DEVELOPMENT OF A GEOVISUAL ANALYTICS ENVIRONMENT USING PARALLEL COORDINATES WITH APPLICATIONS TO TROPICAL CYCLONE TREND ANALYSIS

By

Chad Allen Steed

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Computer Science
in the Department of Computer Science and Engineering

Mississippi State, Mississippi

December 2008

DEVELOPMENT OF A GEOVISUAL ANALYTICS ENVIRONMENT USING PARALLEL COORDINATES WITH APPLICATIONS TO TROPICAL CYCLONE TREND ANALYSIS

By

Chad Allen Steed

Approved:

J. Edward Swan II
Associate Professor of Computer
Science and Engineering
(Major Professor and
Co-Dissertation Director)

T. J. Jankun-Kelly
Assistant Professor of Computer
Science and Engineering
(Co-Dissertation Director)

Robert J. Moorhead II
Professor of Electrical and
Computer Engineering
(Committee Member)

Patrick J. Fitzpatrick
Associate Research Professor,
Northern Gulf Institute
(Committee Member)

Edward B. Allen
Associate Professor of Computer
Science and Engineering
(Graduate Coordinator and
Committee Member)

Sarah A. Rajala
Dean of the Bagley College
of Engineering

Name: Chad Allen Steed

Date of Degree: December 12, 2008

Institution: Mississippi State University

Major Field: Computer Science

Major Professor: Dr. J. Edward Swan II

Title of Study: DEVELOPMENT OF A GEOVISUAL ANALYTICS ENVIRONMENT USING PARALLEL COORDINATES WITH APPLICATIONS TO TROPICAL CYCLONE TREND ANALYSIS

Pages in Study: 221

Candidate for Degree of Doctor of Philosophy

A global transformation is being fueled by unprecedented growth in the quality, quantity, and number of different parameters in environmental data through the convergence of several technological advances in data collection and modeling. Although these data hold great potential for helping us understand many complex and, in some cases, life-threatening environmental processes, our ability to generate such data is far outpacing our ability to analyze it. In particular, conventional environmental data analysis tools are inadequate for coping with the size and complexity of these data. As a result, users are forced to reduce the problem in order to adapt to the capabilities of the tools. To overcome these limitations, we must complement the power of computational methods with human knowledge, flexible thinking, imagination, and our capacity for insight by developing visual analysis tools that distill information into the actionable criteria needed for enhanced decision support.

In light of said challenges, we have integrated automated statistical analysis capabilities with a highly interactive, multivariate visualization interface to produce a promising approach for visual environmental data analysis. By combining advanced interaction techniques such as dynamic axis scaling, conjunctive parallel coordinates, statistical indicators, and aerial perspective shading, we provide an enhanced variant of the classical parallel coordinates plot. Furthermore, the system facilitates statistical processes such as stepwise linear regression and correlation analysis to assist in the identification and quantification of the most significant predictors for a particular dependent variable. These capabilities are combined into a unique geovisual analytics system that is demonstrated via a pedagogical case study and three North Atlantic tropical cyclone climate studies using a systematic workflow. In addition to revealing several significant associations between environmental observations and tropical cyclone activity, this research corroborates the notion that enhanced parallel coordinates coupled with statistical analysis can be used for more effective knowledge discovery and confirmation in complex, real-world data sets.

Key words: geovisual analytics, multidimensional multivariate data visualization, parallel coordinates, tropical cyclone, hurricane, climate study, visual interaction techniques, statistical analysis, exploratory data analysis, geovisualization, stepwise regression, correlation analysis

DEDICATION

I dedicate this dissertation to my wife, Jessica, and my children, Julia and Blake.

ACKNOWLEDGMENTS

First and foremost, I thank my Lord and Savior, Jesus Christ, for renewing my strength daily and for teaching me patience through adversity during this journey. I thank God for His amazing love manifested by the death, burial, and resurrection of His Son, Jesus Christ, which paved the way for not only my salvation, but the salvation of all who believe and follow Him.

> But they that wait upon the LORD shall renew their strength; they shall mount up with wings as eagles; they shall run, and not be weary; and they shall walk, and not faint.
>
> –Isaiah 40:31

> For God so loved the world, that He gave his only begotten Son, that whosoever believeth in Him should not perish, but have everlasting life.
>
> –John 3:16

I also thank Dr. Swan, my major professor, for offering excellent academic and career guidance as well as a wealth of encouragement and inspiration at just the right times along the way. Dr. Swan's enthusiasm always motivated me to rise early and work late in order to deliver my very best. I greatly appreciate Dr. Swan's character, humility, easygoing demeanor, and generosity as well as the example of loving dedication he displayed for his own family. Insomuch as I have been blessed by knowing Dr. Swan, I hope that I can be an equal blessing to others in the future.

I also thank the extraordinary group of professors who served on my committee. I thank Dr. Jankun-Kelly for the inspiring InfoVis lectures, informal meetings, and critiques that helped me form the research ideas described in this dissertation. I have been continually amazed by his knowledge of visualization techniques and I have the highest regard for his feedback on my own efforts. Furthermore, I greatly appreciate his humor and for introducing me to the wonderful world of Mac. I thank Dr. Fitzpatrick for his abundant enthusiasm and energy, for teaching me the art of statistical climate analysis, and for his expert insight and guidance that formed the core of my climate study evaluations. I truly admire Dr. Fitzpatrick for the tenacity and fortitude that he demonstrated as he rebuilt in New Orleans after Hurricane Katrina. I thank Dr. Moorhead who, in addition to teaching me and being a great role model, always had time to meet with me and shake hands with my guests. My Ph.D. journey really began in 1999 when Dr. Moorhead took time out of his busy schedule to discuss opportunities for MSU graduate study over a fine lunch at the Stennis Space Center cafeteria. I thank Dr. Allen for helping me form a solid plan of study, which made all the difference in maximizing my training time in Starkville. I also want to thank Dr. Allen for his guidance and encouragement during this journey.

I thank Travis Atkison for being the best friend I have ever had in my life and teaching me the game of racquetball. It is no coincidence that we started our Ph.D. pursuits at the same time as we helped each other through. Travis, his wife, Rebekah, and sons, Anthony and Zachary Atkison, will always be very dear friends of our family.

I also thank my pastor at Faith Baptist Church in Starkville, Dr. Blaine Allen, for his friendship and spiritual guidance that sparked much needed maturity in my walk with the

Lord. I am also thankful for the wonderful, genuine, and loving congregation at Faith. I also thank my pastor at First Baptist Church in Purvis, MS, Rev. Bill Wright, for his words of encouragement and Godly example. I am forever indebted to many people in these churches for their prayers and support.

I acknowledge the support of the Naval Research Laboratory and the Select Graduate Training program. My division head, Dr. Herbert Eppert, my branch head, Mike Harris, and my supervisor, Will Avera, have always given me generous support, advice and encouragement. I thank my NRL co-workers and friends including John Breckenridge, Dr. Paul Elmore, Dr. John Sample, Dr. Dave King, Kevin Shaw, Maura Lohrenz, Les Baham, Dianne James, Carolyn Gilroy, and Peggy Fayard. I also thank my former manager at Lockheed Martin Scientific Systems, Larry Wilson, for first encouraging me to take full advantage of opportunities to advance and further my education.

I thank my project sponsors at PEO-C4I PMW-120, Dr. Edward Mozley, Marcus Speckhahn, and Kim Koehler, for having the patience to endure my academic pursuits while I also worked on their research programs. I thank Dr. Phil Klotzbach of Colorado State University's Tropical Meteorology Project for providing the Atlantic tropical cyclone predictors data set that was used in the climate studies. I acknowledge the Earth System Research Laboratory of the National Oceanic & Atomspheric Administration (NOAA) for providing access to the Atlantic Meridional Mode (AMM) Index data set.

I would like to dearly thank my parents for raising me in a safe and loving environment, putting me through school, and instilling within me a drive to excel. I thank my loving mother, Penny Steed, for her cheerfulness, sacrificially giving of herself to help out when

trouble or sickness arose, and for bringing me to church in the days of my youth. I thank my father, Terry Steed, for working so hard to provide for and protect our family, yet always dedicating time to bring me fishing or hunting—I will always cherish these times. I also thank my sister, Gail Clark, for her love, humor, and for being an inspiration to everyone; I will always be proud to say that I am your little brother. I thank my two nieces, Emily and Elizabeth Clark, for their tender smiles and hugs. I thank my father-in-law and mother-in-law, Larry and Paula Miley, for giving me their daughter's hand in marriage and for their generous love and support throughout. In addition, I thank my delightful grandmothers, Donnie Busbea and the late Bernice Steed, for their love and affection. What a wonderful family!

Last but certainly not least, I want to thank my beautiful wife, Jessica, for sticking with me since the seventh grade and loving me unconditionally, despite my *many* imperfections. Jessica kept me grounded and focused on my mission while keeping our household in order when I was engrossed in my work. Jessica, I simply could not have made it this far without you. I thank my sweet and beautiful daughter, Julia (my *Princess*), who always melts my heart with her smile, laughter, and enchanting green eyes. I also thank my thoughtful yet strong-willed son, Blake (my *Chief*), whose dedication and infectious laughter give me a peace that is beyond words. The joy, love, and laughter so freely offered by my wife and children nourishes my soul and I love them, dearly.

> I firmly believe that any man's finest hour, the greatest fulfillment of all that he holds dear, is the moment when he has worked his heart out in a good cause and lies exhausted on the field of battle—victorious.
>
> – Vince Lombardi

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In 1854, a devastating cholera epidemic was raging in London. In response, Dr. John Snow conducted an elaborate investigation on the effect of the water supply on the spreading cholera. In his search for clues, he placed dots at the locations of the recorded deaths on a map of the neighborhood that also showed the drinking water well locations with crosses (see Figure 1.1). From this map, the concentration of dots in the area surrounding one of the wells (the well located at the center of the map on BROAD STREET) was remarkable. He persuaded the local council to remove the handle of this well pump and the epidemic stopped, suggesting that the disease was being transmitted by contact with the handle [52, 144]. Although this same insight might have also been discovered via some calculation, the use of a graphical analysis technique proved to be more efficient.

Over 250 years later, scientists and researchers face similar problems that require sifting through data to find associations among a set of inter-related parameters. For instance, medical researchers still strive to isolate causes and effects from disease outbreaks, intelligence analysts search for patterns in complex data streams to protect national security, and weather scientists seek to find correlations and patterns in environmental measurements to better understand and predict weather phenomena.

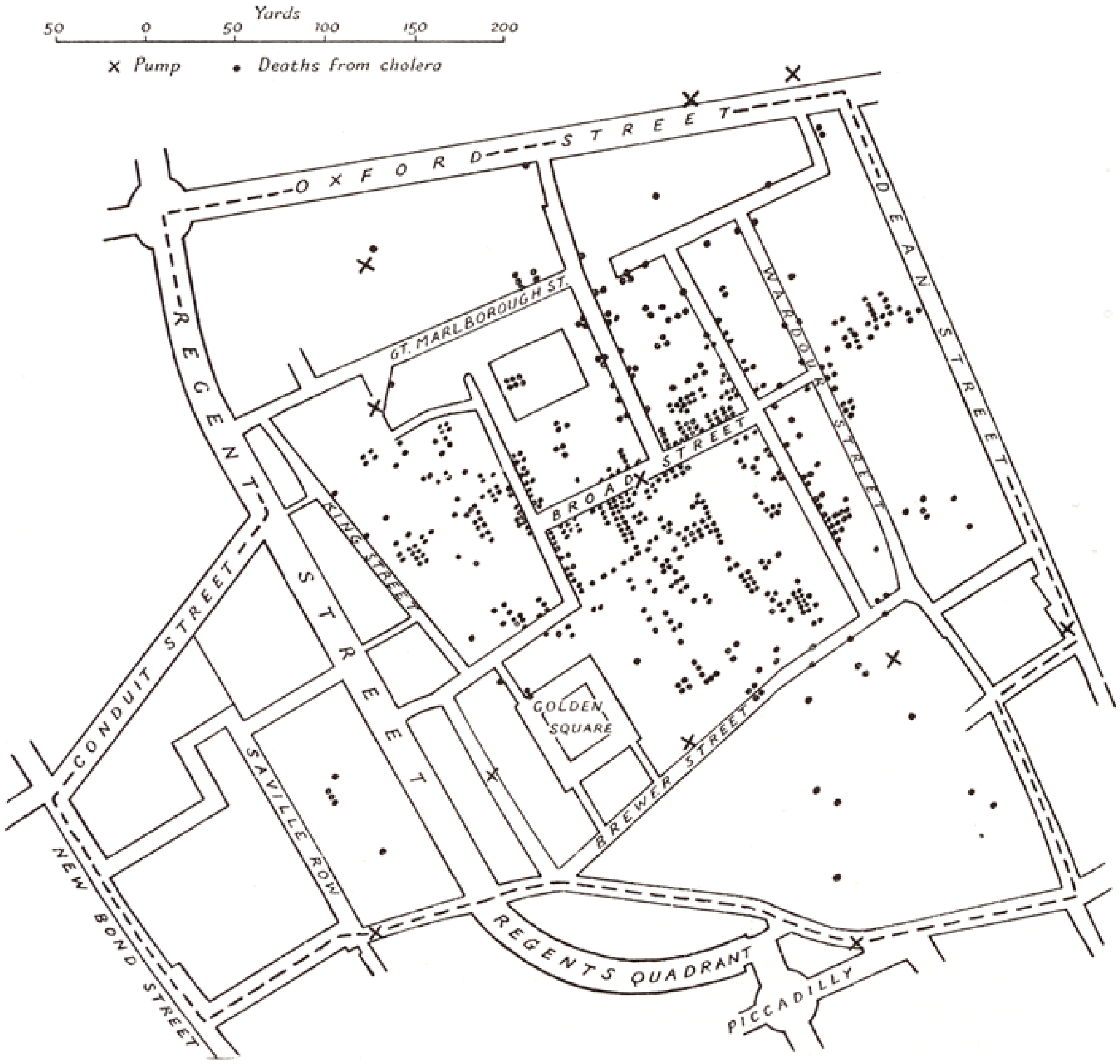


Figure 1.1

Snow's map of the 1854 London cholera epidemic (from Tufte [144]).

However, contemporary problems are exacerbated by unprecedented growth in both the quality and quantity of data due to the convergence of several technological advances. In 2002, Lyman and Varian [116] estimated that five exabytes of new information was produced from print, film, magnetic, and optical storage media, which is equivalent to the information contained in 37,000 new libraries with book collections the size of the Library of Congress. Based on a world population of 6.3 billion, approximately 800 megabytes (MB) of recorded information is produced annually per person [116]. The FedEx Corporation claims an average daily volume of more than 7.5 million shipment transactions [1]. On an average business day, 16.0 petabytes (PB) of data traffic crosses AT&T's global networks [2]. Furthermore, MacEachren and Kraak [118] estimate that up to 85% of all information has a spatial component, which further complicates data analysis.

Although the data explosion is not restricted to any single domain, let us consider, in greater detail, the situation regarding environmental data. The U.S. Naval Oceanographic Office (NAVOCEANO) surveys the world's oceans 24 hours a day, 7 days a week, and at least 10 months of each year with the world's finest-equipped ocean survey fleet. In 2002, Depner et al. [32] projected a 22-fold increase in the amount of bathymetric data to be processed by NAVOCEANO in the future due to the installation of new data collection equipment. This increase translates into about 2.75 terabytes (TB) per year versus the 2002 level of 125 gigabytes (GB) per year. If imagery and side scan sonar data are included, the figure rises to 2400 times the 2002 data collection quantity which is about 300 TB per year [32].

There are many organizations, like NAVOCEANO, that face the challenge of developing new methods to effectively handle the exponential increase in the volume, quality, and resolution of environmental data. Perhaps the most significant contributors to this increase are in the American and international remote sensing programs. For instance, the National Aeronautics and Space Administration's (NASA's) Earth Observing System[1] produces 194 GB of data per day while Landsat 7[2] produces an additional 150 GB per day. Considering the higher level products produced from these raw data to allow scientists to work with easily understandable variables (e.g. surface temperature, ocean productivity), the volume of data produced by these two satellite series alone is about one TB per day [30].

Although these data hold tremendous potential for revealing unknown and valuable insight to help us understand complex systems and phenomena [12, 34, 58], our ability to generate such data is far outpacing our ability to analyze it [30]. In this dissertation, we corroborate the notion that new visual data analysis approaches are a key component to understanding these new data. More specifically, we focus on utilizing visual analytics to conduct climate trend analysis under the premise that the resulting technologies will yield more rapid and accurate analysis. Existing tools in this domain are inadequate because they were not designed for the increasing quality, quantity, and number of different parameters; they cannot provide dynamic interactions with the data behind the visualizations; and they are not linked to decision support algorithms. Methods to display and scrutinize the data in useful and meaningful ways have not kept pace with recent advances

[1] The Earth Observing System is a coordinated set of polar-orbiting and low inclination satellites.

[2] Landsat 7 is a series of satellites that provide global land surface images.

in climate modeling and measurement, thereby forcing scientists to reduce their problems in order to adapt to the tools. Therefore, in order to effectively understand today's environmental data we need to follow Dr. Snow's example and bring to bear innovative, practical visual analysis techniques for enhanced knowledge extraction.

1.1 Motivation

In 1973, Hamming [60] said that "the purpose of computation is insight, not numbers." Likewise, visualization should focus on insight (discovery, decision making, and explanation) rather than pictures. In fact, Card et al. [24] define visualization as "the use of computer-supported, interactive, visual representations of data to amplify cognition" where cognition refers to the acquisition or use of knowledge. More recently, Munzner et al. [121] likened visualization to statistics where the focus is on the analysis and interpretation.

In scientific visualization, the representations tend to be based on physical data that are inherently spatial. When the goal is to visualize nonphysical information—such as financial data, business statistics, documents, or abstract conceptions—there is a basic problem of mapping nonspatial abstractions into an effective visual form. The mass and complexity of such abstract information has been a key factor for extending visualization into the abstract realm [24]. This situation has resulted in the emergence of the field of information visualization, which is loosely defined as the use of interactive visual representations of abstract data to enhance cognition. Ware [153] characterized information visualization as the fusion of various elements from the fields of computer graphics, human-computer interaction, perception, and neurology.

The importance of visualization in dealing with today's data explosion is amplified by the recent U.S. National Institutes of Health (NIH) and U.S. National Science Foundation (NSF) Visualization Research Challenges Report [92]. In this report, visualization researchers are encouraged to "collaborate closely with domain experts who have driving tasks in data-rich fields to produce tools and techniques that solve clear real-world needs [92]." Recently, Stephen Few [42] remarked that a wide chasm separates information visualization researchers from software vendors that could benefit from their work. Researchers are responsible for at least half the chasm due to three failures: much information visualization research has no practical application, much information visualization research produces incomprehensible visualizations and ineffective functionality, and much information visualization research is not presented in an understandable and compelling manner. According to Few, researchers should respond to the actual needs of the people.

Insomuch as the power of visualization is manifest in our innate ability to identify patterns in graphical forms very quickly [113], the usefulness of today's environmental data will be determined by the insight that can be obtained from it. Applying new visualization techniques will help scientists explore these data by reducing and refining the data stream. By exploiting the high bandwidth human visual perception channel through visualization interfaces, we can help scientists understand this data orders of magnitude faster than looking at raw numbers or text. The insight provided through visualization can help scientists find and create new hypotheses, techniques, and methods that can improve our daily lives [121].

Traditionally, maps have been utilized to present, synthesize, analyze, and explore environmental information. But due to the complexity and volume of information, maps alone cannot address today's challenges [106]. That is, when very large and complex multivariate data sets have to be visualized on traditional maps or in Geographic Information System (GIS) environments, the current methods reach their limits [142]. Consequently, the cartographic community is exploring new approaches to mapping which provide flexible interfaces to interact with the data behind the visualization and encourage exploration. As stated by Kraak [106], the new approach is used to "stimulate visual thinking about geospatial patterns, relationships and trends." These new systems are operating in the realm of geovisualization, where cartographic and geographic information representation techniques are integrated with recent advances in scientific visualization, information visualization, exploratory data analysis, and image analysis [117].

From its beginnings, geovisualization has emphasized a visual analysis approach, but the more recently introduced term 'visual analytics' makes this goal more explicit. In general, visual analytics refers to the science of analytical reasoning facilitated by visual interfaces. These tools and techniques are used to synthesize information and derive insight from large, ambiguous, and dynamic data, detect the expected and discover the unexpected, provide timely, defensible, and understandable assessments, and communicate assessments efficiently for action. The need for visual analytics is driven by an ever increasing amount of data to analyze, increasing complexity and uncertainty in the data, decreasing amount of time to analyze data, and a lack of methods, technology, or tools available today or perceived on the horizon [140]. Although the demand was initiated by the needs of the U.S. Department of Homeland Security, it was quickly echoed by other domains such as human health and commerce [141]. The same challenges are evident in environmental data analysis.

A sub-area of visual analytics is called geovisual analytics, which focuses on visual analytic tasks and tools involving geographic and temporal aspects of data. In this dissertation, we extend interactive geovisual analytics to generate representations of multidimensional, multivariate data and evaluate the effectiveness of the resulting capabilities in the study of tropical cyclone trends.

1.2 Statement of Hypothesis

The following hypotheses motivated this dissertation:

1. The development of an advanced geovisual analytics approach using parallel coordinates and statistical techniques reveals a deeper level of understanding than traditional methods when applied to the task of finding complex multivariate trends in environmental data sets. With new ways to creatively explore the data, the approach offers a more effective visual interface to glean new insight about the data behind the visualization.

2. The effectiveness of the geovisual analytics approach is necessarily explored in the context of practical environmental studies, which are grounded in real-world data sets instead of invented or abstract data sets, in close collaboration with domain experts. The discovery of new associations and the confirmation of known patterns by domain experts will validate the promise of this new approach in environmental data analysis.

1.3 Contributions

The main contributions of this research are:

- Developed a parallel coordinates based visualization interface for representing multidimensional, multivariate data by integrating several previously introduced and some new extensions in a single interface.
- Extended dynamic visual query techniques to provide enhanced access to the data behind the visualization.
- Investigated statistical data analysis techniques that help identify and quantify significant predictor variables for a single dependent variable.
- Fused the parallel coordinates interface, dynamic visual query techniques, and statistical analytics into an innovative geovisual analytics application.
- Evaluated the effectiveness of using new geovisual analytics in tropical cyclone climate analyses in close collaboration with a hurricane expert, Dr. Fitzpatrick (who also served on this dissertation graduate committee), using a systematic workflow.
- Identified the strengths and weaknesses of using geovisual analytics in multivariate analysis over traditional environmental analysis techniques.
- In regard to North Atlantic tropical cyclone activity, discovered several significant associations in the AMM data set and confirmed many known trends in the CSU data set, thereby corroborating the notion of enhanced analysis using our approach.

1.4 Organization

The remainder of this dissertation is organized as follows. In order to evaluate the hypotheses presented in this chapter, we conducted a comprehensive investigation of previously published parallel coordinates and other multivariate visualization techniques. The results of this investigation, which are described in Chapter 2, were then used to build a unique parallel coordinates geovisual analytics system for multivariate data analysis. The system combines several fundamental parallel coordinates capabilities and variants of more advanced techniques from prior works. The system also offers new interactive functionality with parallel coordinates: dynamic axis scaling using mouse wheel movement and continuous aerial perspective shading of polylines. These techniques are then used in Chapter 4 to demonstrate the enhanced visual data analysis capabilities in four separate case studies. The new insight obtained from these evaluations in collaboration with our hurricane expert, Dr. Fitzpatrick, validate the promise of this approach in environmental data analysis. Specifically, we developed a deeper level of understanding about the physical associations of global signals for seasonal North Atlantic tropical cyclone activity in the latter three case studies.

CHAPTER 2

BACKGROUND

2.1 Multidimensional Multivariate Data

Statisticians, psychologists, accountants, physicists, etc. have been studying multidimensional, multivariate visualization long before the field of computer science emerged. The introduction of low-priced personal computers in the 1980s opened the door to more advanced graphical data analysis, which, in turn, led to the expansion of the quest for more effective multidimensional multivariate visualization approaches [157]. Recent technological advances have dramatically increased the need for new multidimensional multivariate visualization techniques, especially in the realm of environmental analysis. These advances have increased not only the quality and the number of samples that are analyzed, but also the number of different variables collected. Consequently, data sets with high dimensionality are becoming increasingly common in many domains (e.g. climate analysis, computer science, finance, medical, social sciences) [154].

In 1997, Wong and Bergeron [157] published a review covering 30 years of multidimensional multivariate visualization advances in which they note that much of the terminology in the area is ill-defined, especially the term *dimensionality*. Although the mathematician considers the dimension to be the number of independent variables in an algebraic equation, the engineer regards dimension as measurements of any sort (e.g. breadth, length, height, area). In addition, the prefix *multi* is often interchanged with the prefix *hyper*. In statistics, the prefix *multi* refers to two or more, while *hyper* refers to three and four (or beyond) [157].

Furthermore, Wong and Bergeron [157] note that there is also a difference between multidimensional *objects* and multidimensional *data*. The multidimensional *object* is a spatial object where the goal is to understand its geometry. Commonly realized as two dimensional images or three dimensional volumes, they are best described within n-dimensional Euclidean spaces $\mathbb{R}^n$. In contrast, multidimensional *data* refer to the study of relationships among multiple parameters which can be classified as either *dependent* or *independent* categories. In statistics, the terms *factor* and *response* are often used instead. The *dependent* variable y is a function of the *independent* variable x. This relationship is described by the equation $y = f(x)$. Wong and Bergeron [157] used the convention that the term *multidimensional* describes the dimensionality of the independent variables, while the term *multivariate* refers to the dimensionality of the dependent variables.

2.2 Parallel Coordinates Mathematical Background

Inselberg [72] originally introduced parallel coordinates in the early 1980s. In general, the parallel coordinates technique yields a two-dimensional representation of a multidimensional multivariate data set. That is to say, the N-dimensional data tuple is represented as a polyline where its N-points pass through N parallel y-axes. The resulting visualization provides a compact two-dimensional representation of even large multidimensional multivariate data sets [135].

By mapping high-dimensional data onto two dimensions, the parallel coordinate plot breaks the limitation of dimension representation in the Euclidean space that is generally restricted to only three. Furthermore, the technique helps the viewer observe trends, patterns, and correlations in a multidimensional multivariate data set. It can also be used to visualize hyper-geometrical features such as multidimensional lines, planes and envelopes [73,80,81].

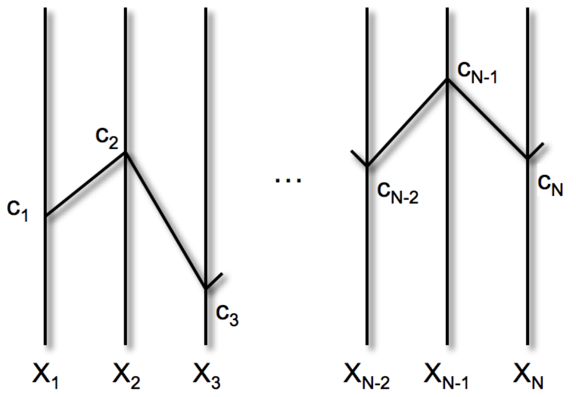


Figure 2.1

The polyline in parallel coordinates maps the point $C \in \mathbb{R}^N$ to $\mathbb{R}^2$.

The parallel coordinates technique provides a one-to-one mapping between subsets of $\mathbb{R}^N$ and subsets of $\mathbb{R}^2$ yielding a systematic approach to the analysis of multidimensional multivariate data. In $\mathbb{R}^2$ with xy-Cartesian coordinates, N vertical axes are placed equidistant and perpendicular to the x-axis to graphically represent N variables, all having the same orientation as the y-axis and perpendicular to the x-axis (see Figure 2.1). The N-dimensional point C with coordinates $(c_1, c_2, \ldots, c_N)$ is represented by points on the N axes which are joined with a polygonal line whose N vertices are on the X_i-axis for $i = 1, \ldots, N$ and have xy-coordinates $(i-1, c_i)$. This establishes the correspondence between points in $\mathbb{R}^N$ and polygonal lines with vertices on $X_1, X_2, \ldots, X_N$ which is the fundamental duality between the Cartesian and parallel coordinate systems [73, 80, 81].

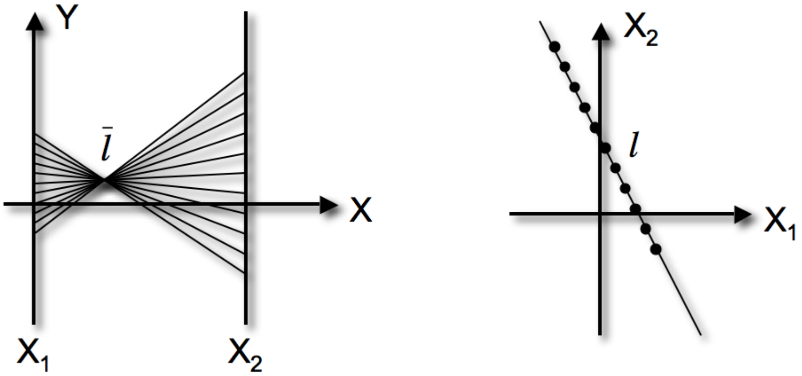


Figure 2.2

Illustration of the point line duality for parallel coordinates.

As shown in Figure 2.2, points in the Cartesian system are represented by lines in the parallel system. In two dimensions, a point in Cartesian coordinates corresponds to an edge in parallel coordinates, whereas a line in Cartesian coordinates, represented by

$$l : x_2 = mx_1 + b, \tag{2.1}$$

corresponds to a point

$$\bar{l} : (d/(1-m), b/(1-m)) \tag{2.2}$$

in parallel coordinates where d is the distance between the parallel axes. The point is also called the envelope of the collection of points in the line l with coordinates $(a, ma + b)$.

Another characteristic of the parallel system is the duality between rotation and translation with the Cartesian system. Rotation of a line about a point in the Cartesian system is shown in the parallel system by the translation of the corresponding point along a line representing the point of rotation in the Cartesian system. Inselberg [81] provides comprehensive descriptions of this duality, the point line duality and other transformation characteristics for parallel coordinates.

2.3 Statistical Analysis Techniques

In our research, we utilized a number of statistical methods to enhance cognition in visual data analysis. In this section, we describe the following statistical analysis techniques that we found to be the most beneficial to climate analysis: descriptive statistics, correlation analysis, simple linear regression, and multiple linear regression.

2.3.1 Descriptive Statistics

Descriptive statistics are the most basic form of statistics, which describe patterns and general trends in data sets. Key descriptive statistics include the mode, median, mean, central tendency, variation, range, variance, standard deviation, skewness, interquartile range (IQR) and kurtosis. In this research, we have focused on the median, mean, standard deviation, and IQR measures. These statistics provide important measures of central tendency and variability.

Central tendency measures, such as the mean and median, give a single description of the typical value in a data set. Although it is simply a numerical average, the mean, or sample mean, is the most widely used central tendency measure. It is computed as the sum of all the values divided by the number of values [151]. Suppose that the observations in a sample are given by $x_1, x_2, \ldots, x_n$. The sample mean is given by

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}. \tag{2.3}$$

The mean is often used to characterize the "typical" value in a data set and it corresponds to what scientists call the "center of mass." If we consider the horizontal axis of a histogram as a bar that has all the samples piled on it, where each sample has an equal weight, the mean is the point at which the bar balances [25].

Another measure of central tendency is the median which essentially cuts the distribution exactly in half; an equal number of values are larger than the median as there are smaller than the median. By definition, this value is also called the 50th percentile [151].

The median is computed by sorting the data values from smallest to largest; the median is the middle element. If there are an even number of elements, there is no single middle sample, so we split the difference between the two middle samples and call the resulting value the median [25]. It is important to note that the median may not be an actual or possible value in the data set [151]. If $x_1, x_2, \ldots, x_n$ represent observations in a sample of size n, arranged in increasing order of magnitude, then the sample median is defined by

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{x_{n/2}+x_{(n/2)+1}}{2}, & \text{if } n \text{ is even.} \end{cases} \tag{2.4}$$

One advantage of the median is that it is easy to compute if the number of observations is relatively small. On the other hand, the mean value will not vary as much from sample to sample as will the median when dealing with samples from populations. Consequently, if we wish to estimate the center of a population based on a sample value, the mean value will yield more stable results than the median. That is, a sample mean is more likely to be closer to the population mean than the sample median [151]. Another advantage of the mean is that it can be manipulated algebraically which makes it easier to use in equations than the median [69].

A disadvantage of the mean is that it can be sensitive to extreme values, which are also known as outliers, since it is calculated by adding up all the samples in the distribution. Conversely, the median is not influenced by extreme values and it gives a better center of the data values [69].

Another important set of statistics are the variability measures that quantify how dispersed the values in a data set tend to be. In particular, we are interested in the interquartile range (IQR) and standard deviation range. The IQR is the difference between the 25th and the 75th percentile scores; it is essentially the range of the middle 50% of the data values. For a set of data, the 25th percentile is the value for which 25% of the data are less than that value. This value is the same as the median of the data that are less than the overall median. The 25th percentile is also known of as the first quartile, low quartile, lower quartile, and Q1. Similarly, the 75th percentile is the number for which 75% of the data values are less than that number. This value is the same as the median of the part of data that is greater than the median. The 75th percentile is also known of as the third quartile, high quartile, higher quartile, and Q3 [25]. The IQR is used as a robust measure of scale and can be as an alternative to the standard deviation for quantifying variability. Because it is based on the median values, the IQR is less affected by extremes than the standard deviation and it is the measure of scale used in the box plot technique [145].

The most common measure of statistical dispersion is the standard deviation, which measures in the same units as the data the average amount by which the values in a data set differ from the mean value. Formally, the standard deviation is the root mean square deviation of values from their arithmetic mean. If the data set has many values close to the mean, then the standard deviation is small; if many values are far from the mean, then it is large. When all the values are equal, then standard deviation is zero [151]. If $x_1, x_2, \ldots, x_n$ represent the sample of size n, and $\overline{x}$ is the mean of the sample, then the sample standard deviation is defined by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.5}$$

2.3.2 Correlation Analysis

Correlation analysis attempts to measure the strength of the relationships between two variables by using a single number called the correlation coefficient. The measure of the correlation between two variables X and Y is estimated by the sample correlation coefficient, r, which is also called the Pearson product-moment correlation coefficient. This quantity is named after Karl Pearson who developed the method to do agricultural research [151]. 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

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i{}^2 - (\sum x_i)^2][n\sum y_i{}^2 - (\sum y_i)^2]}}. \tag{2.6}$$

By facilitating a measure of how related two variables are, correlation analysis allows us to make predictions about one variable based on what is known about another. There are two directions or types of correlation: positive and negative. With a positive correlation, as values of one variable increase, values of the other variable also increase. Likewise, as the values of one variable decrease, the values of the other variable will also decrease [110]. The right side of Figure 2.3 shows a positive correlation between the *HP* and *Displacement* axes in a scatterplot and a parallel coordinate plot.

With a negative correlation, as the values of one variable increase, the values of the other variable decrease. Likewise, as values of one variable decrease, the values of the other variable increase. This is an inverse correlation which is also called negative to describe the direction of the correlation [110]. The left side of Figure 2.3 shows a negative correlation relationship between the *HP* axis and the *MPG* axis in a scatterplot and a parallel coordinate plot [110].

Both positive and negative correlations range in strength from weak to strong. A positive correlation is a number between 0 and 1 where 0 means no correlation and 1 is a perfect positive correlation. As the correlation coefficient gets closer to 1, it is getting stronger. That is, .8 is stronger than .6 but .6 is stronger than .4. On the other hand, a negative correlation is a number between 0 and -1 where 0 means no correlation and -1 is a perfect negative correlation. As the correlation gets closer to -1, it is getting stronger. Then, $-.7$ is stronger than $-.5$, but $-.5$ is stronger than $-.3$ [110]. In practice, r is rarely perfect as it usually lies somewhere between -1 and $+1$ [151].

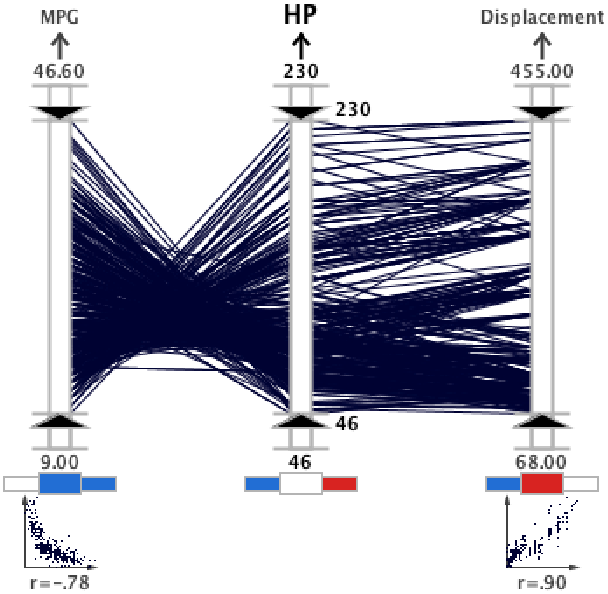


Figure 2.3

Positive and negative correlations represented in parallel coordinates.

When an independent variable is highly correlated with several other independent variables, the variable has *collinearity*. This condition is also described as *multicollinearity* or *ill conditioning*. The variable has much in common with several other variables and may have little information unique to itself [69]. This condition results in a loss in power and makes interpretation more difficult in the results of a regression model.

The advantage of correlation analysis is that we can make predictions about the behavior of one variable based on what we know about another variable if the two variables are correlated. The disadvantage of this approach is that one must remember that correlation does not measure cause. That is to say, correlation tells us that two variables are related but it does not tell us that one causes the other. Consequently, it is important to avoid drawing conclusions about cause and effect using correlation analysis [110].

2.3.3 Simple Linear Regression

The term *regression* comes from Sir Francis Galton (1822–1911) and his work on heredity. While studying the seeds sizes in mother and daughter pea plants, Galton noted that the seed sizes tend to regress to the mean from mother to daughter plants. Mother plants with very large seeds would produce daughter plants with seeds that are above average but closer to the mean size—an effect Galton termed the "regression to mediocrity." [25].

In modern usage, the purpose of regression is to develop an equation to predict one variable based on the knowledge of another. We have a single dependent variable or response Y, which depends on one or more independent variables, $x_1, x_2, \ldots, x_k$. The fit of this relationship is characterized by a prediction equation called the regression equation. With a single Y and a single x, we have a regression of Y on x. For k independent variables, we have a regression of Y on $x_1, x_2, \ldots, x_k$ [151]. The term simple linear regression (SLR) refers to regression procedures that involve a single regressor variable.

The regression equation is given by

$$Y = a + bX + \varepsilon. \tag{2.7}$$

In this equation Y is the dependent variable, a is the y intercept, b is the gradient or slope of the line, X is the independent variable and ε is a random error term. In our SLR analysis, we utilize the SLR r^2 term, which comes from the linear correlations and signifies the portion of variation in Y that is explained by the straight-line regression [151].

2.3.4 Multiple Linear Regression

In most problems, more than one independent variable is necessary for the regression model. When the model is linear in the coefficients, it is called a multiple linear regression (MLR) model. We now have multiple predictors and we want to predict Y on the basis of our knowledge of all the predictors, an extension of SLR [69]. Formally stated, the multiple regression problem is about finding the regression equation to predict Y on the basis of k predictors, $x_1, x_2, \ldots, x_k$. The MLR equation is given by

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k. \tag{2.8}$$

In this equation b_0 represents the intercept and b_1, b_2, $\ldots$, b_k represents the regression coefficients for predictors $x_1, x_2, \ldots, x_k$ [69]. In our research, we employed stepwise regression with a “backwards glance” to identify the most relevant predictors for Y. This method selects the optimum number of most important variables using a predefined significant value. Stepwise regression helps find a model that does a good job of predicting the dependent variable with as few independent variables as possible. Reducing the number of independent variables is helpful because it simplifies the interpretation and it usually means cheaper data collection and analysis [25].

The idea of stepwise procedures is to start with an initial model and add or delete variables step by step, one at a time, to make the model better. The procedure stops when no appreciable improvement is gained by making another step. Although the resulting model may not be the best of all possible models, it is generally one of the best [25].

Chalmer [25] provides the following steps for the stepwise regression procedure.

1. The first step is to find the independent variable that has the strongest correlation with the dependent variable. If the correlation is significant, the variable is added to the model.
2. The second step is find the independent variable, of those not yet added to the model, that provides the greatest increase in the coefficient of multiple determination, R^2, if added to the model. If the increase is significant, and the variable tolerance is not too small, it is added to the model.
3. The next step is to re-evaluate the independent variables in the model and determine if any should be removed. The variable with the highest p-value is removed conditioned on the p-value being larger than some pre-determined level. This step distinguishes stepwise regression from forward selection.
4. The second and third steps are repeated until there are no more variables that meet the criteria for addition or deletion.

Stepwise regression is dangerous when it is used with a large number of variables and a small sample size. Howell [69] suggested at least ten observations for every predictor and noted that others have suggested the number of observations, N, should exceed the number of predictors, p, by at least fifty or $N \geq p+40$. Even when these conditions are met, the results should be considered preliminary and should be confirmed by additional data [25].

With the MLR analysis used here, extra steps are taken to ensure the proper selection of variables. The initially chosen variables are examined for multicollinearity using an automatic filter; if any variables are correlated with each other by more than the significant correlation threshold, one is removed. In this way, the chosen variables are truly independent of each other.

In addition, a normalization procedure is also used in our MLR analysis so that equal comparison between the variables can be executed. Denoting σ as the standard deviation of a variable, y as the dependent variable, $\bar{x}$ as the predictor mean, and $\bar{y}$ as the dependent variable mean, a number k of statistically significant predictors are normalized by the following regression:

$$(y-\bar{y})/\sigma_y = \sum_{i=1}^{k} b_i(x_i-\bar{x}_i)/\sigma_i. \tag{2.9}$$

The advantage of this approach is that the importance of a predictor may be assessed by comparing regression coefficients b_i between different variables, and that the y-intercept becomes zero. In addition, $\bar{x}_i$ may be interpreted (to a first approximation) as a threshold value which distinguishes between positive and negative contributions (for $b_i > 0$), and the opposite for negative b_i.

We also evaluate R^2 as a criterion to illustrate the adequacy of a regression model. The R^2 value indicates what proportion of the total variation in the response Y is explained by the model [151]. The equation for R^2 is

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{2.10}$$

The regression sum of squares, *SSR*, reflects the amount of variation in the y-values explained by the model. The *SST* is the total corrected sum of squares of y. Here $\hat{y}$ is used to distinguish between estimated or predicted values given by the regression equation and an actual observed value y for some value of x [151]. The R^2 term is similar to the r^2 term discussed in the preceding section on SLR. However, the r^2 term is the proportion of variability of one variable that is explained by the linear relationship to another variable. If we have a linear model and one independent variable, the R^2 is the same as the square of the correlation between the dependent and independent variables, r^2 [25].

