A Visual Analytics Approach for Correlation, Classification, and Regression Analysis

Chad A. Steed[a], J. Edward Swan II[c], Patrick J. Fitzpatrick[b], and T.J. Jankun-Kelly[c]

[a]Oak Ridge National Laboratory, Computational Sciences and Engineering Division, Oak Ridge, TN, USA, 37831;

[b]Northern Gulf Institute, Mississippi State University, Stennis Space Center, MS, 39529;

[c]Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS, 39762.

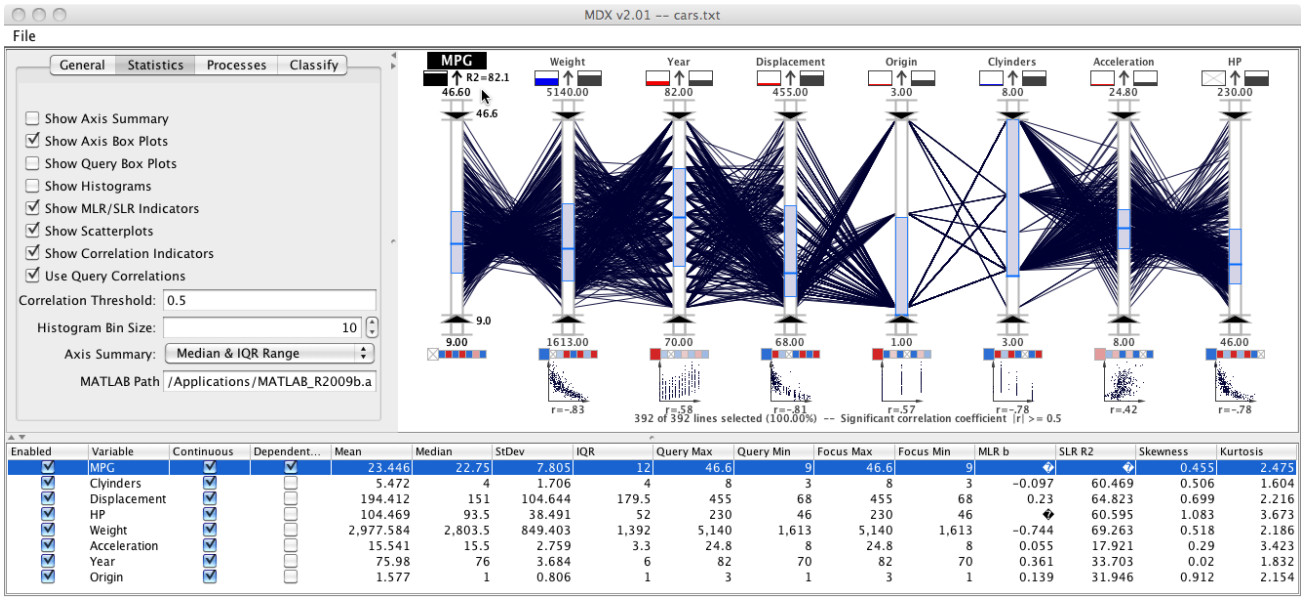
Abstract

New approaches that combine the strengths of humans and machines are necessary to equip analysts with the proper tools for exploring today's increasing complex, multivariate data sets. In this paper, a visual data mining framework, called the Multidimensional Data eXplorer (MDX), is described that addresses the challenges of today's data by combining automated statistical analytics with a highly interactive parallel coordinates based canvas. In addition to several intuitive interaction capabilities, this framework offers a rich set of graphical statistical indicators, interactive regression analysis, visual correlation mining, automated axis arrangements and filtering, and data classification techniques. The current work provides a detailed description of the system as well as a discussion of key design aspects and critical feedback from domain experts.

Keywords: visual analytics, parallel coordinates, correlation mining, classification, visualization, java

1 Introduction

A byproduct of continued technological advances is increasingly complex multivariate data sets which, in turn, yield information overload when explored with conventional visual analysis techniques. The



Enabled	Variable	Continuous	Dependent...	Mean	Median	StDev	IQR	Query Max	Query Min	Focus Max	Focus Min	MLR b	SLR R2	Skewness	Kurtosis
☑	MPG	☑	☑	23.446	22.75	7.805	12	46.6	9	46.6	9	[illegible]	[illegible]	0.455	2.475
☑	Clyinders	☑	☐	5.472	4	1.706	4	8	3	8	3	-0.097	60.469	0.506	1.604
☑	Displacement	☑	☐	194.412	151	104.644	179.5	455	68	455	68	0.23	64.823	0.699	2.216
☑	HP	☑	☐	104.469	93.5	38.491	52	230	46	230	46	[illegible]	60.595	1.083	3.673
☑	Weight	☑	☐	2,977.584	2,803.5	849.403	1,392	5,140	1,613	5,140	1,613	-0.744	69.263	0.518	2.186
☑	Acceleration	☑	☐	15.541	15.5	2.759	3.3	24.8	8	24.8	8	0.055	17.921	0.29	3.423
☑	Year	☑	☐	75.98	76	3.684	6	82	70	82	70	0.361	33.703	0.02	1.832
☑	Origin	☑	☐	1.577	1	0.806	1	3	1	3	1	0.139	31.946	0.912	2.154

Figure 1: The Multidimensional Data eXplorer (MDX) consists of a settings panel (upper left), a data table panel (bottom), and an interactive parallel coordinates panel (upper right).

ability to collect, model, and store information is growing at a much faster rate than our ability to analyze it. However, the transformation of these vast volumes of data into actionable insight is critical in many domains (e.g. climate change, cyber-security, financial analysis). Without the proper techniques, analysts are forced to reduce the problem and discard layers of information in order to fit the tools. New techniques and approaches are necessary to turn today's flood of information into opportunity.

One of the most promising solutions for this challenge lies in the continued development of techniques in the rapidly growing field of visual analytics. Visual analytics, also known as visual data mining, combines interactive visualizations with automated analytics that help the analyst discover and comprehend patterns in complicated, heterogeneous data sets. In general, visual analytics can be described as "the science of analytical reasoning facilitated by interactive visual interfaces." [Thomas and Cook, 2005]. Visual analytics seeks to combine the strengths of humans with those of machines. While methods from knowledge discovery, statistics, and mathematics drive the automated analytics, human capabilities to perceive, relate, and conclude strengthen the iterative process.

In the current work, a novel visual data mining framework–called the Multidimensional Data eXplorer (MDX)–is presented that utilizes statistical analysis and data classification techniques in an interactive multivariate representation to improve knowledge discovery in the complex multivariate

data sets that characterize today's data (see Figure 1). In addition to intuitive interaction capabilities, this framework introduces a rich set of graphical statistical indicators, automated regression analysis, visual correlation indicators, optimal arrangement techniques, and data classification algorithms. These capabilities are combined into a parallel coordinates based framework for enhanced multivariate visual analysis.

The current work features an expanded version of MDX that builds on recent efforts in which MDX was applied to tropical cyclone climate studies. In Steed et al. [2009b], the initial version of MDX, which lacked integrated statistical processes, was introduced and the system was demonstrated in a case study with a set of tropical cyclone predictors. Follow-on work [Steed et al., 2009a,c] presented an enhanced version of MDX that included statistical analytics and deeper analysis of the previously analyzed tropical cyclone predictors, as well as analysis of a new set of predictors. In the current work, the MDX visual data mining and knowledge discovery capabilities are featured. In addition to presenting new features that facilitate visual correlation mining and automated axis arrangements, the new contributions in this work are new data classification capabilities (see Section 4.4), a novel regression analysis interface (see Section 4.5) that facilitates interactive model development and confirmation, and a detailed description of the visual and automated correlation mining capabilities (see Sections 4.2 and 4.3).

The remainder of this paper is organized as follows. To begin, a survey of related work is given in Section 2. In Section 3, the cars data set–used in the examples throughout this paper–is described. In Section 4.1, the graphical indicators of descriptive statistics are presented. In Section 4.2, a discussion is provided on the interactive correlation analysis indicators and interaction features that are offered in the latest version of MDX. Then, the automated correlation analysis algorithms are described in Section 4.3 and the new automated data classification capabilities are discussed and demonstrated in Section 4.4. In Section 4.5, the details of the enhanced visual regression capabilities are described including the closing of the iterative regression analysis loop. In Section 4.6, the optimal axis arrangement capabilities are described. Then, significant findings from the development and use of MDX, visual design criteria, and domain expert feedback are given in Section 5. Finally, conclusions and future work are discussed in Section 6.

2 Related Work

As demonstrated by Wong and Bergeron [1997], there have been many approaches to the visual analysis of multivariate multidimensional data over the years. However, the techniques employed in operational systems are generally constrained to non-interactive, basic graphics using methods developed over a decade ago; and it is questionable whether these methods can cope with the complex data of today. For example, analysts often rely on simple scatter plots and histograms which require several separate plots or layered plots to study multiple attributes in a data set. However, the use of separate plots is not an ideal approach in this type of analysis due to perceptual issues described by Healey et al. [2004] such as the extremely limited memory for information that can be gained from one glance to the next. These issues are illustrated through the so-called change blindness phenomenon (a perceptual issue described by Rensink [2002]) and they are exacerbated when searching for combinations of conditions.

