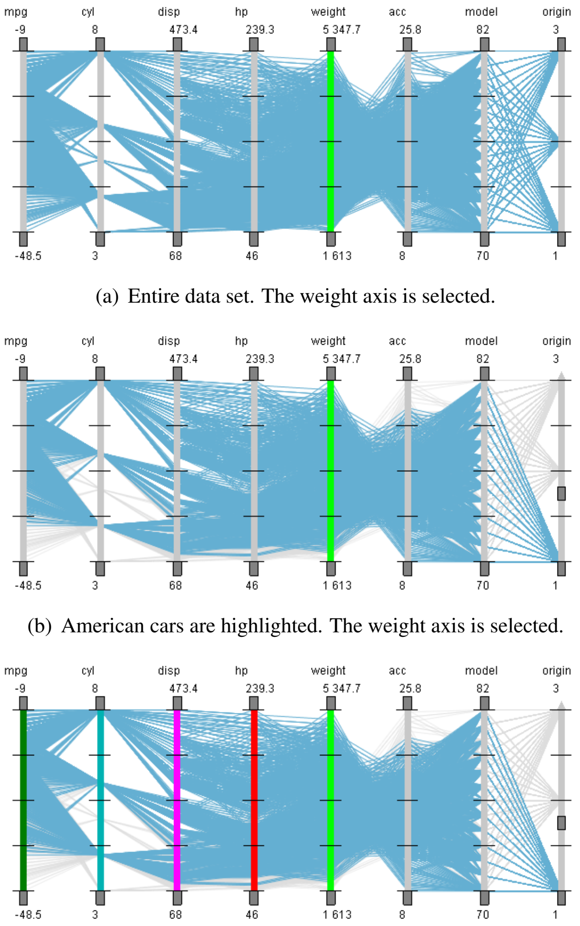
4 Visualization evaluation

We have evaluated our application and ideas using data sets from three different areas. The first two were originally of moderate size (up to 500 items in 8 dimensions), but one has been synthetically extended to hold more objects. These have been collected from [19]. The third one is a data set of stocks information of approximately 12000 items in 6 dimensions. A summary of the three data sets is shown in figure 13.

4.1 Cars

The first data set is a modified and extended version of the standard data set of a collection of cars. The data consists of 3136 observations and each entity has 8 variables: miles per gallon, number of cylinders, displacement, horsepower, weight, acceleration, model and origin. All of the data items are numeric. The origin variable is a number between 1 and 3 where 1 is America, 2 is Europe and 3 is Japan.

Our task is to examine how the *weight* variable is affected by brushing the *origin* axis in the parallel coordi-

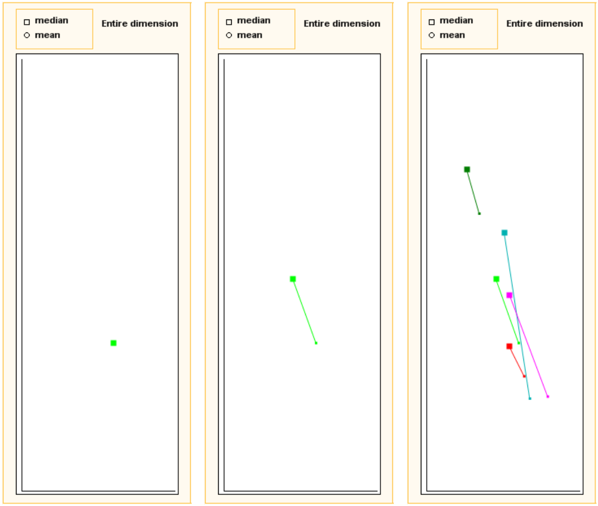


(a) Entire data set. The weight axis is selected.

(b) American cars are highlighted. The weight axis is selected.

(c) American cars are highlighted. Multiple axes are selected.

Figure 14. Parallel coordinate representations of the cars data set, showing the effect of brushing the *origin* axis and selecting multiple axes.



(a)

(b)

(c)

Figure 15. VDMD representation of statistics from the cars data set. X-axis: Coefficient of variance. Y-axis: Median. (a): Statistics of the *weight* axis when the entire data set is included. (b): Statistics of the *weight* axis after the brushing in figure 14(b). (c): Statistics of the selected axes in figure 14(c).

nates. Are American cars in general heavier than European or Japanese cars?

If we keep just the American cars, we can see in figure 14(b) that some objects are removed from the very bottom of the weight axis. This would indicate that the cars with the lowest weight are not produced in America, but we still know little about how the distribution along the axis has changed after the brushing. Figure 14 demonstrates how the parallel coordinate representation is changed when the *origin* axis is brushed.