2.4 Tropical Cyclone Trend Analysis

In climate studies, scientists are interested in discovering which environmental factors influence significant weather phenomena. A prominent weather feature is a *tropical cyclone*, defined as a warm-core non-frontal synoptic-scale cyclone, originating over tropical or subtropical waters, with organized thunderstorms and a closed surface wind circulation about a well-defined center. Tropical cyclones begin as a *tropical depression*, with sustained 10-meter winds less than 17 $\frac{m}{s}$. Most intensify into *tropical storms* which have sustained winds between 17 $\frac{m}{s}$ (39 mph or 34 knots) and 32 $\frac{m}{s}$ (73 mph or 63 knots). About 56% of tropical cyclones reach winds of at least 33 $\frac{m}{s}$ (74 mph or 64 knots), and are then designated with regional terms such as *hurricanes* in the Atlantic basin, and *typhoons* in the Western North Pacific Ocean. When sustained 10-meter winds reach 49 $\frac{m}{s}$ (111 mph or 96 knots), they are called *intense hurricanes* in the Atlantic [6].

Most people first think of the winds associated with these systems as the main source for destruction. When winds exceed design specifications, structures fail and the debris sent flying into the air compound the damage. Winds also cause downed trees and power lines which cause prolonged power outages. In addition to the steady winds from these systems, wind gusts, tornadoes, downbursts from thunderstorms, and sometimes mesoscale vortices at the boundary of the eye and eye wall generate pockets of enhanced winds which amplify the destruction. Flooding from tropical cyclones is also quite destructive and is currently the leading cause hurricane-related fatalities in the United States. Remnants of tropical systems can also cause inland rain accumulation that may lead to flooding and even mudslides. Historically, the most deadly aspect is the storm surge which is defined as the abnormal rise of the sea along the shore [46].

The variability and destructiveness of recent hurricane seasons have escalated efforts to forecast hurricane activity. Tropical cyclone activity in each ocean basin can vary on a yearly scale as well as a multidecadal scale due to large-scale atmospheric influences and climate forcing. Scientists are developing procedures to forecast whether an upcoming tropical cyclone season[1] will be above normal, normal, or below normal.

Annual hurricane activity forecasting began in the early 1980s when Dr. William Gray at Colorado State University began to study how to predict, months in advance, the number of tropical storms and hurricanes for the upcoming Atlantic hurricane season. In 1984, Gray started publicly predicting how active Atlantic hurricane seasons would be before they started. Through this research, Gray and his colleagues have discovered many global

[1] The North Atlantic tropical cyclone season begins on June 1 and ends on November 30 each year.

signals that affect hurricane activity such as El Niño Southern Oscillation (ENSO), African rainfall, pressure and temperature difference between the western African coast and the Sahel region during the previous February – May period, Caribbean sea surface pressure, Quasi-Biennial Oscillation, Caribbean wind shear, Atlantic Ocean water temperature, and the strength of the Azores high-pressure system. Gray predicts above average hurricane activity for a season when more of these predictors are favorable for hurricane activity than unfavorable. On the other hand, if more are unfavorable, a quiet season is predicted. When the same number of predictors have positive and negative influences, an average season is predicted[2]. However, care should be taken when attributing the activity to any one feature because most of the features are interrelated. For example, El Niño is associated with strong Caribbean wind shear [46].

Gray generates quantitative predictions using statistical techniques and intuition. Between 1984–2004, Gray correctly predicted above or below average activity in sixteen of the twenty-one seasons. Even the four incorrect years have led to improvement of the forecast methods. For example, the 1989 forecast led to the discovery of the importance of African rainfall since this was the only non-drought year of that decade. Starting in 1990, African rainfall was included in the forecasts [46].

Gray's seasonal forecasts draw a great deal of public attention. Furthermore, insurance companies use the predictions to prepare for an upcoming season by buying insurance for themselves (a practice known as reinsurance). Gray's work has also inspired other

[2] An average hurricane season for the Atlantic has about nine tropical storms, six of which will become hurricanes and two of which will be intense.

universities and organizations to issue seasonal forecasts using various statistical methodologies [46]. Others are studying causes of multidecadal cycles of activity, and whether anthropogenic global warming is also an influence [109].

One particularly useful method for predicting seasonal hurricane activity is based on the idea that there are predictors of the main dynamic parameters that affect storm activity which can be observed up to a year in advance. Using historical data, their importance is estimated using statistical regression techniques similar to those described by Vitart [150]. Although sometimes complicated to establish, these techniques provide an ordered list of the most important predictors for dynamic parameters. Klotzbach et al. [101] use this technique for forecasting North Atlantic tropical cyclone activity. Similarly, Fitzpatrick [44, 45] applied stepwise regression analysis to the prediction of tropical cyclone intensity. Scientists gain additional insight in these studies by evaluating descriptive statistics and performing correlation analysis.

In conjunction with statistical analysis, researchers have relied on simple scatterplots (see Figure 2.4) and histograms which require several separate plots or layered plots to analyze multiple variables. Using separate plots, however, is not an optimal approach in this type of analysis due to perceptual issues such as *change blindness* (a phenomenon described by Rensink [130]), especially when searching for combinations of conditions. Change blindness results in the inability of the low-level human perceptual system to recall detail outside the viewing area. A more useful technique employed by statisticians to uncover patterns in multivariate data is the scatterplot matrix (SPLOM), which contains all the pairwise scatterplots of the variables on a single display in a matrix configuration;

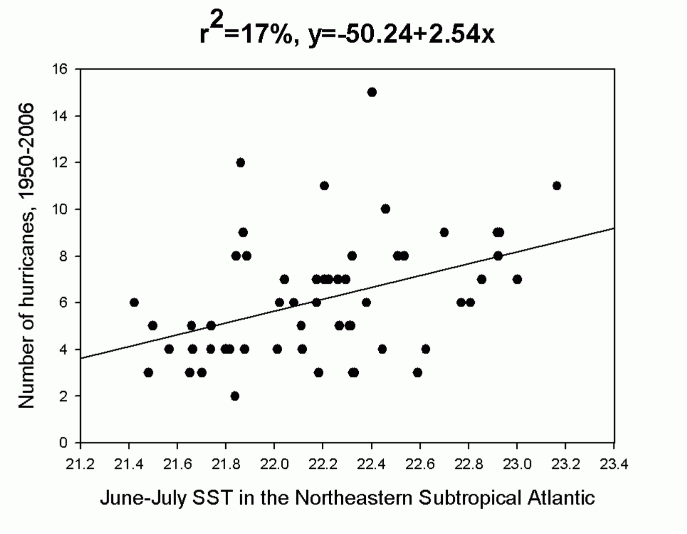


Figure 2.4

Scatterplot showing sea surface temperature versus hurricane activity.

but it requires a large amount of screen space and forming a multidimensional association from a set of two-dimensional displays is mentally challenging.

Although layered plots condense the information into a single display, there are significant issues due to occlusion and interference as demonstrated by Healey et al. [64]. Furthermore, the geospatial data used in climate studies are usually displayed in the context of a geographical map; although certain important patterns (those directly related to geographic position) may be recognized in this context, additional information may be discovered more rapidly using non-geographical information visualization techniques. Due to the multivariate nature of climate study data, researchers need interactive visualization techniques that can accommodate the simultaneous display of many variables.

2.5 Literature Review

To cope with the demands of environmental data analysis, basic research efforts are underway to identify new visualization techniques using guidelines from human perception. While some researchers are developing illustrative visualization solutions to represent multidimensional multivariate data sets, other efforts draw inspiration from advances in the domain of information visualization. Solutions to these problems will emerge from these fields and from other areas such as Exploratory Data Analysis (EDA) and visual interaction methods. That is, the challenges of environmental data analysis require solutions that fuse aspects from each of these areas in a manner that feels seamless to the analyst. In this section, we present a review of significant research related to both these topics and our approach.

2.5.1 Exploratory Data Analysis

In general, EDA is an attitude or philosophy about the execution of data analysis [4]. Introduced by Tukey [145], EDA relies heavily on graphical techniques to investigate data and discover significant hypotheses. EDA complements conventional statistical tools for testing hypotheses, a process Tukey calls confirmatory data analysis (CDA). In 1977, Tukey believed that too much emphasis was being placed on CDA and that, although it is necessary, there is no need to begin with it. Instead, EDA and CDA can and should proceed side-by-side. Unless EDA reveals new insight (usually quantitative), there is likely no need for CDA (except for planning and experiments). Likewise, even though EDA is the foundation stone (first step) it can never be the entire story; CDA goes further than EDA, accessing the strengths of evidence [145].

EDA places an emphasis on the use of images that yield rapid insight about data rather than on the quality of the graphics [24]. Tukey believed that pictures based on data exploration should force their messages upon the analyst and pictures that tell us what we already know are mostly a waste of space. On the other hand, pictures that require intense review and concentration are wasteful of time and inadequate in effect. Tukey postulated that the greatest value of an image in EDA is "when it forces us to notice what we never expected to see [145]." This principle has been carried over into more recent research in information visualization and illustrative visualization whereby unconventional representation techniques are utilized to foster creative thinking.

During analysis, while the data is manipulated to reveal its structural secrets, the analyst must remain ready to grasp new, often unsuspected, insight. When insight is discovered, the underlying structure of the data is revealed. In addition to knowing what is in the data, it is often equally important for the analyst to understand what is *not* in the data [4]. With these goals in mind, EDA systems should provide specific items that an analyst will want to exhaust such as a good fitting economical model, a list of outliers, a sense of robustness of conclusions, estimates for parameters, uncertainty for estimates, a ranked list of important factors, conclusions about statistical significance of factors, and optimal settings [4].

Recently, Gatalsky et al. [51] presented a EDA technique, called the "space-time cube", to analyze events in spatio-temporal data. They were interested in visual EDA techniques that allowed the analyst to understand how the events are distributed in time and space. The "space-time cube" supports the representation of time as an additional spatial dimension. Using earthquake data, the authors demonstrated the promise of the "space-time cube" in detecting events that occur closely in space within short time intervals as shown in Figure 2.5. In this figure, the vertical positions of the circles correspond to event times and the circle sizes and colors reflect event characteristics. While evaluating this system, they also determined that a linked map view is crucial to the usefulness of the cube display and incorporation of an automated cluster detection capability would make it more useful by reducing manual searches [51].

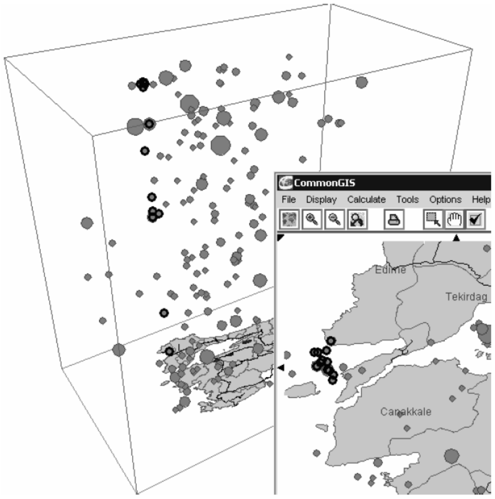
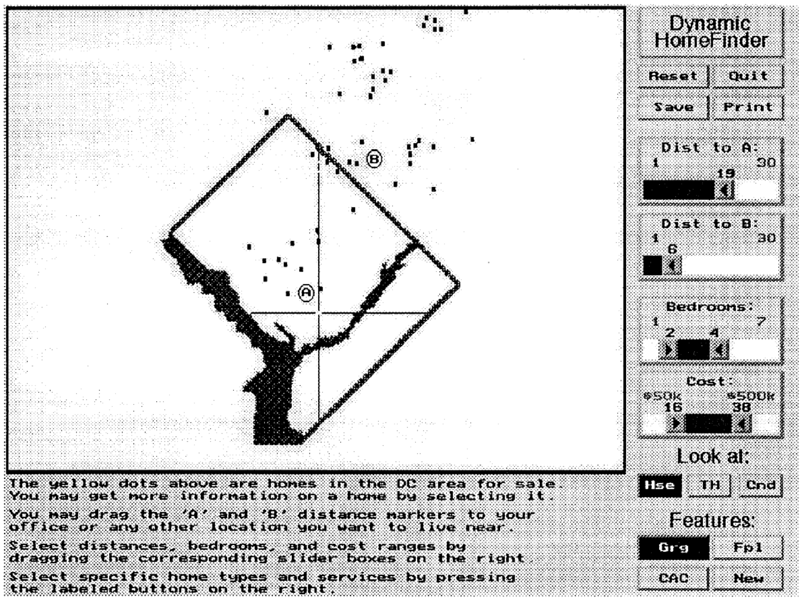


Figure 2.5

Spatio-temporal analysis with the "space-time cube" by Gatalsky et al. [51].

2.5.2 Dynamic Interaction Techniques

In modern EDA systems, the capability to dynamically interact with the data behind the visualization is crucial to its usefulness. A good example of this interactivity was described in 1994 when Ahlberg and Shneiderman introduced the concept of rapid, dynamic queries using visual widgets in the HomeFinder [133] and FilmFinder applications [8]. With this concept, the visualization is continuously updated as users adjust sliders and buttons to ask simple questions and find patterns or exceptions in the data. This approach benefits the novice and expert alike because the novice can learn the visual query mechanics faster and the expert can use the same interface to formulate more complex queries.



The yellow dots above are homes in the DC area for sale.
You may get more information on a home by selecting it.

You may drag the 'A' and 'B' distance markers to your office or any other location you want to live near.

Select distances, bedrooms, and cost ranges by dragging the corresponding slider boxes on the right.

Select specific home types and services by pressing the labeled buttons on the right.

Figure 2.6

Ahlberg and Shneiderman's [8] Dynamic HomeFinder

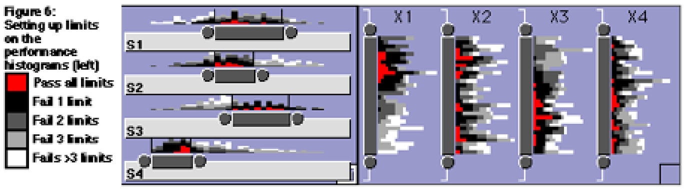


Figure 2.7

The Influence Explorer application introduced by Tweedie et al. [146].

Shneiderman introduced the Dynamic HomeFinder application (see Figure 2.6) to demonstrate the dynamic query concept in a geographic scattermap. The application is designed to help real estate brokers and clients find home listings. The user controls the information that is displayed by manipulating sliders and buttons (in the right panel) that control factors such as price, number of bedrooms, and distance to work [133]. A similar system called the FilmFinder [8] provides an exploratory interface to a movie database. After Ahlberg and Shneiderman introduced these concepts, Tweedie et al. [146] introduced the Influence Explorer which extended the double slider concept to include frequency information between the slider handles (see Figure 2.7).

2.5.3 Parallel Coordinates

The parallel coordinates visualization technique provides the core functionality in our system. This multidimensional multivariate information visualization technique has a rich history since its introduction in the 1980s. In the remainder of this section, we will provide a brief history of parallel coordinates by examining the prior works in three areas: theory, extensions, and applications.

2.5.3.1 Parallel Coordinates Theory

While studying Euclidean geometry as a Ph.D. student, Alfred Inselberg was frustrated by the lack of visualization techniques for multidimensional geometry and experimented with placing the coordinate axes parallel to one another. In 1959, his professors encouraged him to pursue the idea. In 1977, Inselberg revisited the concept after being challenged by his linear algebra students to show some multidimensional spaces. This activity lead to the subsequent formulation of the parallel coordinate theory. Later, Inselberg collaborated with Bernard Dimsdale who made many critical contributions [71].

Inselberg was originally motivated by a challenge to construct planar diagrams of N-variable representations ($R^N \rightarrow R^2$). The resulting polygonal lines in parallel coordinates reveal a particular set of linear dependent vectors [73]. Inselberg first published the parallel coordinate methodology in 1981 [72] and later refined the concept in 1985 [73] and 1987 [79]. These initial papers covered the theory of parallel coordinates with specific application to the visualization of hyper-geometrical surfaces. Inselberg provides a thorough mathematical description of the concept in several subsequent publications [74–78,80–82]. In 1987, Inselberg and Chomut [79] published a paper focused on the convexity algorithm for parallel coordinates. Later, Fiorini and Inselberg [43] discussed the potential of parallel coordinates for use in configuration space representation of a mechanical arm in order to demonstrate the potential of parallel coordinates for cases where Cartesian coordinates cannot be used.

In 1990, Inselberg and Dimsdale [80] expanded the parallel coordinates discussion to detail its use in visualizing analytic and synthetic multidimensional geometry. In addition to a formal, mathematical definition, the application of parallel coordinates to air traffic control, robotics, computer vision, computational geometry, statistics, and instrumentation is also mentioned [80].

Later that same year, Wegman [155] employed parallel coordinates to the analysis of high-dimensional data. Motivated by the poor performance of traditional representation techniques (such as scatter diagrams) and their inability to work well beyond three dimensions, Wegman decided to draw the *n*-dimensional axes as parallel instead of orthogonal. Although parallel coordinates may seem complex to comprehend at first glance, Wegman [155] notes that the rich intuition about the appearance of structures in Cartesian coordinates has been developed over many years and similar intuition about parallel coordinates must also be developed.

Wegman introduces several statistical interpretations for parallel coordinate polyline configurations. He notes that correlation structures can be easily diagnosed; for example, highly positively correlated data tend to have lines that do not intersect between the axes. Wegman also mentions that rank-based statistics are expected to have an intimate relationship to parallel coordinates because of scale invariance. Moreover, linear relationships are easy to spot in parallel coordinate plots, particularly negative linear relationships, because the eye seems to quickly spot the crossover effect. Detecting linear structure is important to understanding the data, especially in linear regression models. Furthermore, clustering is easily diagnosed and the clustering may occur in multiple dimensions. Wegman observed that the mode, the location of the most intense concentration of observations, will be represented as the most intense bundle of broken line paths [155].

Wegman presents parallel coordinate plots from a data set of seventy-four 1979 model-year automobiles which has five dimensions of data to demonstrate the effectiveness of this method at representing several relationships in this data set. Wegman also discusses some extensions to parallel coordinates such as a density plot feature to help with over-plotting in large data sets, and the use of color histograms to replace the lines in the plot [155].

According to Inselberg's 1997 paper, parallel coordinate plots are designed to

- Exhibit low representational complexity,
- Work for any *N*,
- Treat every variable uniformly,
- Represent the display object's projective transformations,
- Intuitively convey information about the object it represents, and
- Stand on a methodology that is based on rigorous mathematical and algorithmic results [74].

Inselberg also presents a scenario for the knowledge discovery process, guidelines for using parallel coordinates in data mining, and construction and use of visual models for multivariate relations. He uses a VLSI chip production data set which contains measures for sixteen process parameters [74].

The first guideline Inselberg mentions is to not be intimidated by the picture. It is also important to understand the objectives of the system and use them to obtain visual cues prior to embarking in the exploration process. The viewer should also carefully scrutinize the picture and test assumptions, especially those you are very confident of. For each of these guidelines, the author gives a working example using the chip production data set [74].

Later, Inselberg and Avidan [77,78] describe a classification algorithm that is designed to automate some of the discovery process for high-dimensional data using parallel coordinates. The algorithm provides comprehensible and explicit rules, does dimensionality selection (minimal set of original variables required to state the rule) and it orders the variables to optimize the clarity of separation between the designated set and its complement. The first data set used is a satellite image data set with thirty-six variables. Also a vowel recognition data set is used which has about ten variables of eleven vowels and a monkey neural data set is used which has 600 samples of about thirty-two variables. Using the discovery algorithm with the monkey neural data set, they found that two particular variables showed the greatest separation. Using traditional approaches would have required the inspection of a scatterplot with 496 pairs to identify this relationship [77,78].

In 2001, Inselberg [76] described the application of parallel coordinates to visual and automatic knowledge discovery. A VLSI chip production data set and case studies illustrating the use of parallel coordinates for process control are presented in this publication; these data and a similar case study were mentioned previously in Inselberg's 1997 paper [74]. The *Classifier* automation feature mentioned in this publication [76] is also described in the author's prior works [77,78].

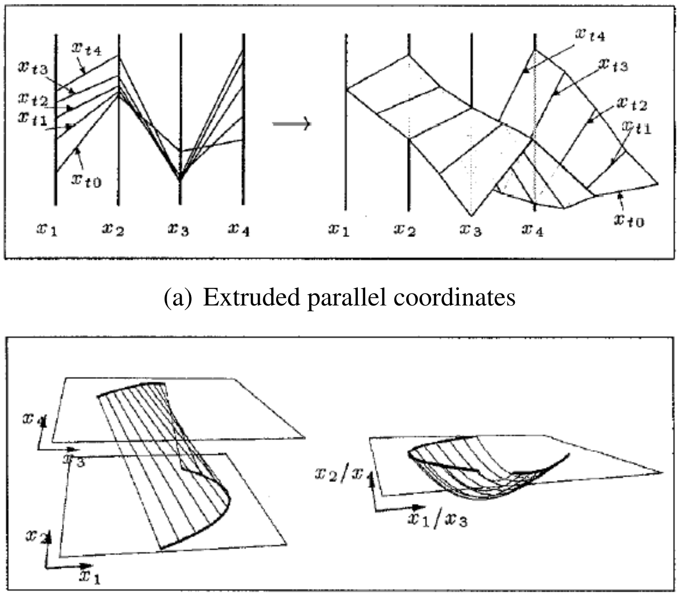
In another publication, Inselberg [75] discussed the use of parallel coordinates for collision detection in air traffic control. More recently, Hung and Inselberg [70] utilized the technique to represent families of planes and hyperplanes and to visualize the properties of complex surfaces like the winding helicoid.

2.5.3.2 Parallel Coordinates Technique Extensions

Since Inselberg's [72] introduction of the parallel coordinates concept, many innovative extensions have been applied to the technique. In 1995 and 1996, Lee et al. [113, 114] discussed multidimensional visualization techniques based on parallel coordinates. They developed a system called *WinVis* that differs from standard parallel coordinate representations in four ways. First of all, the user can toggle between having a line represented as a tuple (database record) or several tuples satisfying a specified attribute value. To reduce the complexity of lines with large amounts of data, group bars appear on the vertical axes in the place of attribute values. Users can also group subjects of interest into classes to see how the data set represents them and how they correlate with other attributes in the data. Horizontal histograms appear to the right of the group bars when the data is divided into classes. Furthermore, the application also includes the capability to partition a continuous axis into ranges and it provides both unsupervised (clustering) and supervised machine-learning techniques [113, 114].

Martin and Ward [119] investigated the brushing operation—the mechanism that allows for interactive selection of subsets of data for highlighting, deletion, or masking—on a variety of display types. Traditionally, these interactions have been implemented as painting or rubberband rectangles, but they describe a new method that provides *N*-dimensional brushing capabilities in a multivariate data visualization tool called *XmdvTool*. This tool provides four display types, one of which is parallel coordinates. In the parallel coordinates display, the user can brush lines as filled regions across all axes using the mouse. Five brushing operations are presented for all the display types: linking (between the difference displays), masking (deleting), moving average, and quantitative presentation. Of these techniques presented, only highlighting is discussed for the parallel coordinates display. In addition to these brushings, the authors present a limited user study to evaluate the system's usability [119].

In 1997, Gröller et al. [56] used parallel coordinates to visualize the behavior of dynamical systems whose temporal evolution from some initial state is governed by a set of rules. After formally introducing the mathematical concepts of dynamical systems, several visualization techniques are described for analyzing these systems. In addition to classical parallel coordinates, two new variants are described: extruded parallel coordinates and three-dimensional parallel coordinates [56]. This work is described in more detail by Wegenkittl et al. [154]. The authors begin with a formal description of multidimensional data and several common methods for visualizing high-dimensional data such as attribute mapping, geometric coding, sonification, reduction of dimension, and parallel coordinates. The authors introduce the concept of extruded parallel coordinates (see Figure 2.8(a)) in which the parallel coordinates system is moved along the third spatial axis instead of using the same coordinate system for each sample. Figure 2.8(a) shows a discrete sampled trajectory in parallel coordinates (left) and a three-dimensional extruded surface defining the same trajectory (right). The polyline of the sample can be viewed as cross sections of a moving plane with a complex surface that defines the trajectory. Correlation and clustering can be detected visually and rotating the surface reveals the evolution of the trajectories over time without any animation methods that would be otherwise necessary. The authors also note that convergence or divergence can be observed by slightly modifying the starting coordinates. Moreover, the authors introduce another concept called three-dimensional parallel coordinates (see Figure 2.8(b)) in which the third dimension is used as opposed to the two-dimensions used in standard parallel coordinates. Using this concept, the information resides in separate two-dimensional spaces (planes) where two dimensional trajectories are shown. These planes are combined in three-space and linked by surfaces that connect the separated projections of trajectories [154].



(a) Extruded parallel coordinates

(b) Three-dimensional parallel coordinates

Figure 2.8

Two variants of parallel coordinates introduced by Wegenkittl et al. [154].

In 1998, Ankerst et al. [13] described a systematic approach for arranging dimensions according to their similarity so that dimensions showing a similar behavior are positioned next to each other. The similarity algorithm is a heuristic algorithm since the one and two dimensional similarly problem is NP-complete. In computer science, NP-complete problems make up the class of the most difficult computational problems for which no efficient solution algorithm has been discovered. In addition to describing the similarity algorithm, an experimental evaluation of the technique is presented using parallel coordinates, circle segments, and recursive patterns [13].

In 1999, Chou et al. [28] addressed the identification of clusters with parallel coordinates. The authors describe a system that helps determine whether or not a set of *n*-dimensional points are close to a certain line using a scan-line algorithm implemented in the Java programming language [28]. Hoffman et al. [68] described a graphic primitive called the *dimensional anchor* which is an attempt to provide a unified framework for a variety of visualizations, including parallel coordinates.

Fua et al. [49] focus on interactive visualization of large multivariate data sets (data sets that contain 10^6 to 10^9 elements or more) by using extended parallel coordinates. Specifically, multi-resolution views via hierarchical clustering are utilized to present the parallel coordinates display and convey information about the clusters. A suite of navigation and filtering tools implemented as OpenGL extensions to *XmdvTool 3.1* are also provided to navigate the resulting structure. The software provides *drill-down* and *roll-up* operations to view data at a level of increasing detail and decreasing detail, respectively. The system also provides a proximity-based color scheme and a feature called *dimensional zooming* which is a distortion operation that allows the user to scale up each dimension independently with respect to the extents of the brushed subspace [49]. The authors provide more details on the system and describe two cases studies in a subsequent publication [50].

Motivated by limited interaction capabilities in conventional parallel coordinates, Siirtola [135] introduced two unique techniques for dynamic interaction. Siirtola focuses on direct manipulation capabilities which are defined as manipulating a visual control and getting a response within 0.1 seconds. The first technique, polyline averaging, facilitates the dynamic summarization of a set of polylines to reduce computational requirements. The other technique provides an interactive visualization of correlation coefficients between subsets of polylines. Furthermore, Siirtola describes direct queries that allows users to select ranges from the *y*-axes for highlighting. These ranges can be combined with other ranges using the logical connectors AND, OR, and XOR which corresponds to a database management system query. In essence, these capabilities demonstrate the intuitive method parallel coordinates offer for doing visual data mining [135].

Siirtola's parallel coordinates user interface (UI) provides many powerful features. In addition to displaying the minimum and maximum values for the axes, the percentage of lines selected are shown beneath each axis. The application can also display the quartiles for an axis as a box plot—a useful feature for dividing a set of polylines into subsets. Siirtola's system can also represent the correlation coefficients graphically (see Figure 2.9 and Figure 2.10) [135]. In Figure 2.9 a subset with positive correlation is shown on the left and a subset with negative correlation is shown on the right. In Figure 2.10 a subset is shown where better mileage has a positive correlation coefficient with higher horsepower.

Siirtola discusses a capability called *dimensional zooming* whereby the selected axis is scaled up according to a selected range. Hierarchical clustering is also discussed but the author states that this technique is too computationally intense for interactive use. For

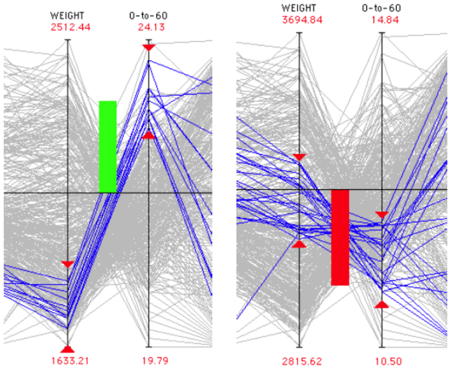
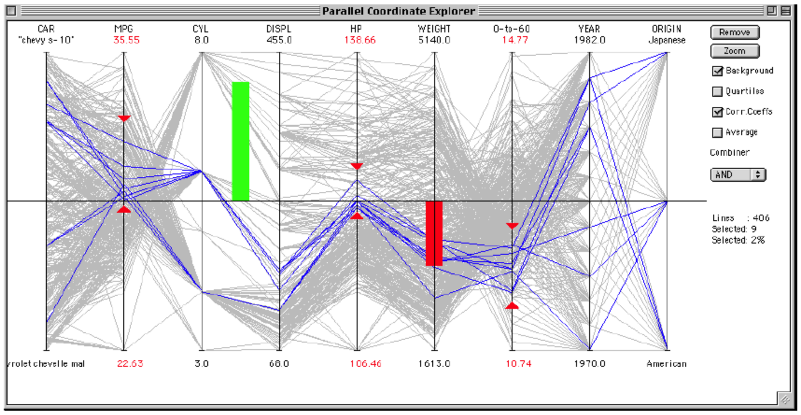


Figure 2.9

Correlation analysis in Siirtola's [135] *Parallel Coordinate Explorer*.



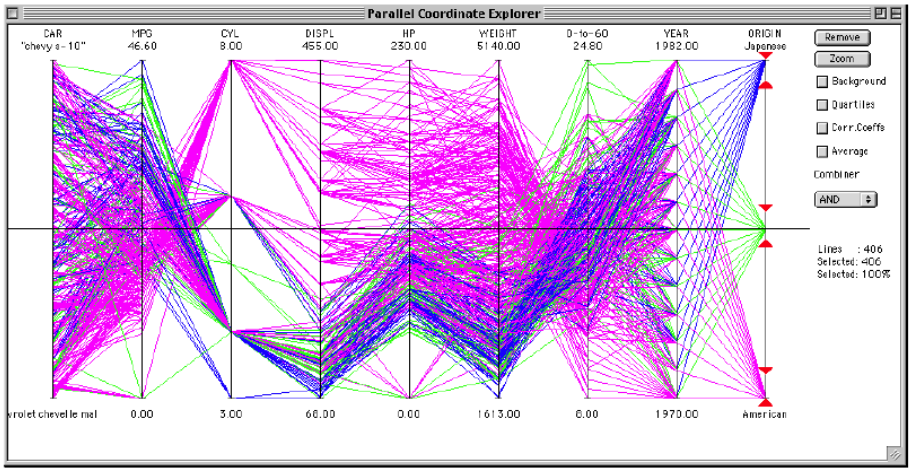
☑ Background
□ Quartiles
☑ Corr.Coeffs
□ Average

Combiner
AND

Lines : 406
Selected: 9
Selected: 2%

Figure 2.10

Correlation indicators in parallel coordinates by Siirtola [135].

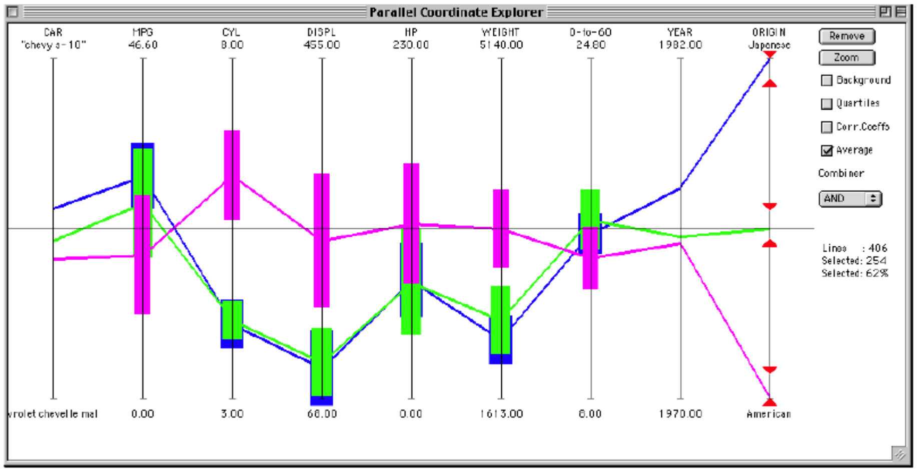


□ Background
□ Quartiles
□ Corr.Coeffs
□ Average
Combiner
AND

```
Lines   : 406
Selected: 406
Selected: 100%
```

Figure 2.11

Siirtola [135] parallel coordinates without polyline averaging.



□ Background
□ Quartiles
□ Corr.Coeffs
☑ Average
Combiner
AND

```
Lines   : 406
Selected: 254
Selected: 62%
```

Figure 2.12

Siirtola's [135] parallel coordinates with polyline averaging.

interactive summarization, Siirtola suggests simple averages of a set of lines to get a quick overview. The standard deviation can also be shown (graphically) to reveal information hidden by the averaging. An example is briefly discussed comparing American, European, and Japanesse cars using normal lines (see Figure 2.11) and averaged lines (see Figure 2.12) for the popular cars data set. Another interesting example is shown by animating the performance and efficiency of American cars in 1970–1982 [135].

In 2001, Andrienko and Andrienko [10] used parallel coordinates dynamically linked to an interactive map by simultaneously highlighting corresponding objects (see black lines in Figure 2.13 that represent smallest percentages of old population). Using data from the 1991 Census of Portugal, the authors focus on the visualization of several comparable attributes. The parallel coordinates plot is linked to the maps is two ways: mouse-over linking and durable selection. The authors extend the parallel coordinates display to provide common axis scales (for the comparable attribute) and vertical alignment of the medians and quartiles of the attributes (a similar plot can be created using the mean and standard deviation). Data is shown in several choropleth maps and a parallel coordinates display. It is noted that with parallel coordinates, prevailing parallel lines indicate positive correlation while diagonal lines may mean negative correlation [10]. Later, Chen and Wang [27] introduced a dimension reduction technique that is based on a genetic algorithm called quad-tree mapping to overcome overcrowded parallel coordinates and other multidimensional technique displays.

In 2002, Heyden et al. [67] integrated parallel coordinates with principal component analysis (PCA) and N-way PCA to analyze large multi-response experimental designs.

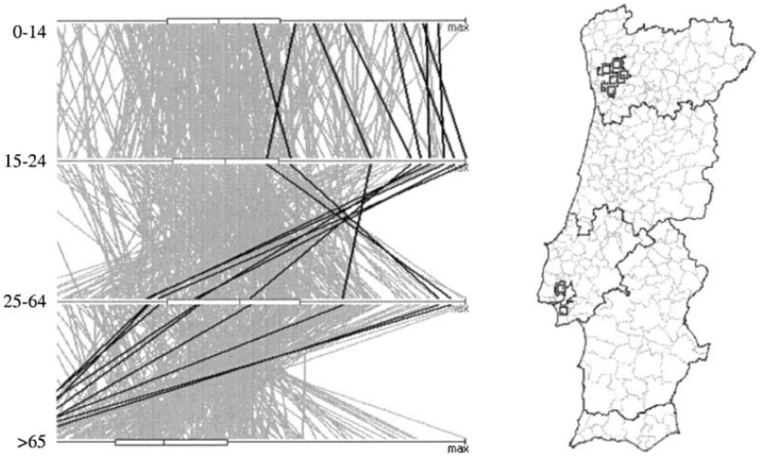


Figure 2.13

Andrienko and Andrienko's [10] parallel coordinates system.

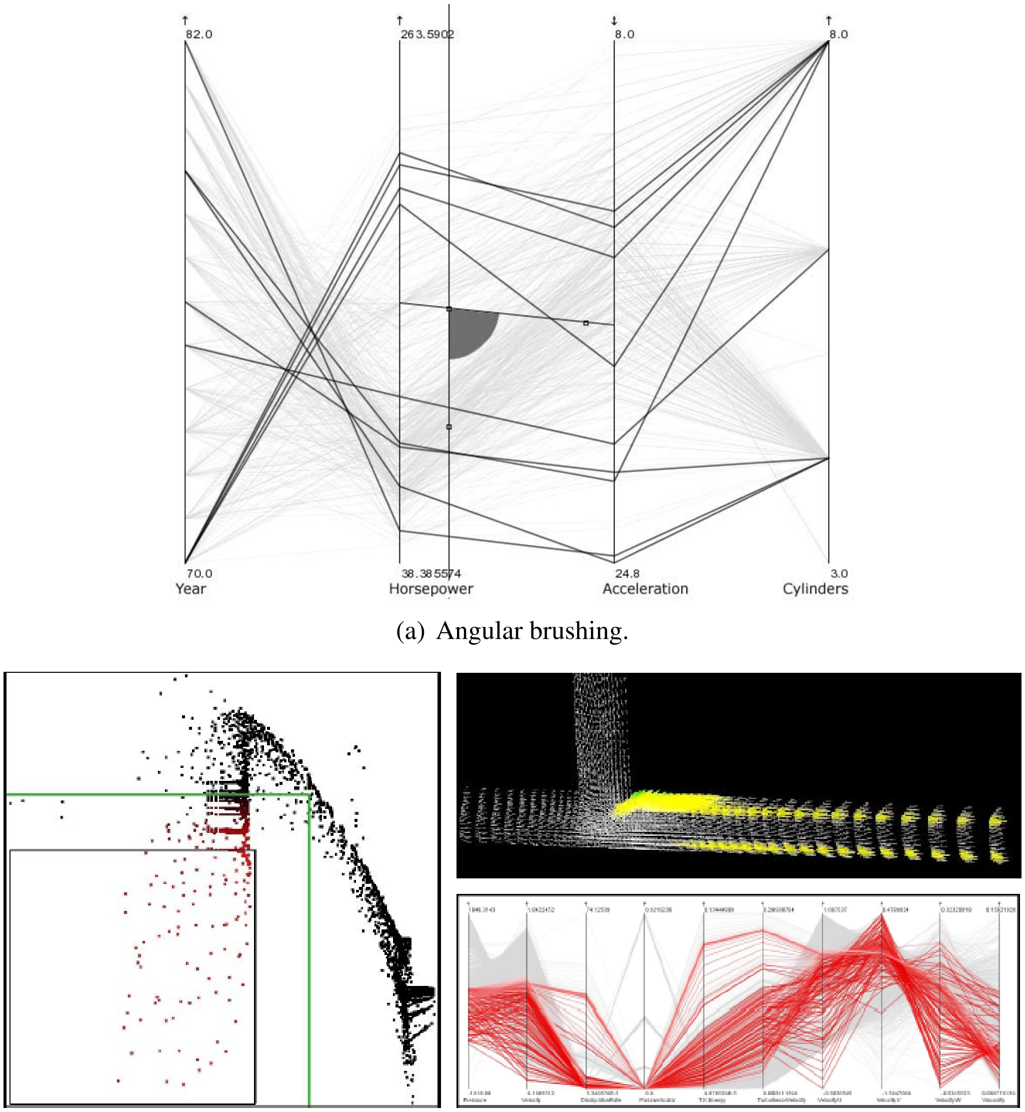
The approach is compared to the traditional method of calculating factor effects by multiple linear regression. The parallel coordinates technique is presented as a useful and complementary aid to interpretation of large multi-response experimental design data. The techniques add a multivariate dimension to the more traditional univariate analysis of such data.

In Hauser et al. [63], an angular brushing interaction and a smooth brushing technique for degree-of-interest functions (focus+context) are described. With angular brushing (see Figure 2.14(a)) the space between the axes can be used to select lines as opposed to the standard brushing techniques that work on the parallel axes themselves. The user can specify a subset of slopes which yields the data points that are to be highlighted; this method can also be combined with composite brushing. Furthermore, brushing and linked displays are discussed in the work. Figure 2.14(b) shows an example of the display in a brushing and linking case study with a scatterplot (left), a linked three-dimensional view (top right), and parallel coordinates (bottom right).

In 2003, Yang et al. [161] introduced a new approach for dealing with high dimensional data called Visual Hierarchical Dimensions Reduction (VHDR). This technique provides hierarchical dimension ordering, spacing, and filtering that is based on the similarities among the dimensions. The technique is also discussed in Yang et al. [160] and Peng et al. [125]. The technique can be used with parallel coordinates and other multidimensional visualization techniques (such as star glyphs and pixel-oriented techniques). Similarly, Zhao et al. [164,165] use an edit-distance based technique to rearrange variables in parallel coordinates to enhance the discovery of interesting patterns.

In research by Graham and Kennedy [55], the traditional polylines in parallel coordinates are replaced with a collection of smooth curves across the axes to overcome line crossovers and the ambiguity they may cause. Furthermore, a focus+context technique that involves the spreading out of points on axes with a few discrete values is introduced. The smooth curves are meant to resolve the difficulty of following lines that share common points on axes—the cross-over problem. The spreading technique is meant to aid in differentiating lines that are bunched close together along their paths [55]

Dykes and Mountain [34] presented an innovative system that provides several linked views of the spatio-temporal data in a geovisualization interface. One display method used in the system is geocentric parallel coordinates. These plots are used to analyze quantitative, multivariate data and relate the statistical and spatial distributions with other dynamically linked views.



(a) Angular brushing.

(b) Linked views.

Figure 2.14

Angular brushing and linked views by Hauser et al. [63].

Brodbeck and Girardin [22] developed an interactive visualization system called *SurveyVisualizer* which aids in the visual analysis of customer satisfaction survey information. The system uses a visual encoding called the Parallel Coordinate Tree which combines the advantages of a familiar tree layout with parallel coordinates. The system also uses a bifocal lens as a distortion effect in the parallel coordinates view [22].

Johansson et al. [91] presented a novel extension to parallel coordinates that uses a classification approach, the self-organizing map (an unsupervised learning algorithm), to create an initial clustering of the data. The authors developed this extension to help with the visualization of large data sets. A zoomable interface and linked views form the foundation of the system. An example molecular data set with 1000 tuples and 16 data items is used to illustrate the approach [91].

Barlow and Stuart [16] introduced a software system that animates the movement of parallel coordinate objects in time. The animated parallel coordinates represent the changing positions of objects within the multidimensional space but viewed in two-dimensional space.

Andrienko and Andrienko [11] describe several modifications to classical parallel coordinates for visually exploring object classes resulting from cluster analysis. Several other multidimensional visualization techniques (scatterplots, table lens, histograms) are also evaluated for use in the visualization system. The paper mentions statistical-based scaling of the axes. In addition, classes are also represented in parallel coordinates with ellipses instead of lines and also as “striped” envelopes. Both methods are based on the partitioning of the value ranges of the attribute into equal frequency intervals. Extensive examples with a demographic data set are given stressing the effectiveness of the proposed extensions.

Artero et al. [14] used parallel coordinates to visualize large data sets which usually result in overcrowded displays. In this publication, a synthetic data set with 7,500 five-attribute records was used for analysis. Algorithms are presented that identify clusters of information using frequency and density information from the data set. Running times for the algorithms and results are presented and the techniques are compared to prior clustering techniques that have been used with parallel coordinate [14].