One approach often used by statisticians to overcome this issue is to use the scatterplot matrix (SPLOM), which represents multiple adjacent scatterplots for all the variable comparisons in a single display with a matrix configuration [Wong and Bergeron, 1997]; but the SPLOM requires a large amount of screen space and forming multivariate associations is still challenging. Wilkinson et al. [2006] used statistical measures for organizing both the SPLOM and parallel coordinates plots to guide the viewer through an exploratory analysis of high-dimensional data sets. Although the organization methods improve the analysis, the previously mentioned perceptual issues with SPLOMs remain to some degree. Another alternative is to use layered plots, which condense the information into a single display; but there are significant issues due to layer occlusion and interference as demonstrated by Healey et al. [2004].

Parallel coordinates is arguably one of the most popular multivariate visualization techniques and it is the basis of the highly interactive canvas in the MDX framework. The parallel coordinates technique was initially popularized by Inselberg [1985] as a novel approach for representing hyper-dimensional geometries, and later demonstrated in the direct analysis of multivariate relationships in data by Wegman [1990]. In general, the technique yields a compact two-dimensional representation of even large multi-dimensional data sets by representing the N-dimensional data tuple C with coordinates $(c_1, c_2, \ldots, c_N)$ by points on N parallel axes which are joined with a polyline (see Figure 2 whose N vertices are on the

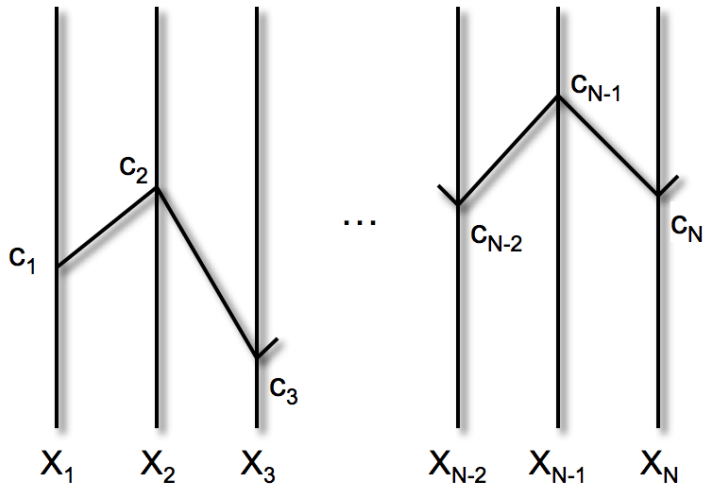


Figure 2: The polyline in parallel coordinates maps the N-dimensional data tuple C with coordinates $(c_1, c_2, \ldots, c_N)$ with points on N parallel axes which are joined with a polyline whose N vertices are on the X_i-axis for $i = 1, \ldots, N$.

X_i-axis for $i = 1, \ldots, N$ and have xy-coordinates $(i - 1, c_i)$) Inselberg [2009]. In theory, the number of attributes that can be represented in parallel coordinates is only limited by the horizontal resolution of the display device. But in a practical sense, the axes that are immediately adjacent to one another yield the most obvious information about relationships between attributes in the classical parallel coordinates plot. In order to analyze attributes that are separated by one or more attributes in the plot, intelligent interactions and graphical indicators are required. In light of this limitation, several innovative parallel coordinates extensions that improve interaction and cognition have been described in the visualization research literature since the introduction of the classical technique. For example, Hauser et al. [2002] described a histogram display, dynamic axis re-ordering, axis inversion, and some details-on-demand capabilities for parallel coordinates. In addition, Siirtola [2000] presented a rich set of dynamic interaction techniques (e.g., conjunctive queries) and Johansson et al. [2005] described new line shading schemes for parallel coordinates. Furthermore, several focus+context implementations for parallel coordinates have been introduced by Fua et al. [1999], Artero et al. [2004], Johansson et al. [2005], and Novotńy and Hauser [2006]. More recently, Qu et al. [2007] introduced a method for integrating correlation computations into a parallel coordinates display. The MDX system described in the following section utilizes variants of these extensions to the classical parallel coordinates plot.

The MDX system enhances the classical parallel coordinates axis by providing cues that guide and refine the analyst's exploration of the information space. This approach is akin to the concept of the *scented widget* described by Willett et al. [2007]. Scented widgets are graphical user interface components that are augmented with an embedded visualization to enable efficient navigation in the information space of the data items. The concept arises from the information foraging theory described by Pirolli and Card [1999] which models human information gathering to the food foraging activities of animals. In this model, the concept of information scent is identified as the "user perception of the value, cost, or access path of information sources obtained by proximal cues" [Pirolli and Card, 1999].

The scented axis widgets are also assisted by automated data mining processes that reduce the knowledge discovery timelines. In Seo and Shneiderman [2005], a framework is used to explore and comprehend multidimensional data using a powerful rank-by-feature system that guides the user and supports confirmation of discoveries. Recently, Piringer et al. [2008] expanded this rank-by-feature approach with a specific focus on comparing subsets in high-dimensional data sets. The MDX system is designed to support a similar rank-by-feature framework with subset selection capabilities using stepwise regression, correlation mining, and interactive visual analysis.

3 Cars Data Set

The 1983 ASA automobile data set[1] is used throughout the current work to illustrate the MDX capabilities. This popular data set includes 8 variables on 406 different cars and was used in the 1983 ASA Data Exposition. The variables included in this data set are MPG (miles-per-gallon), number of cylinders, engine displacement (cubic inches), horsepower, vehicle weight (lbs.), time to accelerate from 0 to 60 (sec.), model year, and origin of car (1-America, 2-European, and 3-Japanese). The cars data set contains 14 records with null values which are ignored by MDX reducing the number of records analyzed to 392.

[1]The 1983 ASA automobile data set is available from the Carnegie Mellon University StatLib website http://lib.stat.cmu.edu/datasets/.

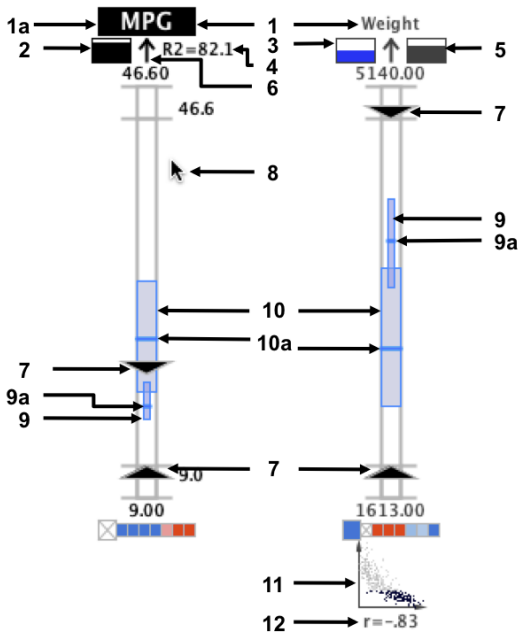


Figure 3: The scented axis widgets in the parallel coordinates display are augmented with graphical indicators of key descriptive statistical quantities, correlation measures, and regression analysis outputs. In this annotated figure, the numbered callouts highlight specific features of the axis widgets which are described in detail in the remainder of this paper.

4 Visual Data Mining and Analysis Techniques

In essence, the parallel coordinates panel in MDX (see Figure 1) is a highly interactive canvas that visually presents many multivariate associations in a manner that facilitates multifaceted exploratory data analysis. In addition to providing several fundamental capabilities such as relocatable axes, axis inversion, and details-on-demand, the MDX canvas also provides several novel visual interaction techniques such as axis scaling (focus+context), aerial perspective shading, and dynamic visual queries. These interaction capabilities are described in detail in prior publications of this ongoing research [Steed et al., 2009a,b,c].

In conjunction with these interactive visual query capabilities, MDX provides several data mining techniques that facilitate more rapid, creative, and comprehensive statistical data analysis than conventional systems. As the analyst interacts with the system, several key statistical quantities are calculated on-the-fly and mapped to visual features within the parallel coordinates display (see Figure 3) to augment the polyline configurations. The statistical indicators guide the analyst in the identification and

quantification of the key features and associations in the data set. In addition to graphical indicators of descriptive statistical quantities, the framework offers graphical indicators for correlation measures, tunable data classification methods, regression analysis, an automatic multicollinearity filter, and automatic axis arrangement capabilities. In the remainder of this section, these techniques are presented along with several evaluations of these techniques on the cars data set.

4.1 Graphical Indicators of Descriptive Statistics

By providing visual summaries of patterns and general trends in data sets, the graphical statistical indicators in MDX support visual data mining in harmony with the EDA philosophy introduced by Tukey [1977]. Each parallel coordinate axis is represented by a scented widget that includes visual representations of several key descriptive statistics. Referring to Figure 3, the median (9a and 10a), interquartile range (IQR) (9 and 10), and frequency information (see Figure 4) are calculated for the data in the focus area of each axis and presented graphically as modified box plots within the interior of the widget. Alternatively, the analyst can switch this display to use the mean and standard deviation range in the box plots.