Even if the parallel coordinates display is too cluttered for us to see changes in density along the axis, using the VDMD will still help us to draw conclusions about the data. We let the square symbol display the median of the *weight* axis, taking the whole data set into consideration. If we then keep just the American cars by brushing the *origin* axis, we will see how the median changes to a higher value, see figure 15(b). This indicates that American cars are, in general, heavier than European and Japanese cars. Looking at the zoom display of the VDMD in figure 16, we can see that it is mainly the two lower regions that have caused the change of the median. These observations would be an indication that most American cars have a higher weight than the average. If the pinpoints were set to represent all cars in the data set, we could now select any other axis to see how it has been affected by the brushing of the *origin* axis, see figure 15(c). Even if we did not initially intend to concentrate on changes to the other axes, we can now easily select arbitrary axes to see how they were affected. We can for example draw the additional conclusions that engines of American cars seem to have more horsepower and bigger displacements than the average car. From figure 14 it is hard to determine how many cars with three, four or five cylinders have been shaded. Figure 15(c) shows how the VDMD can help us with this task. The longest vector, representing the number of cylinders shows that the median has increased significantly after the brushing, indicating that American cars generally have engines with more cylinders than average.

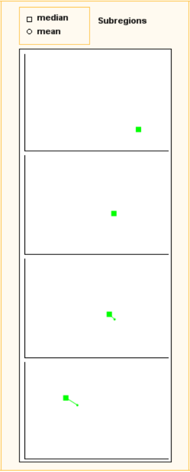


Figure 16. The zoom display of the VDMD shows which parts of the weight axis has been affected the most by the brushing in figure 14(b).

4.2 Pollution

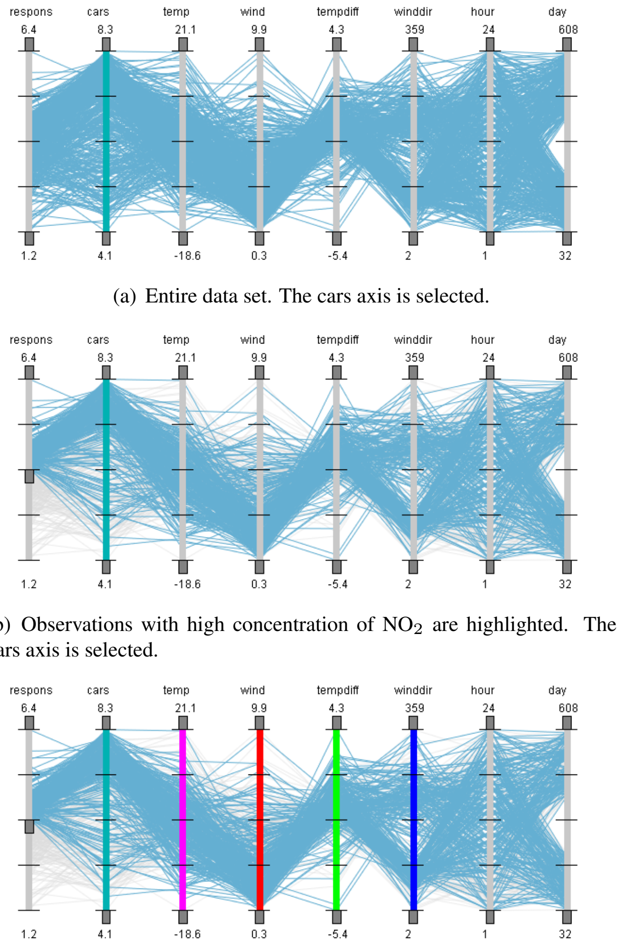
This data set originates in a norwegian study where air pollution is measured and compared with traffic volume and meteorological data. It contains 500 items in 8 dimensions.

The first column represents the logarithm of the concentration of NO_2 at each observation, and column 2 is the logarithm of the number of cars per hour. Columns 3 to 8 represent temperature, wind speed, temperature difference between 25 and 2 meters above ground, wind direction (degrees between 0 and 360), hour of the day and a continuous day number.

Using this data set, we first want to know if there is any correlation between high 'response' value of NO_2 and the number of cars passing. Figures 17(a) and 17(b) demonstrate how the parallel coordinates representation changes when we highlight the observations on the upper half of the *response* axis. We can observe that the density of observations on the lower half of the *cars* axis appears to decrease. This can be confirmed by looking at the changes of the VDMD. Figure 18(b), shows how the mean value of the *cars* axis is increased. From figure 18(c) we can see that high concentration of NO_2 also seem to be correlated with lower wind speeds and lower temperatures than the average observation.