Notsu et al. [122] developed a visualization technique called *Time-tunnel* to display parallel coordinates. Using data-wings, this tool intuitively displays multiple representations of data. *Time-tunnel* consists of three main types of representation: a *data-wing*, a time-plane, and time-bar. As shown in Figure 2.15, a *data-wing* is a box that has a shape like a sheet. The upper parallel coordinate plot in this example is converted into a *Time-tunnel* plot with three *data-wings*. For multidimensional data, the user can use multiple *data-wings* [122].

Johansson et al. [86] presented a new approach for simultaneously exploring the relationships of a single dimension with many others in the data set (see Figure 2.16). This method is achieved by extending standard parallel coordinates to a 3-dimensional, clustered, multi-relational representation. With this technique the axes are placed on a circle with a focus axis in the center. Furthermore, a technique called relation spacing is used to position the axes according to how interesting the different relations are. A number of interaction techniques are proposed also and the K-means clustering algorithm is used in the

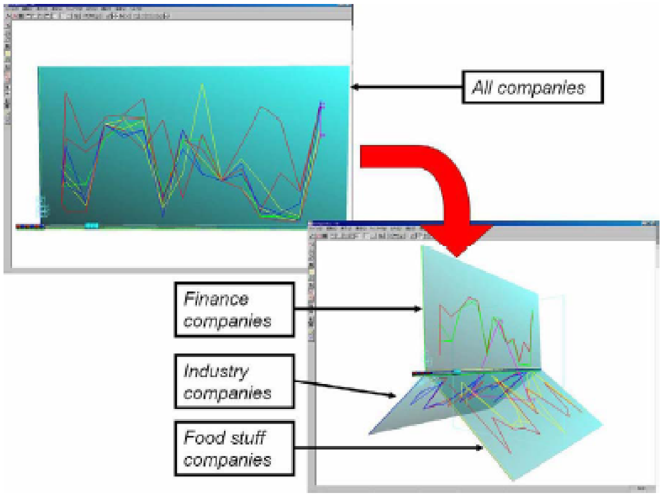


Figure 2.15

The *Time-Tunnel* technique introduced by Notsu et al. [122].

application. Plots in the paper are given based on housing, pollution, and meteorological data sets [86].

Johansson et al. [89] also addressed the visualization of a large number of data elements in parallel coordinates. With parallel coordinates, the overcrowded display is usually too cluttered to find patterns, trends, and relationships. The authors try to solve this problem by using clusters and high-precision textures to represent them more effectively in parallel coordinates. A number of interaction techniques are also discussed to allow investigation of the structure in the clusters. A transfer function is introduced that is used with the parallel coordinates to allow for a more powerful and customized analysis process. In addition, a feature animation technique is described that facilitates the visualization of statistical properties in the clusters. This technique helps the viewer see the skewness and the variance of a cluster which can be used as guidance for starting the analysis. The system

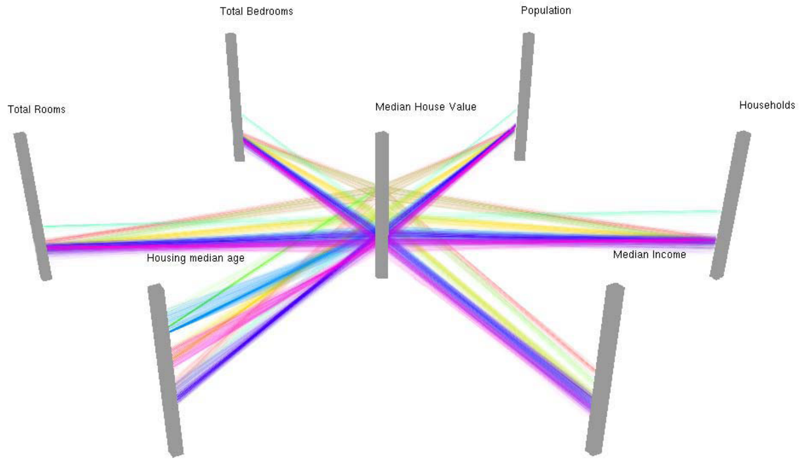


Figure 2.16

A unique parallel coordinates variant introduced by Johansson et al. [86].

also has an outlier enhancement feature that uses the IQR to define an item as an outlier or not; items in the upper and lower 25 percent of the data are considered outliers.

In 2006, Johansson et al. [90] combined the clutter reduction and multi-relational three-dimensional parallel coordinates to address the dense parallel coordinates display. They also described a feature animation techniques to aid in the presentation of cluster statistics. Later, Forsell and Johansson [47] compared multi-relational, three-dimensional parallel coordinates to standard two-dimensional parallel coordinates in a comprehensive user study. Based on the results of this study, they recommend standard two-dimensional parallel coordinates for tasks concerning relationships between data items and multi-relational three-dimensional parallel coordinates for establishing relationships between variables. The results showed some advantages of three-dimensional parallel coordinates, such as the ability to have more relations with a particular axis simultaneously visible, at the cost of having line patterns distorted by perspective effects. The study results suggested that the perspective effects should not be a hindrance to the viewer [47]. Later, Johansson et al. [87] expand this study and from their results concluded that the three-dimensional parallel coordinates with eleven axes is as efficient as standard two-dimensional parallel coordinates in terms of the noise level in the data.

Ericson et al. [38] addressed over-plotting and clutter in parallel coordinates with multiple linked views. The system interface has handles on each axis to crop the data as a form of brushing. Classification of the selected data as clusters or histograms can be turned on and off as layers over the parallel coordinates display. Clusters are calculated using the K-means technique and cluster centroids are re-calculated after every change in the data selection. As shown in Figure 2.17, a linked statistical display is provided which shows statistics for the selected dimensions. The left side of this figure shows the system before brushing and the right side shows it after brushing. In the statistical display, the *y*-axis is the median value and the *x*-axis is the mean value. When selection changes are made, the symbols are animated to the new positions so the user can follow the changes more intuitively. The application is developed in Visual Basic.Net with the OpenViz visualization library and all calculations are performed in real-time using the MATLAB® mathematical programming environment [38].

Bertini et al. [21] described a system called *SpringView* which offers simultaneous viewing of parallel coordinate plots with a new visualization technique called *radviz*. The *radvis* is a two-dimensional visualization in which data elements are drawn on a normalized circle that presents the data dimensions uniformly spaced on its circumference. The

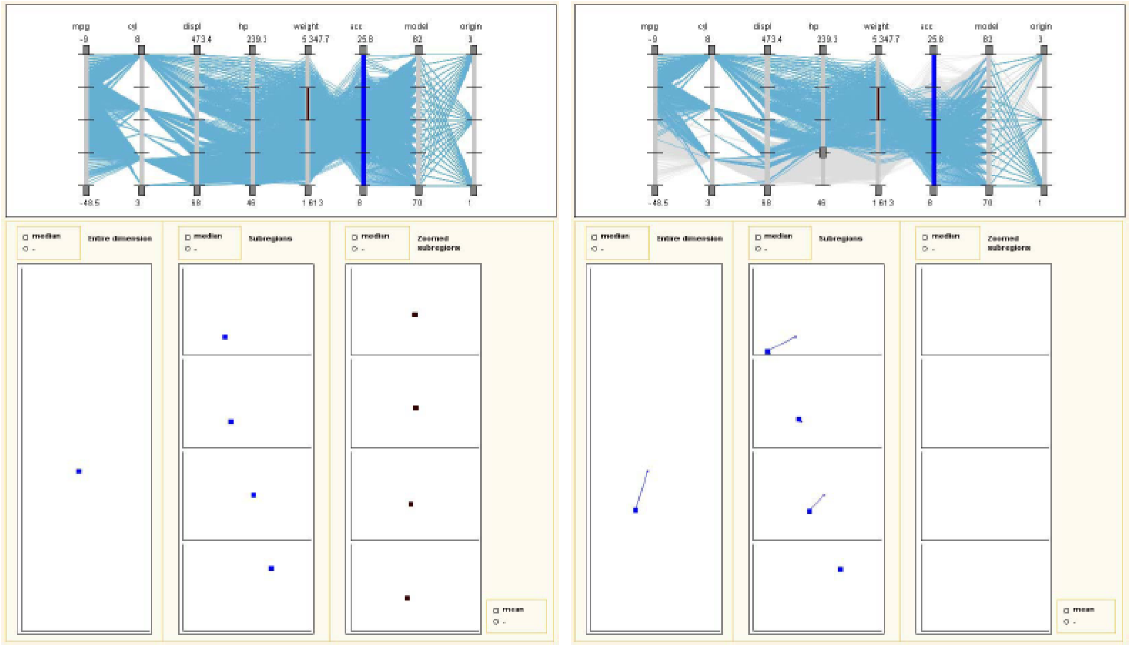


Figure 2.17

The parallel coordinates system interface by Ericson et al. [38].

views are dynamically linked to one another for user interaction. Two data sets were used to assess the effectiveness of the system. The focus of this effort is to address cluttered views in multidimensional visualization. Furthermore, an automated technique is presented for showing similarities or clustering in the data [21].

Fanea et al. [40] combined parallel coordinates and star glyphs in such a way that the advantages of both representations are used to offset their deficiencies when used separately. This new technique is called *Parallel Glyphs* and it extends two-dimensional parallel coordinates into the third dimension and naturally connects them with star glyphs. In addition, color scales are applied to the three-dimensional *Parallel Glyphs* to support comparison and selection tasks in three-dimensions. The authors also demonstrate a number of interesting interaction methods with the new technique to enhance data comparison. One of the interactions is a lens interaction technique [40]. Later, Bendix et al. [19] and Kosara et al. [103] described the concept of *Parallel Sets* which is essentially a modification of parallel coordinates tailored specifically for categorical data (see Figure 2.18).

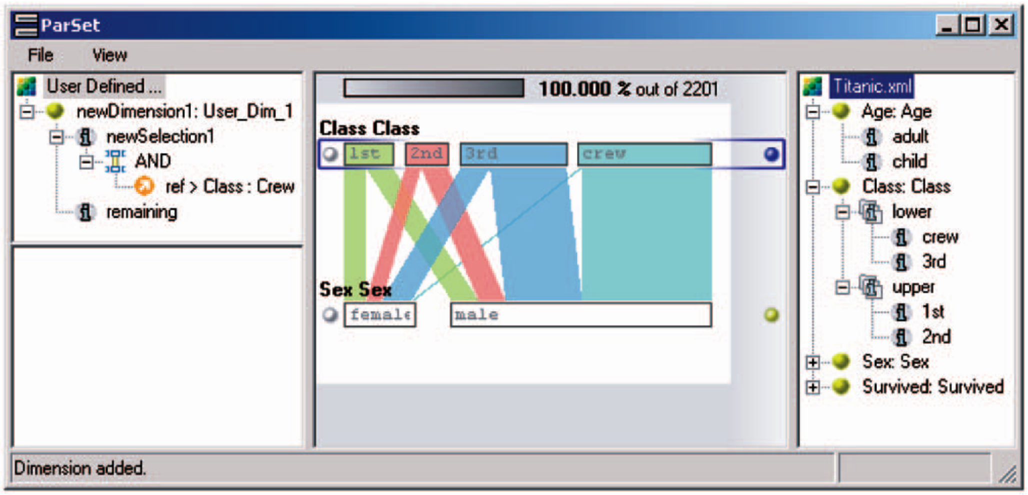


Figure 2.18

Parallel Sets by Bendix et al. [19].

In 2006, Albazzaz and Wang [9] modified parallel coordinates with dimension reduction and upper and lower limits for separating abnormal and normal data in the plots. The dimension reduction transforms the original variables to a smaller number of variables. The upper and lower limits (like percentiles) are drawn as darker polylines in parallel coordinates [9].

Novotný and Hauser [123] offered a new approach to focus+context visualization in parallel coordinates that focuses on being truthful to outliers. The technique enables the context visualization at several levels of abstraction for representing outliers and trends. The technique is applied to data sets with up to three million data records and up to fifty dimensions and an illustration of the workflow of the system is provided [123].

Ellis and Dix [36] described several ways to measure occlusion in parallel coordinates using an implementation called *Sampling Lens*. Three methods are discussed for calculating occlusion: the raster algorithm which rasterizes the lines on a grid and counts the number of plotted points at each grid cell to get an estimate of the over-plotted percent; the random algorithm which treats every plotted point as if it were randomly placed in the viewable pixels and calculates the over-plotted percent using probability; and the lines algorithms that estimates the intersection volumes of all lines crossing the lines. The random algorithm is identified as the best option in terms of accuracy and efficiency [36].

Sifer [134] presented a new visualization method that uses parallel dimension axes. Called *SGViewer*, this technique is compared to parallel coordinates and table-based interfaces in several case studies. As shown in Figure 2.19 the technique is derived from parallel coordinates and *Parallel Sets* [103]. The main difference from *Parallel Sets* is that the color paths have been dropped in an effort to remove clutter. A screen capture from the *SGViewer* is shown in Figure 2.20 with sales data [134].

Guo et al. [58] introduced a geovisualization system, called *VIS-STAMP*, for understanding spatio-temporal and multivariate patterns. As shown in Figure 2.21, the approach involves a self-organizing map, parallel coordinates, several forms of reorderable matrices, a geographic small multiple display, and a two-dimensional cartographic color design method. After clustering the data set with the self-organizing map and assigning colors,

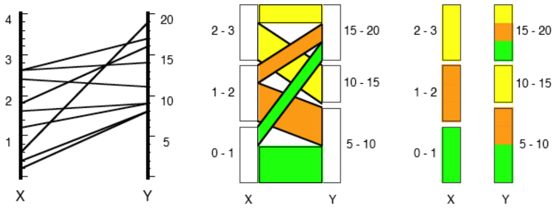


Figure 2.19

Formulation of the *SGViewer* by Sifer [134].

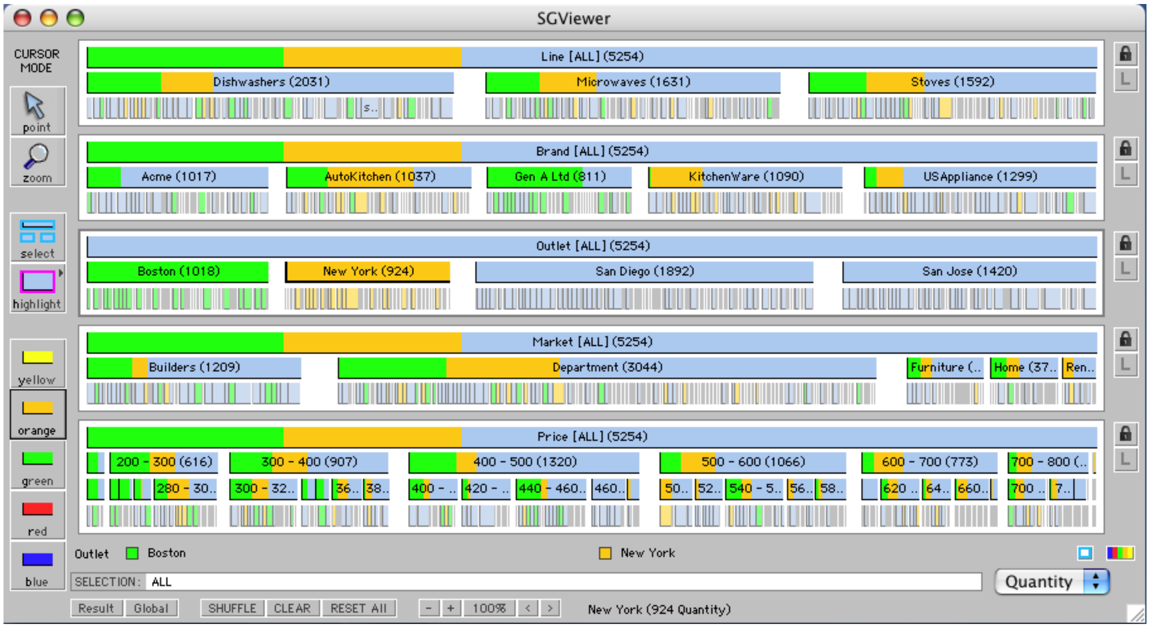


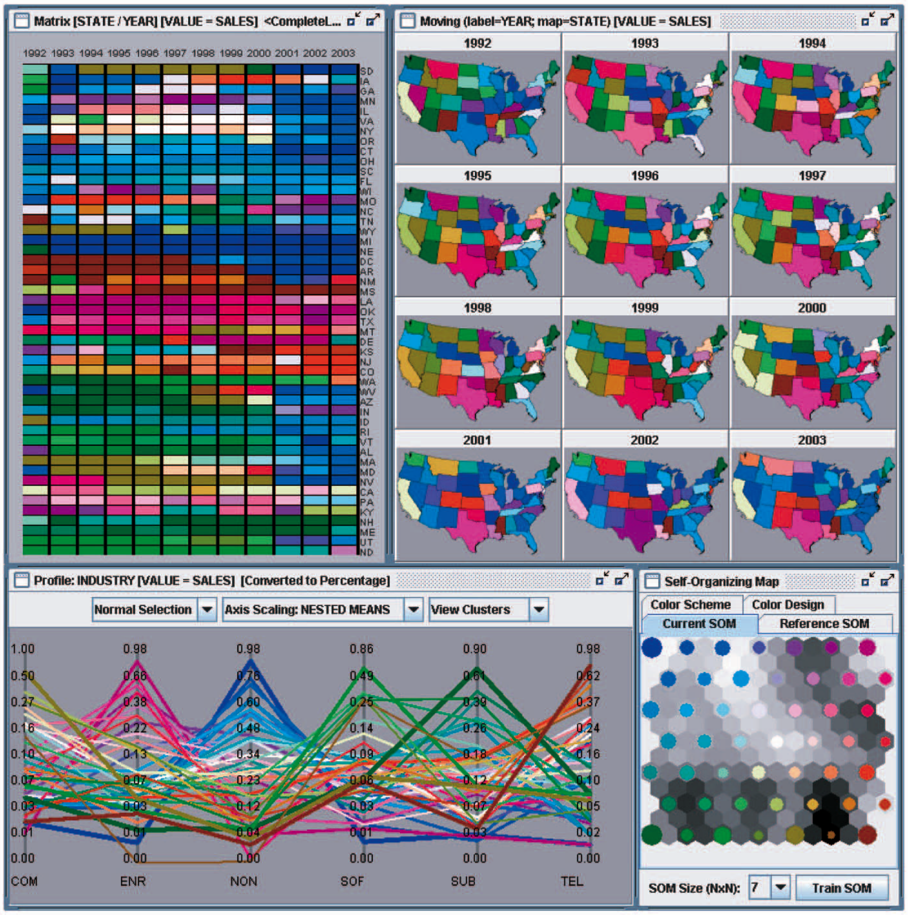
Figure 2.20

The *SGViewer* by Sifer [134].

the parallel coordinates plot is used to show the cluster of multivariate profiles. The parallel coordinates implementation has a number of interesting features. First, the plot uses a nested-means scaling on each axis to alleviate the problem of overlapping lines. Nested-means is a nonlinear scaling method that recursively calculates a number of mean values (and submeans) and uses these values as break points to divide each axis into equal-length segments. Also the thickness of the lines indicates frequency information for the cluster. The application also supports several other scaling methods on the axes. The axes support min-max scaling (using the minimum and maximum data values to linearly scale the axis), cell min-max scaling (using the minimum and maximum cluster mean values) and global min-max scaling (using the minimum and maximum for all variable values) [58].

Xu et al. [158] discussed a visualization technique based on the scatter plot matrix, star diagrams and parallel coordinates concepts. The three coordinate-based geometrical visualizations are combined into a single visualization called the *parallel dual* plot. The new method is created by transforming a scatter plot into a star glyph, then the star glyph is presented with parallel coordinates. The authors claim this technique overcomes the over-plotting problem in standard parallel coordinates [158].

Johansson et al. [88] introduced temporal density parallel coordinates and depth cue parallel coordinates as extensions of two-dimensional parallel coordinates for the analysis of temporal data. The temporal density technique reveals the density of a specified time period and is based on a density map that can be updated. The depth cue method reveals where in time actual data values or changes occur. These techniques use polygons



Normal Selection

Axis Scaling: NESTED MEANS

View Clusters

SOM Size (NxN): 7

Figure 2.21

The *VIS-STAMP* system introduced by Guo et al. [58].

instead of lines to represent temporal changes and have been implemented on the GPU for interactive performance with thousands of data items [88]

Hao et al. [61] described a technique called Intelligent Visual Analytics Query (*IVQuery*) that fuses visual interaction and automated techniques to help the analyst discover special patterns, properties, or relations in the data. The benefits of the technique are demonstrated over traditional zoom and filter techniques using several real world data sets. The *IVQuery* concept is demonstrated with parallel coordinates, visual maps, and scatter plot representations. The technique incorporates ordering by correlation coefficients (Pearson Correlation), offers similarity measures (normalized Euclidean distance), k-means clustering, and nearest neighbor classification. These measures are used in the application to rearrange the visual layout. A case study is presented for each representation technique. For parallel coordinates, the case study demonstrates the pair-wise correlation coefficient calculation of a selected attribute with the other attributes and the automatic reordering of the axes with the results. The reordering places attributes with positive correlations on the right of the selected attribute and negative correlations on the left with descending order [61].

Elmqvist et al. [37] presented a visualization application called *DataMeadow* that is composed of several compact representations called *DataRose*. The *DataRose* provides interactive representations of multiple large-scale multivariate data sets. As shown in Figure 2.22, the *DataRose* is a parallel coordinate starplot of selected data columns with dynamic query sliders integrated into each axis. With the starplot, the parallel coordinates system is folded into polar space and each axis is mapped to the radius of a circle. The

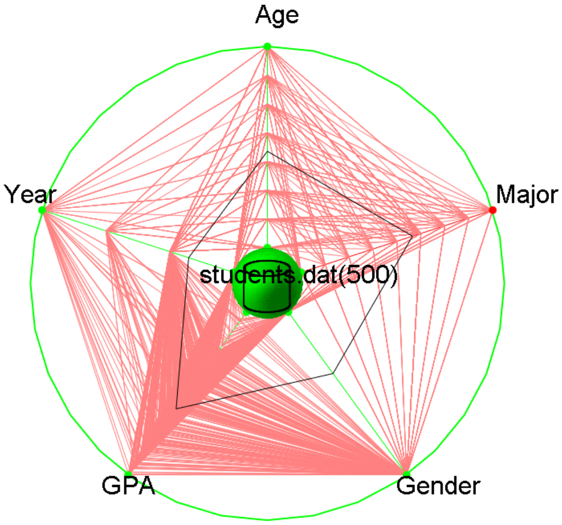


Figure 2.22

The *DataRose* concept by Elmqvist et al. [37].

DataRose offers three representation modes: color histogram, opacity band, and parallel coordinates [37].

Qu et al. [129] used parallel coordinates and polar systems to visualize weather data related to air pollution. The authors present extensions to the standard display axis in parallel coordinates to overcome the inadequacy of a straight axis for encoding wind direction. As shown in Figure 2.23, a S-shaped axis is used to alleviate these problems and attract attention to the wind direction variable which is very important to air pollution analysis. In addition, the system combines parallel coordinates with scatterplots above each axis for accurate quantitative analysis. It is noted that lines in the parallel coordinates become points in the scatterplots. The system uses a weighted complete graph as a guide map for displaying the relationships between dimensions. In the graph, the node is weighted with the correlation coefficients between variables. In addition to providing a way to visualize the relationships, the graph is also used to generate an optimal axis order for parallel coordinates in either an interactive or an automatic mode. The correlation coefficients are also indicated with color in the visualization where red is used for positive values and blue is used for negative values [129].

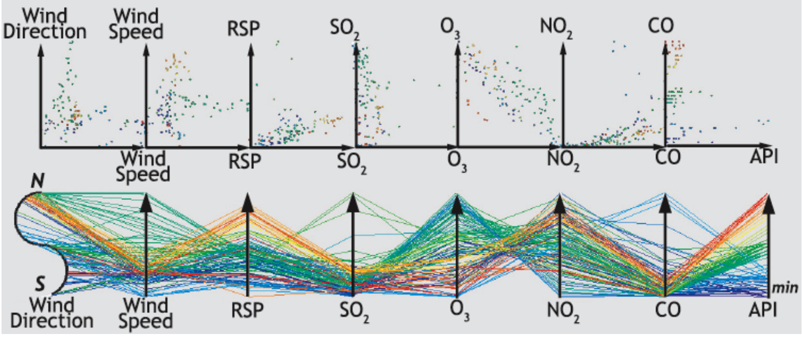


Figure 2.23

The S-shaped parallel coordinates axes introduced by Qu et al. [129].

Kumasaka and Shibata [108] introduced a new variant of parallel coordinates called the textile plot. In the textile plot, the ordering, locations, and scales of the axes are chosen automatically so that the polylines are as horizontal as possible. The suitability of this technique is explored for numerical and categorical data, or a mixture of these different types [108].

Shearer et al. [132] described the application of animation to parallel coordinates with non-time-varying data. The approach is described for a particle physics simulation data set which contains a large number of points [132]. Haroz et al. [62] described the analysis of time-variant cosmological particle data using parallel coordinates. They utilized multiple views for interactive exploration and selection of important features to overcome limited visualization dimensions and facilitate uncertainty visualization in the correlations between variables [62].

2.5.3.3 Applications of Parallel Coordinates

Since its introduction as a method for representing hyper-geometrical surfaces [72], the parallel coordinates technique has been applied to many different domains. In 1995, Martin and Ward [119] used three data sets to evaluate their brushing operations for parallel coordinates: a data set containing information about American, European, and Japanese automobiles manufactured between 1970–1983; a crime statistics data set for Detroit in the years 1961–1973; and a data set with 8000 ore grade values and their positions [119].

In 1998, Ankerst et al. [13] presented an similarity-based axis arrangement approach for parallel coordinates using a data sets of eight different stock prices. In 1999, King and Harris [97] utilized parallel coordinates to visualize pulmonary capillary exchange data. This is one of the first practical uses of parallel coordinates in the medical domain [97].

Goel et al. [53] introduced a visualization tool to aid aircraft designers during the conceptual design stage. In this stage, the design of an aircraft is defined by a vector of 10–30 parameters. The goal of the system is to find the vector that minimizes an objective function while meeting a set of constraints. The tool, called *VizCraft*, allows the viewer to switch between a visualization of the aircraft and a view of the design in the form of a parameter set. The system's parallel coordinates display capabilities allow the designer to compare one design with another using human pattern recognition capabilities [53].

Fua et al. [50] presented two case studies using their hierarchical parallel coordinates technique and tree-maps with a five-dimensional data set with 16,000 elements. This data set was formed by combining Satellite Pour I'Observation de la Terre (SPOT) satellite imagery, magnetics, and radiometrics remote sensing data sets from the Grant's patch region of Western Australia [50].

Hall and Berthold [59] employed parallel coordinates to visualize fuzzy data. Fuzzy data may consist of fuzzy rules, which can be viewed as cutting a swath through an *n*-dimensional space. Three examples are given to show the utility of this application. The data set used is the well-studied Iris plant database which consists of 150 examples each with four features which describe three types of Iris classes [59]. Berthold and Hall [20] later revisited the fuzzy parallel coordinates system using the same flower data set and a new ocean satellite image data set.

Siirtola [135] used the American Statistical Association (ASA) cars data set which has nine dimensions (two dimensions are categorical), 406 polylines, and 3,654 data items. The cars data set stores all cars road tested by the *Consumer Reports* magazine between 1971 and 1983 [135].

In 2001, Lee et al. [115] investigated the use of two different visualization techniques to analyze real-world, clickstream data for online retail sales. Session data are handled with parallel coordinates and product performance data are visualized using starfield graphs [115]. Later, Spraragen and Podlaseck [139] used an extended version of parallel coordinates to browse many hundreds of musical works. In addition to describing the system's interface design, the authors present an initial usability study [139].

Falkman [39] introduced a new three-dimensional parallel coordinates technique, called *The Cube*, to support clinicians in daily diagnostic work. The technique is represented by three-dimensional parallel diagrams with a linked statistics panel [39]

In 2002, Hauser et al. [63] described applications of parallel coordinates to computational fluid dynamics and the cars data set. In 2003, Yang et al. applied their *VHDR* system to a real world data set derived from part of the unweighted PUMS census data of the Los Angeles and Long Beach area for the years 1970, 1980, and 1990. The data set has 42 dimensions and 20,000 elements [161].

Siirtola [136] integrated parallel coordinates with a visualization technique called the *Reorderable Matrix* in a linked display. These two display techniques are analyzed and the results are reported of an experiment that compares the participants' task performance with the two views, with and without linking. The experiment showed that the view linking slows the task performance but it accelerates learning and is well received by the users [136].

Unwin et al. [148] utilized parallel coordinates for exploratory modeling analysis—evaluating and comparing many models simultaneously. The target data set comes from a trial on a treatment for primary biliary cirrhosis of the liver. There were 418 patients and seventeen potentially explanatory variables. Using standard interactive parallel coordinates, the authors described an in-depth study of how the tool is used to discover patterns in the data set [148].

Friendly and Kwan [48] discussed a framework for ordering information in visual displays. The authors developed several principles for ordering information and applied them to parallel coordinates and several other display techniques. The authors noted that parallel coordinates work well for well-structured data but it is sometimes disappointing with real data. A number of parallel coordinate plots are shown in this paper, most of which come from a crime statistics data set [48].

Edsall [35] discussed implementations of interactive parallel coordinates showing spatial and spatio-temporal data with linked maps and scatterplot views. The plots are demonstrated as an effective data exploration tool through case studies in two different problem domains: climate modeling and analysis and epidemiology. These domains are similar because researchers in both areas search for patterns and trends across space and time. The climate study system was constructed in Tcl/Tk and linked to IBM's Data Explorer for use with climatological data. The epidemiology system was developed with ArcView's GIS system using its scripting language Avenue [35].

Potts et al. [128] and, later, Tory et al. [143], applied parallel coordinates to the layout of the parameter space used for volumetric rendering. The variables include camera orientation, transfer functions for color and opacity, zoom and translation of the view, a volumetric data file, and a rendering technique. In this layout, all parameters are explicitly represented to illustrate the space of available options for volume rendering. The system also features a history bar to allow users to backtrack to previous states and quickly scroll to see when options have been tried. Usability testing has shown that the tool is promising for the exploration of volume data [128, 143]. Crider et al. [31] described a port of the interface introduced by Tory et al. [143] that allowed it to be controlled by physical sliders on a mixer board instead of graphical widgets. Based on a user study described in the paper, the mixing board with motorized sliders seems like a promising interaction device for a variety of visualization applications [31].

Zhao et al. [164, 165] demonstrated their parallel coordinates tool, called *V-Miner*, along with a case study about Motorola engineers' use of the tool to find significant patterns in their product test and design data. The data set used in this analysis is from an extensive set of tests by Motorola engineers on a new type of mobile phone and it consists of over 100 test variables that characterize the performance of the mobile phone. The tool was used by Motorola engineers for about three months and they reported that it helped them find important patterns and information about the test data [164, 165].

In 2004, Schneidewind et al. [131] used parallel coordinates as one of several visualization techniques to analyze image retrieval results. The system provides several linked views where the parallel coordinates display is the preferred technique to analyze features in detail [131].

Later, Johansson demonstrated a parallel coordinates extension by analyzing a molecular data set with 1000 tuples and sixteen data items [91]. Barlow and Stuart tested a parallel coordinates system with neurophysiological research data sets [16]. In addition, Andrienko and Andrienko provided examples of parallel coordinates plots from a demographic data set [11].

Wang et al. [152] used parallel coordinates in a visual data mining tool, *VisDM-PC*, that can interactively perform data classification, relativity analysis, association rule analysis and implement the roll-up and drill-down function. The system functions were tested using several data sets. A data set of Chinese 2000 CET (College English Test) scores which contains six attributes was used in the examples. The data set consisted of 246 scores. The application features are introduced along with patterns in the data set to demonstrate its effectiveness [152].

In 2005, Yang [162] visualized association rules using smooth polylines in standard parallel coordinates. Later, Krasser et al. [107] applied parallel coordinates to the analysis of network traffic information.

Lanzenberger et al. [111] compared the stardinates and parallel coordinates techniques for visualizing psychotherapeutic data. A comparative study with twenty-two participants revealed that stardinates were a more appropriate method for interpreting highly structured data in detail and parallel coordinates showed advantages for gaining insight on the first glance [111].

Johansson et al. [86] presented extended parallel coordinate plots based on housing, pollution, and meteorological data sets. Soon thereafter, Grünfeld [57] used parallel coordinates and scatterplot display techniques to visualize data that describes the contents of the metals Cu (copper), Ni (nickel), Pb (lead), V (vanadium) and Zn (zinc) in mosses within an area of 300 $\times$ 300 km in southern Sweden, sampled in 1985 (177 samples), 1990 (156 samples), and 1995 (188 samples). The freeware visualization package *XmdvTool* was used in the visual analysis. In addition to highlighting several interesting findings in the data, the author discussed the advantages and limitations of the visualization techniques compared to histograms, quartile plots, and proportional symbol maps [57].

Feldt et al. [41] described a multiple linked view application (one view was parallel coordinates) and its use in the analysis of a statistical database for Sweden. The application was called the *GeoWizard* and it was a distributed application that provided a parallel coordinates display for an overview along with a scatter plot matrix. In addition, a choropleth map and the 2D scatter plot provided more detailed views of the data [41].

Ericson et al. [38] evaluated extended parallel coordinates with three different data sets: the cars data set, a pollution data set, and a stock market data set. For each data set, plots were shown and an evaluation of the use of the interface was given for investigating specific features of the data set. A good workflow was given that demonstrates patterns in the data and how the capabilities of the application helped in the discovery of the patterns [38].

Matković et al. [120] addressed the visualization of data from injection system simulations in an innovative system called *ComVis*. *ComVis* supported multiple linked views and common information visualization displays such as scatterplots, histograms, parallel coordinates, etc. [120]. The *ComVis* system was expanded later by Konyha et al. [102] and an additional case study of a road traffic data set was described. The focus of this paper was on the visualization of data sets that include families of function graphs [102].

Fanea et al. [40] presented a case study of their technique called *Parallel Glyphs* using a set of one hundred generations of plants generated by a genetic algorithm, each having five attributes. Later, Bendix et al. [19] and Kosara et al. [103] used sales and marketing data and a demographic data set from the Titanic disaster for case studies of their *Parallel Sets* techniques that was derived from parallel coordinates.

In 2006, Karki et al. [95] applied a new visualization system that uses parallel coordinates with star plots, scatter plots, and polygon-surface rendering techniques for exploring large collections of mineral elasticity data. Dwyer et al. [33] discussed the analysis of network-structured data using centrality analysis. The authors presented three methods for exploring and comparing centrality measures within a network: three-dimensional parallel coordinates, orbit-based comparison, and hierarchy-based comparison. These methods were demonstrated in case studies using biological and social network data sets [33]. Later, Albazzaz and Wang [9] demonstrated parallel coordinates extensions in statistical process control.

Pillat and Freitas [126] discussed a multiple coordinated view interface that was designed to visualize multidimensional data. The *InfoVis Toolkit* was used in the system implementation and one of its views provided a parallel coordinates display. The authors described an informal experiment and presented plots using the cars data set [126]. Lawrence et al. [112] presented a visualization system called *exploRase* which provides parallel coordinates and other methods in interlinked displays for the exploratory analysis of systems biology data.

Jern and Franzèn [84] introduced a parallel coordinates display integrated with time series and trend graph displays that serve as a visual control panel for the *GeoAnalytics* application. The authors presented a case study using Sweden's statistical databases, which contains economic, social, and demographic information. In addition to the standard parallel coordinates view, the system provided five other linked views of the data set [84].

Kraak [105] focused on the new role of maps in the analysis of today's complex data; a new breed of map visualizations that operate in the realm of geovisualization. A mock geovisualization was developed to represent the events of Napoleon's campaign. In the initial exploration and discovery phase, the author suggested the use of a parallel coordinates display for an overview of the data set [105].

Ye and Lin [163] used parallel coordinates to speed up the convergence rate of simulated annealing for moderate and high dimensional optimization problems. Rather than polygonal lines, smooth curves were used in the parallel coordinates display and several numerical studies were presented to demonstrate the improved performance of the technique [163].

Bair et al. [15] presented information on perceptually optimal visualizations of layered three-dimensional surfaces. The authors presented guidelines for generating texture patterns that minimize confusion in depth discrimination and maximize the ability to find distinct features. The authors used parallel coordinates in conjunction with Analysis of Variance, Linear Discriminant Analysis, and Decision Trees to analyze the data from human in the loop experiments. The parallel coordinate plot was identified as a convenient way to visually assess hypotheses about relationships among the parameters [15]. Later, Sifer presented a parallel coordinates based visualization method using a sales data set and a network traffic data set [134].

Singer et al. [138] used parallel coordinates to analyze the structure and functionality of communication networks. A network of n nodes were transformed to n points in an n-dimensional space. The authors began with visualization of subgraphs of the Internet AS graph and continued with the visualization of networks at the IP level. Using parallel coordinates, the authors showed how they can identify network properties such as stability in instances of node-link failures, node back-up, node interdependence, and unique topological patterns common in networks [138]. Guo et al. [58] presented a case study of their system called *VIS-STAMP* using a sales data set which includes sixteen industry types for forty-nine states and twelve years.

Siirtola and Räihä [137] surveyed interaction techniques for parallel coordinates and compared them to established visualization design guidelines. The authors also described their experiences with several prototype parallel coordinates applications and an experiment to study the usability of parallel coordinates. In this experiment, information technology professionals were asked to answer questions about the ASA cars data set using SQL query tools and a parallel coordinates application. The results of this study suggested that parallel coordinates user interfaces were not as difficult to use as generally believed. Although the accuracy of the answers was about the same between the two methods, the parallel coordinates method was substantially faster in solving the set of tasks.

In 2007, Park and Martin [124] utilized parallel coordinates to assess the reusability of waste (any solid, liquid, or contained gaseous substance arising from the application of a process) using several case studies. Later, Caat et al. [23] addressed the visualization of time-varying multichannel electroencephalography (EEG) data. The system uses a tiled organization in the form of a two-dimensional row-column presentation instead of the one-dimensional arrangement of columns that is used in classical parallel coordinates. There was one tile for every electrode and each tile displayed a combined minmax plot, density map, and parallel coordinates plot. The method was compared to existing EEG visualization methods based on a number of criteria. Also, a user evaluation was described in which the method was compared to traditional EEG visualization techniques. For the experiment task, the new visualization method was about 40% faster than the standard visualization method. The speed gain was without loss of information and the new method used less space than the standard visualization method. The new method did not show decreases in speed with increased amounts of information but the standard visualization method had a decrease in speed with an increase of data. The new method was recommended for use in studying healthy people and in clinical settings [23].

Xu et al. described the use of a parallel coordinates extension called *parallel dual* plots to the analysis of a vegetable oil data set [158]. Later, Xu et al. [159] focused on the capability of parallel coordinates to assist in visual pattern recognition and classification using the same vegetable oil data set. The system described in this paper provides three interaction operations: fading or exposing, translation, and zooming [159].

Jern et al. [85], described their use of the *GeoAnalytics* toolkit which includes a parallel coordinates display with standard interaction capabilities. A sample application was described called *GeoWizard* which used parallel coordinates with embedded visual inquiry methods that serves as the visual control panel for dynamically linked and coordinated views [85].

Jern [83] also described an approach called *Visual Space Management* which incorporated multiple linked views and explored retail data related to space performance. That is, the tool was intended to help retailers gain a better understanding of each store's layout in relation to its capacity and performance. The system integrated several information visualization techniques and a three-dimensional interactive layout of the store floor plans with retail data sources. Built on the *GeoAnalytics* toolkit, the parallel coordinate plots provided a multivariate visual control panel in the coordinated system [83].

Henley et al. [66] utilized parallel coordinates and scatter plots to compare nucleotide sequences for genomic study. A user study was conducted to evaluate the advantages and disadvantages for each visualization technique. All subjects indicated that parallel coordinates caused confusion from the crossing lines and subjects also noticed that the parallel coordinates showed long common sequences well. Overall the subjects in this study preferred the scatter plot to the parallel coordinates display [66].

A.Godinho et al. [7] presented a system called *PRISMA* that explored the use of multiple coordinated views to visualize multidimensional data sets. The system provided treemap, scatterplot, and standard parallel coordinates displays. The authors presented the results from a usability study which involved eleven users [7].

Elmqvist et al. [37] presented a case study of the *DataRose* technique using U.S. Census data. In this case study, a convincing workflow was described that followed the use of the tool by a fictitious analyst searching for new patterns in the data. Also the authors presented the results of a user study (a qualitative expert review) that involved two visualization researchers. The study followed a think-aloud protocol and the authors noted the feedback from the researchers [37].

Hao et al. demonstrated the *IVQuery* technique using real world data sets such as data warehouse performance, product sales, and server performance [61]. Qu et al. [129] presented a case study of their extended parallel coordinates system using an air pollution data to demonstrate the improved effectiveness at correlation detection, finding similarities and differences, and time series trends. The authors found that the effectiveness of parallel coordinates depends on the axis order. They also noted that parallel coordinates worked well for showing general correlations and the polar system worked well for more quantitative analysis. They also briefly mentioned some feedback on the system from domain scientists [129].

Chang et al. [26] introduced an interactive urban information visualization tool that provided continuous levels of abstraction. The system also provided multiple data views, one of which is a parallel coordinates display that was color-coded to match a matrix view. A user study was described, feedback on the system was reported, and a case study was presented with demographic data from the 2000 U.S. Census [26]. In 2008, Jones et al. [93] described a data exploration system for time-varying, multivariate, point-based data from gyrokinetic particle simulations. In this system, a parallel coordinates view provided a global overview of the data [93].

CHAPTER 3

APPROACH

In this research, geovisual analytics have been brought to bear in the domain of environmental data analysis to facilitate more effective knowledge extraction than traditional approaches—the type of research that has long been considered the central promise of visualization. The research approach is motivated by the following hypotheses:

1. The development of an advanced geovisual analytics approach using parallel coordinates and statistical techniques reveals a deeper level of understanding than traditional methods when applied to the task of finding complex multivariate trends in environmental data sets. With new ways to creatively explore the data, the approach offers a more effective visual interface to glean new insight about the data behind the visualization.

2. The effectiveness of the geovisual analytics approach is necessarily explored in the context of practical environmental studies, which are grounded in real-world data sets instead of invented or abstract data sets, in close collaboration with domain experts. The discovery of new associations and the confirmation of known patterns by domain experts will validate the promise of this new approach in environmental data analysis.

Traditional climate study workflows use statistical processes to automatically discover significant predictors using historical data. However, the lack of adequate visual analysis tools in the realm of environmental data analysis forces the scientist to reduce the problem to fit the tools. We have discovered new insight by expanding the tools to facilitate exploratory visual analysis of the associations in the data. Moreover, the coupling of statistical data analysis processes directly into the visual interface enables faster, more accurate discovery and confirmation of patterns in the data.