The wide overall box plot on each axis (see 10 in Figure 3) represents the central tendency and variability for all the axis samples while the more narrow query box plots (see 9 in Figure 3)–drawn over the overall box plots–capture these statistics for only the samples that are selected with the axis query sliders (see 7 in Figure 3). Within the box plots, the thicker horizontal lines (see 9a and 10a in Figure 3) that divide the box vertically represent the median or mean value in the IQR mode or standard deviation range mode, respectively.

The axis query sliders are double-ended and can be manipulated with the mouse cursor to dynamically adjust which lines are highlighted (queried) in the parallel coordinates display. The sliders facilitate the so-called "pinching" query capability described by Inselberg [2009]. Lines that are "pinched" between the slider limits for all the axes in the display are rendered in a more prominent manner giving the user the ability to perform rapid Boolean AND selections.

These graphical central tendency and variability indicators provide a geometrical shape that indicates the typical value and how "spread out" the samples are in the distribution, respectively. For

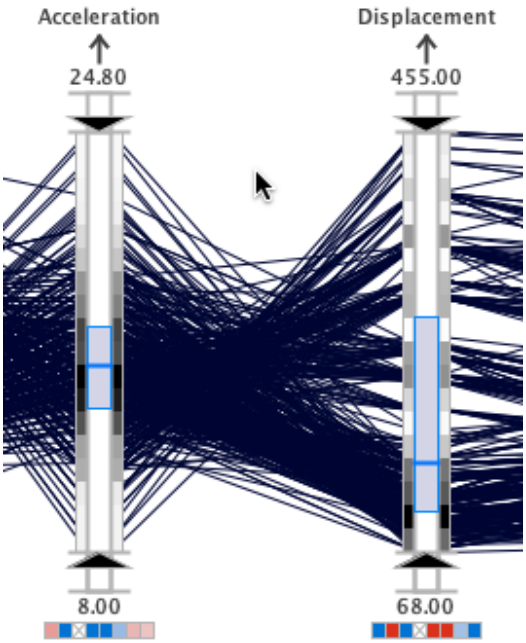


Figure 4: The values on the **Acceleration** are less dispersed than the adjacent **Displacement** axis. The dispersion of the axes can be perceived visually via a comparison of the overall box plots on each axis. Furthermore, the frequency information, which is displayed as histogram bins shaded according to the number of polylines passing through each bin region, provides a more detailed summary of the dispersiveness of each axis.

example, in Figure 4 the overall box plots on the **Acceleration** axis indicate its values are less dispersed than the adjacent **Displacement** axis. The dispersiveness of the samples for a particular axis is also shown in more detail in the histogram bins on either side of the axes that encode the frequency information with shading based on the number of polylines that pass through the bin regions. The dispersion of samples for an axis can be a key indicator of the predictability of an attribute. For this reason, these indicators are key elements in such activities as multivariate sensitivity analysis.

The query box plots provide a mechanism to compare subsets of the data with the overall tendencies in the data. In Figure 5, the records with above normal fuel economy are queried using by "pinching" the regions on the **MPG** axis. The query boxplots on the **Displacement** and **Weight** reveal that the more fuel efficient car models tend to have lower displacement and weight. This example also highlights a single car record–the queried line connected to the upper range of the **Displacement** axis box plot–with good fuel economy, but significantly higher engine displacement than the other queried records. Without effective highlighting, such anomalous records can be difficult, at best, to find in a densely packed parallel coordinates display.

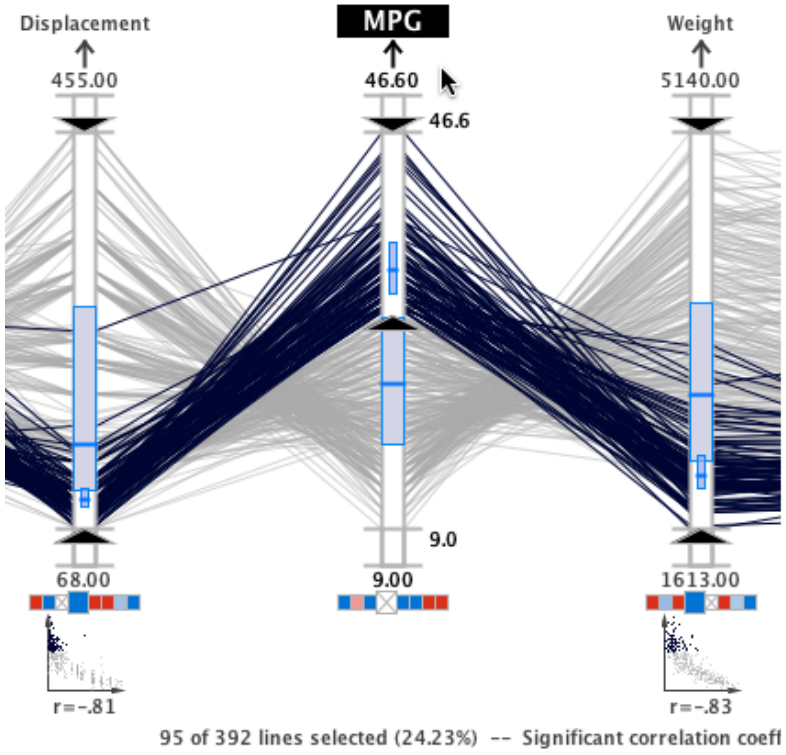


Figure 5: Using the query sliders for the **MPG** axis, the car records with above normal fuel economy are "pinched" between the sliders to highlight the polylines of interest. The wider boxplots characterize the entire set of axis values while the narrow boxplots characterize the current subset of "pinched" values. In this example, the query shows a single record with good fuel economy and significantly higher engine displacement than the other 94 records that are currently highlighted with the query sliders.

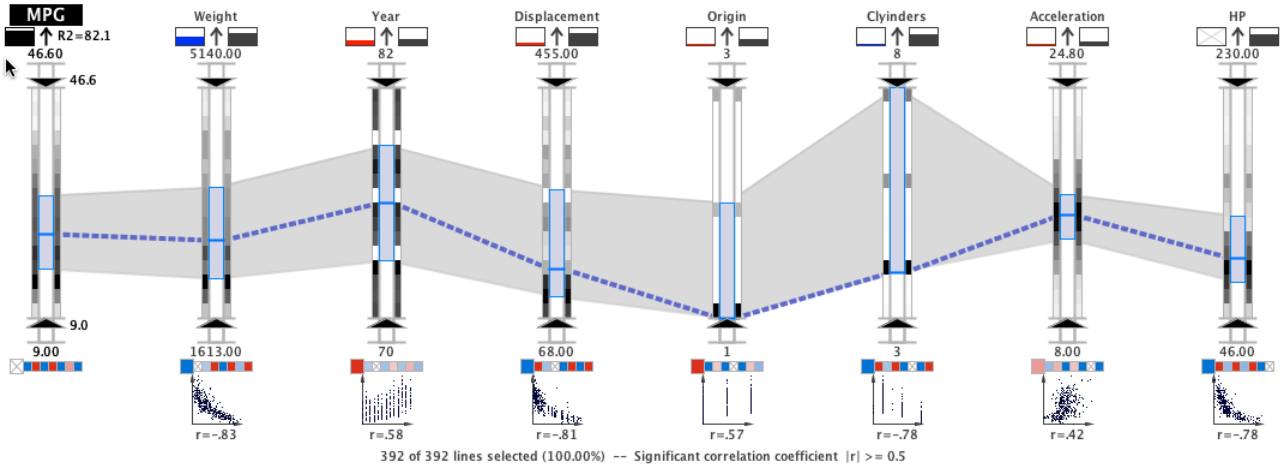


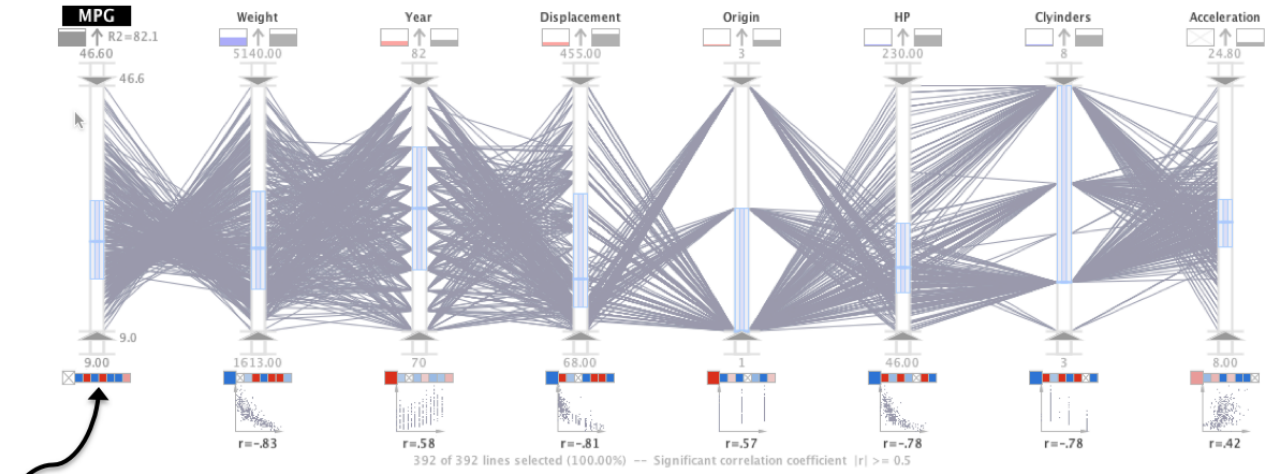
Figure 6: The user can modify the display settings to enable an axis summary feature and vertical histograms. The axis summary connects the overall central tendency and variability measures with a gray polygon connected between the axes and a blue-gray dashed line. When the data set polylines are not shown, as this figure shows, the summaries and histograms can be combined to explore general trends in large data sets without loss of interactivity.