4.3 Stocks

The last data set to be tested contains approximately 12000 entries of stocks information, from which we have no initial knowledge. The first two columns represent the company and the date. Columns 3 to 6 show the opening, highest, lowest and the closing rates, respectively. The last column represents the volume of the stock. Figure 20(a) shows the entire data set represented with parallel coordinates. We want to know if the overall closing rates have been higher or lower than average over the last time period. In figure 20(b), the entries with dates from the upper quartile are highlighted. It can be observed that entries with the highest closing rates have been removed. This would tempt us to believe that the overall closing rates have decreased. Looking closer though, we can see that many of the removed objects have had lower closing rates too. From these observations, we are not able to draw any conclusions at all.

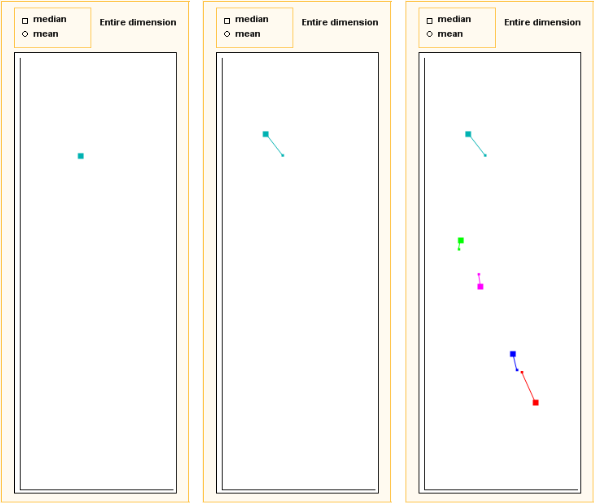


(a) Entire data set. The cars axis is selected.

(b) Observations with high concentration of NO_2 are highlighted. The cars axis is selected.

(c) Observations with high concentration of NO_2 are highlighted. Multiple axes are selected.

Figure 17. Parallel coordinate representations of the pollution data set, showing the effect of brushing the *response* axis and selecting multiple axes.



(a) (b) (c)

Figure 18. VDMD representation of statistics from the pollution data set. X-axis: Coefficient of variance. Y-axis: Median. (a): Statistics of the *cars* axis when the entire data set is included. (b): Statistics of the *cars* axis after the brushing in figure 17(b). (c): Statistics of the selected axes in figure 17(c).

Extending the parallel coordinates with the VDMD, we can observe that the overall median is in fact higher during this period (see figure 19). This is an example where a cluttered parallel coordinate representation may not be able to convey changes in the data correctly. The VDMD gives the opposite, but correct, impression of the change in the distribution.

5 Conclusions and future work

We have introduced the Visual Data Mining Display (VDMD), a method to extract and display statistics from multivariate data presented in parallel coordinates. The method is suitable for analysis of correlations and patterns in multivariate data, especially when the number of data objects is too large to be visualized in parallel coordinates without clutter. It has proved helpful in giving an overview of large data sets, as well as in observing changes in the distributions, making use of animation and vector plotting. The method has advantages both with continuous and discrete data. Large sets of continuous data may appear very cluttered in parallel coordinates, whereas multiple line segments of discrete data sets may appear as one single data item.

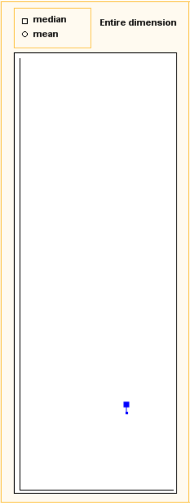
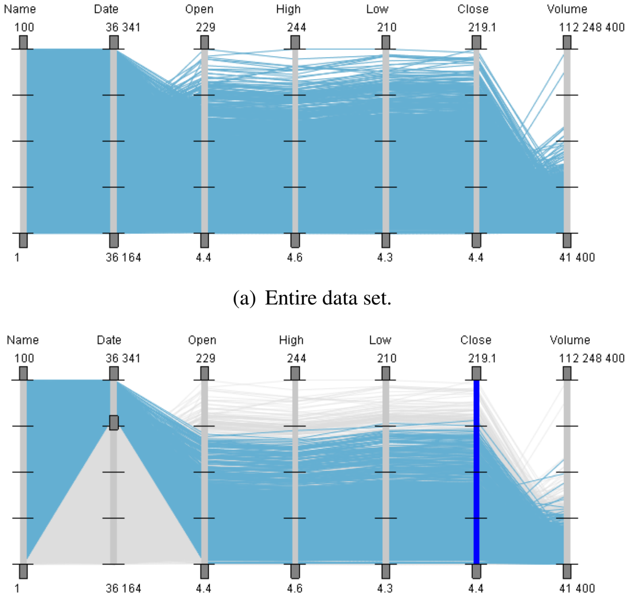


Figure 19. VDMD representation of the *closing rate* axis. The y-axis represents the median.