In the remainder of this chapter, we provide a detailed overview of the various capabilities of the new geovisual analytics systems, which was produced after our comprehensive investigation of prior approaches presented in Chapter 2. The new system combines several fundamental parallel coordinates capabilities and variants of more advanced techniques from prior works. The system also offers new interactive functionality with parallel coordinates: dynamic axis scaling using mouse wheel movement and continuous aerial perspective shading of polylines. These techniques are used in Chapter 4 to demonstrate the enhanced visual data analysis capabilities in four separate case studies. The new insight obtained from these evaluations validated the promise of this approach in environmental data analysis. Specifically, we developed a deeper level of understanding about the physical associations of global signals for seasonal North Atlantic tropical cyclone activity in the latter three case studies.

3.1 System Overview

An innovative geovisual analytics system called MDX has been developed that combines interactive parallel coordinates with automated statistical processes to provide a practical tool for analyzing multivariate data sets. MDX was developed using the Java Development Kit version 1.6 and its Advanced Windowing Toolkit and Swing class libraries. From a software engineering perspective, the system consists of 6,766 Total Lines of Code[1] and 5,360 Method Lines of Code[2] in thirty-five classes. The system provides interactive performance on a laptop computer with a 2.33 GHz Intel Core 2 Duo processor, three GB Random Access Memory (RAM), and an ATI Radeon X1600 graphics card with 256 MB Video RAM.

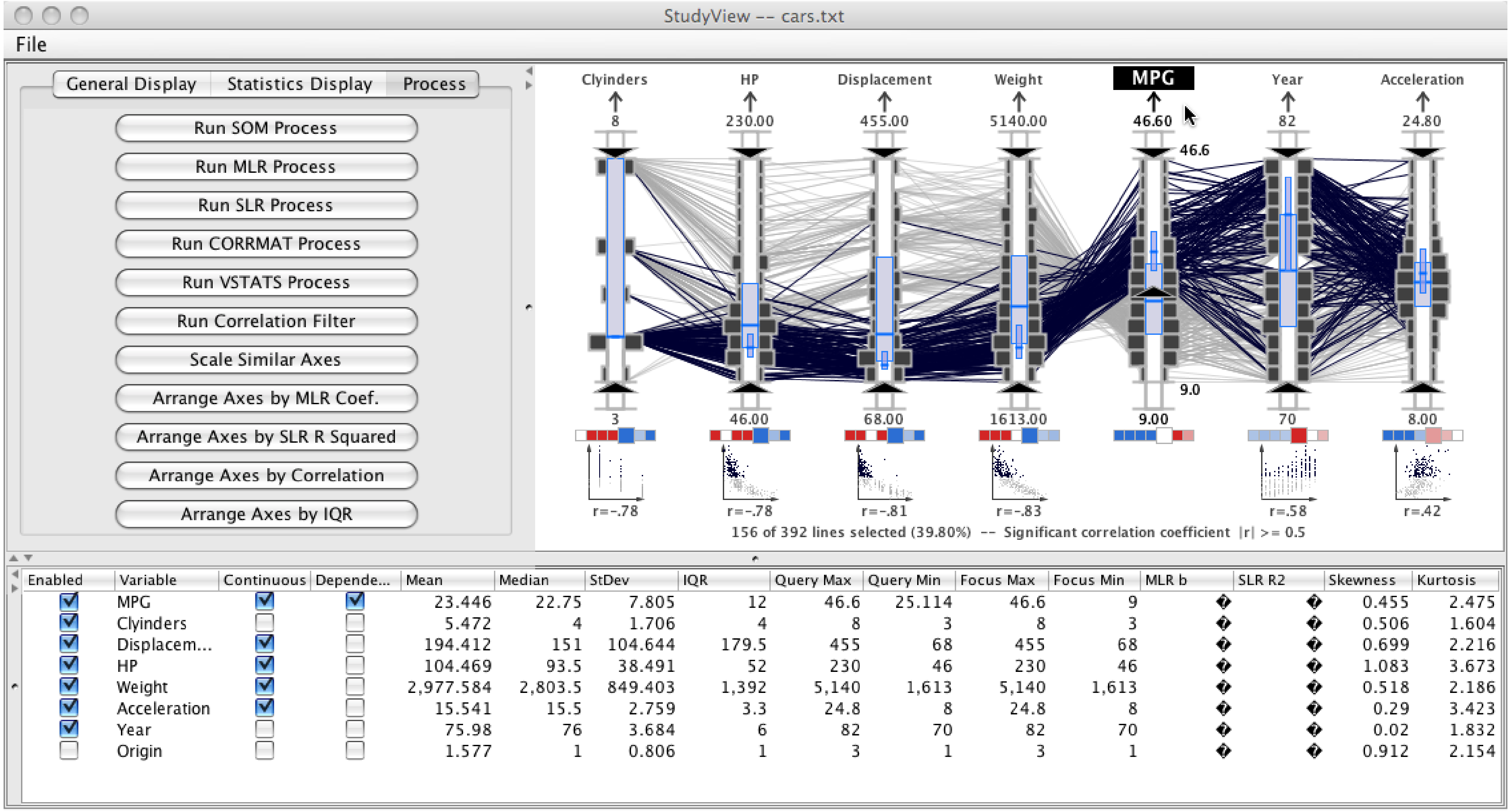
As shown in Figure 3.1, MDX provides an efficient graphical user interface (GUI) that offers a settings panel (upper left panel), an interactive table view of axis settings and statistics (lower panel), and an enhanced parallel coordinates view (upper right panel). Although the table and settings panels are critical for the usability of our system, the parallel coordinates panel is the heart of the visual analysis capabilities. In this panel, the classical parallel coordinates plot is extended with dynamic interaction capabilities that provide access to the data behind the visualization. Furthermore, the parallel coordinates view is dynamically linked with statistical indicators and automatic statistical processes to provide an ideal environment for exploratory data analysis.

3.2 Visualization Capabilities

The visualization capabilities of the system are contained in the parallel coordinates panel. In addition to many fundamental parallel coordinates capabilities such as relocatable axes, axis inversion, and details-on-demand, this panel provides several innovative interaction capabilities such as axis scaling (focus+context), aerial perspective shading,

[1] The TLOC counts non-blank and non-comment lines in a compilation unit.

[2] The MLOC counts non-blank and non-comment lines inside method bodies.



156 of 392 lines selected (39.80%) -- Significant correlation coefficient |r| >= 0.5

Enabled	Variable	Continuous	Depende...	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	25.114	46.6	9	�	�	0.455	2.475
☑	Clyinders	☐	☐	5.472	4	1.706	4	8	3	8	3	�	�	0.506	1.604
☑	Displacem...	☑	☐	194.412	151	104.644	179.5	455	68	455	68	�	�	0.699	2.216
☑	HP	☑	☐	104.469	93.5	38.491	52	230	46	230	46	�	�	1.083	3.673
☑	Weight	☑	☐	2,977.584	2,803.5	849.403	1,392	5,140	1,613	5,140	1,613	�	�	0.518	2.186
☑	Acceleration	☑	☐	15.541	15.5	2.759	3.3	24.8	8	24.8	8	�	�	0.29	3.423
☑	Year	☐	☐	75.98	76	3.684	6	82	70	82	70	�	�	0.02	1.832
☐	Origin	☐	☐	1.577	1	0.806	1	3	1	3	1	�	�	0.912	2.154

Figure 3.1

MDX user interface.

Table 3.1

MDX visualization capabilities.

Capability Name	Description
Dynamic Axis Scaling	Interactively scale (zoom into) an axis.
Aerial Perspective	Encode proximity with line shading.
Dynamic Visual Queries	Conjunctive queries with UI widgets.
Statistical Indicators	Visually encode analysis results.
Interactive Scatterplots	Precise analysis of variable relations.

and dynamic visual queries. In this section, we highlight these visualization capabilities which are also summarized in Table 3.1.

3.2.1 Classical Parallel Coordinates Interaction Capabilities

Over the years, there have been many innovative extensions to the original parallel coordinates plot that have greatly increased its usefulness. Several of these extensions were integrated into the MDX system. Perhaps the most fundamental extension is the ability to quickly reorganize the axes [135]. This capability allows the viewer to quickly rearrange the axes to explore new relationships which is particularly beneficial as the dimensionality of the data increases. In MDX, the viewer can click the left mouse button on the axis name to select and drag the axis to a new position. When the viewer releases the mouse button, the axes will be automatically reordered so that the axis being moved is inserted at the new location and the other axes are rearranged appropriately.

Axis inversion is another fundamental feature in the parallel coordinates display [63]. In our system, the viewer can invert an axis by left clicking on the arrow at the top of the

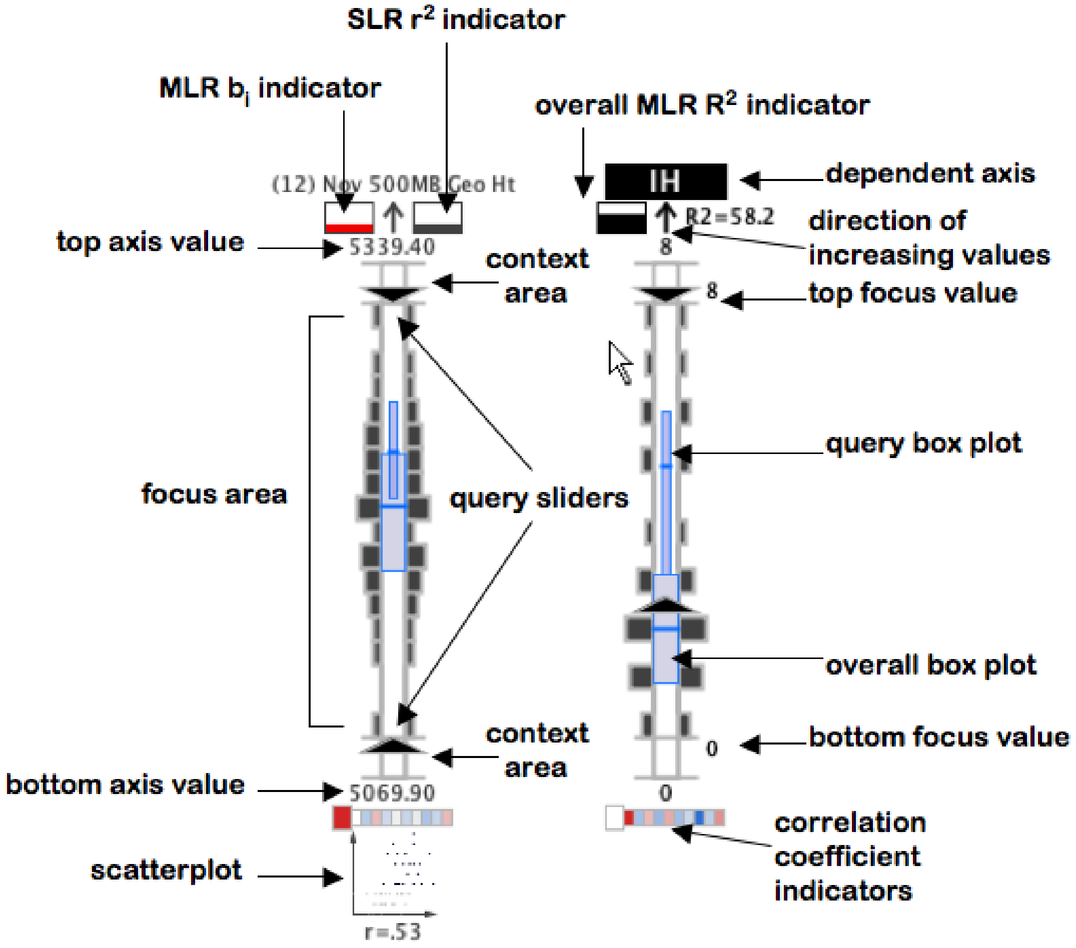


Figure 3.2

Annotated view of the interactive MDX axis widgets.

axis (see Figure 3.2). When this arrow is clicked the top and bottom values for the axis are switched and the display is regenerated. In addition to serving as a button for switching the axis between ascending and descending order, the arrow also indicates the direction of increasing values on the axis.

As illustrated in Figure 3.3, MDX also provides a basic details-on-demand capability that gives the viewer the ability to click on an axis with the middle mouse button to display the axis value under the mouse [63]. In this example, the query reveals a value of 39.661903. The right side of Figure 3.2 shows an additional details-on-demand feature whereby the application displays the values for the top and bottom of the focus area and applies a more prominent highlight color to the axis whose area is intersected by the mouse position.

Additionally, our system's display can be customized through an intuitive pop-up menu. That is, when the viewer clicks the right mouse button in the display, a pop-up menu is revealed which can be used to control many features such as the display of statistical indicators, color schemes, display of tick marks, or performing screen captures. Most of these same settings and functions can be accessed via the settings panel (see Figure 3.1).

3.2.2 Dynamic Visual Queries

Since the viewer is often interested in grouping subsets of data, our application also provides a method to select lines using double-ended sliders. As shown in Figure 3.2, each axis has a pair of sliders which define the top and bottom range for the query area. This capability is an extension of prior research on dynamic interaction techniques [8,

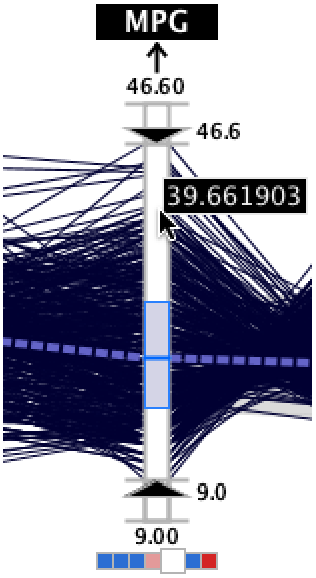


Figure 3.3

MDX's details-on-demand capabilities.

133, 146], particularly those focused on parallel coordinates [63, 135]. The viewer can drag these sliders to dynamically adjust which lines are highlighted effectively giving the viewer the capability to perform rapid, conjunctive queries. Lines within the query area of each visible axis are rendered with a more prominent color while the remaining lines are rendered with a less prominent shade of gray.

In Figure 3.4, an example visual query is shown using with a seasonal tropical cyclone statistics data set. In this example, the *Named Storms (NS)* and *Intense Hurricanes (IH)* axis sliders have been set to highlight the two years with an above normal number of named storms and a below normal number of intense hurricanes for data between 1950 and 2006. In Figure 3.1, another visual query example is shown with the popular American

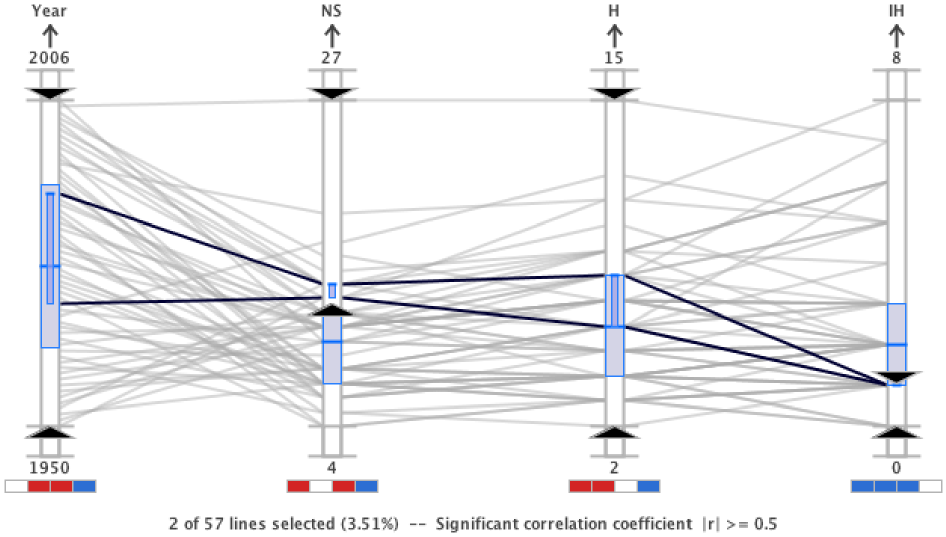

Year
2006
1950
NS
27
4
H
15
2
IH
8
0
2 of 57 lines selected (3.51%) -- Significant correlation coefficient |r| >= 0.5

Figure 3.4

Dynamic conjunctive query capabilities using MDX.

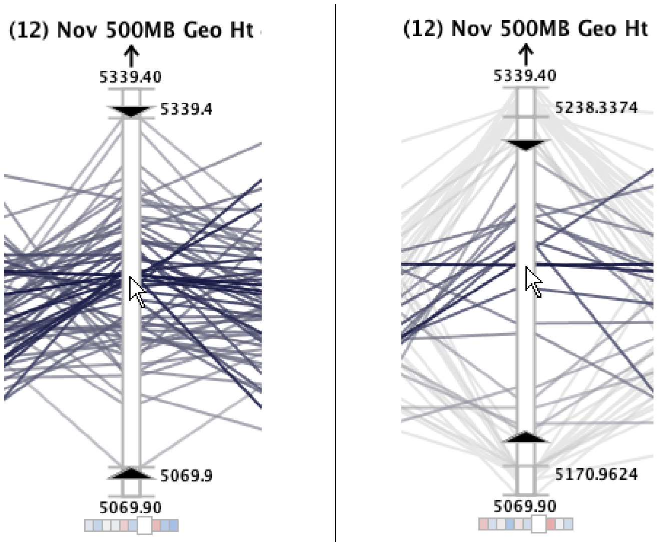
Statistical Association (ASA) cars data set[3]. In this figure, the sliders on the *Year* axis have been adjusted to highlight the more recent model year records.

3.2.3 Axis Scaling (Focus+Context)

MDX's dynamic axis scaling capability provides a method to interactively tunnel through the data until a smaller subset of the original data is in focus. MDX allows the user to modify the minimum and maximum focus area values for a selected axis using mouse wheel movement. This capability builds on other parallel coordinates focus+context implementations presented in research literature [14, 50, 89, 123].

As shown in Figure 3.2, each axis is partitioned into three sections delineated by horizontal tick marks: the central focus area and the top and bottom context areas. When the

[3] The ASA cars data set is available online at http://stat.cmu.edu/datasets



(a) (b)

Figure 3.5

Screen captures of an MDX axis before (a) and after (b) dynamic scaling.

mouse is hovering over the focus area, an upward mouse wheel motion expands the display of the focus area outward and pushes outliers into the context areas (see Figure 3.5).

A downward mouse wheel motion causes the inverse effect: focus region compression. Alternatively, the user may use the mouse wheel over either of the two context areas to alter the minimum or maximum values separately. The user may also manually enter the minimum and maximum values by typing them in appropriate fields of the table view panel (see Figure 3.1). This intuitive capability helps to free space and reduce line clutter, thereby making it easier to analyze relation lines of interest.

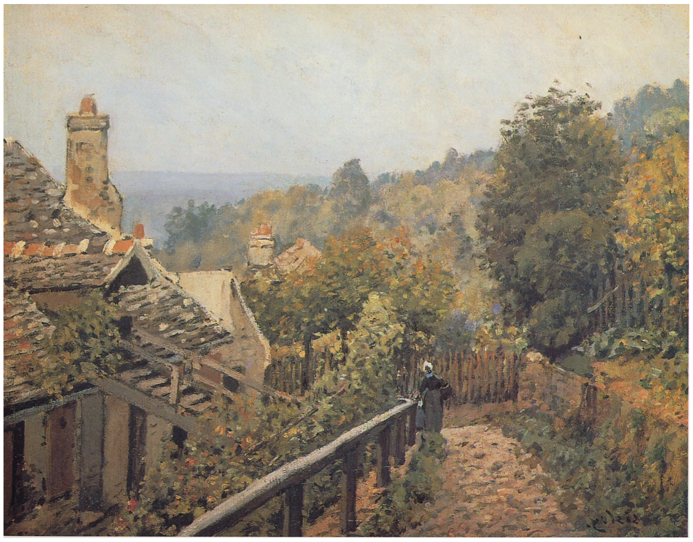


Figure 3.6

Alfred Sisley's 1873 painting *Sentier de la Mi-cote, Louveciennes*.

3.2.4 Aerial Perspective Shading

MDX also offers an innovative line shading scheme that is useful for rapidly monitoring trends due to the similarity of data values over multiple dimensions. This shading scheme simulates the human perception of aerial perspective whereby objects in the distance appear faded while objects nearer to the eye seem more vivid. The technique is a fundamental technique used in painting, especially landscapes (see Figure 3.6). As the distance between the viewer and an object is increased, the contrast between the object and background decreases. Background colors are less saturated and shift toward the background, which is usually blue.

In our implementation, aerial perspective shading can be used in either a discrete or continuous mode. In the discrete mode, the lines are colored according to the axis region that they intersect. If any point of a relation line is in the context (non-focus) area of at least one axis, the line is shaded with a light gray color and drawn beneath the non-context lines. If all the points on a relation line fall within the query area of each axis (the area between the two query sliders), the line is colored using a dark gray value that attracts the viewer's attention and the remaining lines (non-query and non-context) are colored a shade of gray that is slightly darker than the context lines but lighter than the query lines. The resulting effect for discrete aerial perspective shading is illustrated in Figure 3.1 and Figure 3.4.

In the continuous mode, non-context lines go through an additional step to encode the distance of the line from the mouse cursor. As shown in Figure 3.5 and Figure 3.7, query lines that are nearest to the mouse cursor receive the darkest value while lines farthest from the mouse cursor are shaded with a lighter gray. The other query lines are shaded according to a non-linear fall-off function that yields a gradient of colors between said extremes. Consequently, the lines that are nearest to the mouse cursor are more prominent to the viewer due to the color and depth ordering treatments and the viewer can effectively use the mouse to quickly interrogate the data set. In addition to the query lines, the shading scheme is also applied to the small scatterplots that are displayed beneath each axis (see Figure 3.7).

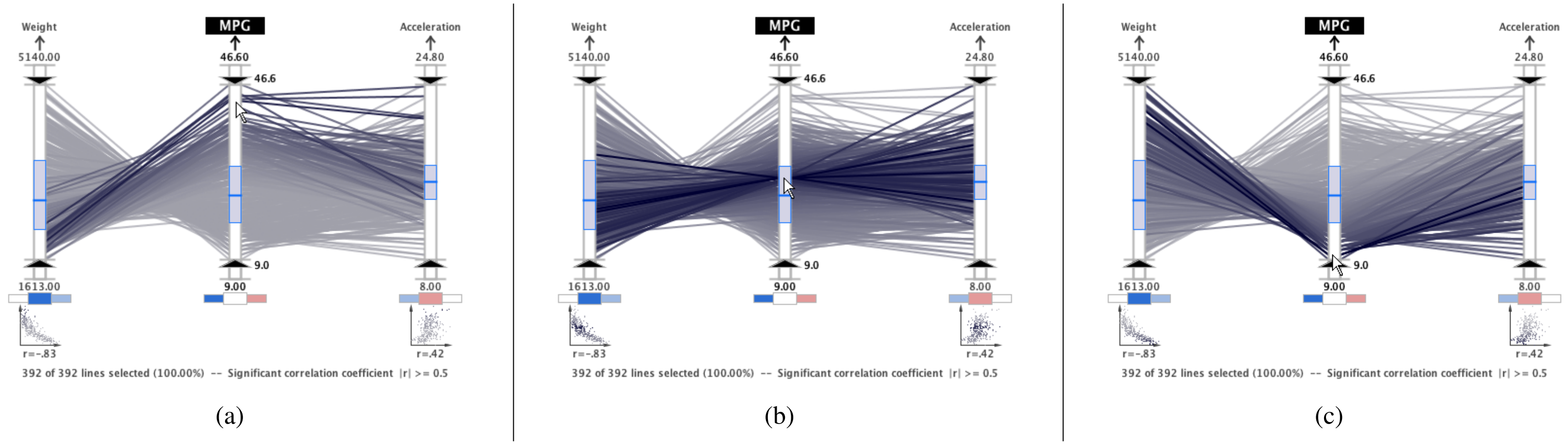

Weight
MPG
Acceleration
5140.00
46.60
24.80
46.6
9.0
1613.00
9.00
8.00
r=-.83
r=.42
392 of 392 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5
(a)
Weight
MPG
Acceleration
5140.00
46.60
24.80
46.6
9.0
1613.00
9.00
8.00
r=-.83
r=.42
392 of 392 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5
(b)
Weight
MPG
Acceleration
5140.00
46.60
24.80
46.6
9.0
1613.00
9.00
8.00
r=-.83
r=.42
392 of 392 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5
(c)

Figure 3.7

Continuous aerial perspective shading image sequence.

Table 3.2

MDX analysis capabilities.

Capability Name	Description
Stepwise Regression	Ranks predictors for a specific dependent variable.
Simple Regression	Quantifies individual predictor significance.
Correlation Analysis	Calculate and display correlation indicators.
Multicollinearity Filter	Hide highly correlated predictors.
Optimal Axis Arrangement	Arrange axes by one of several statistical measures.
Descriptive Statistics	Calculates descriptive statistics on-the-fly.

3.3 Analysis Capabilities

MDX also offers many analysis capabilities that help identify and quantify significant relationships in the data. In addition to graphical indicators of key descriptive statistical quantities, our system provides correlation and regression analysis, an automatic multicollinearity filter, and automatic axis arrangement capabilities. In the remainder of this section, these capabilities, which are summarized in Table 3.2, are described in detail.

3.3.1 Descriptive Statistical Indicators

To support the interactive analysis capabilities of MDX, each axis offers visual representations of key descriptive statistics that are identified in Figure 3.2. The median, interquartile range (IQR), and the frequency information are calculated for the data in the focus area of each axis. Alternatively, the user can configure the system to display the mean and standard deviation range. These central tendency and variability measures provide a numerical value that indicates the typical value and how “spread out” the samples are in the distribution, respectively. As shown in Figure 3.8, the wide overall box plots represent the descriptive statistics for all the axis samples while the more narrow query box plots, which are draw over the overall box plots, capture the descriptive statistics for only the samples that are selected with the axis query sliders. The thick horizontal lines that divide the box plots vertically represents the median value in the IQR mode and the mean in the standard deviation range mode.

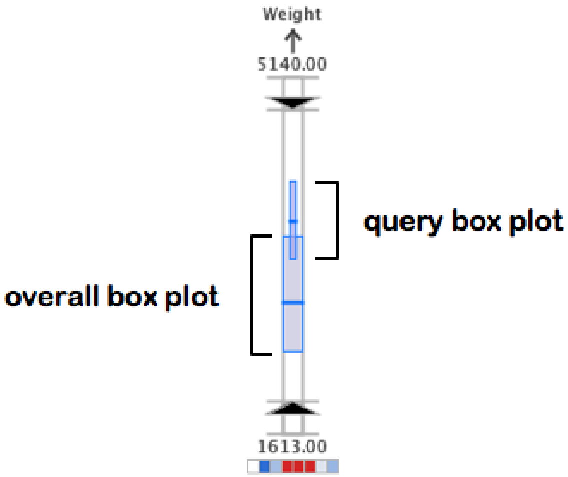


Figure 3.8

The axis box plots represent the variability statistics for the data values.

Alternatively, the viewer 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 3.9). 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 viewer can use this feature for quickly summarizing an axis during analysis. For example, if the data set is very large, the individual polyline drawing can be disabled and the axis summary enabled to dramatically increase the rendering speed of the system. The user can perform all statistical analysis processes and evaluate the descriptive statistics in this summary mode. When a detailed plot is desired, the individual polyline rendering can be reactivated.

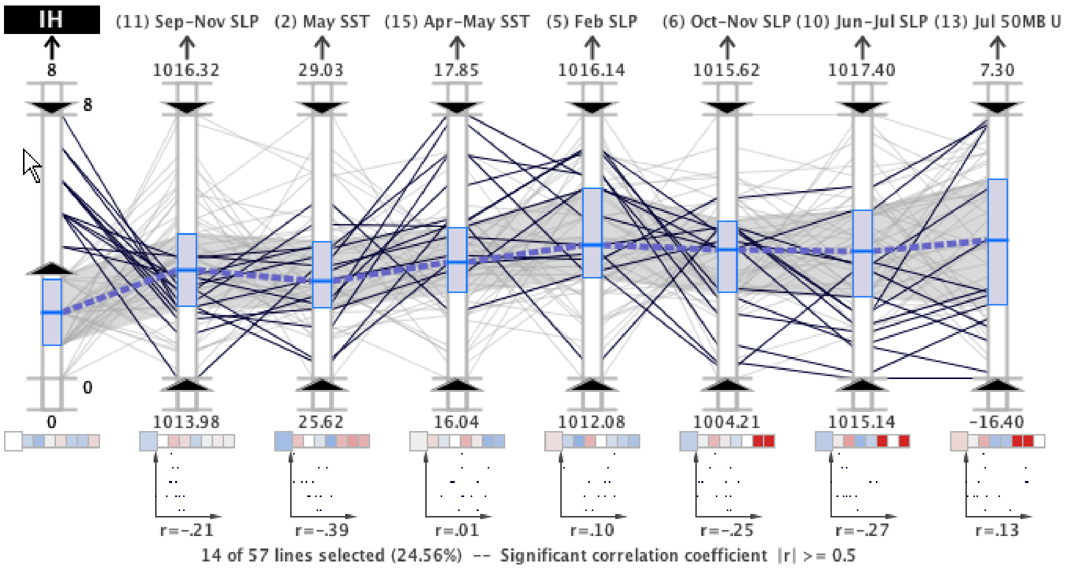


Figure 3.9

The axis summary lines in MDX.

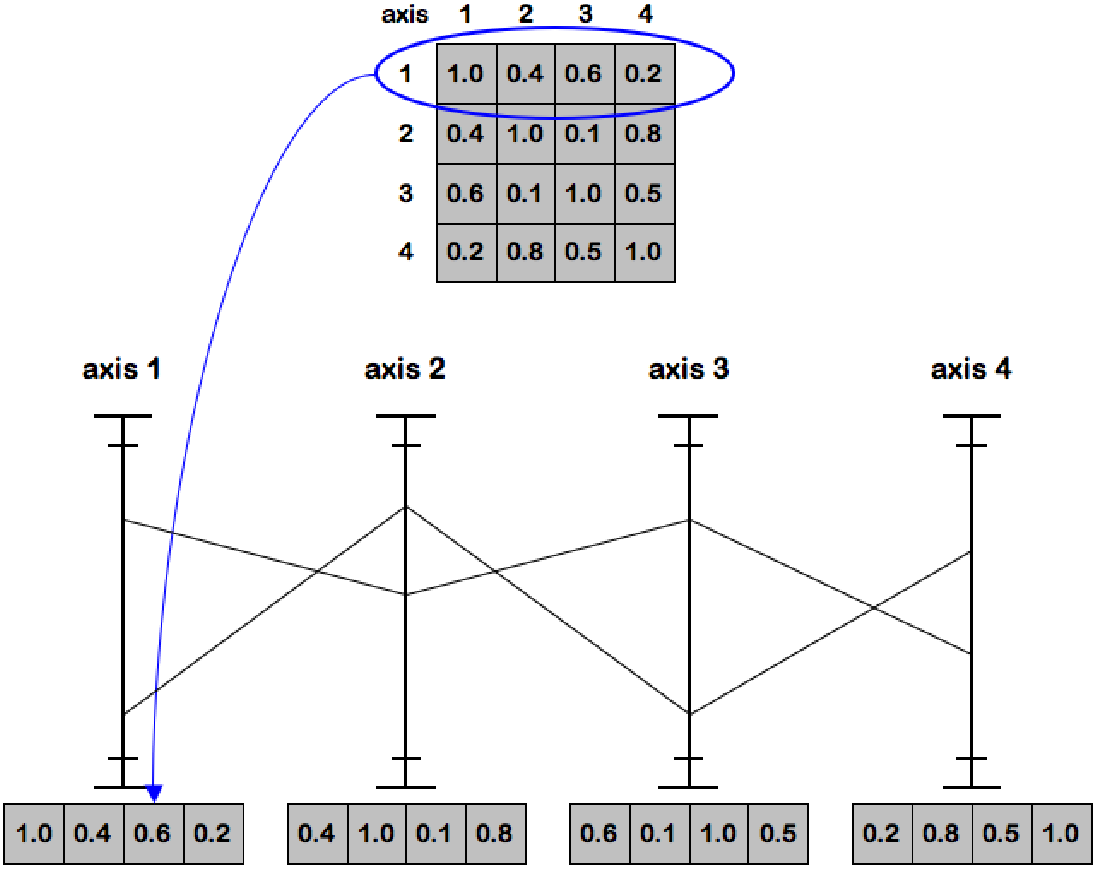
On each axis bar interior, the frequency information can also be displayed by representing histogram bins as small rectangles with widths that are indicative of the number of lines that pass through the bin's region (see Figure 3.2). That is, the widest bins have the most lines passing through while more narrow 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.

3.3.2 Correlation Analysis Capabilities

In statistics, correlation analysis attempts to measure 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 can be quantified using a single number, r, that is called the correlation coefficient [151]. Our system uses the Pearson product-moment correlation coefficient (also called the sample correlation coefficient) to measure the correlation between the axes visible in the parallel coordinates panel.

For each pair of axes in the display, our system computes r which results in a correlation matrix. As shown in Figure 3.10, the rows from this correlation matrix are displayed graphically beneath each axis as a series of color-coded blocks (see Figure 3.2). Each block uses color to encode the sample correlation coefficient between the axis directly above it and the axis that corresponds to its position in the set of blocks. For example, the first block in the correlation indicators under each axis in Figure 3.11 represents the correlation strength between the axis above it and the first axis, the *MPG* axis. When the mouse hovers over an axis in the parallel coordinates panel, the axis is highlighted and the correlation coefficient blocks corresponding to it below the other axes are enlarged (see Figure 3.11).

The blocks are colored blue for negative correlations and red for positive correlations. The stronger the correlation, the more saturated the color so that stronger correlations are



axis	1	2	3	4
1	1.0	0.4	0.6	0.2
2	0.4	1.0	0.1	0.8
3	0.6	0.1	1.0	0.5
4	0.2	0.8	0.5	1.0

Figure 3.10

Construction of the MDX axis correlation indicators.

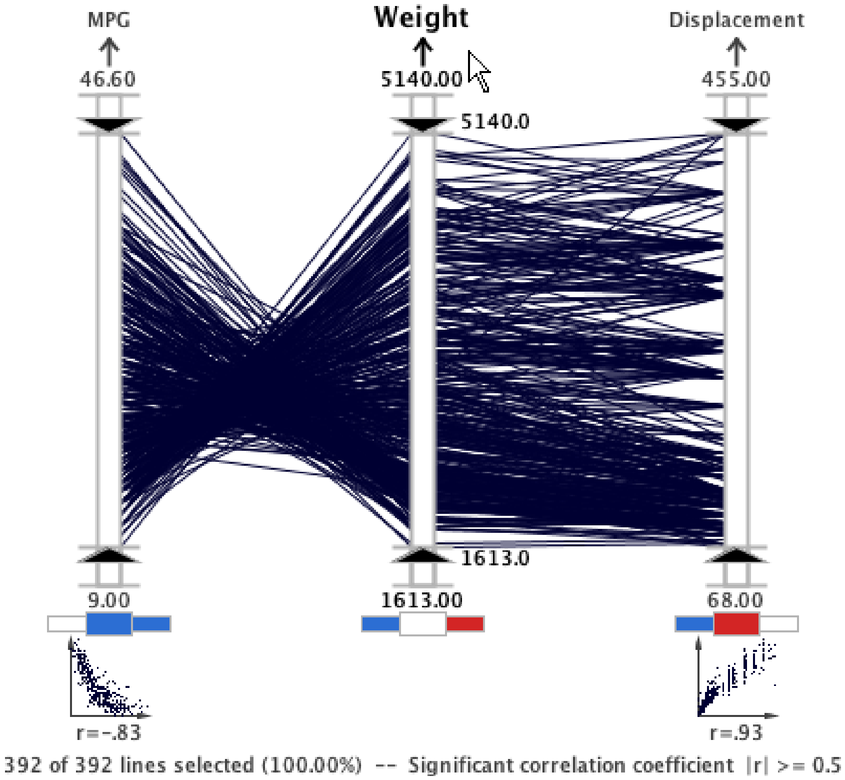

MPG
Weight
Displacement
46.60
5140.00
455.00
5140.0
1613.0
9.00
1613.00
68.00
r=-.83
r=.93
392 of 392 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

Figure 3.11

Axis correlation indicators in MDX.

more prominent. The correlation indicator color scale is shown in Figure 3.12. An axis' *r* value with itself is always equal to one and the corresponding indicator block is colored white. What's more, when the absolute value of a correlation coefficient is greater than or equal to the significant correlation threshold, the block is colored with the fully saturated color. The significant correlation threshold is a user-defined value that is displayed at the bottom of the parallel coordinates plot (see Figure 3.1). The correlation threshold can be adjusted via the settings panel.

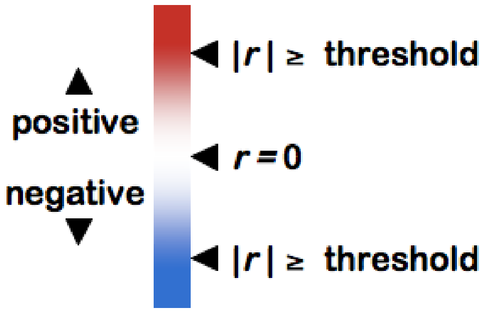


Figure 3.12

Axis correlation indicator color scale.

In addition to the sample correlation coefficient indicators, the system also displays small scatterplots below the correlation indicators for each axis when an axis is highlighted (see Figure 3.11). These scatterplots are created by plotting the points with the highlighted variable as the *y* axis and the variable directly above the scatterplot as the *x* axis. Each scatterplot also shows the numerical *r* value associated with the pair of axes below the scatterplot. The scatterplots provide a visual means to quickly confirm the type of correlation (positive or negative) and the strength of the correlation. It is important to note that the type of correlation is also visually detectable in the line configuration of the parallel coordinates plot. As shown in Figure 3.11, parallel coordinate lines that cross in an 'X' pattern are characteristic of a negative correlation while lines that appear to be more parallel indicate a positive correlation.

Unlike the other correlation indicators, the scatterplot is useful for discovering nonlinear relationships between variables. For example, a nonlinear relationship can be observed in a scatterplot even if the correlation coefficient is zero. In Figure 3.1, nonlinear relationships are illustrated in the scatterplots beneath the second, third, and fourth axes.

3.3.3 Multicollinearity Filter

Our system provides an automatic multicollinearity filter 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; if any axes are correlated with each other by more than the significant correlation threshold, one axis is removed from the display. The filter removes the axis that has a lower *r* with the dependent axis. In this way, the remaining axes are truly independent of each other.

The user can reduce multicollinearity manually by using the correlation indicators to identify and filter correlated axes within a predefined threshold. However, the filter provides an automatic way to ensure independence that can be performed at the click of a button. Removing the strongly correlated independent axes will ultimately improve subsequent MLR analysis by avoiding over-fitting the data.

3.3.4 Regression Analysis Capabilities

Regression analysis is often employed to identify the most relevant relationships in a particular data set. Such techniques are effective for screening data and providing quantitative associations. 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 [151]. Stepwise regression can complement multivariate visualization by isolating the significant variables in a quantitative fashion. As illustrated in Figure 3.13, our system executes a MATLAB®[4] script and captures output from the MATLAB® "regress" and "stepwisefit" utilities that perform simple and stepwise regression, respectively. The MATLAB® output stream is then parsed and displayed graphically within the parallel coordinates panel.

In our MLR analysis, a normalization procedure is used so that the y-intercept becomes zero and the importance of a predictor may be assessed by comparing regression coefficients, b_i, between different predictors. As shown in Figure 3.2, the system visually encodes b in the parallel coordinates panel using the box 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 encodes the r^2 output from the SLR process. In

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

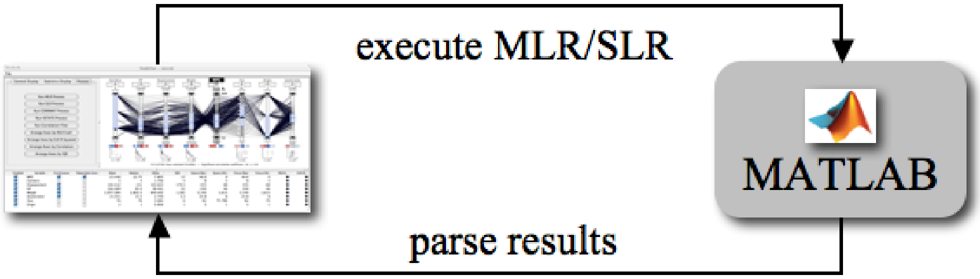


Figure 3.13

Integration of MATLAB® with MDX.

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. The box beneath the dependent variable axis name encodes the overall R^2 value from the MLR analysis (see Figure 3.2).

When these boxes are filled with a light gray 'X' (see Figure 3.14), 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, the reverse shading of the other axes. The dependent axis shading is illustrated by the *IH* axis on the right in Figure 3.2.

3.3.5 Axis Arrangement

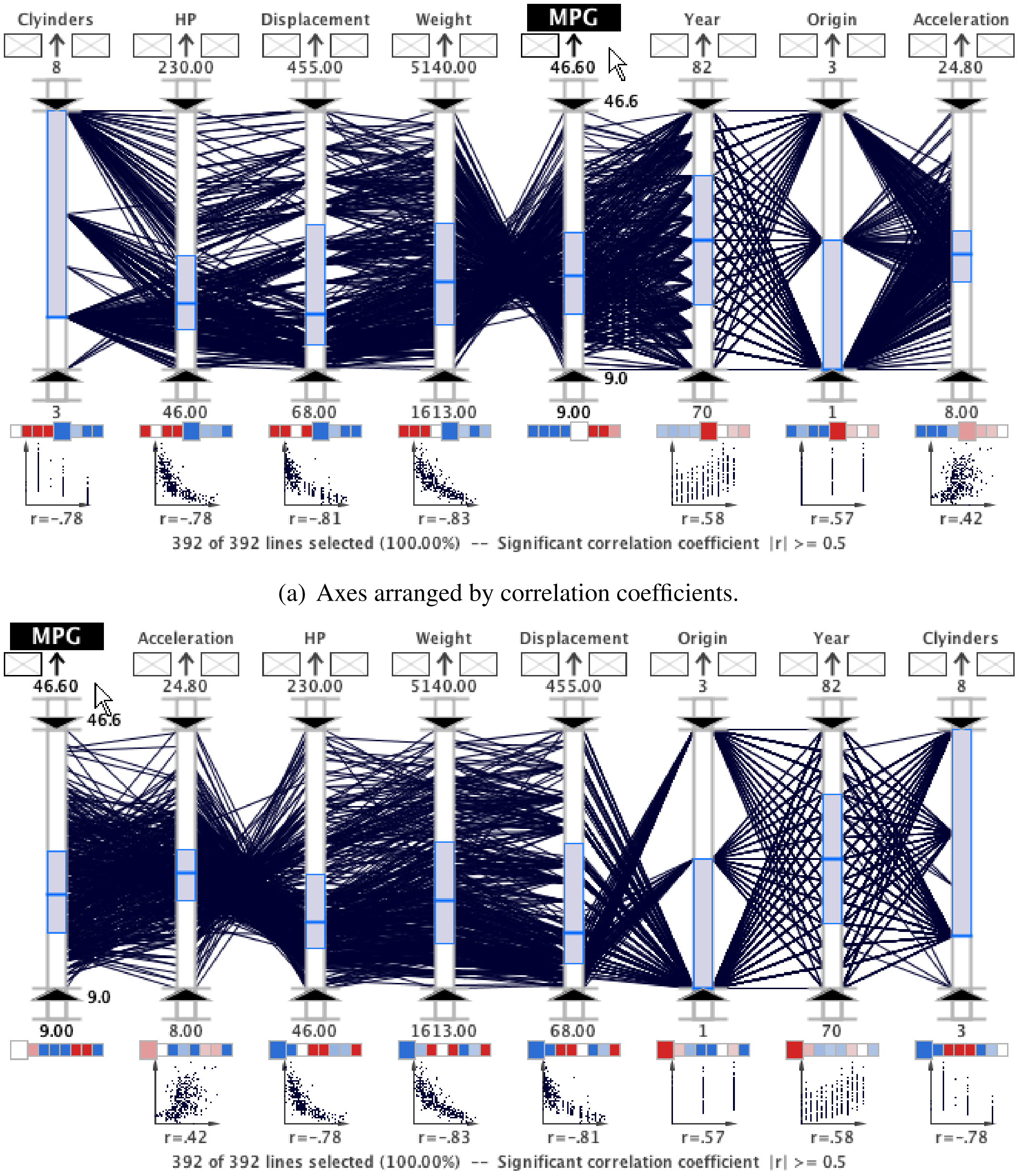
MDX can automatically arrange the axes in the parallel coordinates panel using one of the precomputed statistical measures previously mentioned. The user can choose to sort the axes by one of the following statistical measures:

- correlation coefficient (r),
- IQR / standard deviation range,
- MLR coefficient (b), or
- SLR (r^2) value.