On each axis bar interior, the frequency information can also be displayed by representing histogram bins as small rectangles surrounding the axis bar with shades that are indicative of the number of lines that pass through the bin's region (see Figure 4). That is, the darkest bins have the most lines passing through that area of the axis while the lighter bins have less lines. In addition to enabling or disabling the histogram display, the user can also fine tune the frequency display by modifying the histogram bin size in the settings panel.

As an alternative to display each individual record as a polyline in the display, the analyst can modify the display settings to represent the overall central tendency and variability measures using a gray polygon connected between the axes and a blue-gray dashed line, respectively (see Figure 6). The variability polygon is drawn beneath the other polylines in the parallel coordinates display by connecting the IQR or standard deviation range top and bottom limits between the axes. Similarly, the dashed central tendency line is drawn by connecting the median or mean values between the axes. The user can use this feature for quickly summarizing the axes during analysis. For example, if the data set is large enough to reduce interactivity with the individual polylines, the analyst can disable the drawing of all polylines and enable the axis summary to dramatically increase the rendering speed of the system. In addition, the user may enable the display of the frequency indicators to see a more detailed overview of the data. The analyst can then perform all statistical analysis processes, query subsets, and evaluate the descriptive statistics in this summary mode with interactive rendering performance in the display, even with very large data sets. When a detailed plot is desired, the individual polyline rendering can be reactivated in the settings panel.

4.2 Visual Correlation Analysis

Correlation mining is an important data mining technique due to its usefulness in identifying underlying dependencies between variables. The correlation mining process attempts to estimate the strength of relationships between pairs of variables to facilitate the prediction of one variable based on what is known about another. The relationship between two variables X and Y can be estimated using a single number, r, that is called the sample correlation coefficient [Walpole and Myers, 1993]. MDX uses the Pearson product-moment correlation coefficient to measure the correlation between the axes visible in



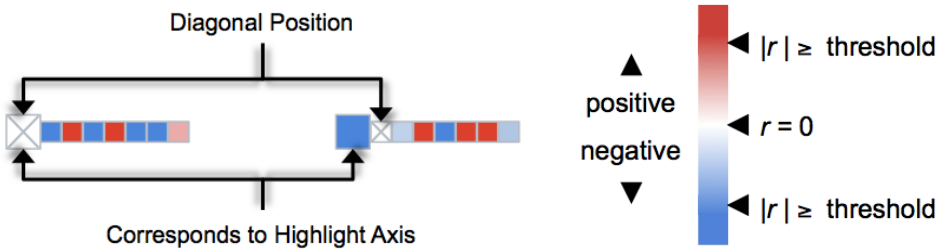
	MPG	Weight	Year	Displacement	Origin	HP	Cylinders	Acceleration
MPG	1.0	-.83	.58	-.81	.57	-.78	-.78	42
Weight	-.83	1.0	-.31	.93	-.59	.86	.90	-.42
Year	.58	-.31	1.0	-.37	.18	-.42	-.35	.29
Displacement	-.81	.93	-.37	1.0	-.61	.90	.95	-.54
Origin	.57	-.59	.18	-.61	1.0	-.46	-.57	.21
HP	-.78	.86	-.42	.90	-.46	1.0	.84	-.69
Cylinders	-.78	.90	-.35	.95	-.57	.84	1.0	-.50
Acceleration	.42	-.42	.29	-.54	.21	-.69	-.50	1.0

Figure 7: The graphical correlation indicator blocks that are displayed beneath each axis in the parallel coordinates plot are color-filled representations of the correlation matrix.

the parallel coordinates panel. Given a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \ldots, n$, r is given by

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum x_i{}^2 - (\sum x_i)^2][n \sum y_i{}^2 - (\sum y_i)^2]}}. \quad (1)$$

For each pair of axes in the display, our system computes r which results in a correlation matrix. The correlation matrix is a $n \times n$ matrix where each i, j element is equal to the value of r between the i and j variables. As shown in Figure 7, the rows from this correlation matrix are exploded and displayed graphically beneath each axis as a series of color-filled blocks. The colors used to fill the blocks are calculated based on the value of r between the axis directly above it and the axis that corresponds to its position in the set of blocks for the particular axis. For example, the first block in the correlation indicators under each axis in Figure 7 represents the correlation strength between the axis above it and the first axis, **MPG**.



(a) Correlation Indicator Blocks (b) Correlation Color Scale

Figure 8: The correlation indicator blocks that correspond to the diagonal elements are the correlation of the axis with itself – a perfect relationship. In (a), these blocks are shaded white and marked with an 'X' symbol. The enlarged blocks indicate the bivariate correlations between the highlighted axis and the other axes. The color scale in (b) is used to shade the blocks red for positive and blue for negative. A saturation scale is applied to encode the strength of the correlations such that correlations above the $r_{threshold}$ value are shaded with the most saturated colors.

The color of each indicator block is calculated using the color scale shown in Figure 8(b) which results in shades of blue for negative correlations and red for positive correlations. The color scale maps the saturation of the color to the strength of r so that the strongest correlations are displayed more prominently. An axis' r value with itself (the diagonal element) is always equal to one and the corresponding indicator block is shaded white with a gray 'X' symbol (see Figure 8(a)). Moreover, when the absolute value of r is greater than or equal to the user-defined significant correlation threshold, $r_{threshold}$, the block is shaded with the fully saturated color (either red or blue). The current value of $r_{threshold}$ is displayed at the bottom of the parallel coordinates plot (see Figure 6) and this value can be adjusted via the settings panel.

When the mouse cursor (see 8 in Figure 3) hovers over an axis in the parallel coordinates panel–the mouse cursor is hovering over the **HP** axis in Figure 9–the axis label (see 1 in Figure 3) is enlarged and the correlation coefficient blocks corresponding to it below the other axes are enlarged (see Figure 8(a)). This focus+context effect helps the user to ascertain the correlation of the highlighted axis with all other axes, at a glance. At the same time, the display shows the full correlation matrix for all pairwise combinations of the axes in the display thereby yielding the correlation context.

In addition to graphical representations of r, the system also displays small scatterplots (see 11 in Figure 3) below the axis correlation indicators blocks when an axis is highlighted with the mouse cursor. For example, in Figure 9 the **MPG** axis is highlighted. These scatterplots are created by plotting the

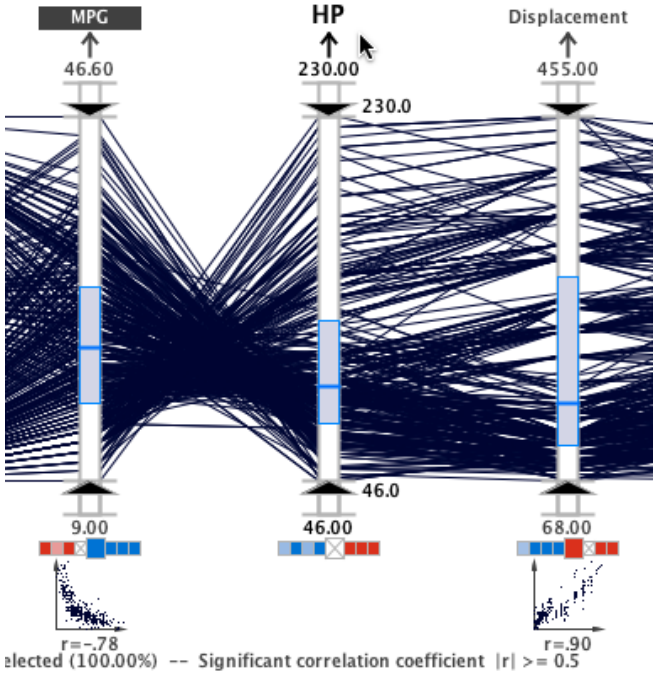


Figure 9: The distinct 'X' shaped polylines crossings between the **MPG** and **HP** axes are characteristic of strong negative correlations. The more horizontal polylines between the **HP** and **Displacement** axes are characteristic of strong positive correlations. Visual correlation mining is facilitated via the parallel coordinate polyline configurations, scatterplots, correlation indicator block colors, and the numeric display of r.

points from the highlighted variable along the y axis and the variable directly above the scatterplot along the x axis. Each scatterplot also shows the numerical r value associated with this pair of axes below the scatterplot (see 12 in Figure 3). The scatterplots in MDX provide a visual mechanism to quickly confirm the type of correlation (positive or negative) as well as the strength of the correlation.