(a) Entire data set.

(b) Stock entries from the latest time period are highlighted.

Figure 20. Parallel coordinate representation of the stocks data set.

Our method can be used both by experienced and novice users of visual data mining, although the user needs to be aware of the nature of the statistical aggregation functions that are used. As an example, the mode works well with discrete data to display the most frequent value. With continuous data it can be very misleading, however, since the mode can appear anywhere where a few objects happen to have exactly the same value. The experienced user is assumed to be aware of such properties. Non-numeric discrete variables may not provide useful statistical information but can be included in the data to give the user the ability to select subsets of the data (such as the *origin* axes of the cars data set described above).

In the current implementation, the sub-regions of the axes are predefined and fixed so that each region corresponds to a quarter of the whole range. One future extension would be to let the user interactively set the ranges, for a more flexible level-of-detail feature.

There are currently only two kinds of glyphs available for displaying statistics in the VDMD: filled squares and circles. These can be arranged in a two-dimensional coordinate system to display two values. For experienced users, more variations in the glyph appearance could be used to enhance the amount of information in the display. Additional features of the glyphs could, for example, include variations in size and orientation, or simply further types of graphical symbols.

There are three available levels of detail in the VDMD. It might be useful to extend the number of levels, to explore data more locally or even to explore single observations.

With more computer power, the interactions could be reflected in the VDMD in a more dynamic way than in the present version. One powerful feature would be to have the symbols move in real-time in response to movement of the handles.

The VDMD could be extended to display multiple vectors for each dimension. The different vectors would then correspond to the various selections and interactions that have been made in the parallel coordinate plot. This would be useful as a journaling tool, especially if clicking the vectors resulted in the related selection being selected in the parallel coordinate plot.

Acknowledgements

This work has been funded by grant A3 02:116, supported by the Swedish Foundation of Strategic Research.

References

[1] G. Andrienko and N. Andrienko. Constructing parallel coordinates plot for problem solving. In *Proc. 1st International Symposium on Smart Graphics*, pages 9–14, 2001.

[2] G. Andrienko and N. Andrienko. Parallel coordinates for exploring properties of subsets. In *Proc. Of SmartGraphics*, 2004.

[3] N. Elmqvist and P. Tsigas. Growing squares: Animated visualization of causal relations. In *ACM Symposium on Software Visualization*, 2003.

[4] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. Technical report, Worcester Polytechnic Institute, 1999.

[5] G. Furnas. Generalized fisheye views. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 16–23, 1986.

[6] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, 2002.

[7] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proc. Of IEEE Conference on vis'90*, pages 361–378, 1990.

[8] J. Johansson, M. Cooper, and M. Jern. 3-dimensional display for clustered multi-relational parallel coordinates. In *Proceedings of the 8th IEEE International Conference on Information Visualisation (to appear)*, 2005.

[9] N. Lesh and M. Mitzenmacher. Interactive data summarization: An example application. In *AVI'04, May 25-28, Gallipoli(LE), Italy*, 2004.

[10] J. D. Mackinlay, G. G. Robertson, and S. K. Card. The perspective wall: detail and context smoothly integrated. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 173–176, 1991.

[11] A. Martin and M. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of IEEE Conference on Visualisation'95*, pages 271–278, 1995.

[12] L. J. S. N. Barlow. Animator: A tool for the animation of parallel coordinates. In *Proceedings of the Eighth International Conference on Information Visualisation (IV04)*, 2004.

[13] Openviz homepage (03/16/2005). http://www.openviz.com.

[14] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *IEEE Symposium on Information Visualization, Austin, Texas, USA*, pages 89–96, 2004.

[15] K. B. Pratt and G. Tschapek. Visualizing concept drift. In *SIGKDD'03*, 2003.

[16] M. Sarkar and M. H. Brown. Graphical fisheye views of graphs. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pages 83–91, 1992.

[17] H. Siirtola. Direct manipulation of parallel coordinates. In *IEEE International Conference on Information Visualization*, pages 373–378, 2000.

[18] H. Siirtola. Combining parallel coordinates with the reorderable matrix. In *Proceedings of The Coordinated and Multiple Views in Exploratory Visualization*, 2003.

[19] Statlib datasets archive homepage. http://lib.stat.cmu.edu/datasets/.

[20] M. Theus. Interactive data visualization using mondrian. Technical report, Department of Computeroriented Statistics and Data Analysis, Augsburg, Germany.

[21] K. Zhao, B. Liu, T. M. Tirpak, and A. Schaller. Detecting patterns of change using enhanced parallel coordinates visualization. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, 2003.