This capability facilitates more rapid statistical comparison of axes. The user can execute the sorting 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 the correlation coefficient, 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 3.14(a), 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. When the axes are sorted in this manner, the user can quickly identify the strongest correlations with the target axis.

The IQR / standard deviation range, MLR b, and SLR r^2 arrangement options all sort the axes in descending order based on the statistical measures. The dependent axis is placed at the leftmost position and the other axes are arranged accordingly. The IQR / standard deviation range arrangement (see Figure 3.14(b)) is useful for examining the dispersion characteristics of each axis. The SLR r^2 arrangement (see Figure 3.15(a)) is useful for observing the individual correlation of axes with the dependent axis. The MLR *b* arrangement (see Figure 3.15(b)) helps to analyze the stepwise regression model results and quantify the most significant axes for the dependent axis.



392 of 392 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

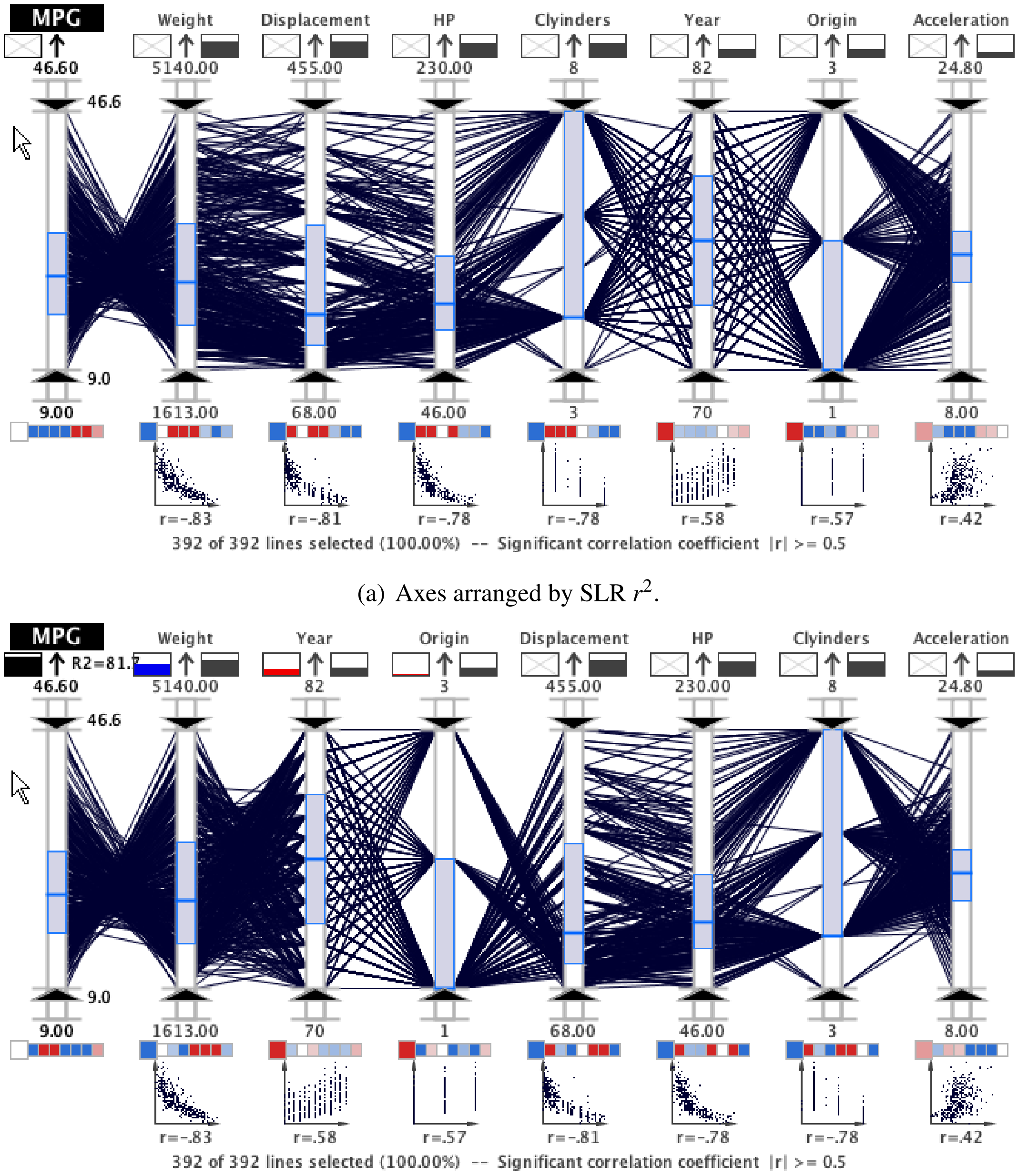
(a) Axes arranged by correlation coefficients.

392 of 392 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

(b) Axes arranged by population variability ranges.

Figure 3.14

Axis arrangement by correlation coefficients and variability ranges.



(a) Axes arranged by SLR r^2.

(b) Axes arranged by MLR b.

Figure 3.15

Axes arranged by SLR r^2 values (a) and MLR b values (b).

CHAPTER 4

EVALUATION

Traditional statistical data analysis, particularly in the realm of environmental study, will benefit immensely by incorporating new visual analytic techniques using a sophisticated system like MDX. In this chapter, we validate the two hypotheses presented in Chapter 1 with four practical case studies. The hypotheses that motivated this dissertation are:

1. The development of an advanced geovisual analytics approach using parallel coordinates and statistical techniques reveals a deeper level of understanding than traditional methods when applied to the task of finding complex multivariate trends in environmental data sets.
2. The effectiveness of the geovisual analytics approach is necessarily explored in the context of practical environmental studies, which are grounded in real-world data sets instead of invented or abstract data sets, in close collaboration with domain experts. The discovery of new associations and the confirmation of known patterns by domain experts will validate the promise of this new approach in environmental data analysis.

These case studies reveal that a geovisual analytics approach enables the scientist to achieve a deeper level of understanding of the data than conventional multivariate analysis techniques. The latter three case studies presented in this chapter address real-world problems and have been conducted in close collaboration with a hurricane expert, Dr. Fitzpatrick, who also serves on the graduate committee for this dissertation. In addition to demonstrating the promise of this approach for multivariate data analysis, these evaluations highlight several significant associations in North Atlantic tropical cyclone predictor data sets—a noteworthy contribution to the understanding of these destructive weather features.

The first case study demonstrates the capabilities of MDX with a pedagogical data set that is used to demonstrate statistical data analysis to students. In the other three case studies, MDX is used to identify and quantify significant associations in tropical cyclone predictor data sets. To this end, the climate studies fulfill the NIH/NSF Visualization Challenges Report recommendation that visualization researchers "collaborate closely with domain experts who have driving tasks in data-rich fields to produce tools and techniques that solve clear real-world needs [92]" through the inclusion of a hurricane expert throughout the analysis. In the remainder of this chapter, we will describe the workflow developed for conducting these case studies, the details of each case study, and the strengths and weaknesses for using geovisual analytics in climate analysis.

4.1 Systematic Workflow for Environmental Analysis

The visualization capabilities and statistical processes offered by MDX provide a rich environment for performing complex multivariate data analysis. During the system development and testing, we formulated a systematic workflow to guide the scientist. In this section, the workflow, that is depicted graphically in Figure 4.1, will be described. A more detailed illustration of this workflow is provided in Figure 4.2. Although this workflow

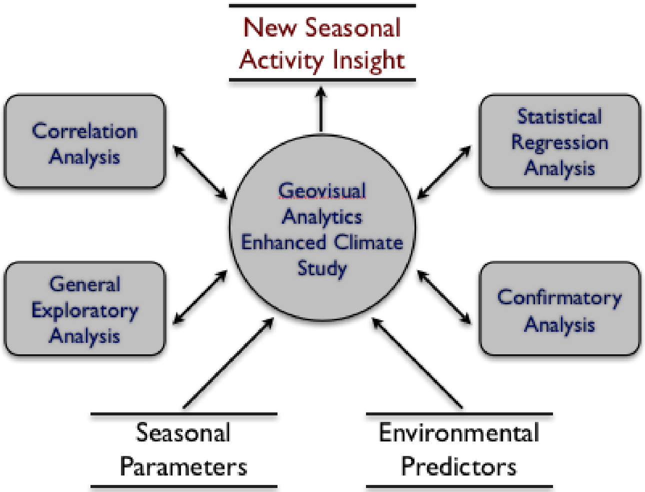

New Seasonal Activity Insight
Correlation Analysis
Statistical Regression Analysis
Geovisual Analytics Enhanced Climate Study
General Exploratory Analysis
Confirmatory Analysis
Seasonal Parameters
Environmental Predictors

Figure 4.1

Climate study system context diagram.

is described in a sequential order, typical analysis involves several iterations and moving between the various processes.

After preparing and loading the data set into the system, the scientist will manually filter the display to remove unnecessary axes. Then, the scientist will manually arrange the variable axes and interact with the display using the previously mentioned visual query techniques. During this initial exploratory analysis, the scientist will acquire a preliminary overview of the entire data set.

Next, the scientist will observe the statistical correlations in the data using the correlation analysis processes and indicators. The system's automated axis arrangement tools can be used in this stage to highlight strong correlations and compare IQR or standard deviation ranges in the data. To prepare for the regression analysis, the scientist can manually

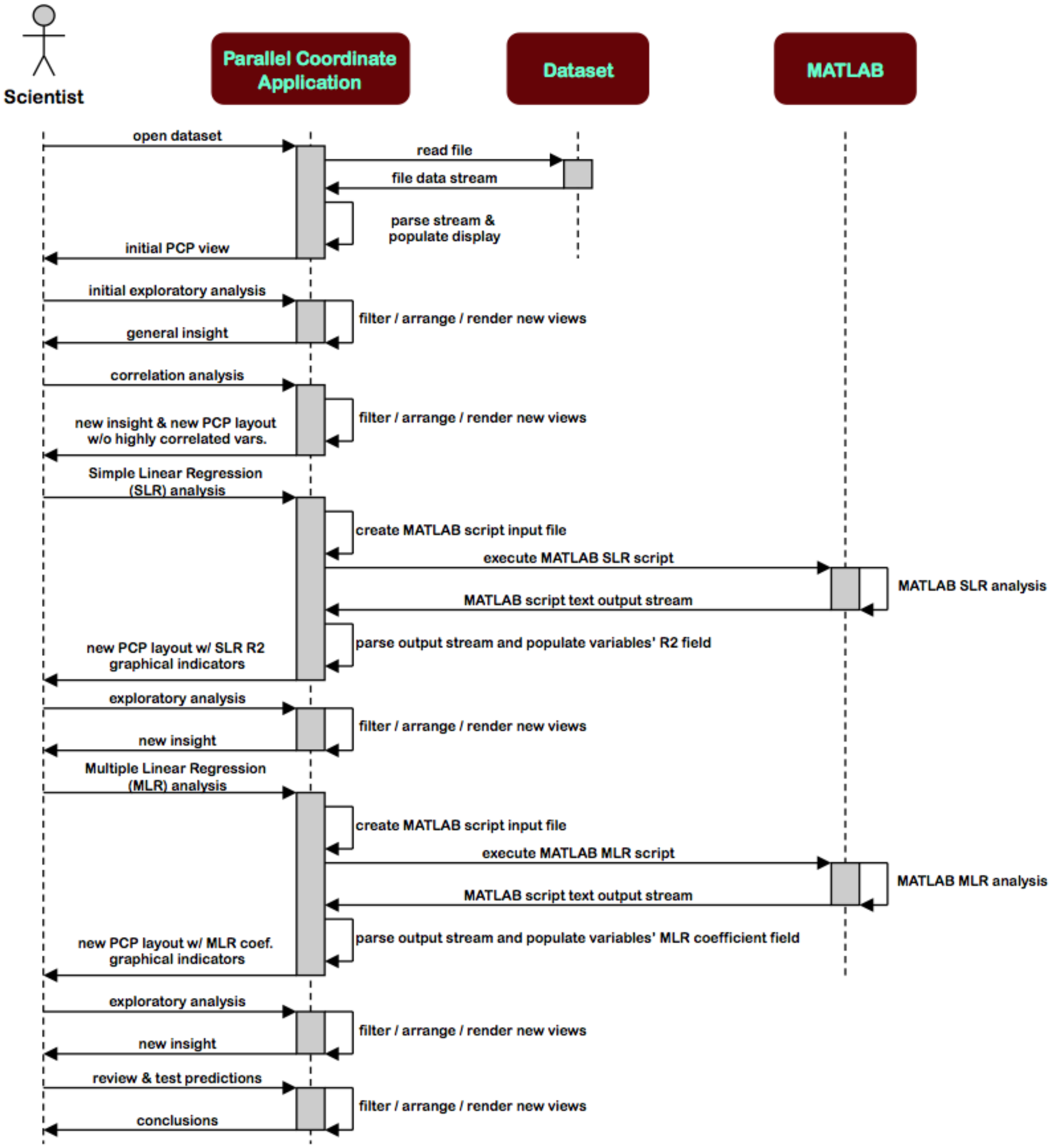

Scientist
Parallel Coordinate Application
Dataset
MATLAB
open dataset
read file
file data stream
parse stream & populate display
initial PCP view
initial exploratory analysis
filter / arrange / render new views
general insight
correlation analysis
filter / arrange / render new views
new insight & new PCP layout w/o highly correlated vars.
Simple Linear Regression (SLR) analysis
create MATLAB script input file
execute MATLAB SLR script
MATLAB SLR analysis
MATLAB script text output stream
parse output stream and populate variables' R2 field
new PCP layout w/ SLR R2 graphical indicators
exploratory analysis
filter / arrange / render new views
new insight
Multiple Linear Regression (MLR) analysis
create MATLAB script input file
execute MATLAB MLR script
MATLAB MLR analysis
MATLAB script text output stream
parse output stream and populate variables' MLR coefficient field
new PCP layout w/ MLR coef. graphical indicators
exploratory analysis
filter / arrange / render new views
new insight
review & test predictions
filter / arrange / render new views
conclusions

Figure 4.2

Sequence diagram for climate study using MDX.

reduce multicollinearity by using the correlation indicators to identify and filter correlated variables using a predefined significance level. The scientist can also utilize the automatic multicollinearity filter to ensure that the predictors are truly independent of one another. Removing the strongly correlated independent variables will ultimately improve the MLR analysis by avoiding over-fitting the data. The scientist will gain additional insight in this phase by observing correlations between the predictors as well as correlations between each predictor and the dependent variable.

After the correlation analysis, the scientist will use the integrated SLR process. This capability provides an alternative indication of the individual associations between the predictors and the dependent variable. The scientist may glean additional insight from this exercise to determine if additional variables should be removed from the view. Then, the scientist is ready to execute the MLR process in order to quantify the significance of the predictors to the dependent variable. The result of this process is a ranked list of the most important variables for the dependent variable. Unlike the SLR and correlation analysis, the MLR analysis considers the contribution in relation to the other predictors.

By following this workflow in our system, the scientist will develop new ideas about how the specific variables can be used to predict the dependent variable. That is, the scientist will have formed hypotheses about the associations between the variables. Then, the scientist can continue to explore the data in the system to attempt to prove or disprove the new hypotheses; a process that Tukey [145] calls confirmatory data analysis. For example, the scientist may discover patterns in the climate data that will help predict the hurricane activity in 2005 based on the analysis of data from 1950 to 2006. If the theory holds after this testing, the scientist may use the new insight to predict future hurricane activity.

4.2 Case Study 1: Exploring Relationships in Body Dimensions

Our first case study uses a data set that is well-suited for statistical data analysis and more complicated multivariate analyses such as regression and discriminant analysis. In fact, Peterson et al. [65] used this data set to illustrate a practical project, which is based on research by the first two authors of the work, for teaching students the art of data analysis. In this project, the authors investigated the correspondence between body build, weight, and girths in a group of physically active men and women, most of whom were within the normal weight range. Body build was characterized by skeletal width and depth measurements at nine well-defined locations. The study affirmed the notion that body build (skeletal) variables and height predict scale weight substantially better than height alone. Using regression, a weight equation based on the body build variables was obtained for the group. Also trunk and limb girths were recorded at twelve well-defined sites and regression analysis was again used to determine the best prediction equations for the measured girths. This data set gives students an opportunity to explore anthropometric, forensic, and ergonomic topics using several analysis techniques [65].

This case study will provide a good demonstration of the value of the MDX capabilities using a data set that exhibits valuable characteristics. In addition to having several strong correlations, the data set has a large number of samples compared to the number of different variables. The following analysis of this data will provide an ideal study to evaluate the effectiveness of MDX, which will precede several more complicated analyses of real-world climate data sets that have more complex associations. We will conduct the analysis of this data set following the same steps as the authors suggest, except we will omit the discriminant analysis and some additional regression tasks which are mentioned in the project.

4.2.1 Body Measurement Data

The body dimension data set contains measurements from a total of 507 subjects—260 females and 247 males. The measurements were taken primarily from individuals in their twenties and early thirties, all physically active. For each individual, 25 measurements, which are listed in Table 4.1 were recorded [65].

The data set includes nine diameter (or skeletal) measurements and twelve girth (or circumference) measurements. At the time of physical maturation, the nine skeletal sites have generally attained their maximum size. Except for the three "boney" girths of the wrist, knee, and ankle, the girth measurements change over the life span. In addition to these measurements, which are recorded in centimeters, each subject had his or her age (years), weight (kilograms), and gender recorded [65].

4.2.2 Data Analysis

As shown in our workflow, the project begins with exploratory analysis of the data using descriptive statistics, such as the mean, median, and standard deviation. The interactive descriptive statistical indicators of MDX facilitate rapid exploration with details

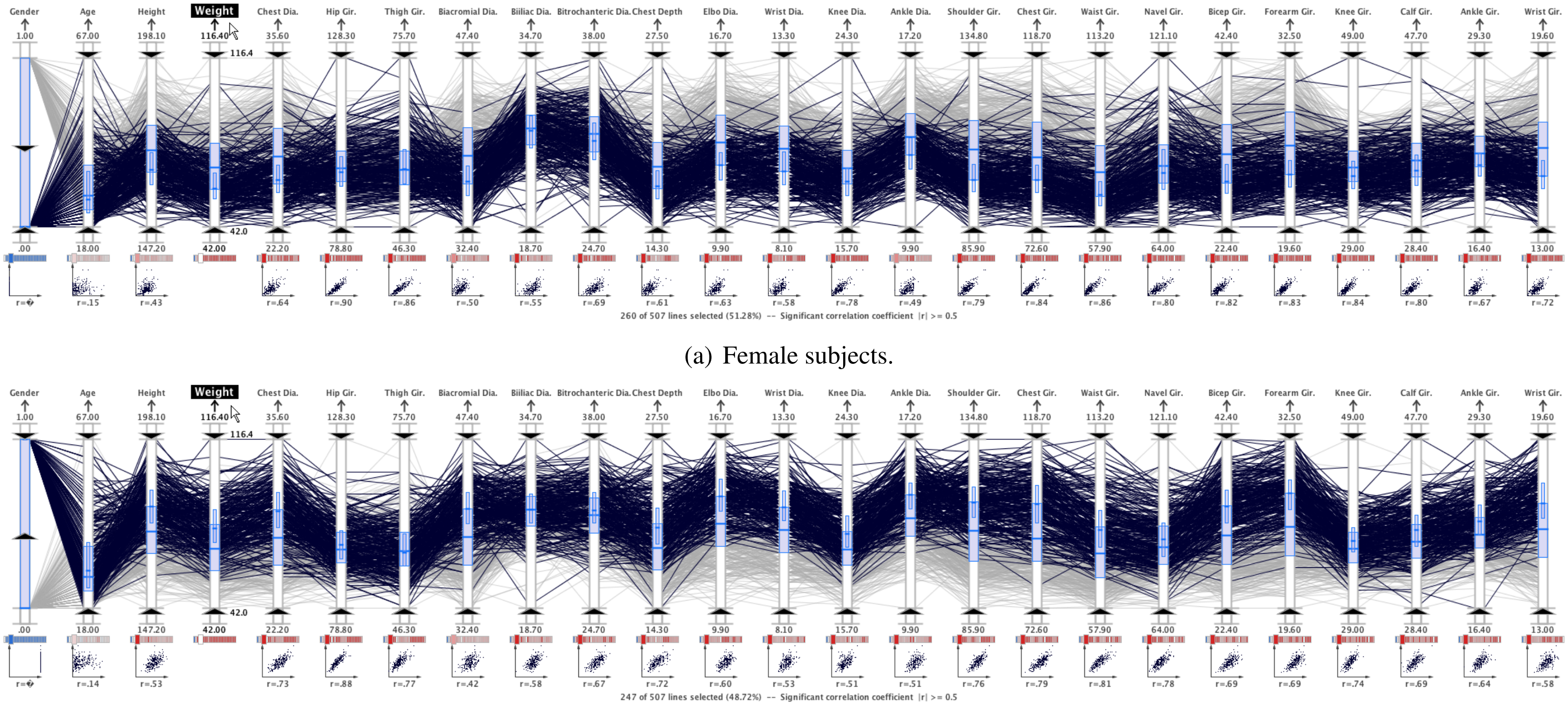
Table 4.1

Body data set measurements and descriptions.

Variable Name	Description
Skeletal Measurements:	
Biacromial diameter	
Biiliac diameter	Also called “pelvic breadth”.
Bitrochanteric diameter	
Chest depth	Between spine and sternum at nipple level, mid-expiration.
Chest diameter	At nipple level, mid-expiration.
Elbow diameter	Sum of two elbows.
Wrist diameter	Sum of two wrists.
Knee diameter	Sum of two knees.
Ankle diameter	Sum of two ankles.
Girth Measurements:	
Shoulder girth	Over the deltoid muscles.
Chest girth	Nipple line in males and just above breast tissue in females, mid-expiration.
Waist girth	Narrowest part of torso below the rib cage, average of contracted and relaxed position.
Navel (or “Abdominal”) girth	At umbilicus and iliac crest, iliac crest as a landmark.
Hip girth	At level of bitrochanteric diameter.
Thigh girth	Below gluteal fold, average of right and left girths.
Bicep girth	Flexed, average of right and left girths.
Forearm girth	Extended, palm up, average of right and left girths.
Knee girth	Extended, palm up, average of right and left girths.
Calf maximum girth	Average of right and left girths.
Ankle minimum girth	Average of right and left girths.
Wrist minimum girth	Average of right and left girths.
Other Measurements:	
Age	Years
Weight	Kilograms
Height	Centimeters
Gender	1 – male, 0 – female

accessible on-demand. One of the first facts that arise from this activity is the statistical differences between genders in the group of subjects. We use the query sliders and discrete aerial perspective shading on the *Gender* axis to observe the measurement differences. The resulting image, which is shown in Figure 4.3, provides an overview of the differences between gender. This plot captures the strong linear correlations of the measurements with the *Weight* axis (except for the *Age* variable). The parallel coordinates plot also shows that several axes have outliers that may be considered for removable from the data set. For example, some subjects are near the age of 60, which is well outside the IQR range for the *Age* axis. For the measurements that change over the life span, these outliers may affect our regression analyses. Furthermore, this image shows a mass of polylines which makes it difficult to use the parallel coordinates plot to discover interesting associations. To overcome this problem, we can utilize the continuous aerial perspective shading (proximity line shading) and axis scaling features. The small scatterplots that are rendered beneath the axes also help with this issue by providing an alternative view of the relationships.

We examine on the *Chest Diameter* axis in Figure 4.4 which shows substantial difference in male and female subjects. That is, the median of the female subjects is nearly at the bottom of the IQR range while the median for the male subjects is almost exactly the same as the top of the IQR range. It is also significant to note that the median value of the male subjects exceed the female subjects in all measurements except for *Thigh Girth*. In Figure 4.5, the *Weight* axis is highlighted and the scatterplots for this variable and the other variables are shown beneath each axis. This figure provides another illustration the gender differences for several measurements. In this figure, the *Gender* query slider is



(a) Female subjects.

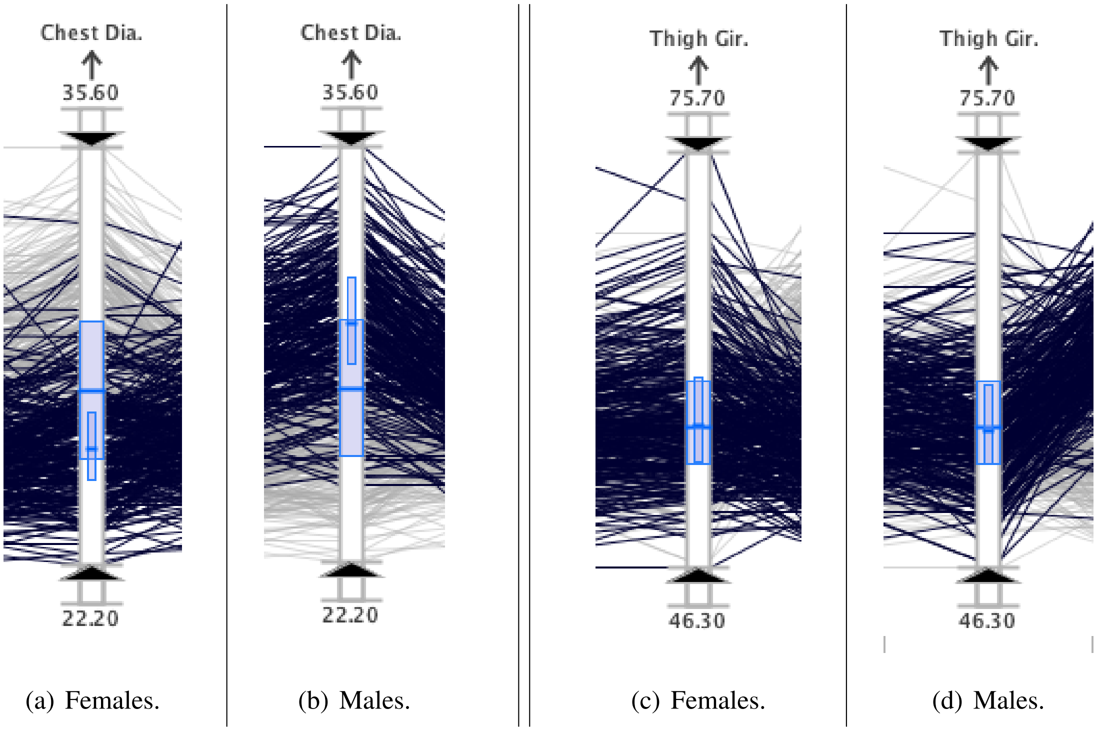
(b) Male Subjects.

Figure 4.3

Gender differences in the body measurements data set.

set to highlight the female subjects. The line shading then helps to visually distinguish between the male and female samples and two distinct groups of points are visible. The two groups are most distinct for the *Thigh Girth* variable.

In Figure 4.4 the query box plots reveal the fact that women typically exceed men in this measurement. While these analyses on gender differences would require several separate plots using traditional techniques, the MDX interface captures these details in a single view.



(a) Females. (b) Males. (c) Females. (d) Males.

Figure 4.4

Gender differences between the *Check Diameter* and *Thigh Girth* variables.

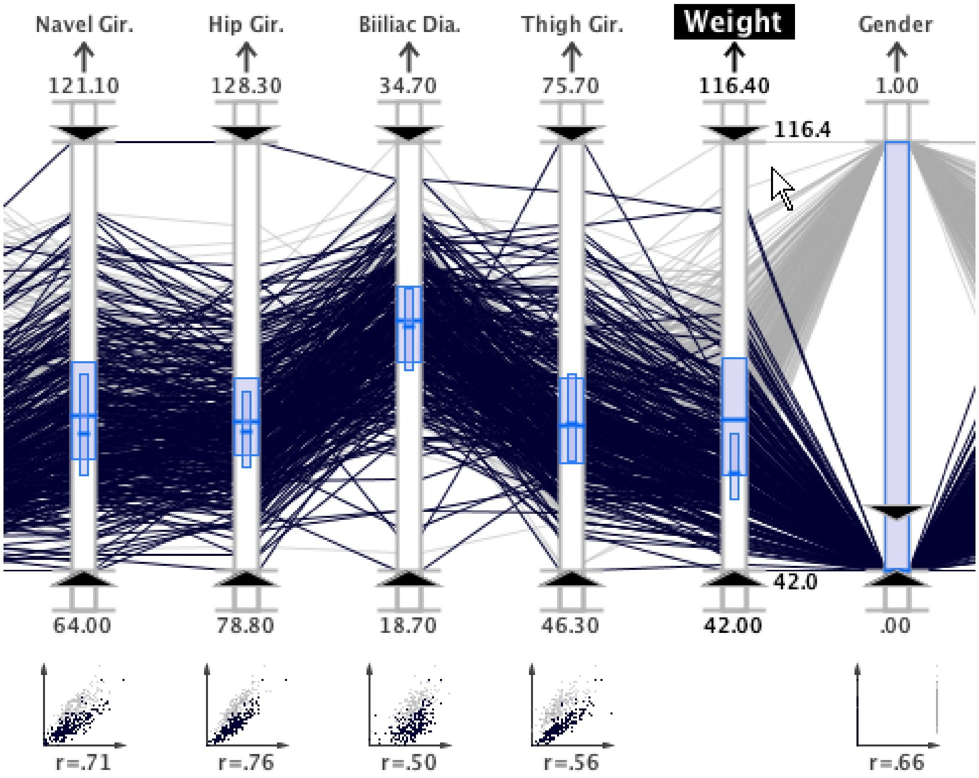


Figure 4.5

Scatterplot displays highlight gender differences in several measurements.

The authors of the body data study also note that the distributions of the data can also be observed graphically via simple histograms. Using the histogram displays in MDX, we can show the frequency information for all the sample measurements simultaneously as shown in Figure 4.6. From this view, we see that the measurements are typically modeled well by normal or gamma distributions. We zoom into this overview and use the axis histogram displays to reveal the approximately normal distributions exhibited by the *Biacromial Diameter* axis for the female (see Figure 4.7(a)) and male (see Figure 4.7(b)) subjects; and a gamma distribution was discovered on the *Waist Girth* axis for female subjects (see Figure 4.7(c)).

As previously noted in the discussion of Figure 4.3, most of the measurements exhibit strong correlations. We can employ MDX's correlation analysis capabilities (parallel coordinates line configurations, scatterplots, and correlation indicators) to explore these associations. As noted by the study authors, similar families of body dimensions are expected to have strong correlations. In Figure 4.8(a), the correlations indicators reveal a strong correlation between *Biiliac Diameter* and *Bitrochanteric Diameter* for all samples. In Figure 4.8(b), the indicators reveal the strong correlation between *Bicep Girth* and *Forearm Girth* in male subjects. In Figure 4.8(c), the strong correlation between *Hip Girth* and *Thigh Girth* in women is shown.

The authors of the body data study also performed multiple regression analysis on the data set. We used MDX's automated multicollinearity filter and stepwise regression capabilities to develop a model of the most significant measurements with *Weight* as the dependent variable. The authors of the body data study did not execute a multicollinearity

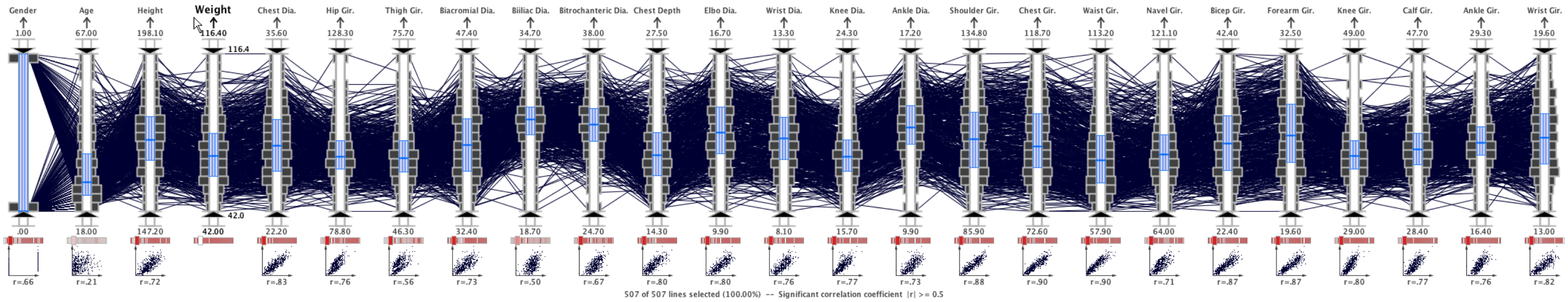
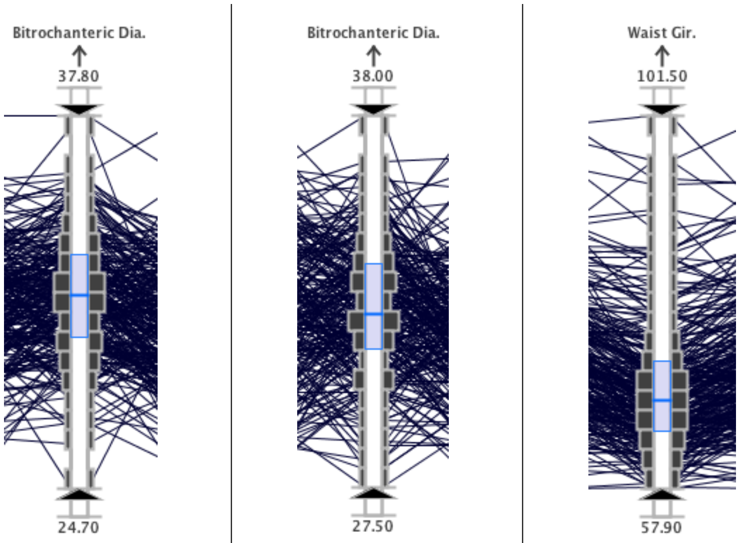


Figure 4.6

Histogram axis displays for all body measurement axes.



(a) Female biacromial diameter.

(b) Male biacromial diameter.

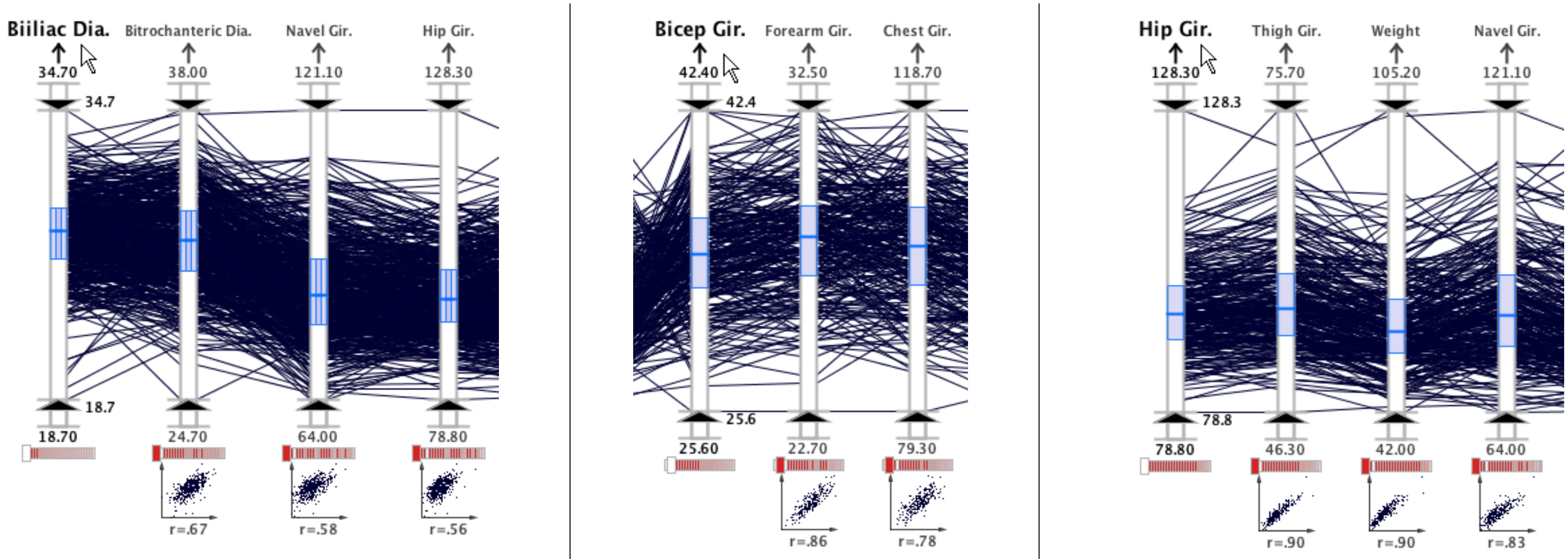
(c) Female waist girth.

Figure 4.7

Detailed histogram view showing normal and gamma distributions.

filter or provide stepwise regression analysis results in the article. In Table 4.2 the results are listed for the regression analyses performed with the *Weight* axis as the dependent variable for all subjects, male subjects, and female subjects.

The first regression model included all 507 subjects (male and female subjects). The multicollinearity filter removed all measurements except the *Waist Girth*, *Thigh Girth*, and *Biiliac Diameter* variables. The resulting regression model yielded an R^2 value of 86% (see Figure 4.9). Based on this model, the most significant measurement for calculating *Weight* in male and female subjects was *Waist Girth*. The tight, linear clustering of lines in the scatterplot and the high correlation coefficient ($r = .9$) provided additional evidence of the strong association between the variables.



(a) Male and female.

(b) Male.

(c) Female.

Figure 4.8

Detailed analysis of strongly correlated body measurements.

Table 4.2

Stepwise regression model for body measurement data set.

Male and Females (507 subjects)
(R^2 is 86%)

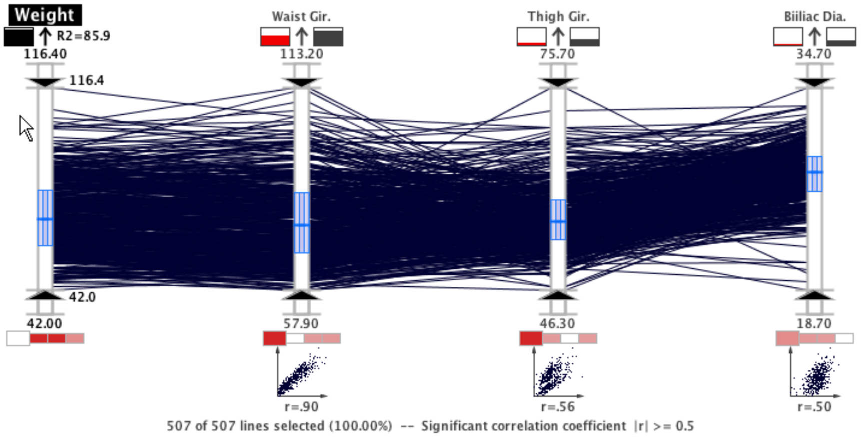
Chosen Variables	Normalized Coefficients b	Sample Mean
Waist Girth	0.791	76.98
Thigh Girth	0.196	56.86
Biiliac Diameter	0.071	27.83

Males (247 subjects)
(R^2 is 85%)

Chosen Variables	Normalized Coefficients b	Sample Mean
Hip Girth	0.728	97.76
Height	0.186	177.74
Elbow Diameter	0.156	14.46
Biacromial Diameter	0.040	41.24

Females (260 subjects)
(R^2 is 86%)

Chosen Variables	Normalized Coefficients b	Sample Mean
Hip Girth	0.779	95.65
Height	0.111	164.87
Wrist Diameter	0.110	9.87
Biacromial Diameter	0.092	36.5



507 of 507 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

Figure 4.9

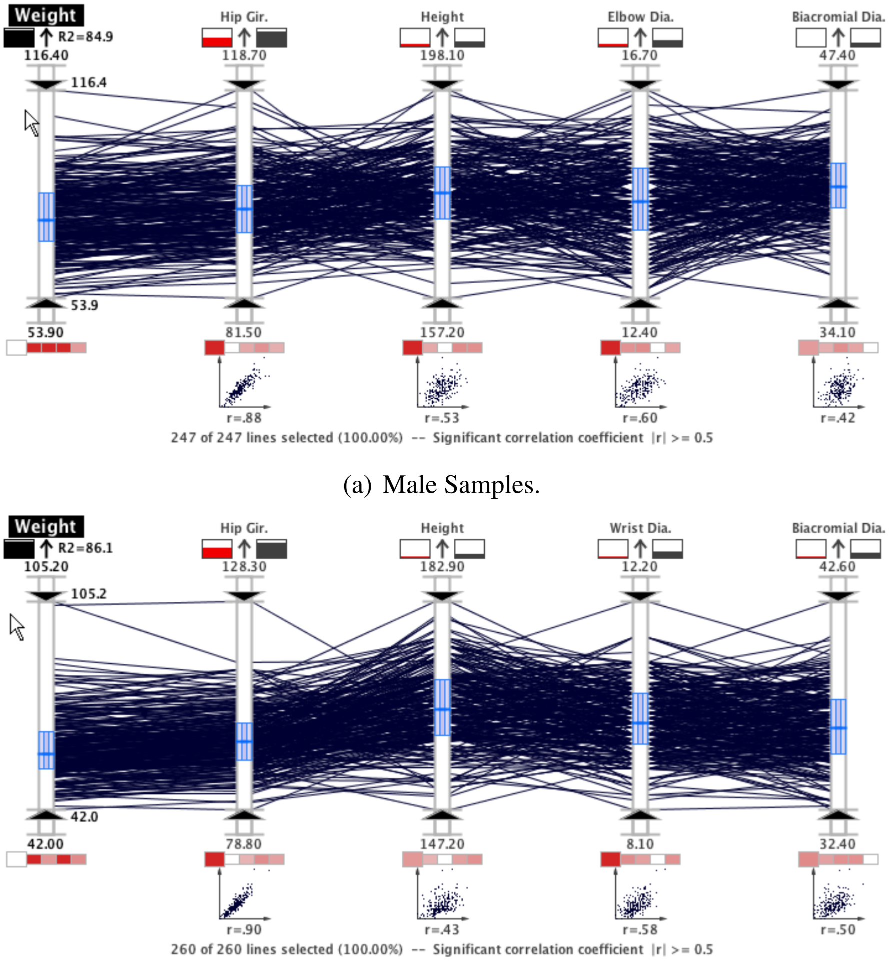
Stepwise regression results including both male and female subjects.