The type of correlation is also visually detectable in the polyline configurations of the parallel coordinates plot. As shown in Figure 9, the parallel coordinates polylines between the **MPG** and **HP** axes cross in an 'X' pattern which is characteristic of a negative correlation. The negative correlation is reinforced by the slope in the scatterplot, the color of the correlation indicator blocks, and the r value display. On the other hand, the polylines between the **HP** and **Displacement** axes appear more horizontal and parallel to one another which indicates a positive correlation. Since the visual patterns for the negative correlations tend to dominate the parallel coordinates display, the user can invert an axis by clicking on the arrow beneath each axis label (see 6 in Figure 3).

Unlike the other correlation indicators, the scatterplot is useful for exploring nonlinear relationships between variables. For example, a nonlinear relationship can be observed in a scatterplot even if the correlation coefficient is zero. In Figure 9, a nonlinear relationship is revealed in the scatterplot showing the **MPG** and **HP** axes. However, the nonlinearity is not apparent in the parallel coordinate polyline configurations. In Figure 6, the scatterplots beneath the **Weight**, **Displacement**, and **HP** axes reveal nonlinear relationships with the highlighted axis, **MPG**.

4.3 Automated Correlation Analysis

The MDX system provides an automatic multicollinearity filter (see Algorithm 1) to ensure the proper selection of axes in subsequent multiple linear regression analysis. This filter examines the visible axes in the parallel coordinates display for multicollinearity[2]; if any axes are correlated with each other by more than the significant correlation threshold, $r_{threshold}$, one axis is removed from the display (see line 11). The filter removes the axis that has a lower r with the dependent axis. In this way, the remaining independent axes are truly independent of each other. The analyst can tune the multicollinearity filter by changing the value of $r_{threshold}$.

The user can reduce multicollinearity manually by using the correlation indicators to identify and filter correlated axes using a predetermined threshold; but the filter provides an automatic way to ensure independence and it can be performed at the click of a button. Removing the strongly correlated independent axes will ultimately improve subsequent regression analysis by avoiding over-fitting the data. Although the filtered axes are removed, they can be re-inserted in the display using the checkbox in the **Visible** column of the table view (see bottom panel in Figure 1).

4.4 Automated Data Classification

Data classification transforms raw data into into classes or groups. Data classification can be useful to help discriminate from many differing elements in displays. MDX provides four algorithms from the

[2]When an independent variable is highly correlated with several other independent variables, the variable has multicollinearity. The variable has much in common with the other variables and may have little information unique to itself.

Algorithm 1 Multicollinearity Filter

Input: Significant correlation threshold, $r_{threshold}$
Input: Array of Axis objects, $axes$
Input: Single dependent axis, $axis_{dependent}$
Output: Truly independent set of axes in display

```
    // Descending sort by r of each axis with axis_dependent
 1: axes_sorted ← SORT(axes)
 2: for Axis object axis ∈ axes_sorted do
 3:   for Axis object axis_compare ∈ axes do
 4:     if axis_compare = axis_dependent then
 5:       continue
 6:     else if axis_compare = axis then
 7:       continue
 8:     else
 9:       r ← CORRELATION(axis, axis_compare)
10:       if r > r_threshold then
11:         remove axis_compare from display
12:       end if
13:     end if
14:   end for
15: end for
```

GeoVista [3] library for classifying the data based on a single attribute (axis): Equal intervals, quantiles, mean-standard deviation, and Jenks' optimal.

With the equal interval classification method, each class occupies an equal interval along the selected classification axis. Although simple to compute and easy to interpret, a major disadvantage of the equal intervals approach is that the class limits do not take into consideration the data distribution along the classification axis. The quantile classification method is also simple to compute, but results in the same percentages of observations per class. With the quantiles method, data are rank-ordered with equal numbers of observations placed in each class and the 50th percentile is logically associated with the classes. Like the equal intervals method, the quantiles does not consider how the data are distributed along the classification axis. By contrast, the mean-standard deviation method does consider how the data are distributed along the classification axis; but it only works well for data that are normally distributed. If the data are normally distributed (or near normal), the mean serves as a good dividing point to facilitate a contrast of values above and below it. The Jenks' optimal

[3]The GeoVista library is a product of the Penn State Department of Geography. The library and its source code and are available online at http://www.geovista.psu.edu/geoviztoolkit/index.html

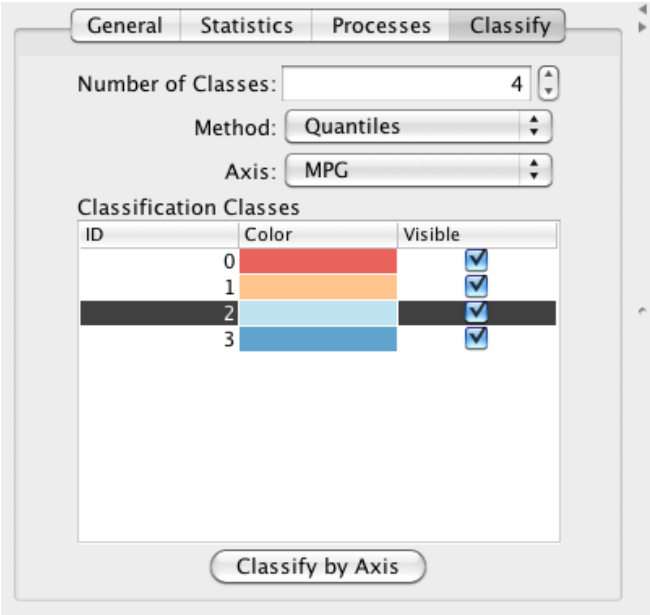
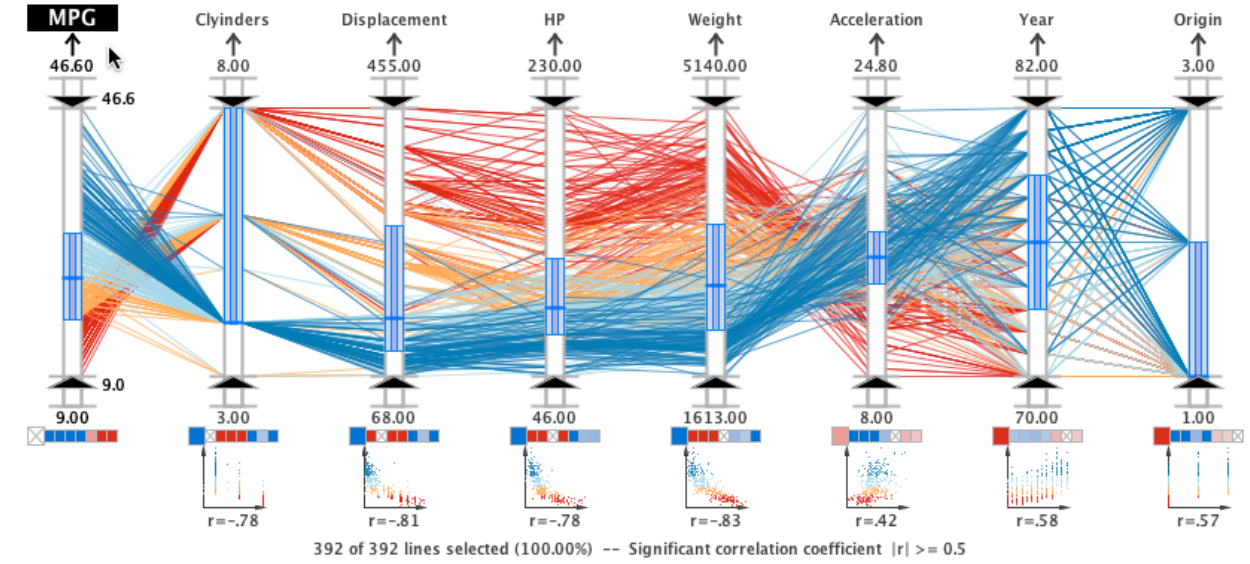


Figure 10: MDX features an automated data classification capability that transforms raw data into classes. The data classification settings panel provides the user with control over the class count, classification method, and classification axis. In addition, the panel displays the information about the individual classes in the panel. In this example, the **Quantiles** classification method was executed on the **MPG** axis to produce 4 classes.