The second regression model included only the 247 male subjects. Prior to running the regression analysis, the multicollinearity filter removed all variables except for *Hip Girth*, *Height*, *Wrist Diameter*, and *Biacromial Diameter*. The stepwise regression model yielded an R^2 value of 85% (see Figure 4.10(a)). The regression model indicated that *Hip Girth* is the most significant measurement for determining *Weight*. The *Hip Girth* variable also exhibited a tight, linear trend in the scatterplot and a strong correlation coefficient ($r = .88$) to support its selection. This regression model included none of the variables that were selected when both the male and female subjects were considered.

The last regression model was generated using only the 260 female subjects. The multicollinearity filter removed all except the *Hip Girth*, *Height*, *Wrist Diameter*, and *Biacromial Diameter* variables. The subsequent regression model then resulted in an R^2 value of 86% (see Figure 4.10(b)). The most significant variable in this case was *Hip Girth*, which also exhibited a very tight, linear trend in the scatterplot and a strong correlation coefficient ($r = .9$). Again, none of the variables from the regression model that included the male and female subjects were included in this model. However, three of the variables in this model are also included in the model for the male subjects. In the male subject model, the variable with the third highest regression coefficient is *Elbow Diameter*, whereas the females subject model had the *Wrist Diameter* as the variable with the third highest regression coefficient. Regression coefficients for the variables that are common to both models are very similar in value.

4.3 Case Study 2: North Atlantic Tropical Cyclone Climate Study

As discussed in Chapter 2 Section 2.4, regression analysis is often employed to identify the most relevant climate relationships for tropical cyclone activity. Such techniques are effective in screening data and providing quantitative associations; but multivariate analysis can be difficult. In this case study, we use the MDX system to evaluate the regression model formed using a set of tropical cyclone predictors for the following categories: number of named storms (*NS*); number of hurricanes (*H*); and number of intense hurricanes (*IH*). This case study will outline how stepwise regression and parallel coordinates can complement each other in such an analysis. That is, the geovisual analytics can visually depict the same association that weather scientists find meaningful and provide a deeper level of understanding when used in conjunction with traditional multiple regression analysis.



247 of 247 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

(a) Male Samples.

260 of 260 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

(b) Female Samples.

Figure 4.10

Male-only and female-only stepwise regression models.

4.3.1 CSU Climate Study Data Set

For the second and third case studies, we analyzed a data set containing potential environmental predictors for a tropical cyclone climate study. This data set was provided by Dr. Phil Klotzbach [99] of the Tropical Meteorology Project[1] at Colorado State University (CSU), and is used to predict the frequency of Atlantic tropical cyclones for the upcoming hurricane season by categories. These categories include:

- Number of named storms (winds 17 $\frac{m}{s}$ or more, at which tropical cyclones receive a “name”)
- Number of hurricanes
- Number of intense hurricanes

These variables have known relationships to Atlantic tropical cyclone activity. For example, Chu [29] described how the North Atlantic basin has fewer tropical cyclones during El Niño Southern Oscillation (ENSO) years, but active seasons in La Niña years. Because of this relationship, scientists use ENSO signals as some predictors of seasonal storm activity. In Table 4.3, variables 1 through 8 are believed to characterize ENSO events. Scientists at the Tropical Meteorology Project issue six forecast reports each year based on statistically significant predictors from this data set.

Table 4.3 lists sixteen potential environmental predictors from the data set along with their geographical region. In Fig. 4.11, the geographical regions where each predictor is measured is shown in a geographic map. In the remainder of this section, the physical relationships of these climate variables to Atlantic tropical cyclone activity are discussed.

[1] The CSU Tropical Meteorology Project website is located at http://typhoon.atmos.colostate.edu/.

Table 4.3

CSU tropical cyclone predictor data set parameters.

	Variable Name	Geographical Region
(1)	June–July Niño 3	5S-5N, 90-150W (eastern equatorial tropical Pacific Ocean)
(2)	May SST	5S-5N, 90-150W (eastern equatorial tropical Pacific Ocean)
(3)	February 200-mb U	5S-10N, 35-55W (equatorial East Brazil)
(4)	February–March 200-mb V	35-62.5S, 70-95E (South Indian Ocean)
(5)	February SLP	0-45S, 90-180W (eastern South Pacific Ocean)
(6)	October–November SLP	45-60N, 120-160W (Gulf of Alaska)
(7)	Sept. 500-mb Geopotential Height	35-55N, 100-120W (western North America)
(8)	November SLP	7.5-22.5N, 125-175W (subtropical northeast Pacific Ocean)
(9)	March–April SLP	0-20N, 0-40W (eastern tropical Atlantic Ocean)
(10)	June–July SLP	10-25N, 10-60W (tropical Atlantic Ocean)
(11)	September–November SLP	15-35N, 75-97W (southeast Gulf of Mexico)
(12)	Nov. 500-mb Geopotential Height	67.5-85N, 50W-10E (North Atlantic Ocean)
(13)	July 50-mb U	5S-5N, 0-360 (equatorial globe)
(14)	February SST	35-50N, 10-30W (northwest European Coast)
(15)	April–May SST	30-45N, 10-30W (northwest European Coast)
(16)	June–July SST	20-40N, 15-35W (northeast subtropical Atlantic Ocean)

SST – Sea Surface Temperature
SLP – Sea Level Pressure

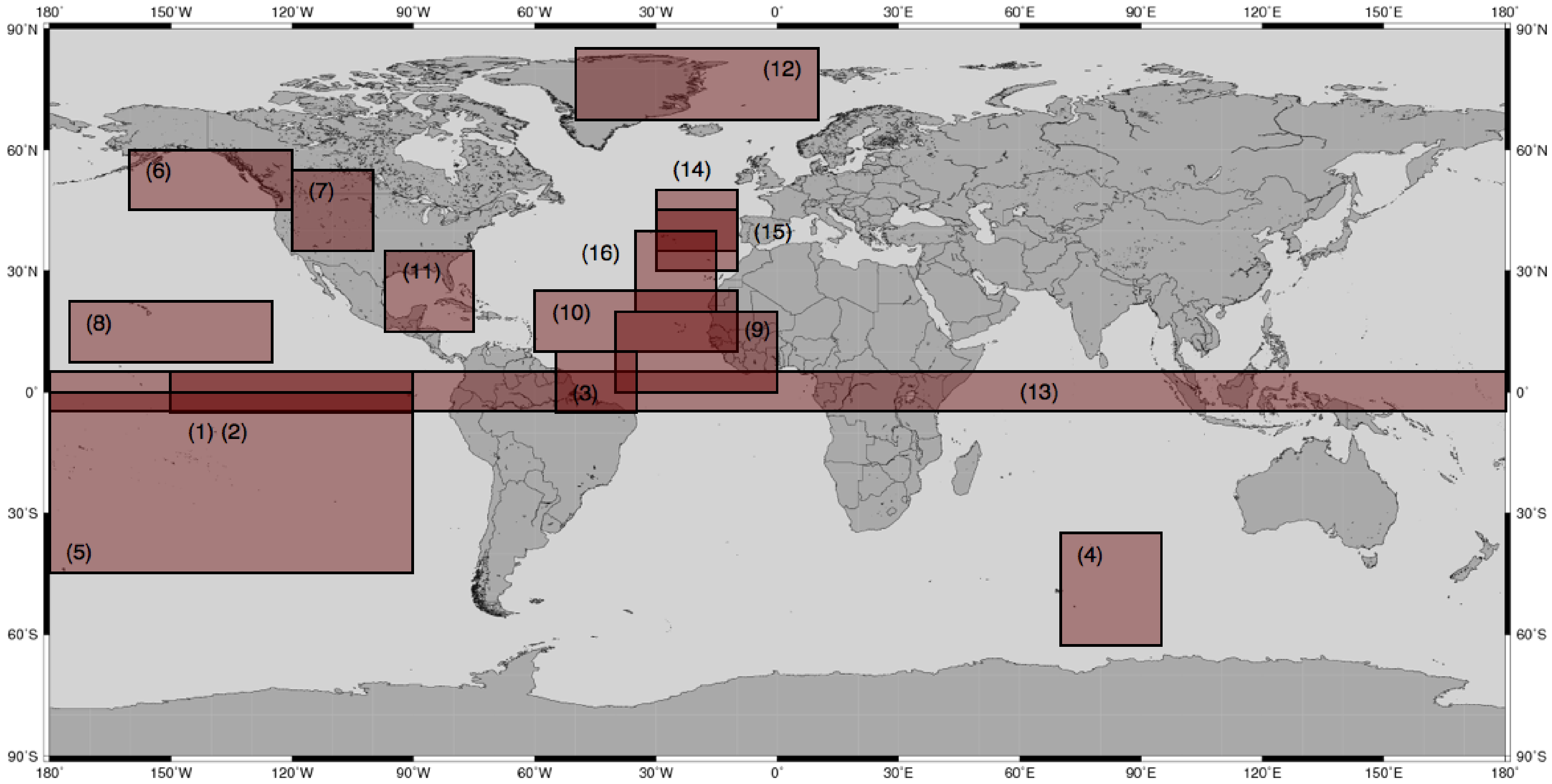


Figure 4.11

Geographic regions for the CSU predictors.

4.3.1.1 El Niño Variables

In a normal year, air rises in the western tropical Pacific (where the water is the warmest as well as slightly elevated) and sinks in the eastern tropical Pacific which is a phenomenon known as the Walker Circulation (see Figure 4.12 for the NOAA Geophysical Fluid Dynamics Laboratory [3]). During an El Niño event, the easterly surface trade winds that cause this water bulge in the western Pacific weaken, and the warm water travels eastward. Furthermore, El Niño conditions shift the upward portion of the Walker Circulation to the eastern Pacific, creating upper-level westerly winds in the Atlantic Ocean as well as subsidence. Both of these factors inhibit tropical cyclone formation and intensification in this region. Opposite conditions (abnormally strong trade winds and colder than normal eastern Pacific water) are called La Niña. La Niña years are associated with weak wind shear and little subsidence in the Atlantic, typically producing active tropical cyclone activity in this basin.

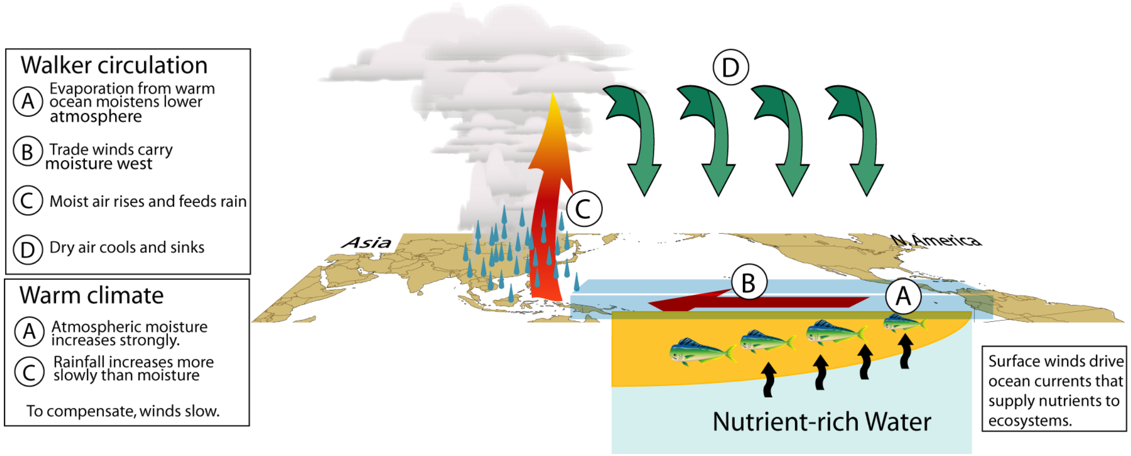

Walker circulation
A Evaporation from warm ocean moistens lower atmosphere
B Trade winds carry moisture west
C Moist air rises and feeds rain
D Dry air cools and sinks
Warm climate
A Atmospheric moisture increases strongly.
C Rainfall increases more slowly than moisture
To compensate, winds slow.
D
C
Asia
N. America
B
A
Nutrient-rich Water
Surface winds drive ocean currents that supply nutrients to ecosystems.

Figure 4.12

Walker Circulation illustration from NOAA.

El Niño events are characterized by several possible variables. The *June–July Niño 3* (1) variable represents sea surface temperature (SST) anomalies of the eastern equatorial tropical Pacific Ocean. Positive values of this variable indicate an El Niño event, and negative represents a La Niña event. *May SST in the eastern equatorial Pacific* (2) represents a similar relationship. The first clues of an impending El Niño can be detected in February by observing three variables. Upper-level westerly (zonal) wind anomalies off the northeast coast of South America imply that the upward branch of the Walker Circulation associated with ENSO remains in the western Pacific and that El Niño conditions are likely to be present in the eastern equatorial Pacific for the next four to six months. This situation is measured by the *February 200-mb zonal wind (U) in equatorial East Brazil* (3). Likewise, anomalous late winter meridional (north) winds at 200-mb in the South Indian Ocean are also associated with El Niño conditions (*February–March 200-mb V in the South Indian Ocean* (4)). Finally, sea level pressure (SLP) in the eastern Pacific south of the equator is a measure of the trade winds whereby weak trade winds (or westerly surface winds) are associated with lower SLP and, therefore, El Niño conditions, while the opposite is correlated to La Niña conditions. Therefore, *February SLP in the eastern South Pacific* (5) is a possible variable. Some fall variables are also correlated to El Niño conditions, such as the *October–November SLP in the Gulf of Alaska* (6), *September 500-mb Geopotential Height in western North America* (7), and *November SLP in the subtropical northeast Pacific* (8).

4.3.1.2 Sea Level Pressure Variables

Pressure in the Atlantic Ocean is also inversely related to tropical cyclone activity, and seems to contain both monthly as well as longer term relationships. Low SLP in the tropical Atlantic implies increased atmospheric instability, moisture, and ascent (more favorable for the genesis of tropical cyclones), and weaker trade winds (which correspond to less wind shear that can tear up the thunderstorms in tropical cyclones). Low SLP in the spring tends to persist through the summer and fall. Therefore, potential variables include *March–April SLP in the eastern tropical Atlantic* (9), *June–July SLP in the tropical Atlantic* (10), and *September–November SLP in the southeast Gulf of Mexico* (11).

4.3.1.3 Teleconnection Variables

The atmosphere is characterized by long-term oscillations which impact global wind patterns, known as teleconnections. Two of these oscillations related to tropical cyclone activity are the Arctic Oscillation and the North Atlantic Oscillation [17]. When these oscillations are in one phase, they cause more ridges in the Atlantic, which corresponds to less wind shear. Also, on decadal timescales, weaker zonal winds in the sub-polar areas are indicative of a relatively strong thermohaline circulation and therefore a warmer Atlantic Ocean. A variable which measures this oscillation is the *November 500-mb Geopotential Height in the North Atlantic* (12).

4.3.1.4 Quasi-Biennial Oscillation Variable

Research has also shown that the Quasi-Biennial Oscillation (QBO) is correlated to tropical cyclone activity. The QBO is a stratospheric (16 to 35 km altitude) oscillation of equatorial east-west winds which vary with a period of about 26 to 30 months or roughly 2 years. These winds typically blow for 12-16 months from the east, then reverse and blow 12-16 months from the west, then back to easterly again. The west phase of the QBO has been shown to provide favorable conditions for development of tropical cyclones, possibly because it reduces wind shear. A variable which measures the QBO is the *July 50-mb Equatorial Wind (U) around the globe* (13).

4.3.1.5 Atlantic Sea Surface Temperature Variables

The Atlantic SST is another major influence on tropical cyclone activity in that basin. Like SLP, winter and spring anomalies tend to persist throughout the season. Therefore, *February SST off the northwest European Coast* (14), *April–May SST off the northwest European Coast* (15), and *June–July SST in the northeast subtropical Atlantic* (16) are potential predictors. In addition, warm SST anomalies also tend to correlate with low SLP.

4.3.2 Climate Analysis Results

Stepwise regression with a backwards glance is used in this analysis, which selects the optimum number of most important variables using a predefined significance value (90% in this study). Stepwise regression can complement parallel coordinates by isolating the significant variables in a quantitative fashion. The interactive MDX system can then be used to develop a deeper understanding of the complex relationships between the variables.

An extra step is taken to ensure the proper selection of variables. The initially chosen variables are examined for multicollinearity; if any variables are correlated with each other by more than 0.5, one is removed and the code rerun. In this way, the chosen variables are approximately independent of each other. A normalization procedure is also executed for equal comparison between the variables and the predictor variable mean, $\bar{x}_i$, can be interpreted (to a first approximation) as a "threshold" value to distinguish between positive and negative contributions. Years when independent variables contain large deviations from the mean could be associated with very active or inactive years, and require closer examination. As will be seen, the geovisual analytics in MDX facilitate the examination of active and inactive Atlantic hurricane seasons.

The sixteen potential variables listed in Table 4.3 are examined in the stepwise regression, yielding several independent variables for each dependent variable. These results show that several climate factors impact tropical cyclone activity. The chosen predictors are shown in Table 4.4 along with their normalized regression coefficient and sample mean. The explained variance (R^2) is shown in the table sub-headings.

The stepwise regression shows only one significant El Niño variable (*late winter South Indian Ocean 200-mb meridional winds* (4)) impacts total number of storms; it is the second most influential predictor. *Late winter northwest coastal European SST* (14) is the leading predictor. The North Atlantic Oscillation (manifested by *500-mb geopotential height in the North Atlantic* (12)) ranks third, and is also the only variable seen in all three

Table 4.4

Stepwise regression models for *NS*, *H*, and *IH* categories.

Number of Named Storms (NS)
(R^2 is 34%)

Chosen Variables	Normalized Coefficients *b*	Sample Mean
Feb. SST (14)	0.302	13.8
Feb.–Mar. 200-mb V (4)	–0.244	2.5
Nov. 500-mb Geopot. Ht. (12)	0.232	5213.0
Sep.–Nov. SLP (11)	–0.175	1015.0

Number of Hurricanes (H)
(R^2 is 42%)

Chosen Variables	Normalized Coefficients *b*	Sample Mean
Oct.–Nov. SLP (6)	–0.284	1009.6
June–July SST (16)	0.259	22.2
Nov. 500-mb Geopot. Ht. (12)	0.258	5213.0
Sep.–Nov. SLP (11)	–0.208	1015.0

Number of Intense Hurricanes (IH)
(R^2 is 54%)

Chosen Variables	Normalized Coefficients *b*	Sample Mean
Nov. 500-mb Geopot. Ht. (12)	0.345	5213.0
June-July SLP (10)	–0.315	1016.2
Sep. 500-mb Geopot. Ht. (7)	0.292	5753.3
Feb. SST (14)	0.235	13.8

tables. This suggests that the presence of a ridge in the Atlantic is conducive to an above average tropical cyclone season. Finally, low *September–November SLP in the southeast Gulf of Mexico* (11) also encourages the formation of tropical cyclones. Note that the coefficient has a negative sign, showing that the lower the pressure, the better the chance of tropical cyclone activity.

For number of hurricanes, the analysis surprisingly shows that *October–November SLP in the Gulf of Alaska* (6) is the most important predictor. The physical role is not clear, although scientists know it is correlated to El Niño activity. *Northeast subtropical Atlantic SST* (16) and *North Atlantic 500-mb geopotential height* (12) are tied for second, and *southeast Gulf SLP* (11) again ranks fourth. The explained variance is 42%—more than the 34% for named storms. This suggests stronger predictor relationships for number of hurricanes.

For intense hurricanes, the variance increases to 54%. In this case, the *North Atlantic November 500-mb geopotential height* variable (12) is the strongest predictor. *Early summer tropical Atlantic SLP* (10) ranks number two, followed by *September 500-mb geopotential height in western North America* (7) and *February SST off the northwest coastal Europe* (14). The higher variance and distinctly different chosen predictors suggests different environmental influences are required for intense hurricanes. This analysis correlates the presence of high pressure in the western U.S. and over the Atlantic, low summer Atlantic SLP, and warm SST as necessary conditions for intense hurricanes.

Because there is unexplained variance and several predictors, can parallel coordinates glean any more insight? To answer this question, the data sets are stratified into below normal, normal, and above normal seasons using MDX's interactive capabilities, and the significant predictors identified by the stepwise regression are analyzed visually. Using the key statistical indicators, the below normal, normal, and above normal seasons are determined by moving the query sliders for the axis of interest to encapsulate the lines above the standard deviation range, within the standard deviation range, and below the standard deviation range, respectively. After setting the query sliders, the aerial perspective shading highlights the relationships of interest for a particular polyline (storm year record), thus enabling rapid analysis of the variables.

Figure 4.13 shows a plot for seasons with below normal named storms (sample size of sixteen). Even though the regression shows *February Atlantic SST* (14) as the most important overall predictor, it is not as effective for discerning inactive seasons. The plot shows considerable scatter, and with only six years of significantly below average SST. The dynamic query capabilities of MDX make these combined queries and subsample analysis a nearly effortless exercise. *September–November Gulf of Mexico SLP* (11) also exhibits much scatter, with a slight majority of years with above normal pressure. However, *February–March 200-mb South Indian Ocean meridional winds* (4)—a surrogate measurement of El Niño—shows fifteen seasons (94%) of strong north winds, tightly clustered in the plots. *This suggests El Niño is the major contributor to inactive Atlantic tropical cyclone seasons.* Note also that below normal *November North Atlantic 500-mb geopotential heights* (12) plays a pivotal role for quiet seasons. Fourteen seasons (87%) contain lower geopotential heights in November, suggesting the presence of upper-level troughs which can shear tropical cyclones. However, this signal is not as strong as the El

Niño predictor. Additionally, many non-query lines exist for positive 200-mb V, showing that other factors besides El Niño contribute to normal and active seasons. In fact, a similar parallel coordinates stratification analysis shows that *November North Atlantic 500-mb geopotential heights* (12) and *September–November Gulf of Mexico SLP* (11) tend to be the critical players for an active tropical cyclone season (not shown).

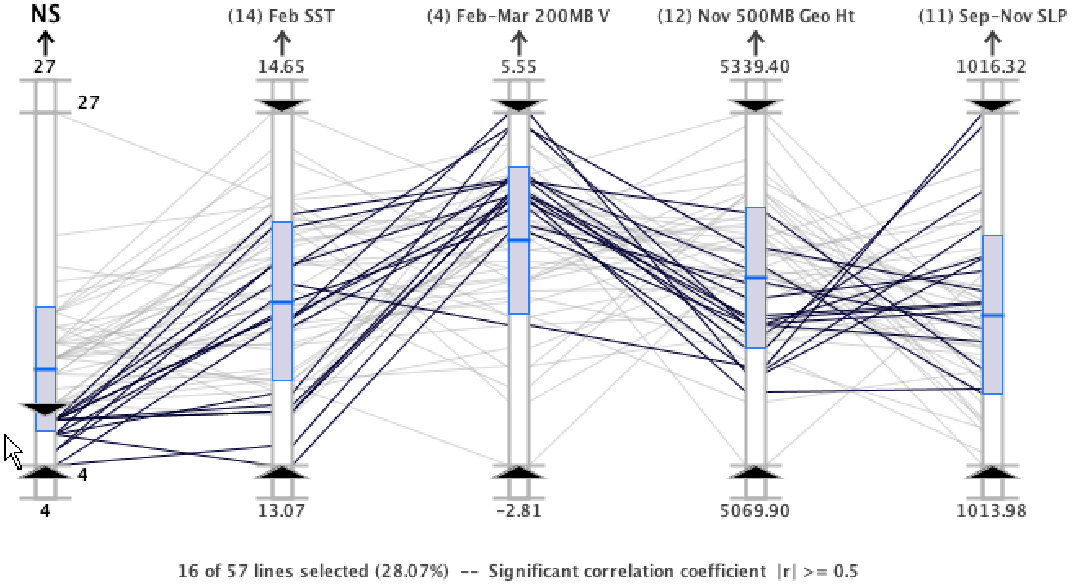


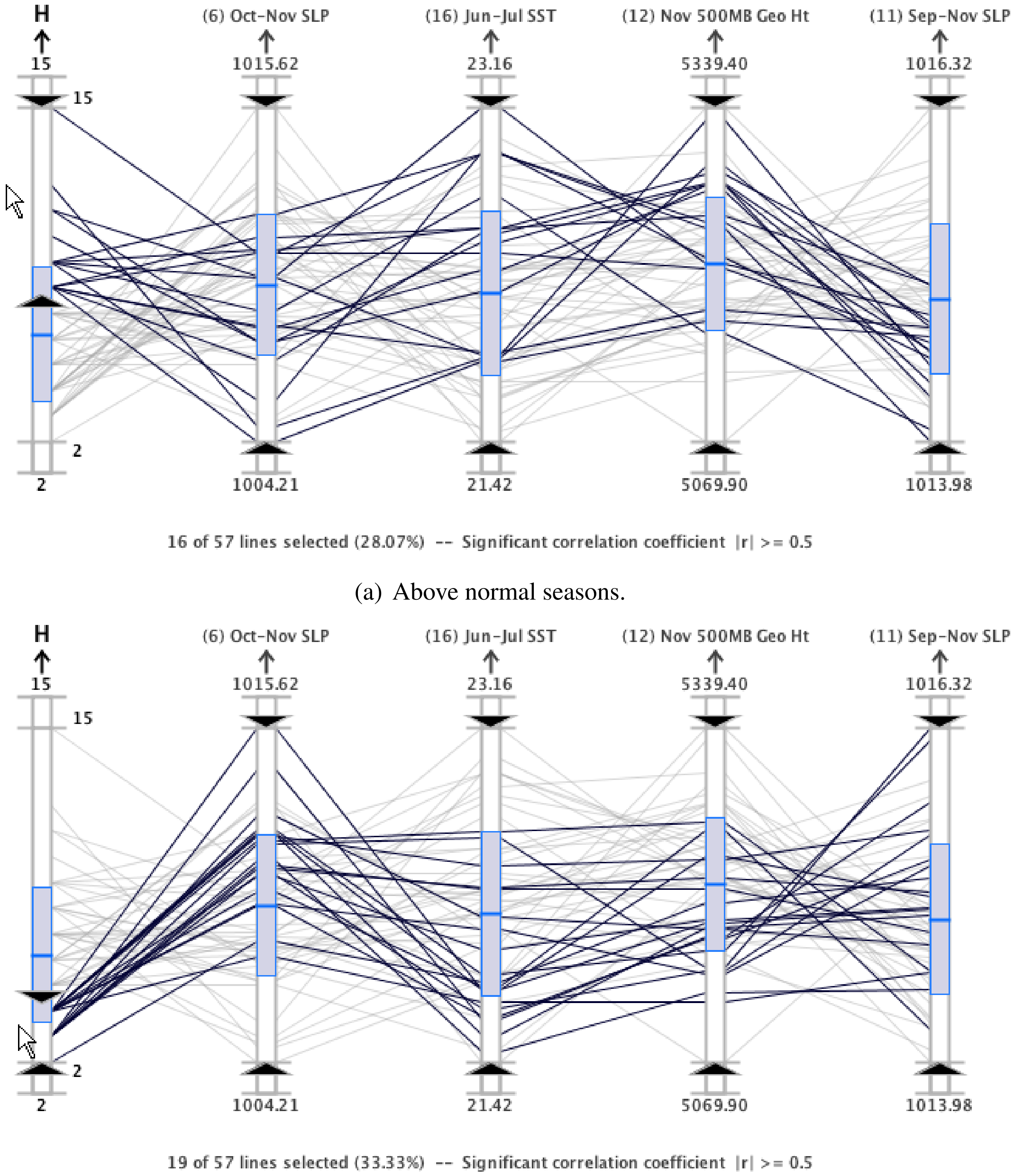
Figure 4.13

NS regression model for below normal seasons (1950 to 2006).

Figure 4.14(b) shows seasons with below normal hurricane activity (nineteen seasons). El Niño again tends to dominate the signal through the *fall Gulf of Alaska SLP* (6) term. However, in contrast to number of named storms, *Atlantic SST* (16) becomes important for number of hurricanes. This suggests that when water temperature is below normal, tropical storms will have difficulty reaching hurricane status. For above normal hurricane activity (see Figure 4.14(a)), *November North Atlantic 500-mb geopotential height* (12) and *Gulf of Mexico SLP* (11) tend to exert dominant roles, with El Niño and *June–July Atlantic SST* a secondary factor.

Intense hurricanes warrant special consideration, since they cause 80% of the economic damage from tropical cyclones. Figure 4.15(b) shows that cold *February Atlantic SST* (14) and high *Atlantic June–July SLP* (10) tend to reduce the number of intense hurricanes, with *November North Atlantic 500-mb geopotential heights* (12) playing a secondary role and *September 500-mb geopotential height in western North America* (7) contributing no role. In contrast, all four predictors have tightly clustered lines showing they all play dominant roles in seasons with above normal intense hurricane activity (see Figure 4.15(a)). These terms are associated with the presence of ridges in the western U.S. and the Atlantic, below average Atlantic SLP, and warm wintertime Atlantic SST off the northwestern European Coast. Ridges are low shear environments, showing that the lack of upper level troughs is an important factor for seasons with many intense hurricanes. Low SLP indicates minimal subsidence. Sinking air suppresses cloud growth and also dries the lower atmosphere, both of which are not conducive to the formation and development of tropical cyclones. Low SLP also could indicate better organized tropical waves (from which many Atlantic tropical cyclones form). Warm wintertime northeast Atlantic water also is a good precursor for above average intense hurricane activity.

We can also use MDX's geovisual analytics to investigate the differences between the extremely busy 2005 season and the slightly below average 2006 season. Figure 4.16 shows the 2005 and 2006 seasons along with the chosen predictors from all three cate-

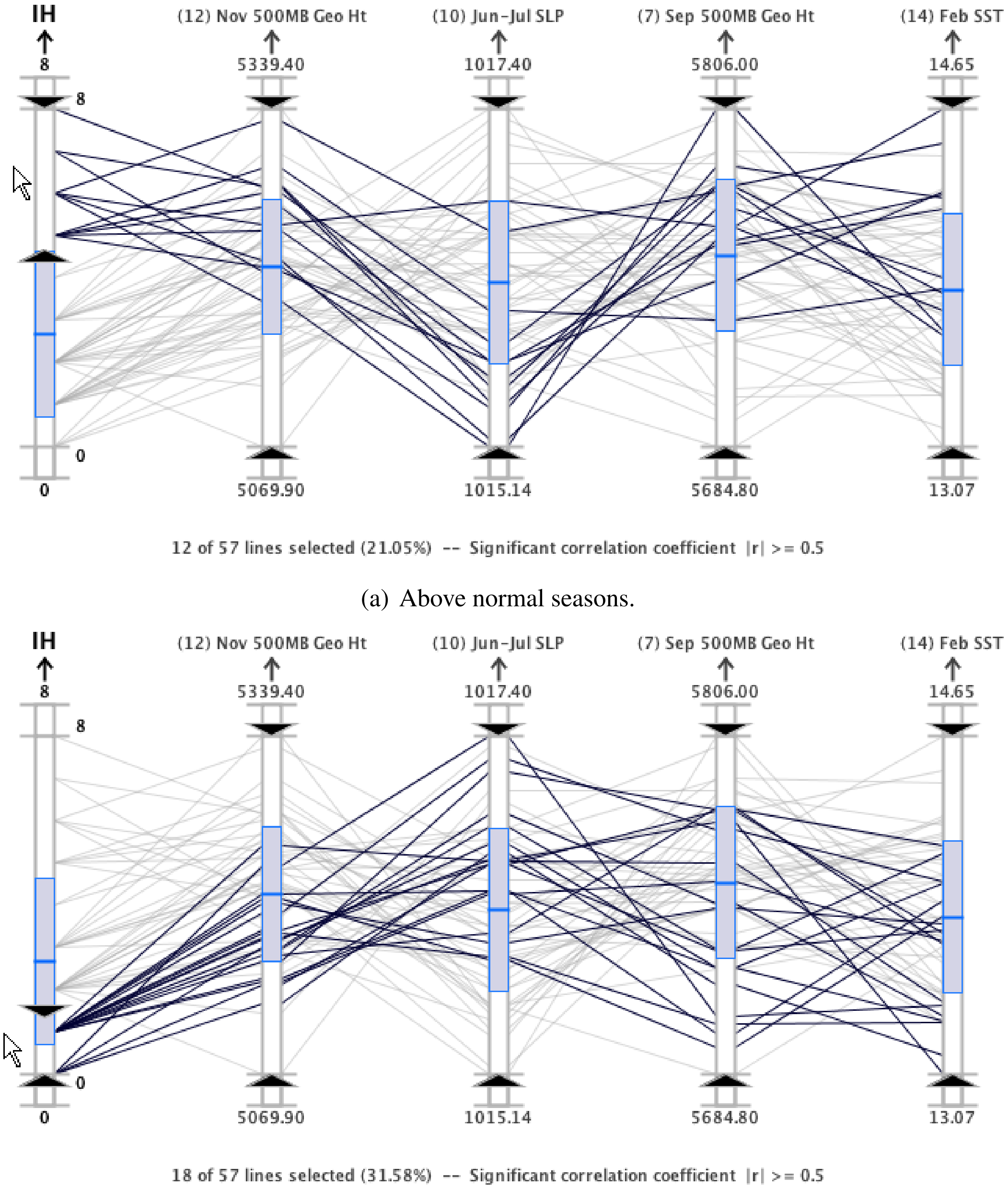


(a) Above normal seasons.

(b) Below normal seasons.

Figure 4.14

H regression model for above and below normal seasons (1950 to 2006).



(a) Above normal seasons.

(b) Below normal seasons.

Figure 4.15

IH regression model for above and below normal seasons (1950 to 2006).

gories (named storms, hurricanes, and intense hurricanes) listed in Table 4.4. This plot reveals that most of the terms are nearly the same except for *October–November SLP in the Gulf of Alaska* (6) (above average in 2005, below average in 2006) and *June–July SLP in the tropical Atlantic* (10) (below average in 2005, above average in 2006). Klotzbach and Gray [100] and Bell et al. [18] demonstrated that the tropical Atlantic was quite dry through most of the 2006 hurricane season due to subsidence associated with the onset of an unusually late ENSO event (indicated by the *Gulf of Alaska SLP* (6) term), as well as frequent outbreaks of African dust storms that year.

4.4 Case Study 3: North Atlantic Intense Hurricane Climate Study

In the third case study, we use the same CSU data set employed in the second case study to demonstrate the utility of an extended version of the MDX system, which includes statistical analysis capabilities. Whereas, in the second case study, the regression analysis is executed externally in a separate process using the IMSL® software[2], the MDX system has been modified for the second case study to provide integrated access to multiple and single linear regression, correlation analysis, and new statistical indicators. The primary objective of this study is to discover the most important predictors for seasonal intense hurricane activity in the North Atlantic to improve forecasting skill. The secondary objective is to identify additional associations between predictors and temporal patterns in the data. In the remainder of this section, the results of this study are described.

[2] The IMSL® numerical libraries are developed by Visual Numerics. More information can be found on the software at http://www.vni.com/products/imsl/.

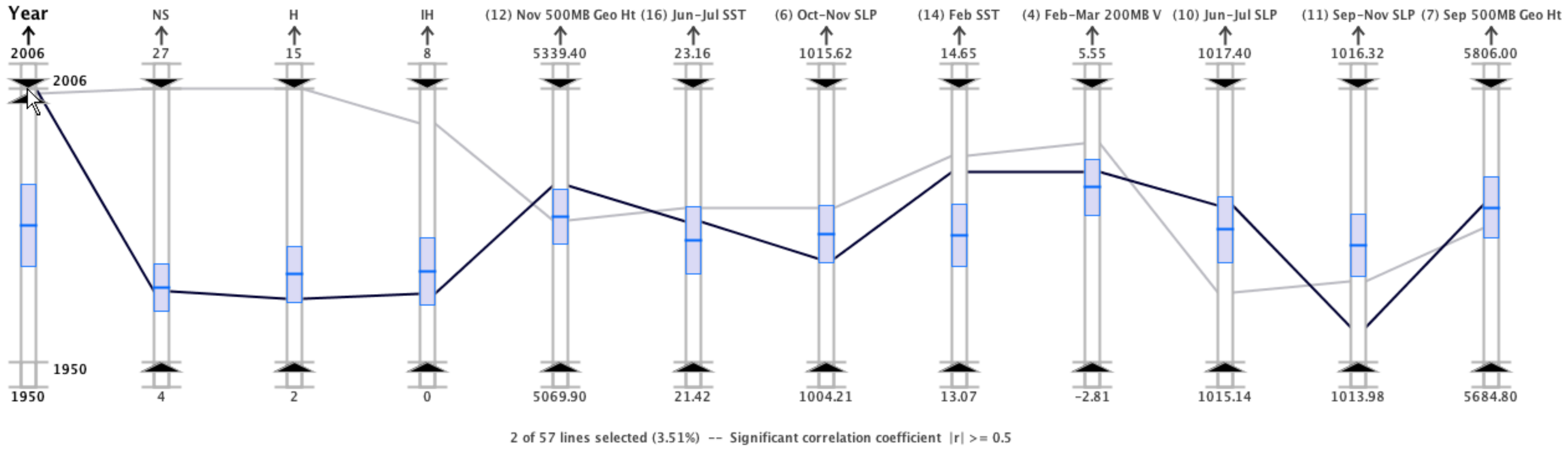


Figure 4.16

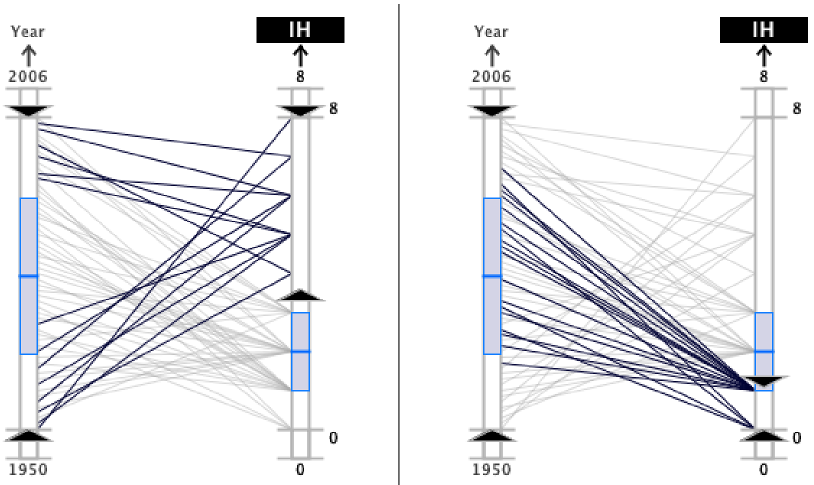
Comparison of predictors in 2005 and 2006 hurricane seasons.

4.4.1 Climate Analysis Results

After loading the predictors and seasonal storm statistics, the visual analysis tools are used to explore the data set and rearrange the axes. In Figure 4.17, the active and inactive *IH* seasons are highlighted. This plot reveals a gap on the *Year* axis (the first axis in Figure 4.17(a)) for the active seasons. From 1960 to 1994, a relatively quiet period is observed—there are no seasons with an above normal number of intense hurricanes. What's more, Figure 4.17(b) shows that the inactive seasons are clustered into this same time of normal or below normal activity. This visual observation agrees with findings published in the weather research literature [54, 100, 101], which suggests a strong multidecadal variability in the number of intense hurricanes per year in the North Atlantic. Another notable observation is that most of the predictors have low variability (evident by the relatively small overall IQRs) except for the *July 50-mb Equatorial Wind (U) around the globe (13)* predictor (the first axis in Fig. 4.17).

4.4.1.1 Correlation Analysis

To prepare for the MLR analysis and to address the secondary objective of the study, the correlations between the predictors are investigated by arranging the sixteen axes by the correlation coefficient with the *IH* axis. In Figure 4.18, the dependent axis, *IH*, is highlighted to show its correlation with each individual predictor. The predictor axes are also arranged in descending order according to each predictor's correlation coefficient with the *IH* axis; negative correlations are arranged on the left and positive correlations are arranged on the right of the *IH* axis. The correlation indicators and axis positions reveal



(a) Active IH seasons.

(b) Inactive IH seasons.

Figure 4.17

Multidecadal variability in *IH* activity shown in parallel coordinates.

that the strongest correlations with the *IH* axis are *June–July SLP in the tropical Atlantic* (10) and *November 500-mb Geopotential Height in the far North Atlantic* (12)—the axes directly to the left and right of *IH* in Figure 4.18, respectively. More specifically, the enlarged color-coded correlation indicator box, 'X'-shaped crossings in parallel coordinates, downward slope in the scatterplot, and numerical display of *r* in this plot reveal that axis (10) has the strongest negative correlation. Likewise, the strongest positive correlation with axis (12) is evident by the correlation indicator, the more parallel polylines in parallel coordinates, the upward slope of the scatterplot, and the numerical display of *r*.

The image sequence shown in Figure 4.19 illustrates the use of the continuous aerial perspective shading capability to investigate a strong negative correlation between *October–November SLP in the Gulf of Alaska* (6) and *November SLP in the Subtropical NE Pacific* (8) axes. The mouse is moved from the top to the bottom of axis (8) in this image se-

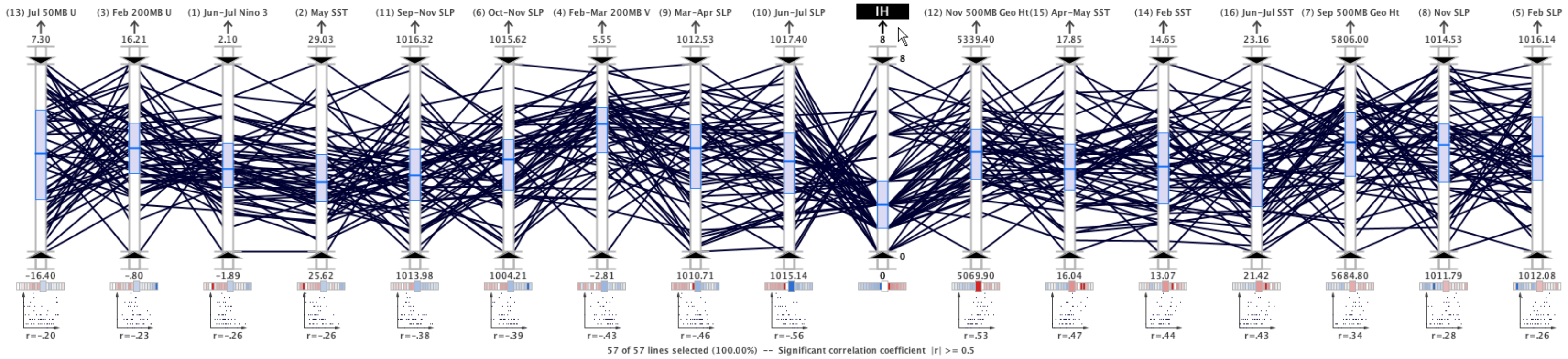
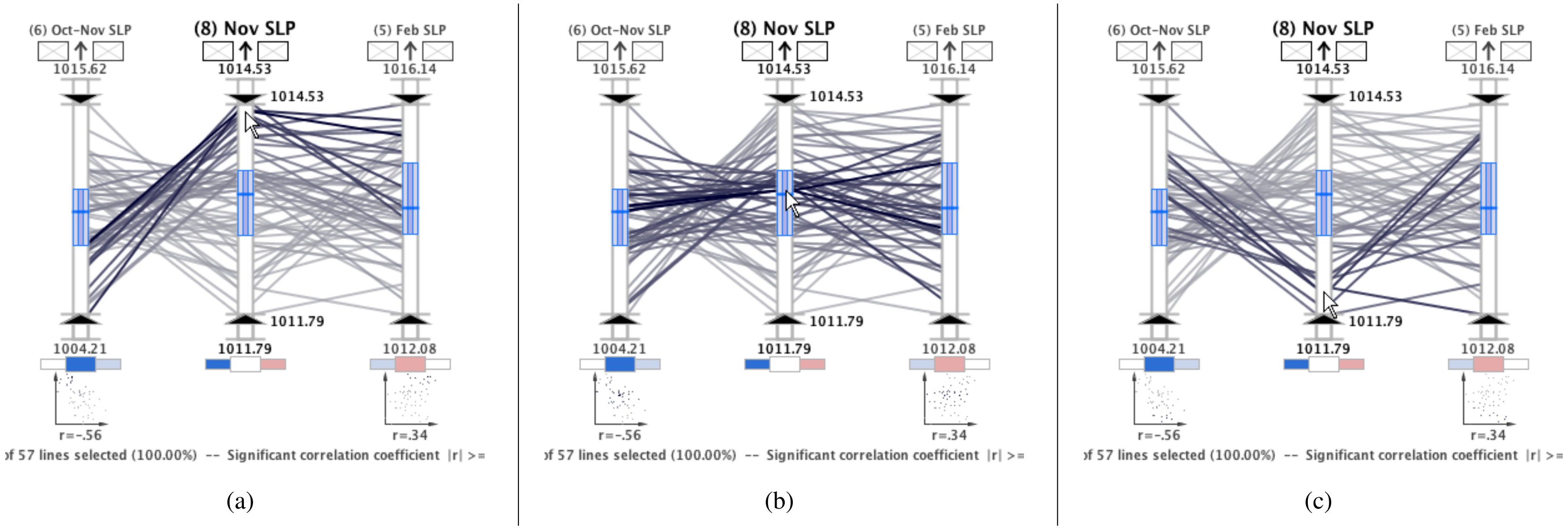


Figure 4.18

Correlation analysis of potential predictors with the *IH* axis.



(a)

(b)

(c)

Figure 4.19

SLP correlation analysis using continuous aerial perspective shading.

quence. This intuitive visual query technique, which shades the polylines according to their proximity to the mouse cursor, highlights the 'X'-shaped polyline crossings between these axes, which is indicative of a negative correlation in parallel coordinates. The negative correlation polyline configuration can be compared to the more parallel polyline configuration between the positively correlated axis (8) and *February SLP in the eastern South Pacific* (5) axis in Figure 4.19.

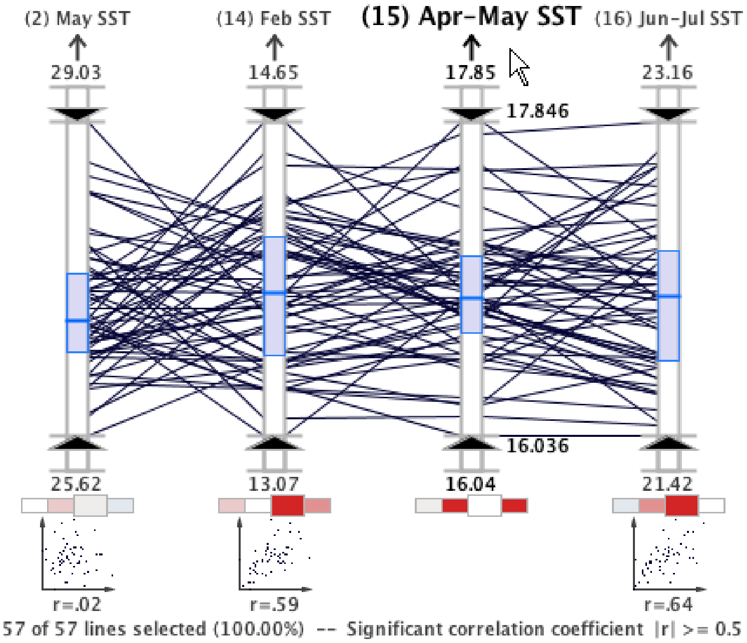


Figure 4.20

Correlation analysis of SST predictors in MDX.

In Figure 4.20, the correlations between three SST predictors and the *April–May SST off the Northwestern European Coast* (15) predictor are shown. In the parallel coordinate plot, strong correlations are identified when $|r| \geq 0.5$, the significant correlation threshold, and visually by a fully saturated correlation indicator. This plot reveals that a relatively strong positive correlation exists between axis (15) and both the *February SST off the Northwestern European Coast* (14) and the *June–July SST in the Northeastern Subtropical Atlantic* (16) axis. Meanwhile, the *May SST in the eastern equatorial Pacific* (2) predictor exhibits almost no correlation ($r = .02$).

To reduce the multicollinearity between the SST predictors, axis (14) and (16) must be removed since they have a strong correlation with axis (15) and axis (15) has a stronger correlation with the *IH* axis (see Figure 4.21). Removing these and any other variables with strong correlations between predictors will ensure the independence of the predictors and thus improve the MLR analysis results.

Before removing axis (14) and (16), the physical relationships between these variables can be considered in order to investigate the cause of the strong correlation. From the geographic extents of these variables listed in Table 4.3 and show in Figure 4.11, we observe that the three SST predictors with strong correlations are all sampled in the North Atlantic Ocean. However, axis (2), which exhibits a very weak correlation, is measured in the Pacific Ocean. Therefore, the strong correlations among axis (14), (15) and (16) can be mostly attributed to the close geographical proximity of the measurements whereas the low correlation of axis (2) can be attributed to the fact that it is measured in the Pacific ocean.

We also discover that the strongest correlation in the data set is between the *June–July Niño 3* (1) and *May SST* (2) predictors where $r = .78$ (Figure 4.22). As indicated by the geographical regions for these predictors (see Figure 4.11 and Table 4.3), a primary

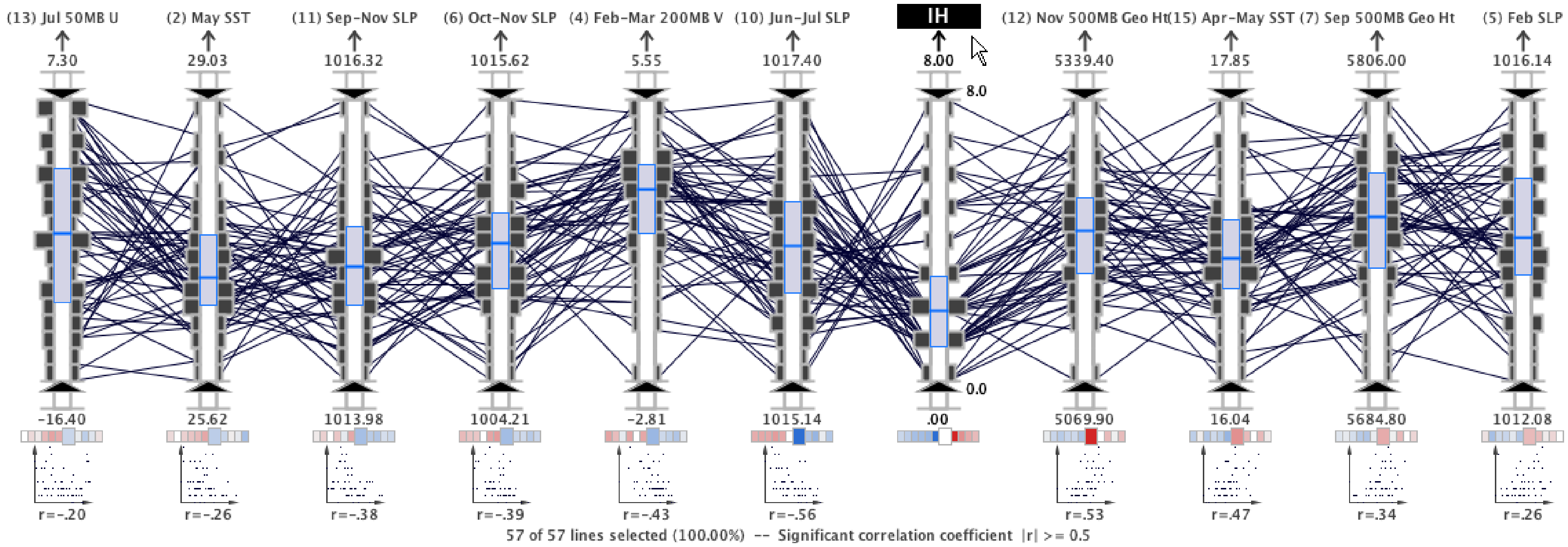


Figure 4.21

Axis configuration after multicollinearity filter execution with *IH* category.

reason for the strong positive correlation between these predictors is due to the common geographical region for measurement. Also, the fact that these predictors are measured at nearly the same time of the year contributes to the association.

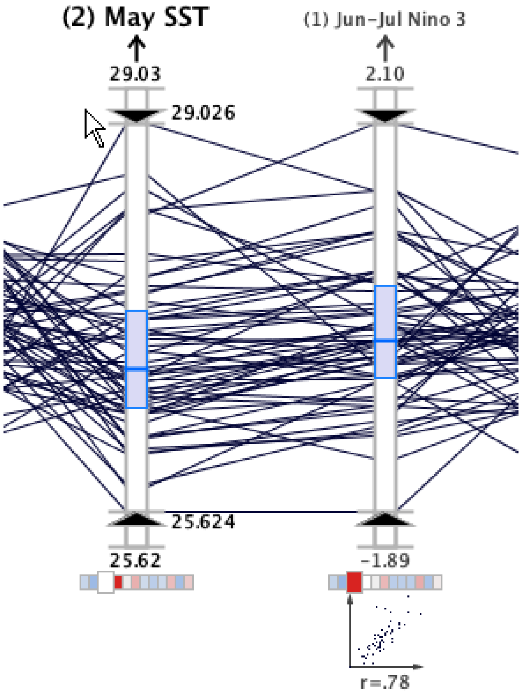


Figure 4.22

Correlation between *June–July Niño 3* (1) and *May SST* (2) variables.

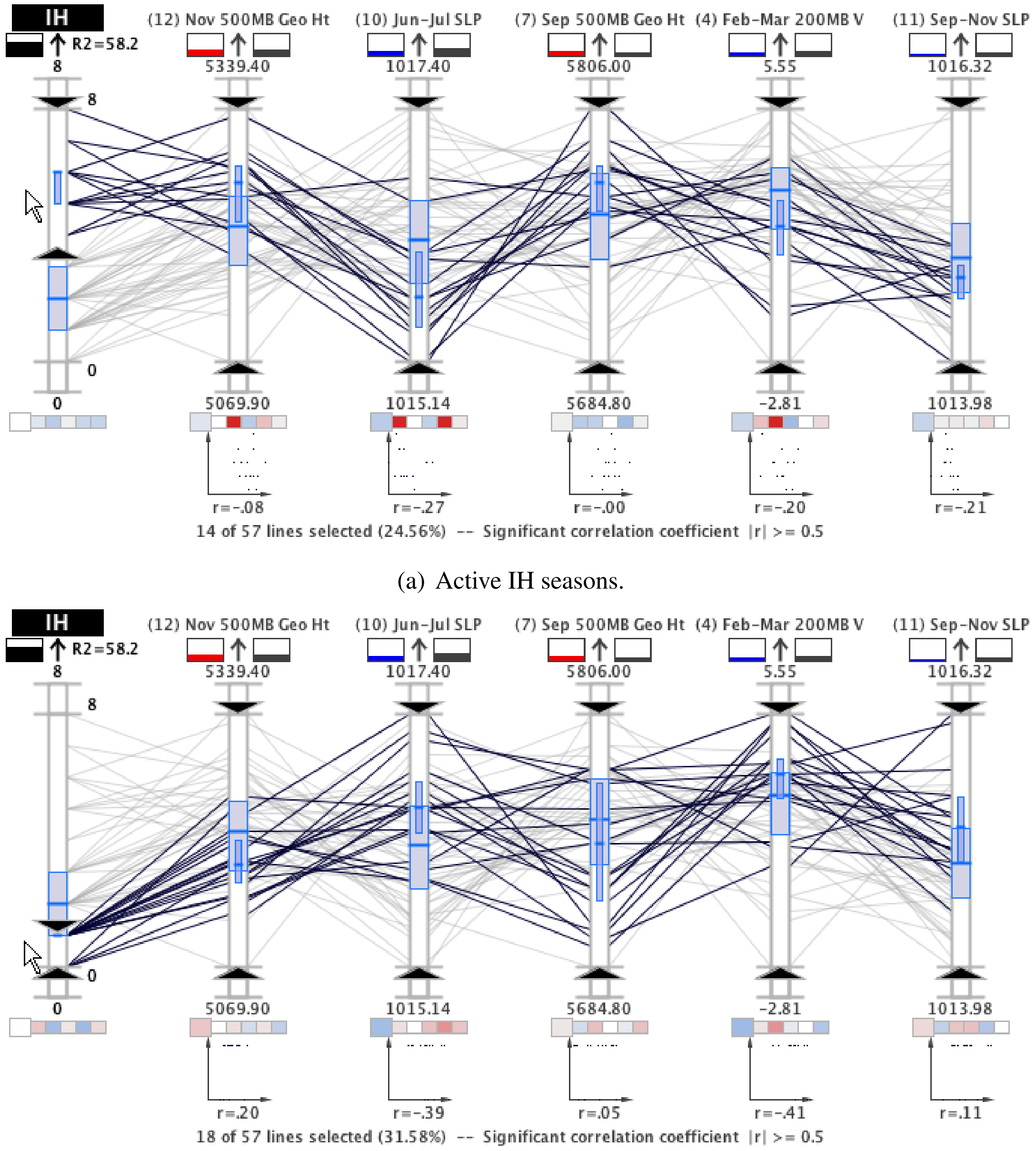
The scientist can continue to employ the correlation indicators to manually find and eliminate the highly correlated predictors, or use the system's automatic multicollinearity filter. Applying this filter to the climate data set removes *March–April SLP in the eastern tropical Atlantic* (9) (because of its strong correlation with axis (10)), axis (14) and (16) (strong correlation with axis (15)), *November SLP in the Subtropical NE Pacific* (8) (strong correlation with *October–November SLP in the Gulf of Alaska* (6)), *June–July Niño 3* (1) (strong correlation with axis (2)), and *February 200-mb zonal wind (U) in Equatorial East Brazil* (3) (strong correlation with *February SLP in the Southeast Pacific* (5)). In Figure 4.21, the resulting axis configuration is shown, arranged by the correlation coefficient with the *IH* axis. In this plot, we find that the only remaining *r* values greater than the significant correlation threshold (visually indicated by the fully saturated fill color in the enlarged correlation indicators) are the two axes on either side of the *IH* axis; but these correlations are with the dependent axis which does not affect the independence between the predictors.

4.4.1.2 Significant Predictors Identified with Regression

Using our system's automatic SLR and stepwise MLR processes, the predictors are automatically analyzed to determine the most important predictors with respect to the number of intense hurricanes in a season. In Figure 4.23, the results of the MLR and SLR analysis are shown. Here the predictors are arranged according to the magnitude of the MLR coefficient, *b*. The significance level in the stepwise regression analyses is 80%.

The numerical results of the regression listed in Table 4.5 and the visual representation shown in Figure 4.23 suggest that the five chosen variables are the most significant predictors for the number of intense hurricanes in a season. Highlighting the active and inactive ranges in Figure 4.23 also reveals how each specific variable behaves in either active or inactive seasons.

In order to validate the selection of these five predictors from a weather science perspective, we can evaluate the physical relationships between each predictor and the de-



(a) Active IH seasons.

(b) Inactive IH seasons

Figure 4.23

IH regression model for active and inactive seasons (1950 to 2006).

Table 4.5

Stepwise regression model for *IH* category.

Number of Intense Hurricanes (IH)
(R^2 is 58%)

Chosen Variables	Normalized Coefficients *b*	Sample Mean
Nov. 500-mb Geopot. Ht. (12)	0.3524	5213.38
June–July SLP (10)	–0.3121	1016.23
Sep. 500–mb Geopot. Ht. (7)	0.2514	5753.33
Feb.–Mar. 200-mb V (4)	–0.1871	2.53
Sep.–Nov. SLP (11)	–0.1431	1014.98

velopment of tropical cyclones. The most significant predictor, axis (12), which, as mentioned earlier measures the the long-term oscillations which impact global wind patterns, known as teleconnections. The MLR results indicate that when predictor (12) is normal or above normal, the environment is more favorable for the development of intense hurricane systems.

Pressure in the Atlantic Ocean is inversely related to tropical cyclone activity; low sea-level pressure in the tropical Atlantic implies increased atmospheric instability, moisture, and ascent (more favorable for the genesis of tropical cyclones), and weaker trade winds (which correspond to less wind shear that can tear up the thunderstorms in tropical cyclones). This relationship explains the selection of axis (11) and axis (10), which are normal or below normal in seasons with above normal *IH* activity.

The MLR analysis also identified two variables that characterize El Niño events which inhibit tropical cyclone formation and intensification in the Atlantic. The first clues of an impending El Niño can be detected in February by observing three variables. The

MLR analysis selected one of these variables, axis (4), which measures the anomalous late winter meridional winds at 200-mb in the southern Indian Ocean (a condition associated with El Niño). As shown in Figures 4.24 and 4.25, normal to below normal values of (4) correspond to more favorable conditions for intense hurricane development. The MLR model includes one Fall variable, axis (7), that is correlated to El Niño conditions for the following year. This predictor is more favorable for hurricane intensification in normal to above normal measurements.

Having identified the most significant predictors and validated their selections, the next step is to determine the range of values for each predictor that corresponds to the above normal *IH* seasons. In Figure 4.24, the query sliders are used to highlight the points with high values on axis (12), low values for axis (16), low values for axis (7), high values for axis (4), and low values for axis (11). This plot reveals that using these axis ranges to predict the intense hurricanes of a season would result in successfully identifying eleven of the fourteen seasons (74%) that had a high number of intense hurricanes between 1950 and 2006. On the other hand, using this technique might result in missing three seasons with above normal activity (with seven, six, and five intense hurricanes). In particular, one of the storm seasons that is not selected by this query is the infamous 2005 hurricane season which had seven intense hurricanes, including the cataclysmic Hurricane Katrina. Using the visual query capabilities, minor adjustments can be applied to the query sliders of the significant predictors to ensure that all fourteen seasons with active intense hurricane activity are captured (see Figure 4.25). Then, the predictor ranges can be used to predict

Table 4.6

Above normal *IH* activity predictor ranges.

Predictor Name	Initial Range	Tweaked Range
Nov. 500-mb Geopot. Ht. (12)	> 5170.00	> 5170.00
June–July SLP (10)	< 1016.60	< **1016.78**
Sep. 500–mb Geopot. Ht. (7)	> 5732.00	> **5730.50**
Feb.–Mar. 200-mb V (4)	< 3.62	< **3.87**
Sep.–Nov. SLP (11)	< 1015.26	< 1015.26

the activity of future tropical cyclone seasons with respect *IH* activity. The significant predictor ranges that are shown in these plots are listed numerically in Table 4.6.

In Figure 4.26, the query sliders are reset on the *IH* axis to include all storm years that fall within the refined predictor ranges. From this plot and the plot in Figure 4.27, we find that an additional eleven seasons are falsely identified as seasons with above normal *IH* activity when, in fact, the seasons show normal activity levels. These eleven seasons represent type I errors (also called false positives) where the test returns a positive result when the actual condition is absent. That is, the query ranges indicated the season has above normal *IH* activity when, in reality, the actual activity is not above normal. From the perspective of human safety, we prefer to have type I errors instead of type II errors (also called false negatives) where we fail to forecast the above normal activity season when, in truth, the activity is above normal. After adjusting the initial query ranges, we eliminate the three type II errors previously mentioned, but the eleven type I errors remain.

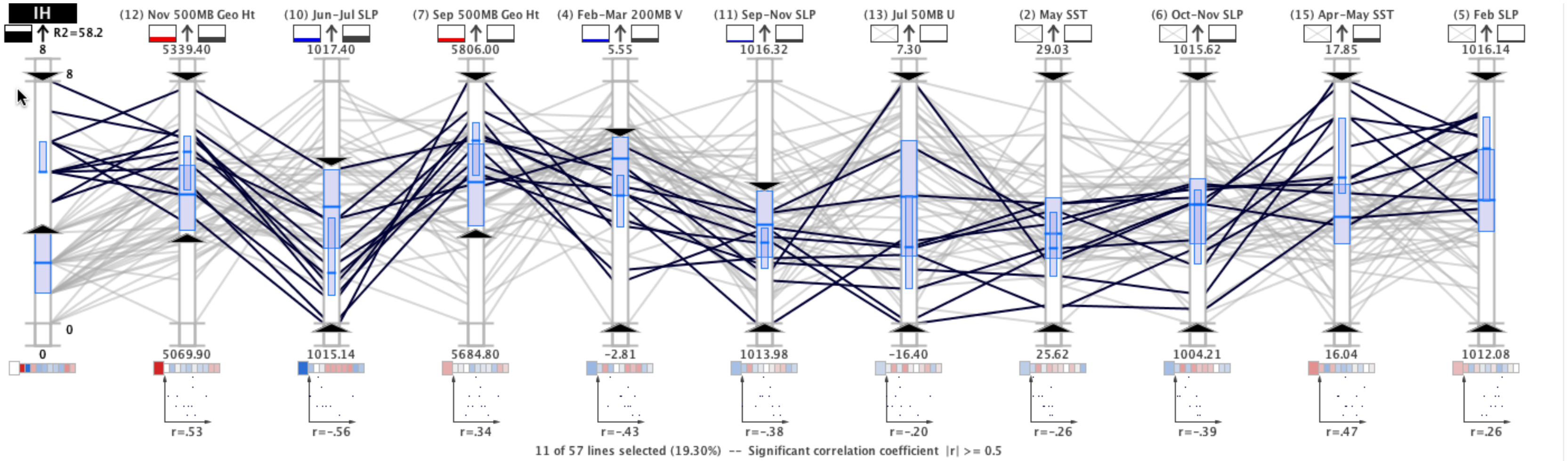


Figure 4.24

Initial query ranges for predicting above normal *IH* activity.

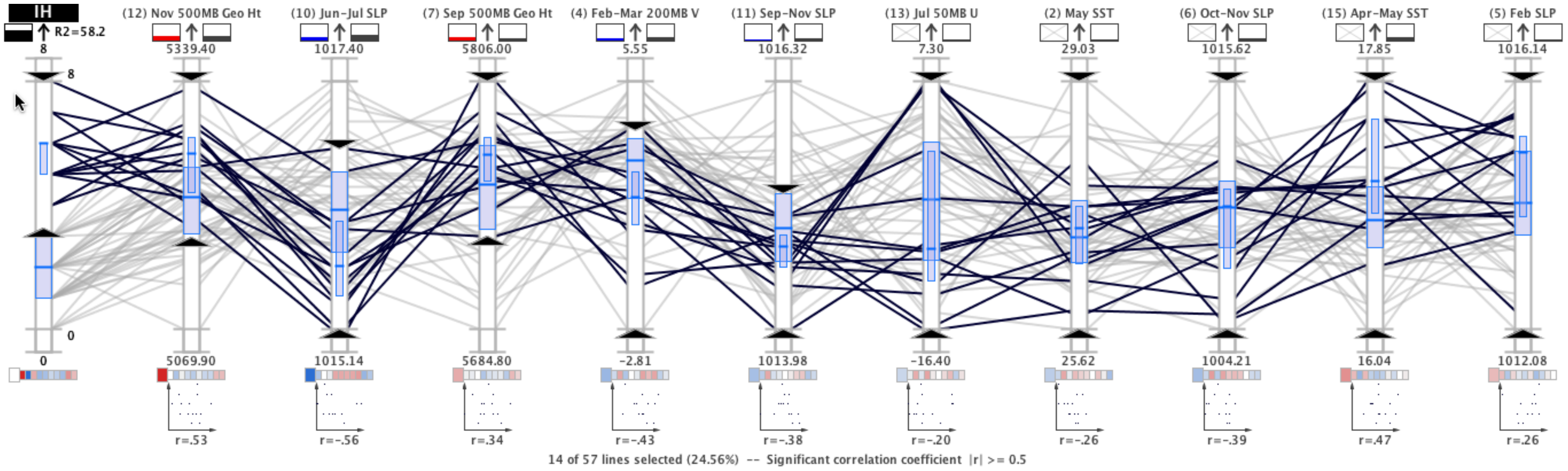


Figure 4.25

Adjusted query ranges for predicting above normal *IH* activity.

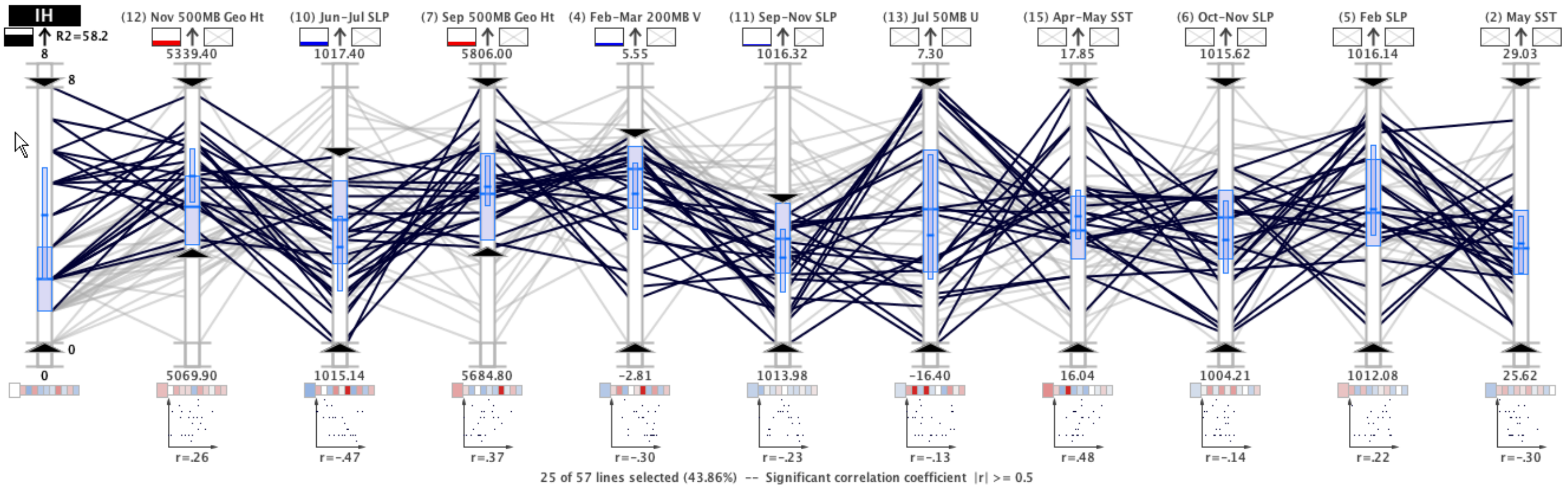


Figure 4.26

Seasons within the refined predictor ranges for above normal *IH* activity.

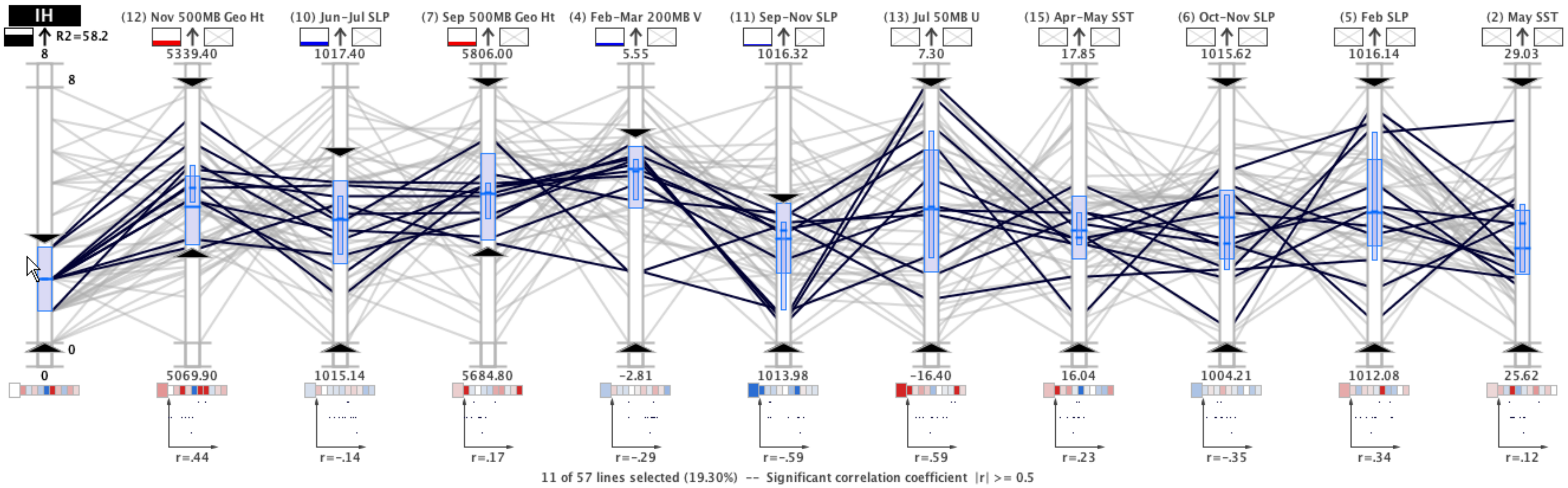


Figure 4.27

Non-active *IH* seasons within predictor ranges for active *IH* seasons.

4.4.1.3 Significant Predictors Identified without Regression

We can also omit the automatic MLR regression analysis and instead rely on the descriptive statistical indicators, correlation analysis capabilities, and visual query capabilities of MDX to find the most important predictors. First we analyze the correlation coefficients with the *IH* axis. As shown in Figures 4.18 and 4.21, axes (10) and (12) have the highest negative ($r = -.56$) and positive correlation ($r = .53$), respectfully. In fact, these are the only predictors with a correlation coefficient over the significant correlation threshold, ± 0.5. Next, the query sliders are used to isolate the seasons with above normal *IH* activity and the axes are sorted by the correlation coefficient for the queried seasons. The resulting configuration reveals that, for the above normal *IH* seasons, the strongest correlation is with the *May SST* (2) axis ($r = -0.39$), followed closely by the *November SLP* (8) axis ($r = -0.38$); but no predictors have a correlation coefficient above the significant correlation threshold. Based on these correlation analyses, no significant conclusions can be formed about the most important predictors of above normal *IH* activity.

Now the descriptive statistics and visual query techniques will be used to search for the most significant predictors. With the above average *IH* seasons still highlighted, the axes are rearranged in MDX based on the IQR ranges (Figure 4.29(a)). In this plot, the *September–November SLP* (11), *February 200-mb zonal winds* (3), and *September 500-mb Geopotential Height* (7) predictors have the three tightest clustering of lines, which manifests visually by the query quartile plots. With the same axis arrangement, the seasons with below average *IH* activity are isolated (see Figure 4.29(b)) to compare the behavior

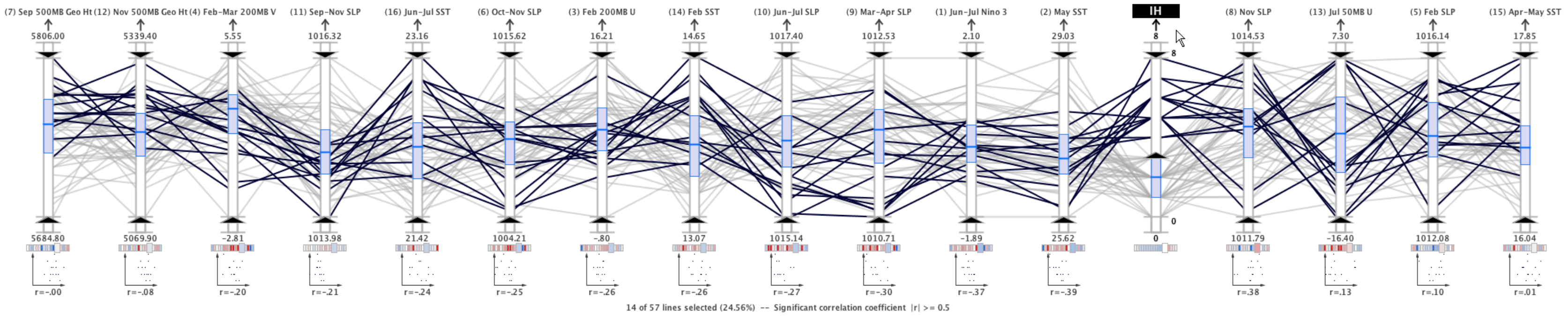
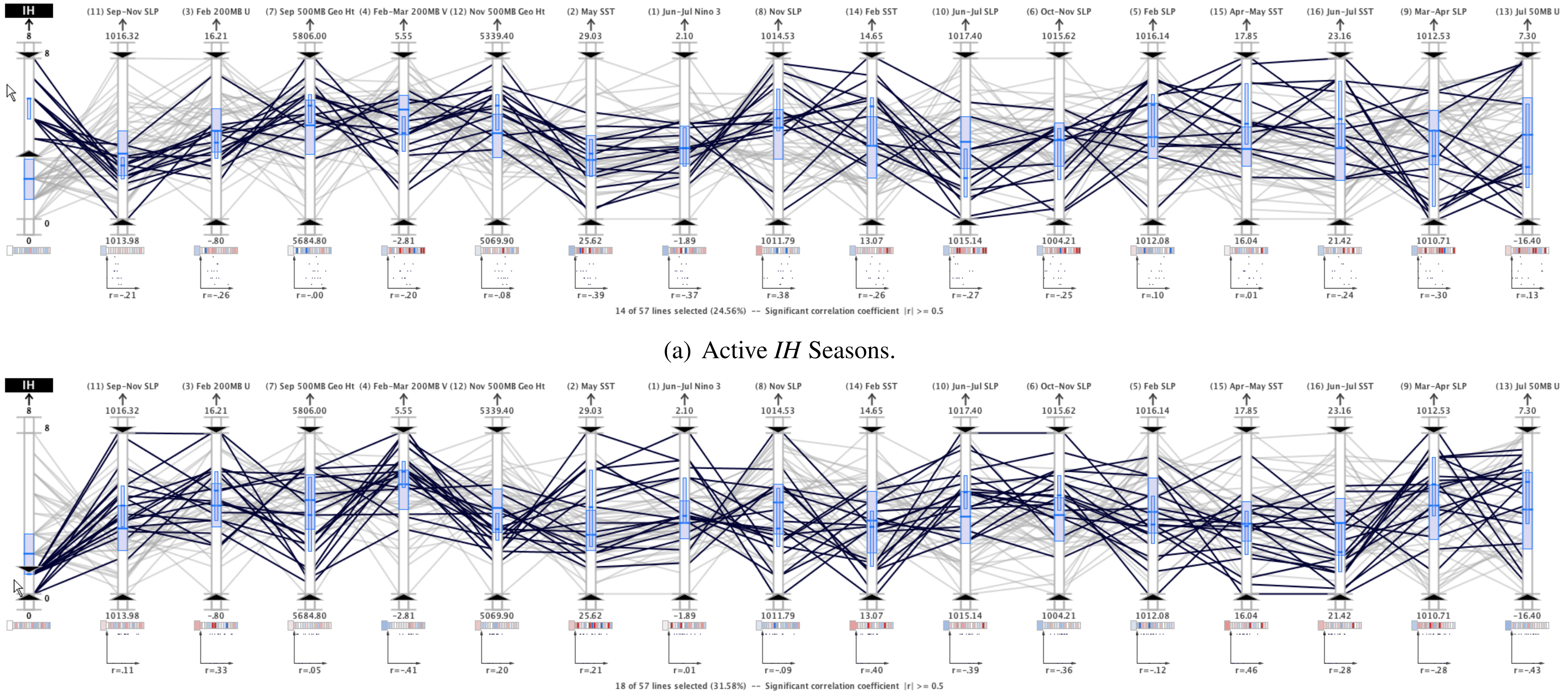


Figure 4.28

Full correlation analysis with the *IH* axis of the predictors. Comparing the predictor behaviors in Figure 4.29, we find that the following axes have the most noticeably different behavior: (11), (4), (12), (10), and (16).

Using these axes, we try to identify predictor ranges that can be used to isolate the seasons with above normal *IH* activity. With the query sliders for the *IH* axis set to isolate the seasons with above normal activity, we use the five predictor query sliders to define the range of each predictor that includes the above normal *IH* seasons. In the process, we discover that the ranges of axes (7) and (1) can be used to significantly improve selection of active seasons; and axis (10) is inconsequential because resetting its query sliders does not affect which seasons are selected. This exploratory analysis reveals the following axes as the most important predictors for determining above normal *IH* activity: (11), (7), (4), (12), (1), (16), and (14). In Figure 4.30(a), the selected axes (those that have their query sliders set) and their query ranges are shown capturing all fourteen above normal *IH* seasons. In Figure 4.30(b), the query sliders on the *IH* axis are reset to show that an additional five seasons are selected in the predictor query ranges that are actually not above normal seasons. Using these query ranges for these seven predictors, the number of type I errors is reduced by six and there are still no type II errors. However, this scheme required two additional predictors. Furthermore, we note that the following predictors were selected in the investigation with and without regression analysis: (11), (7), (4), and (12).

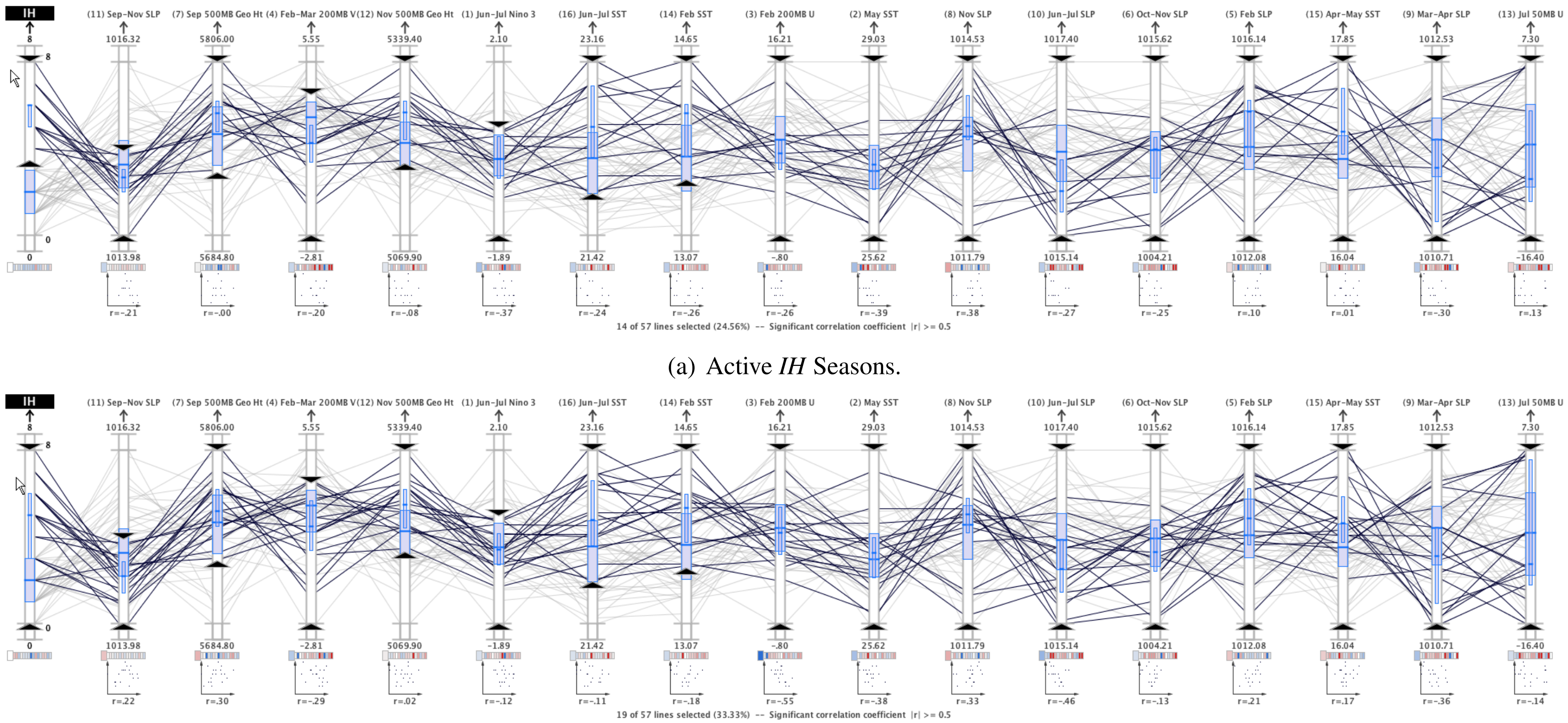


(a) Active *IH* Seasons.

(b) Inactive *IH* Seasons.

Figure 4.29

Analysis of predictors by IQR ranges.



(a) Active *IH* Seasons.

(b) All *IH* Seasons.

Figure 4.30

Query ranges for active *IH* prediction identified without regression.

4.5 Case Study 4: Atlantic Meridional Mode (AMM)

In the fourth case study, we examine the Atlantic Meridional Mode (AMM) data set, which has been shown to have strong associations with North Atlantic tropical cyclone activity in recent research by Vimont [149] and Kossin [104]. In this study, the main goal is to observe the significance of the AMM variables in relation to the CSU predictors. In addition, we will compare the data set to the CSU predictor data and highlight interesting correlations and statistical measurements within the AMM data set. The AMM variables will be examined as predictors for the number of intense hurricanes, number of hurricanes, and number of named storms.