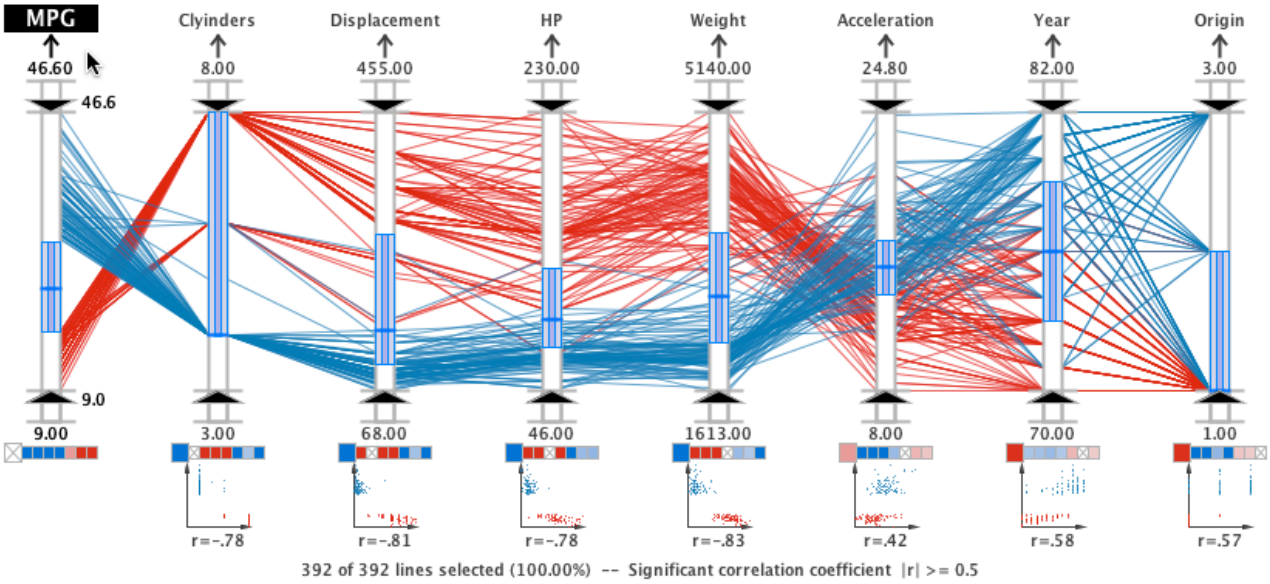
classification method places similar data values in the same class. Although more difficult to interpret, this optimal classification method is a good choice when the intention is to place like values in the same class. There are many criteria to consider in selecting the most suitable classification method for a given data set and the reader is referred to the detailed overview of these classification methods and criteria for selecting the "best" method given by Slocum et al. [2009].

Within MDX, the data classification features are controlled by the fields in the **Classify** tab of the **Settings** panel (see Fig. 10). Within this panel, the analyst can adjust the number of classes that will be produced, the classification algorithm to execute, and the classification axis. When the analyst executes the classification algorithm by clicking on the **Classify by Axis** button, the system will start the classification algorithm and populate the **Classification Classes** table in the **Classify** tab with the information about each class (ID, color, and visibility in the parallel coordinates panel).

In Figure 11(a), the quantile classification method has been applied to the **MPG** axis to create 4 classes. For the two extreme classes, class 0 captures the most fuel efficient car models and class 3 captures the least fuel efficient car models. In addition, the two middle classes (classes 1 and 2) capture



(a) All classes



(b) Extreme classes

Figure 11: The **Quantiles** classification method was executed on the **MPG** axis to produce 4 classes which are indicated by colors in (a). The color legend is shown in Figure 10. The **Visible** column in the **Classification Classes** table of the **Classify** panel gives the user the ability to interactively determine when class polylines are shown in the parallel coordinates plot. This feature provides the ability to perform rapid characterizations and comparisons of a class or a set of classes. In (b), the classes from the upper and lower range of the **MPG** axis are shown in isolation to reveal the patterns for high and low fuel economy, respectively.

the cars with average fuel economy.

The checkbox in the **Visible** column of the class table gives the user control over which polylines are shown in the parallel coordinate plot. Whereas Figure 11(a) shows all the classes, the two middle classes are hidden in Figure 11(b) to facilitate a direct comparison of the extreme upper and lower classes. As a result, only the polylines for class 0 (containing the polylines in the upper range of the **MPG** axis) and 1 (containing the polylines in the lower range of the **MPG** axis) are shown. The resulting visual query reveals that the most fuel efficient car models are those with the lower number of cylinders, displacement, horsepower, and weight. Furthermore, the query also shows that the most fuel efficient car models are generally slower to accelerate, produced in all three countries of origin, and are more common in recent year models. Meanwhile, the class containing the least fuel efficient models have a higher number of cylinders, displacement, horsepower, and weight. As one might expect, acceleration is better in this class. In addition, this class of cars are mostly from the older model years and predominately originate from America (country 1).

The classification capabilities provides an automated way to group similar elements for values on a particular axis and the interaction lets the analyst investigate patterns in class polylines intuitively. The same visual queries facilitated by the classification features can be produced manually using the "pinch" query sliders on the axes. However, the analysis of more than one class of information in a single plot is not possible using the "pinch" query alone. But with the classification method and related interaction capabilities, the user has the ability to rapidly highlight and analyze subsets of the data set in an efficient and iterative manner.

4.5 Visual Regression Analysis

Regression analysis is often exploited to identify the most relevant relationships in a particular data set. Such techniques are effective for providing quantitative associations and obtaining an adequate and interpretable description of how a set of predictors affect the dependent variable in a system. In addition to simple linear regression, MDX offers stepwise multiple linear regression with a backwards glance which selects the optimum number of the most important variables using a predefined significance level [Walpole and Myers, 1993].

The MDX MLR analysis includes a normalization procedure so that the y-intercept becomes zero and the importance of a predictor can be assessed by comparing regression coefficients, b_i, between different predictors. Denoting σ as the standard deviation of a variable, y as the dependent variable, $\overline{x}$ as the predictor mean, and $\overline{y}$ as the dependent variable mean, a number k of statistically significant predictors are normalized by the following equation:

$$(y - \overline{y})/\sigma_y = \sum_{i=1}^{k} b_i(x_i - \overline{x}_i)/\sigma_i. \tag{2}$$

The interactive visual analysis features in MDX complement the stepwise regression capabilities by screening and isolating the significant variables in a quantitative fashion. As illustrated in Figure 12, MDX executes a MATLAB®[4] process and captures output from the MATLAB® *regress* and *stepwisefit* utilities that perform simple and stepwise regression, respectively. The MATLAB® textual output stream is then parsed and relevant statistical information is extracted and represented graphically within the parallel coordinates display.

Referring to 3 in Figure 3, the system graphically encodes b in the parallel coordinates panel using the box that is below the axis label and to the left of the arrow. Like a thermometer, the box is filled from the bottom to the top based on the magnitude of b. The box is colored red if the coefficient is positive and blue if it is negative. The box to the right of the arrow, 5 in Figure 3, encodes the r^2 output from the SLR process. In addition to the coefficients, the MLR analysis returns an overall R^2 value which provides a quantitative indication of how well the model captures the variance between the predictors and the dependent variable. Referring to 2 in Figure 3, the box beneath the dependent variable axis name, 1a in Figure 3, encodes the overall R^2 value from the MLR analysis. The R^2 value is also presented numerically (see 4 in Figure 3).

When these boxes are filled with a light gray 'X', the value is not defined (the SLR or MLR process has not been executed) or, in the case of the MLR analysis, the variable was excluded during the selection process. It is also important to note that the axis corresponding to the dependent variable is indicated by light gray text on a dark gray box for its title (see 1a in Figure 3)–the reverse shading of the other axes.

[4]MATLAB® is an environment and language for mathematical analysis. More information can be found on this product at http://www.mathworks.com.

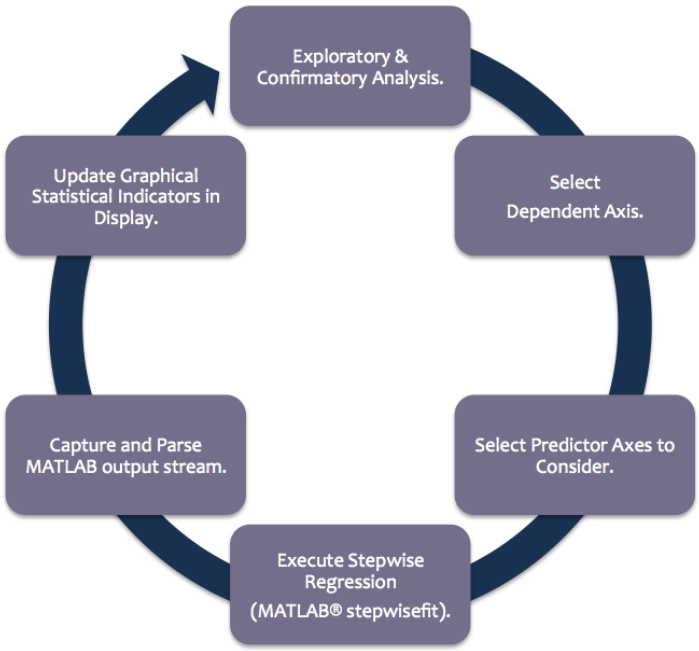
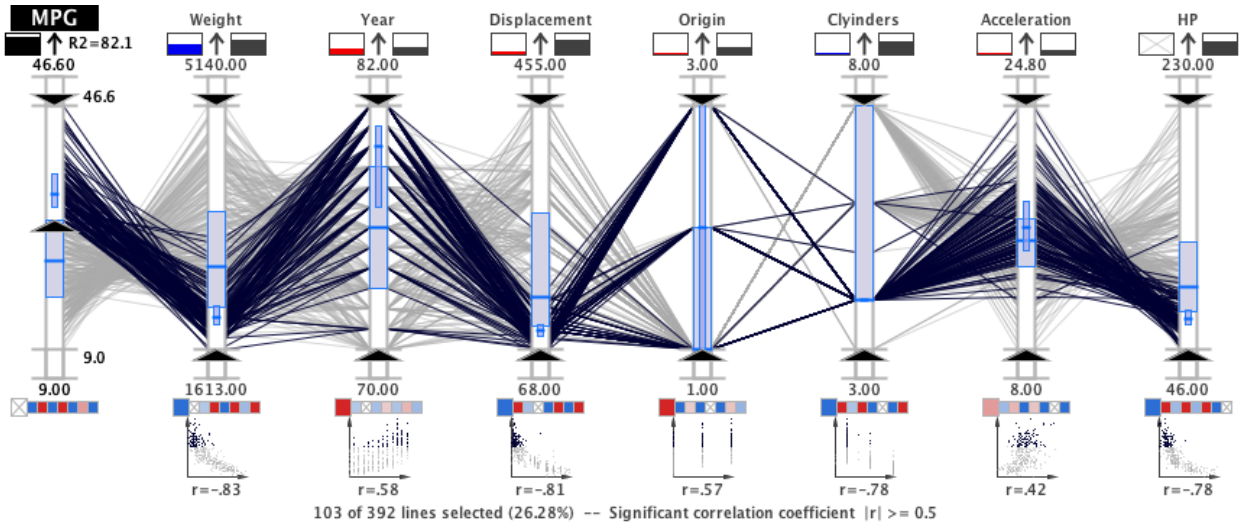


Figure 12: MDX features an interactive stepwise multiple regression analysis capability that allows the user to choose or exclude variables, execute the automated analysis, and examine the results in the augmented parallel coordinates display. The regression capability represents an iterative loop designed to reveal the most significant parameters for the chosen dependent axis.

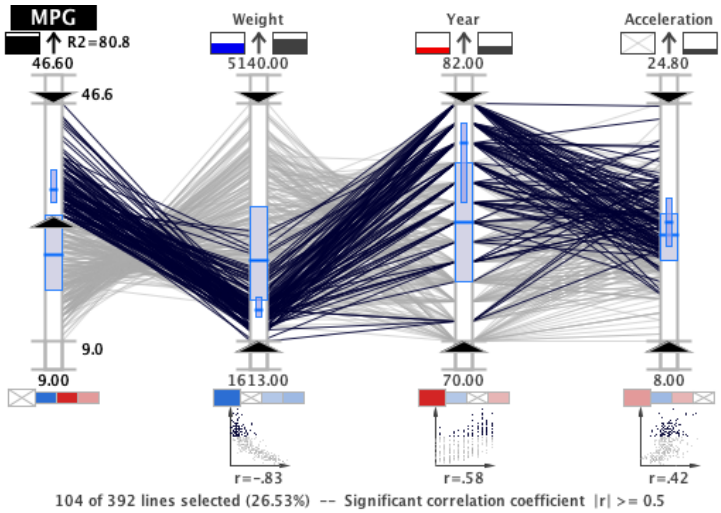
As shown in Figure 12, the stepwise regression process can be represented as an interactive feedback loop in which the analyst can designate the dependent axis, choose which axes to consider in the regression model, execute the regression analysis via MATLAB, and visually examine the results in the augmented parallel coordinates display. Combining the strengths of automated analytics with human intuition, creativity, and flexibility, this approach provides an effective means to discover the most significant set of predictors.

Figure 13(a) shows the resulting model generated by the MDX stepwise regression analysis for the cars data set using **MPG** as the dependent axis. In this example, the axes are sorted in descending order based on the resulting value of b. In addition to the graphical indicators of b, the SLR r^2 graphical indicators convey information about the significance of each variable with respect to the dependent axis. The plot arrangement shows the **Weight** axis is the most significant axis based on the stepwise regression analysis. Furthermore, the "pinch" sliders are used in this example to highlight the most fuel efficient records.

The high percentage of highly saturated colors in the correlation indicator blocks beneath each axis



(a) Regression model without multicollinearity filtering.



(b) Regression model with multicollinearity filtering.

Figure 13: With the **MPG** axis designated as the dependent axis, the regression model shown in (a) was produced by MDX. The high number of highly saturated correlation indicator blocks in this plot indicator the high number of independent variables highly correlated with one another. To reduce this condition, the MDX multicollinearity filter is executed and the regression process is repeated. The output of the new regression (b) provides an adequate model with a fewer number of independent variables than the first attempt (a).

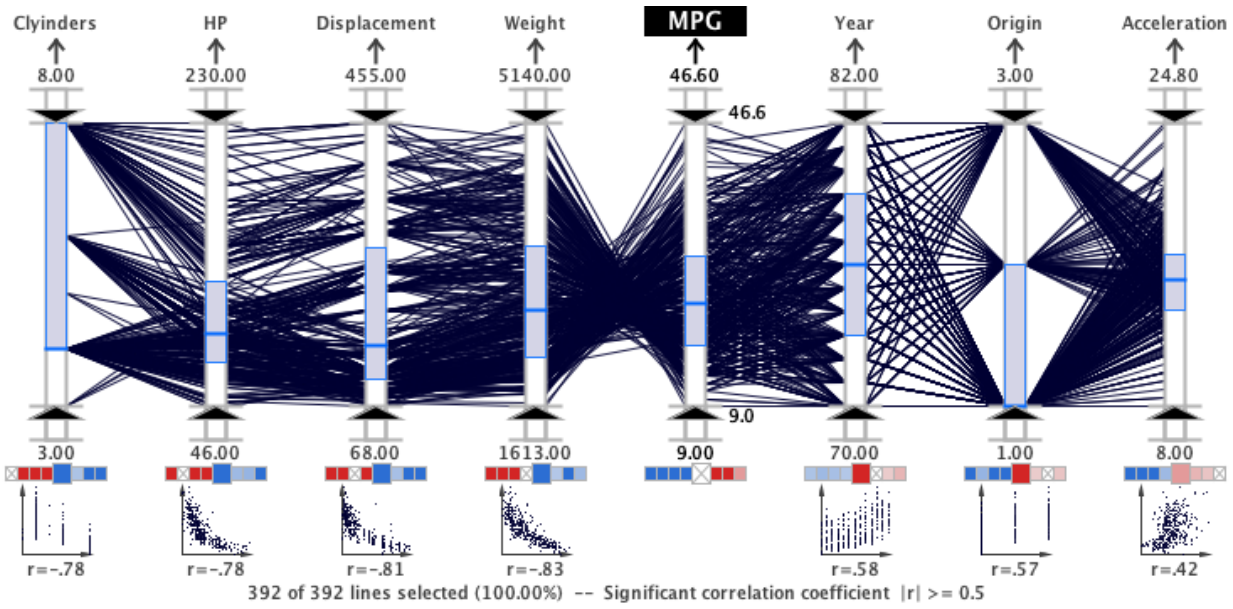


Figure 14: MDX can automatically arrange the display axes according to the correlation coefficients, r, with a particular axis. In this example, the axes are arranged based on the correlation with the **MPG** axis. Axes with negative correlations are arranged on the left of the highlighted axis and positively correlated axes are arranged on the right.

in Figure 13(a) reveal that several of independent variables are highly correlated with one another. This multicollinearity condition should be addressed prior to the regression process execution. Using the MDX automated multicollinearity filter described in Section 4.3, all except the **Weight**, **Year**, and **Acceleration** axes are removed from consideration by the next execution of the regression process. In Figure 13(b), the new regression model is shown. In this model, the **Weight** and **Year** axes were retained but the **Acceleration** axis was not chosen by the regression process. Although the overall R^2 for the model dropped slightly from the non-filtered version, the new model provides an adequate model with a fewer number of independent variables. Reducing the number of predictors is helpful to exploratory analysis because it simplifies interpretation and it usually means cheaper data collection and analysis.

Furthermore, the small boxplot for the **Weight** axis in Figure 13(b) reflects the tight clustering of polylines, which are mostly below the median value. These characteristics are indicators that **Weight** can be used to effectively predict the fuel efficiency of a car. If the lines were mostly dispersed, the small boxplot would be taller and reflect a condition where by the analyst may have difficulty using the predictor to predict the dependent variable. The analyst can continue to utilize the interactive

interface to conduct confirmatory analysis of the resulting regression models and, optionally, iterate to produce new regression models.

4.6 Optimal Axis Arrangement

In the classic parallel coordinates plot, adjacent axes reveal the most information about one other. The correlation indicators and graphical statistical indicators provide one viable way to reveal information between all axes, simultaneously, regardless of the current axis locations. MDX also provides a set of optimal axis arrangement schemes that automatically arrange the axes in the parallel coordinates panel using one of the following precomputed statistical measures:

- Correlation coefficient (r)
- IQR / standard deviation range
- MLR coefficient (b)
- SLR (r^2) value

This capability facilitates more rapid statistical comparisons between the displayed axes. The analyst can execute the arrangement process using the **Process** tab in the settings panel or through the pop-up menu that is displayed when the user right clicks in the parallel coordinates panel.

When the axes are sorted by r, one axis is selected initially as the target axis. The axes are then sorted according to the r value of the target axis and the other visible axes. As shown in Figure 14, the axes with negative correlations are arranged to the left of the target axis in ascending order. Similarly, the axes with positive correlations are arranged to the right of the target axis in descending order. The strongest correlations are placed nearest to the target axis while the weakest correlations are placed farthest away. When the axes are sorted in this manner, the analyst can quickly identify the strongest correlations with the target axis and engage in more effective correlation analysis.

The IQR / standard deviation range, b, and r^2 arrangement options all sort the axes in descending order based on the respective statistical measures. The dependent axis is placed at the leftmost position

and the other axes are arranged accordingly to the right of it. The IQR / standard deviation range arrangement is useful for examining the dispersion characteristics of each axis. The r^2 arrangement is useful as an alternate method for observing the individual correlation of axes with the dependent axis. The b arrangement (see Figure 13(a)) helps to analyze the stepwise regression model results and quantify the most significant axes for the dependent axis.

5 Discussion

In traditional data analysis tools, a collection of separate visualization and analysis tools are usually employed. Furthermore, the visualizations are often comprised of static techniques developed more than a decade ago; and it is questionable whether these techniques can meet the demands of today's data. Consequently, the analyst is afforded limited interactivity with the data, thereby hindering the discovery of new hypotheses. By integrating statistical and visualization processes, MDX gives the analyst rapid, visual query capabilities for faster and more creative knowledge discovery. In case studies that utilized MDX to conduct exploratory climate analysis [Steed et al., 2009a,b,c], MDX was compared to more traditional, static systems. Whereas the more traditional process took days to reach conclusions, analysis with MDX required hours. Perhaps the greatest evidence of the promise of the visual analytics approach came during these climate case studies from a weather science expert, Dr. Patrick Fitzpatrick, who, in addition to authoring several articles [Fitzpatrick, 1997] and books [Fitzpatrick, 1999] on hurricane climate studies, is also a co-author of this paper. Dr. Fitzpatrick indicated that the MDX system facilitated more rapid and comprehensive analysis and validation than traditional static analysis techniques.

The utilization of coordinated multiple views (CMV) in MDX helps the viewer conduct more creative exploratory analysis by offering multiple views of the data where actions in one view are propagated to others according to some visual effect. For example, non-linear relationships are more difficult to discover in parallel coordinates, but straightforward to identify in a scatterplot. On the other hand, the number of variables that can be displayed in a scatterplot is generally restricted to two or three dimensions. Moreover, it is difficult to decipher correlations between axes in all but the extreme cases in the parallel coordinates plot, but the scatterplot is more useful for more subtle cases that are often

encountered in real-world data. Having both the parallel coordinates plot and the scatterplots in MDX gives the analyst access to both views in a complementary fashion, which offsets said deficiencies. Furthermore, the inclusion of parallel coordinates plot in new areas such as climate analysis forces the analyst to consider the data in new ways, which often encourages fresh insight.

As discussed in Section 4.2, extreme negative and positive correlations can be detected by characteristic visual patterns. However, more subtle correlations are not as easily detected and it is impossible to grasp the correlations between all pairs of axes using classical parallel coordinates. With the graphical correlation indicator blocks, the more subtle correlations are conveyed directly using a carefully designed saturation color scale. Also, these correlation indicator blocks capture a holistic overview of all correlations between pairs of axes by exploding the correlation matrix. As Shneiderman notes [Shneiderman, 1996], providing an overview helps the analyst build a mental model of how the data covers the attribute space. In turn, the model helps the analyst formulate new queries on the data [Plaisant et al., 1999]. These linked views provide the level of interactivity and coordination necessary to cope with today's complex, multivariate data.

A significant amount of time has been devoted to formulating an optimal color scheme and layout for the MDX interface. The color scheme and layout is based upon color design principles from fine art and graphic design [Itten, 1970], as well as empirical perceptual studies [Ware, 2004]. For example, muted colors are used in most of the graphical elements reserving the most saturated colors for small portions of the display. This creates a visual balance that is aesthetically pleasing to the viewer. Furthermore, the most vivid colors are placed on the peripheral of the display to further balance the view. The color-coded correlation blocks described in Section 4.2 are a good illustration of the significance of a well-planned color design. The saturation scale directs visual attention to the strongest correlations and the blue and red shades cue the analyst to sign of the correlation. When planned intelligently, the overall color scheme of the application will greatly improve the user experience by reducing fatigue and making important relationships stand out to the viewer. The color scheme can also improve the viewer's confidence in the software's capabilities, at least initially, which is crucial to efficient communication of results.

6 Conclusions

An enhanced version of the MDX framework has been described in this work that offers new visual correlation mining, interactive regression analysis, and interactive, semi-automated data classification. Several illustrations of this new framework have been demonstrated with the popular ASA cars data set to highlight how the approach enhances knowledge discovery in multivariate data sets.

In the future, the MDX system will be expanded to explore additional data sets. The system has already been evaluated with several complex tropical cyclone and oceanographic data sets. In addition to additional climate data sets, new methods for transforming unstructured data into insightful representations within MDX are being investigated now. The MDX system has shown significant promise in several practical evaluations. The results of these evaluations provide compelling evidence that visual data mining solutions can meet the complex challenges of deciphering actionable knowledge from today's increasingly complicated data.

Acknowledgements

This research is sponsored by the U.S. Department of Energy, the U.S. Naval Research Laboratory, the National Oceanographic and Atmospheric Administration (NOAA) with grants NA060AR4600181 and NA050AR4601145, and through the Northern Gulf Institute that is funded by grant NA06OAR4320264. This paper was prepared by the Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT–Battelle, LLC, for the U.S. Department of Energy, under contract DE-AC05-00OR22725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz. Uncovering clusters in crowded parallel coordinates visualization. In *IEEE Symposium on Information Visualization*, pages 81–88, Oct. 2004.

P. J. Fitzpatrick. Understanding and forecasting tropical cyclone intensity change with the Typhoon Intensity Prediction Scheme (TIPS). *Weather and Forecasting*, 12(4):826–846, 1997.

P. J. Fitzpatrick. *Natural Disasters, Hurricanes: A Reference Handbook*. ABC–CLIO, Santa Barbara, CA, 1999.

Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of IEEE Visualization*, pages 43–50, Oct. 1999.

H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of IEEE Symposium on Information Visualization*, pages 127–130, Oct. 2002.

C. G. Healey, L. Tateosian, J. T. Enns, and M. Remple. Perceptually-based brush strokes for nonphotorealistic visualization. *ACM Transactions on Graphics*, 23(1):64–96, 2004.

A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.

A. Inselberg. Parallel coordinates: Interactive visualization for high dimensions. In E. Zudilova-Seinstra, T. Adriaansen, and R. Liere, editors, *Trends in Interactive Visualization*, pages 49–78. Springer-Verlag, London, UK, 2009.

J. Itten. *The Elements of Color*. Van Nostrand Reinhold Publishing, Ravensburg, Germany, 1970.

J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *IEEE Symposium on Information Visualization*, pages 125–132, Oct. 2005.

M. Novotńy and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900, 2006.

H. Piringer, W. Berger, and H. Hauser. Quantifying and comparing features in high-dimensional datasets. In *International Conference on Information Visualization*, pages 240–245, London, UK, Jul. 2008. IEEE Computer Society.

P. Pirolli and S. K. Card. Information foraging. *Psychological Review*, 106(4):643–675, 1999.

C. Plaisant, B. Shneiderman, K. Doan, and T. Bruns. Interface and data architecture for query preview in networked information systems. *ACM Transactions Information Systems*, 17(3):320–341, July 1999.

H. Qu, W. Chan, A. Xu, K. Chung, K. Lau, and P. Guo. Visual analysis of the air pollution problem in Hong Kong. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1408–1415, 2007.

R. A. Rensink. Change detection. *Annual Review of Psychology*, 53:245–577, 2002.

J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, Boulder, CO, Sep. 1996. IEEE Computer Society.

H. Siirtola. Direct manipulation of parallel coordinates. In *Proceedings of the International Conference on Information Visualisation*, pages 373–378, London, England, 2000. IEEE Computer Society.

T. A. Slocum, Robert B. McMaster, F. C. Kessler, and H. H. Howard. *Thematic Cartography and Geovisualization*. Prentice Hall, Upper Saddle River, New Jersey, 3rd edition, 2009.

C. A. Steed, P. J. Fitzpatrick, J. Edward Swan II, and T.J. Jankun-Kelly. Tropical cyclone trend analysis using enhanced parallel coordinates and statistical analytics. *Cartography and Geographic Information Science*, 36(3):251–265, Jul. 2009a.

C. A. Steed, P. J. Fitzpatrick, T.J. Jankun-Kelly, A. N. Yancey, and J. Edward Swan II. An interactive parallel coordinates technique applied to a tropical cyclone climate analysis. *Computers & Geosciences*, 35(7):1529–1539, Jul. 2009b.

C. A. Steed, T.J. Jankun-Kelly, and P. J. Fitzpatrick. Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. In *IEEE Symposium on Visual*

Analytics Science and Technology, pages 19–26, Atlantic City, NJ, Oct. 2009c. IEEE Computer Society.

J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, Los Alamitos, CA, 2005.

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

R. E. Walpole and R. H. Myers. *Probability and Statistics for Engineers and Scientists*. Prentice Hall, Englewood Cliffs, New Jersey, 5th edition, 1993.

C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2nd edition, 2004.

E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.

L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1366–1372, Nov. 2006.

W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, Nov.-Dec. 2007. ISSN 1077-2626.

P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization - Overviews, Methodologies, and Techniques*, pages 3–33. IEEE Computer Society Press, 1997.