4.5.1 AMM Data Set

The AMM is a dynamic mode of variability that is integral to the tropical coupled ocean-atmosphere system. AMM is also strongly related to seasonal hurricane activity on decadal and interannual time series. This connection is due to the AMM's association with several local climate conditions that all influence tropical cyclone activity collectively [149]. AMM is highly correlated with a number of local climatic factors such as SST, shear, low-level vorticity and convergence, static stability, and SLP. These local factors cooperate to increase or decrease Atlantic hurricane activity. Most agree that the AMM is characterized by a meridional SST gradient near the thermal equator location, winds that blow toward warmer water and veer to the right in the Northern hemisphere and to the left in the Southern hemisphere according to the Coriolis force [104].

The AMM data set was obtained from the Earth System Research Laboratory of the National Oceanic & Atmospheric Administration (NOAA). The AMM data set used in this analysis covers the years 1950–2006. For each year, the AMM data set has twelve values, one value for each month. The AMM spatial pattern is determined by applying a Maximum Covariance Analysis to sea surface temperature and the zonal and meridional components of the 10-meter wind field. The data are defined over the region (21S-32N, 74W-15E), and smoothed spatially using three longitude and two latitude points [5].

In this case study, the AMM variables are also analyzed with the CSU predictors that are described in the previous case study. The resulting analysis will help to identify the relative importance of the AMM variables in isolation and relative to the CSU predictors.

4.5.2 Exploratory Analysis

We began by using MDX to conduct exploratory analysis of the AMM data set in isolation. One of the most prominent features that was discovered in this phase is that the monthly AMM variables have strong positive correlations with months that are near in time. For example, Figure 4.31 shows the correlation coefficients with the *AMM-June* variable. This figure shows that exceptionally strong correlations exist with the four nearest months (*AMM-April*, *AMM-May*, *AMM-July*, and *AMM-August*). The figure also shows a weak correlation between the *AMM-June* and *AMM-January* variables. In fact, each AMM variable exhibits this pattern—the correlation strength between the variables falls off as the temporal separation increases. The correlation indicators also show that all of

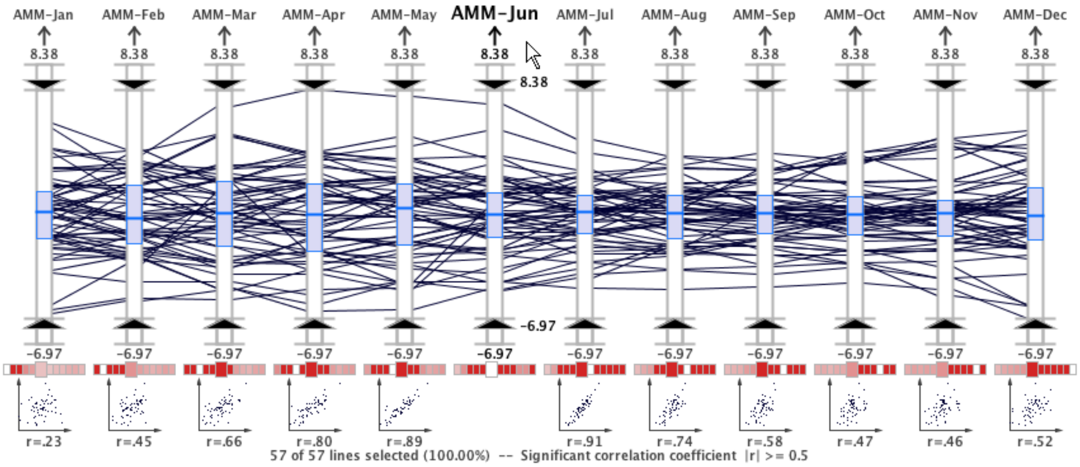
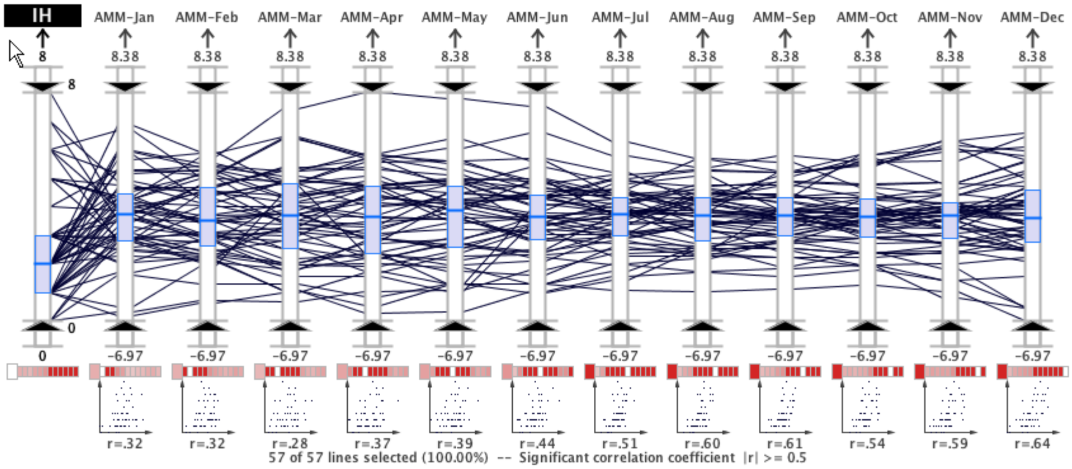


Figure 4.31

Correlation of *AMM-June* with the other eleven AMM variables.

the correlation coefficients are positive since each correlation indicator block is colored red.

When we evaluate the correlations of the AMM variables with the *IH* axis, we observe that the highest correlations occur in the months that historically have the highest intense hurricane activity. In Figure 4.32, the *AMM-August* and *AMM-September* months have the second and third strongest correlation with the *IH* axis. It is not surprising that the weakest correlations with the *IH* activity are observed in the winter and spring months that are outside the North Atlantic season (June 1 – November 30); but it is remarkable that the strongest correlation is with the *AMM-December* axis. The red (positive) correlation indicators and scatterplots beneath each axis also reveal that each AMM variable is positively correlated with the *IH* activity; as the AMM index increases, the number of



57 of 57 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

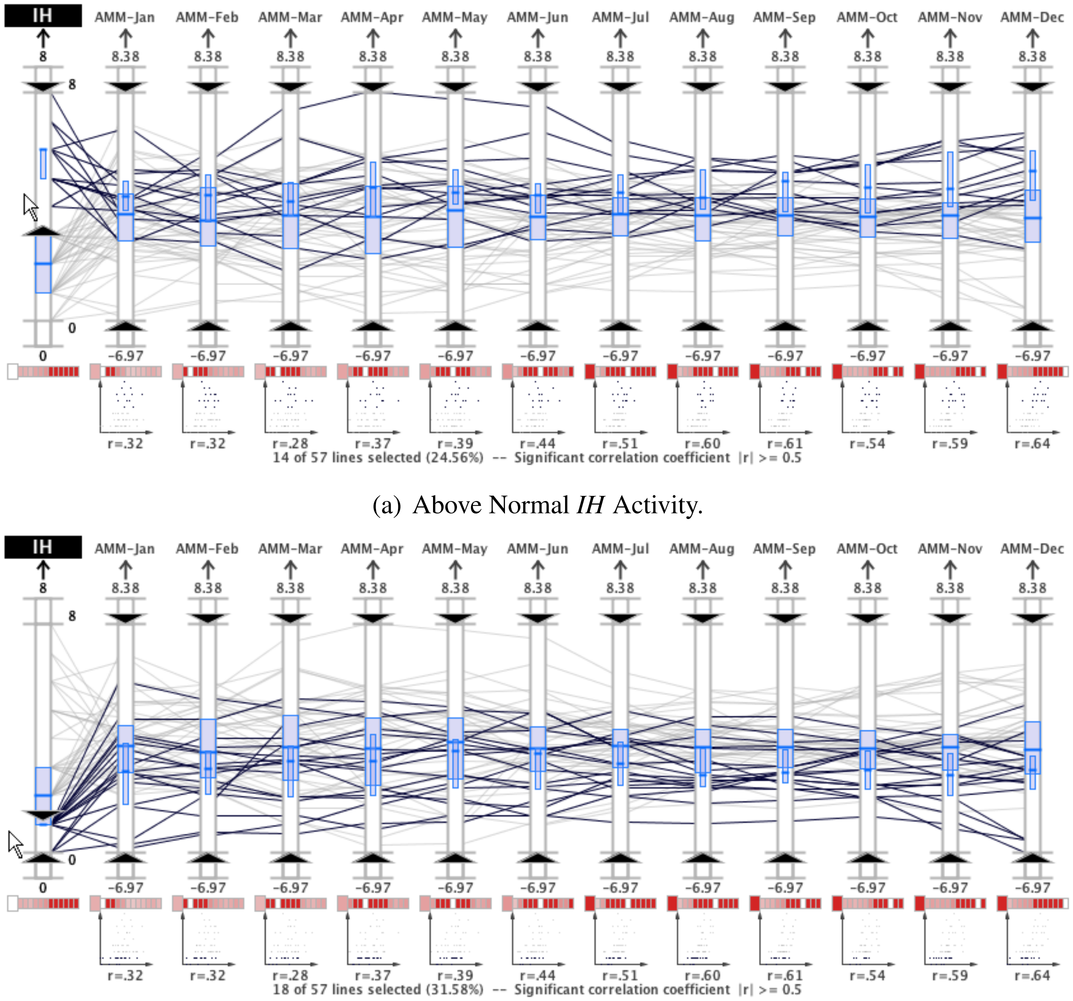
Figure 4.32

Correlation analysis for AMM variables and the *IH* category.

intense hurricanes also increases. These same trends were also found while observing the AMM variable correlations with the *NS* and *H* statistics; but the correlations with the these other seasonal statistics were slightly weaker than with the *IH* statistic.

Using the axis sliders, we compare the AMM variable values for seasons with above normal and below normal *IH* activity in Figure 4.33. We notice in this figure that as the AMM values increase, the *IH* activity also increases. This observation reinforces the positive correlation between the AMM variables and the *IH* variable. The AMM variables were also examined for seasons with above and below normal *NS* and *H* activity and the same positive correlations were observed.

After exploring the AMM variables in isolation, the next phase of our analysis included the CSU data set. In Figure 4.34, the twelve AMM variables are plotted along with the



(a) Above Normal *IH* Activity.

(b) Below Normal *IH* Activity.

Figure 4.33

AMM variable values for above and below normal *IH* activity.

sixteen CSU parameters and the *IH* dependent variable for a total of twenty-nine variables. This figure demonstrates that the number of axes that can be displayed simultaneously in parallel coordinates is only restricted by the horizontal resolution of the display device. In this figure, the AMM variables are sorted by the month measured, which results in more parallel polyline configurations than the CSU parameters since the AMM variables change gradually each month.

We investigated the correlations between the *IH* axis and the independent variables. We used the MDX axis arrangement to order the axes according to the correlation coefficient with the *IH* axis and the resulting display is shown in Figure 4.35(a). From this figure, we see that the four strongest correlations are with AMM variables (*AMM-December*, *AMM-September*, *AMM-August*, and *AMM-November*). Similar correlation analysis is shown in Figure 4.35(b) and Figure 4.35(c) for the *H* and *NS* statistics, respectively. For the *H* statistic, the six strongest correlations are with AMM variables: *AMM-December*, *AMM-November*, *AMM-September*, *AMM-August*, *AMM-October*, and *AMM-July*. For the *NS* statistic, the five strongest correlations are with AMM variables: *AMM-November*, *AMM-December*, *AMM-October*, *AMM-September*, and *AMM-August*.

4.5.2.1 Significant Predictors Identified with Regression

To determine the most significant predictors among the CSU and AMM variables for seasonal tropical cyclone activity, we used the MDX stepwise regression analysis capabilities. We generated three regression models for each of the seasonal statistics: *NS*; *H*; and *IH*. These are the same dependent variables as we used in the second case study. The

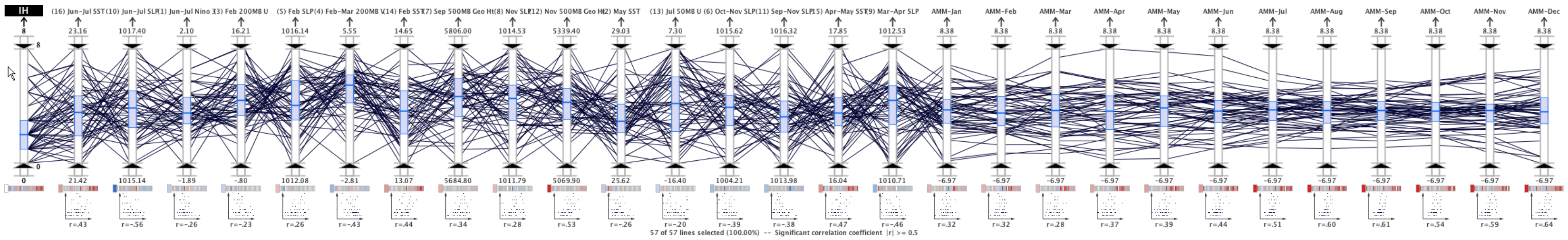
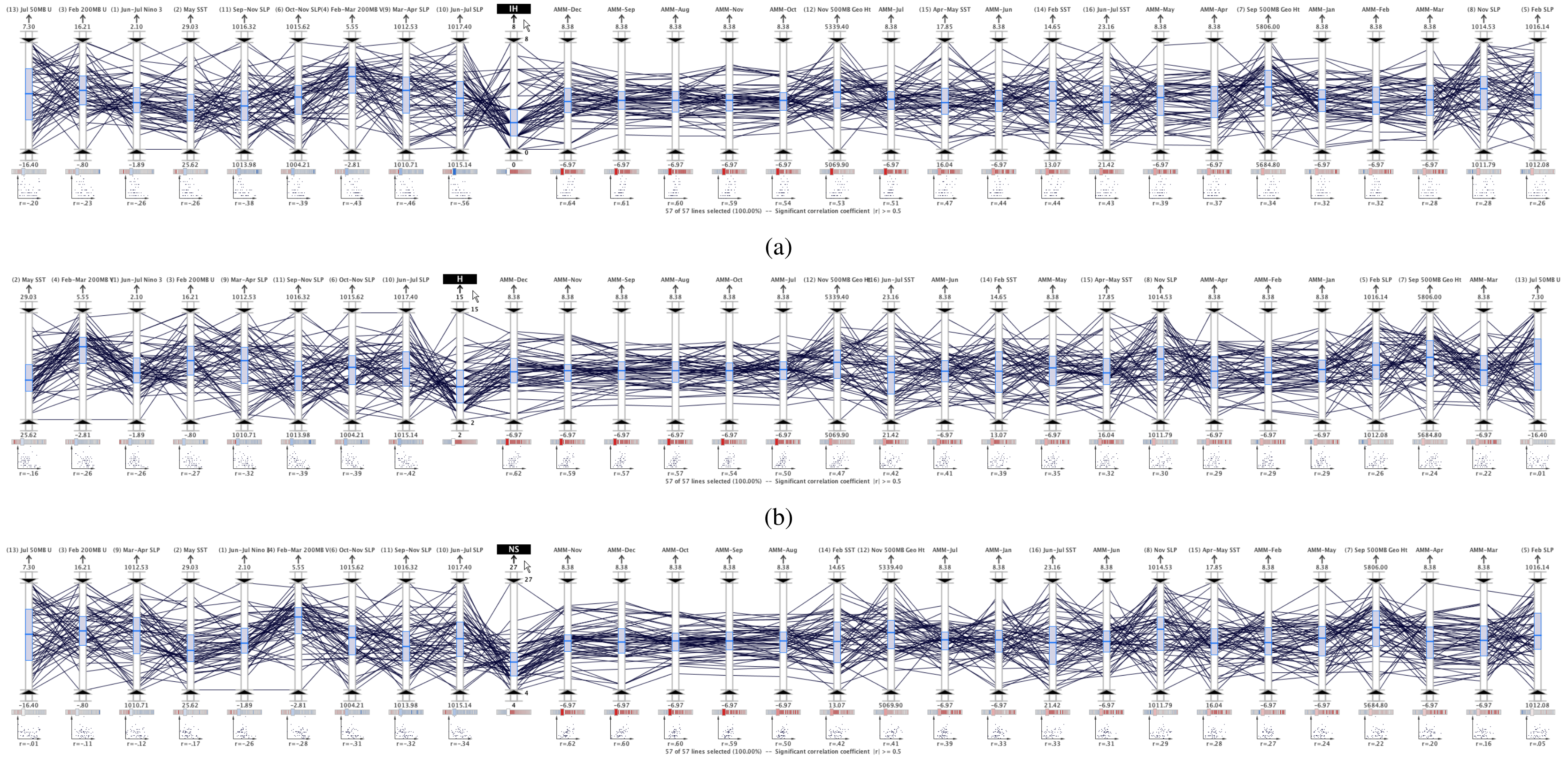


Figure 4.34

All AMM and CSU variables plotted with the *IH* axis.



(a)

(b)

(c)

Figure 4.35

Correlation analysis of *IH*, *H* and *NS* with all CSU and AMM variables.

results of the regression analyses are listed numerically in Table 4.7. In the remainder of the section, we discuss these analyses and how the MDX visual analytics reveal a deeper level of understanding. The significance level in the three stepwise regression analyses that follow is 80%.

4.5.2.2 Multicollinearity Filter

Prior to the regression analyses, the multicollinearity filter was executed for the twenty-eight independent variables with the *IH*, *H*, and *NS* dependent variables. In all three cases, the filter removed seventeen of the independent variables from the axis configuration, leaving eleven variables in the display. For the *IH* and *H* cases, only two AMM variables are included in the new configuration: *AMM-December* and *AMM-May*. In Figure 4.36(a), the unfiltered axes for the *IH* case are arranged according to the correlation coefficient with the *IH* axis. This figure reveals that the *AMM-December* axis has the highest correlation ($r = .64$) of the remaining axes. In Figure 4.36(b), the remaining axes for the *H* case are arranged according to the correlation coefficient with the *H* axis. This figure reveals that once again the *AMM-December* axis has the highest correlation coefficient ($r = .62$)—a significantly stronger correlation than the second highest correlation ($r = .47$) with the *November 500-mb Geopotential Height* (12) axis.

For the *NS* case, three AMM variables remain after the filter, as shown in Figure 4.36(c): *AMM-November*, *AMM-January*, and *AMM-June*. Figure 4.36(c) shows the resulting configuration with the axes arranged according to the *NS* axis correlation coefficient. From this figure, we discover that the highest correlation coefficient is with the *AMM-*

Table 4.7

Regression model for *IH* category with AMM and CSU data sets.

Number of Named (NS)
(R^2 is 45%)

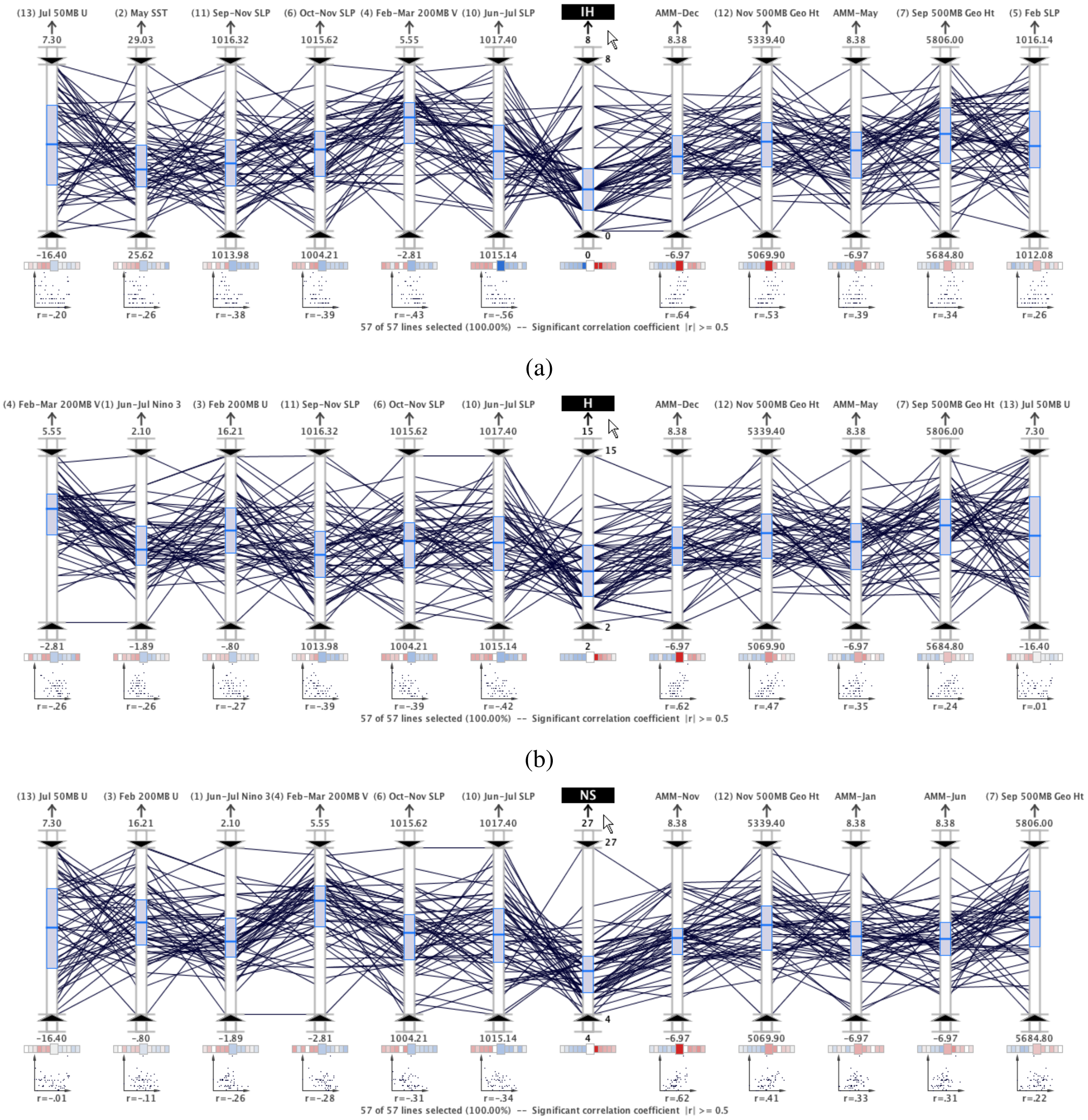
Chosen Variables	Normalized Coefficients *b*	Sample Mean
AMM-Nov	0.5241	0.081
Oct–Nov SLP (6)	–0.2032	1009.56
Nov. 500–mb Geopot. Ht. (12)	0.1540	5213.38

Number of Hurricanes (H)
(R^2 is 56%)

Chosen Variables	Normalized Coefficients *b*	Sample Mean
AMM-Dec	0.5027	0.045
Feb. 200–mb U (3)	–0.2404	8.713
Nov. 500–mb Geopot. Ht. (12)	0.2248	5213.38
Oct–Nov SLP (6)	–0.1893	1009.56

Number of Intense Hurricanes (IH)
(R^2 is 65%)

Chosen Variables	Normalized Coefficients *b*	Sample Mean
AMM-Dec	0.3421	0.045
Nov. 500-mb Geopot. Ht. (12)	0.2951	5213.38
Jun–July SLP (10)	–0.2535	1016.23
Sep. 500–mb Geopot. Ht. (7)	0.2250	5753.33
Feb–Mar 200–mb V (4)	–0.1239	2.53



(a)

(b)

(c)

Figure 4.36

Correlation analysis after multicollinearity filtering for *IH*, *H*, and *NS* axes.

November axis ($r = .62$), which is substantially stronger than the second strongest correlation ($r = .41$) with the *November 500-mb Geopotential Height* (12) axis.

4.5.2.3 Regression Analyses

After executing the multicollinearity filter, the MDX stepwise regression capabilities were used to find the most significant predictors among the remaining eleven variables. With the *IH* variable as the dependent variable, the regression model included the five variables shown in Figure 4.37. From the figure, we see that the R^2 value for this model is 65%, which is better than the 58% obtained using only the CSU predictors in case studies 2 and 3 (see Table 4.4 and Table 4.5). In this model, the *AMM-December* variable had the highest regression coefficient ($b = 0.3421$). The other four variables included in the model are from the CSU data set and they were also included in the models produced in case studies 2 and 3. The only variable not selected in this model that was selected in case studies 2 and 3 is *September–November SLP in the southeast Gulf of Mexico* (11). This variable was not included because the AMM explained more variance and variable (11) could no longer explain any new variance in this particular model.

In Figure 4.38, the values of the regression model variables are compared for above normal *IH* activity and below normal *IH* activity. These figures demonstrate the promise of these variables for forecasting *IH* activity because the typical value for active and inactive seasons is either below or above the overall typical value. Moreover, the IQR is either entirely below or entirely above the overall IQR for all the variables except *September 500-mb Geopotential Height* (7) in below normal seasons.

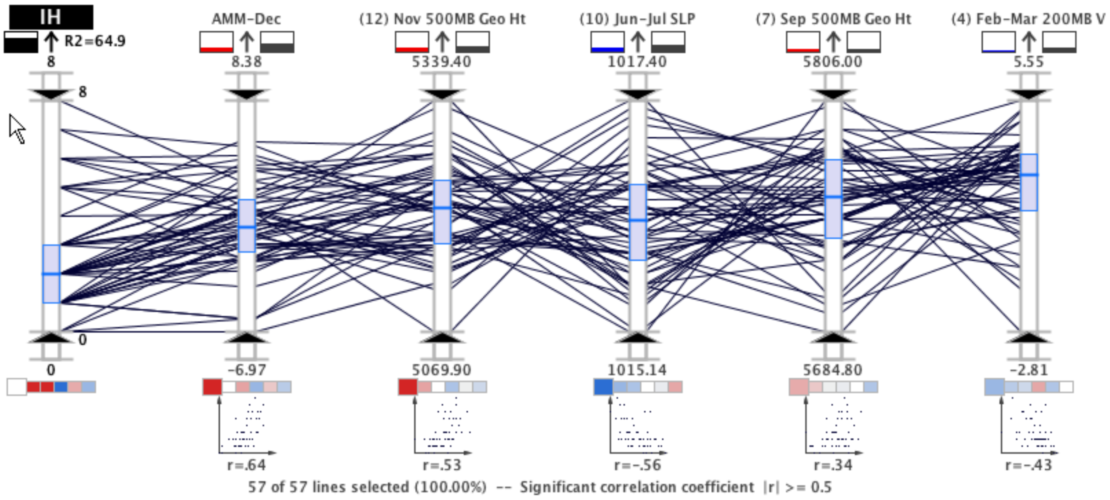
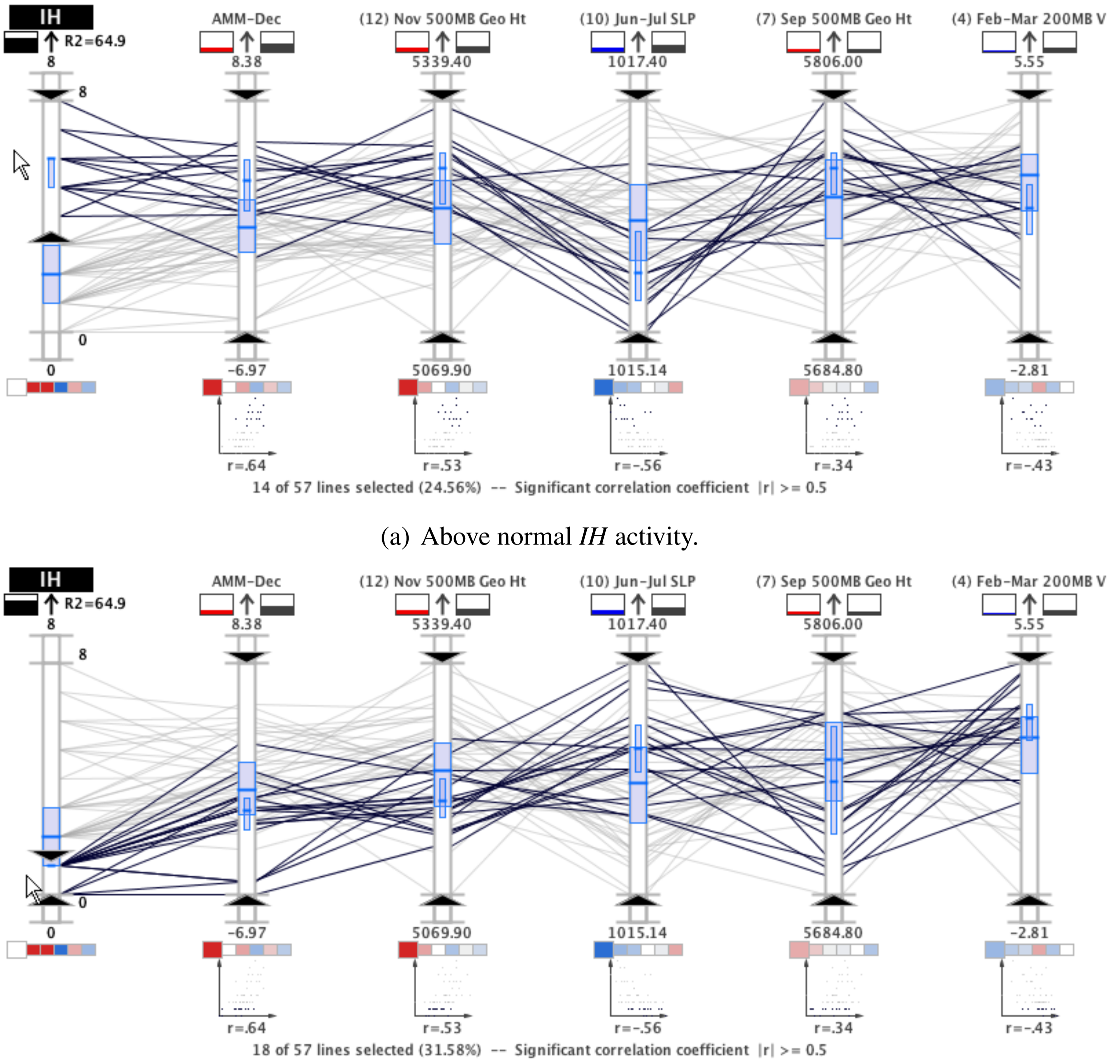


Figure 4.37

IH regression model with AMM and CSU variables (1950 to 2006).

When the *H* variable is set to the dependent variable, the regression model included the four variables shown in Figure 4.39. The R^2 value for this model was 56%, which was significantly better than the 42% from the analysis with only the CSU predictors (see Table 4.4). Like the *IH* regression model, the *AMM-December* variable had the strongest regression coefficient ($b = 0.5027$)—significantly higher than the next strongest coefficient ($b = -0.2404$). Like the case study 2 model, this model included both the *October–November SLP* (6) and *November 500-mb Geopotential Height* (12) variables; but in this model variable (6) had the lowest coefficient instead of the highest. Furthermore, this model dropped the *June–July SLP* (10) and *September–November SLP* (11) variables and included the *February 200-mb zonal wind in equatorial East Brazil* (3) variable. Variables (10) and (11) are dropped for the same reason that the *IH* model dropped variable (11). However, the model kept variable (6) that also measures SLP. The reason variable (6) was



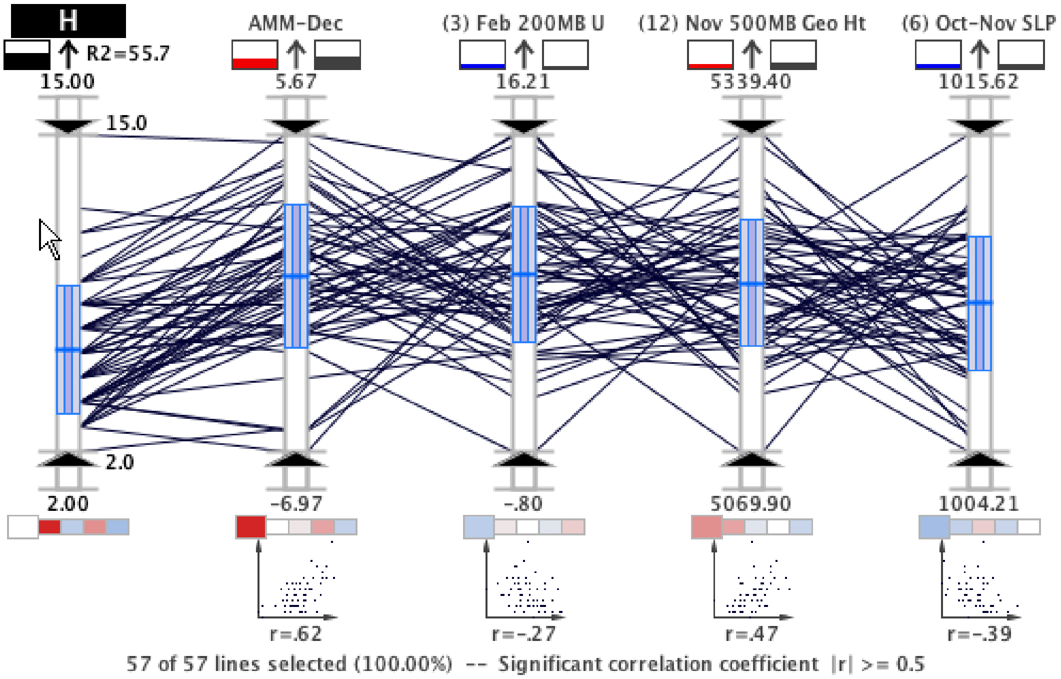
(a) Above normal *IH* activity.

(b) Below normal *IH* activity.

Figure 4.38

Significant axes values for above normal and below normal *IH* activity.

retained and variables (10) and (11) were dropped is likely due to geographic locations—variable (6) was measured in the Gulf of Alaska and variables (10) and (11) were measured in areas that overlap the AMM variables (see Figure 4.11). Since variable (6) had the weakest correlation with *AMM-December* of these three SLP variables (see Figure 4.40), it was not included in the model for the *H* category since there was little new variance to explain.

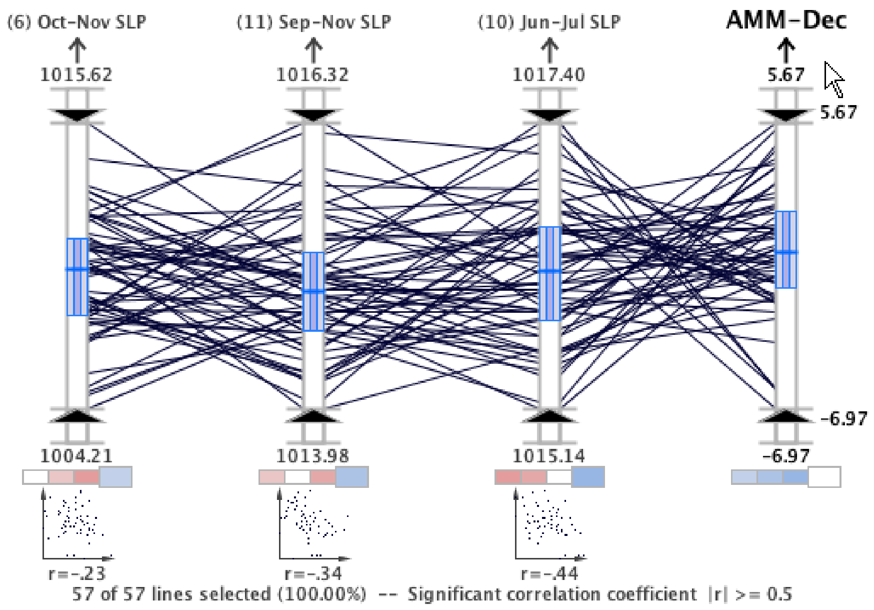


57 of 57 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

Figure 4.39

H regression model with AMM and CSU variables (1950 to 2006).

When the *NS* variable was set to the dependent variable, the regression model included only three variables shown in Figure 4.41. The R^2 value for this model was 45%, which was significantly better than the 34% from the analysis with only the CSU predictors (see



57 of 57 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

Figure 4.40

Correlation analysis between SLP variables and *AMM-December*.

Table 4.4). The only variable included in this model that was also included in the model from case study 2 was *November 500-mb Geopotential Height* (12). Like the *H* regression model, the AMM and CSU *NS* regression model also added the *October–November SLP* (6) variable because it had weak correlation with *AMM-November* since it was measured in the Gulf of Alaska. The *February SST* (14) was removed by the multicollinearity filter process because it was strongly correlated with the *AMM-November* variable ($r = .56$) and it had a weaker correlation ($r = .42$) than the *AMM-November* ($r = .62$) with the dependent variable. Similarly, the *September–November SLP* (11) variable was removed by the filter because it was strongly correlated with the *AMM-January* variable ($r = -.58$) and it had a weaker correlation ($r = -.32$) than the *AMM-January* ($r = .33$) with the dependent variable. Unlike the other two AMM regression models, this model resulted in the exclusion of the *AMM-December* variable, which was also removed during the multicollinearity filter process due to strong correlation with the *AMM-November* variable($r = .92$).

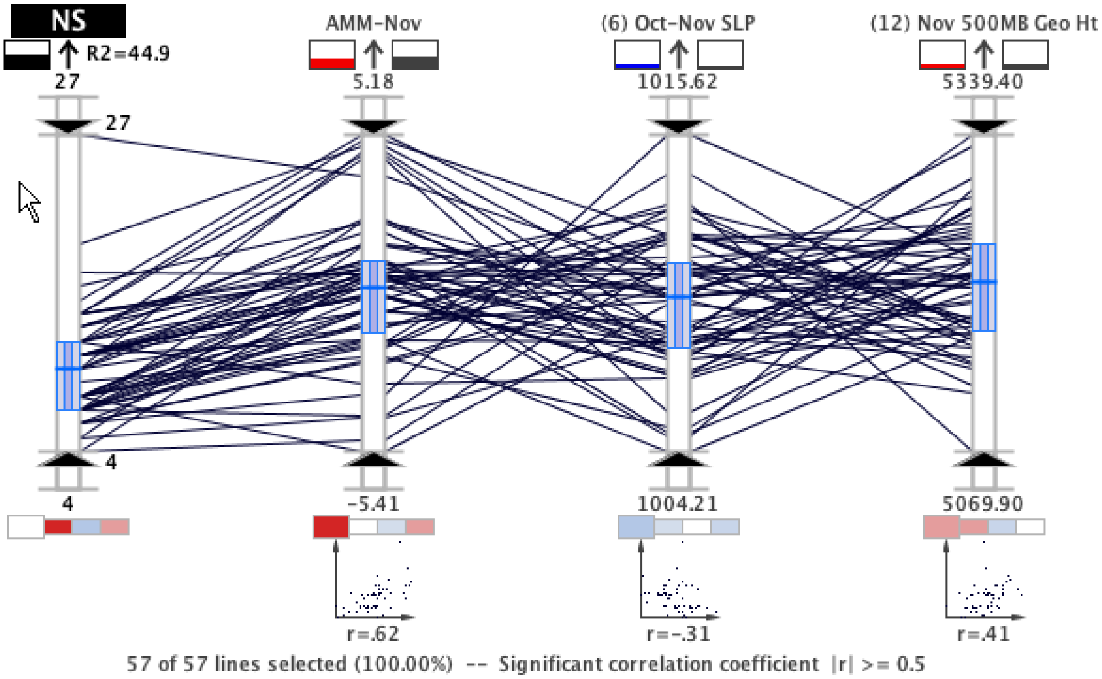


Figure 4.41

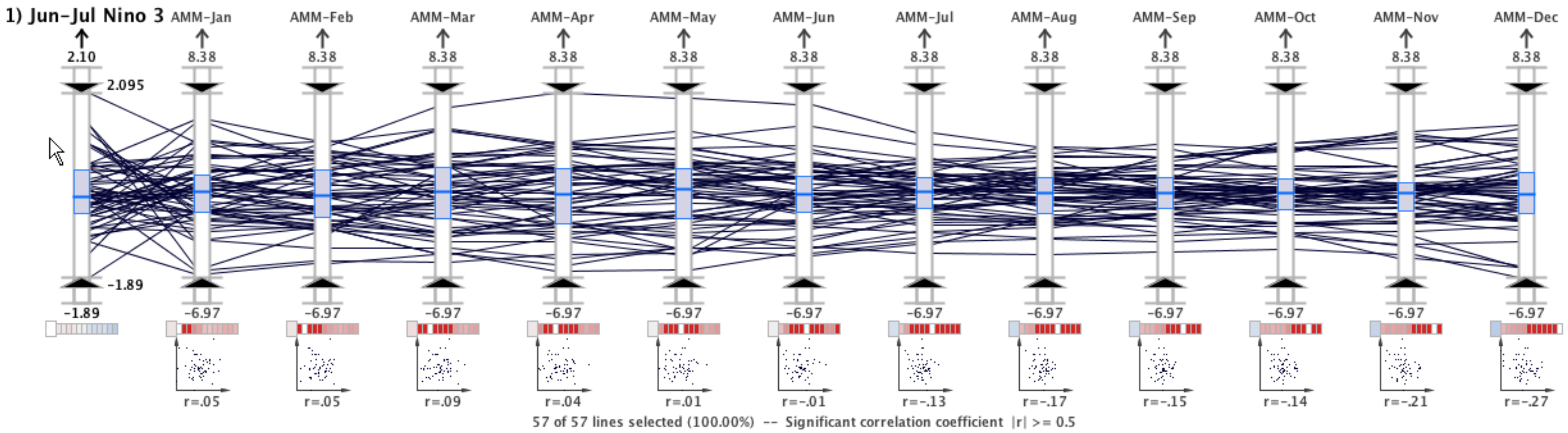
NS regression model with AMM and CSU variables (1950 to 2006).

In the *H* and *NS* models, the *February 200-mb U* (3) and *October–November SLP* (6) variables are CSU predictors that were not included in the corresponding regression models with only the CSU data. Each of these variables are correlated to El Niño conditions. They are likely included in the model because, as Vimont and Kossin [104] state, the AMM is largely independent of ENSO. In Figure 4.42, the independence of these variables can be observed in the lack of correlation between the twelve AMM variables and the *June–July Niño 3* (1) variable.

It is also remarkable that the *November 500-mb Geopotential Height* (12) variable was included in the three AMM and CSU regression models, as well as the three CSU regression models presented in case study three. As described in Section 4.3.1.3, variable (12) measures the long-term oscillations that impact global wind patterns. Specifically, it represents a different climate signal from the AMM that is called the North Atlantic Oscillation. Vimont and Kossin [149] stated that the AMM and the North Atlantic Oscillation are independent during the hurricane season which may explain its inclusion in all three AMM models. Furthermore, the fact that the geographic regions for variable (12) and the AMM do not overlap may contribute to the independence of the variables. Since variable (12) is included in all regression models in case studies 2, 3, and 4, it is highlighted as a very significant signal for tropical cyclone trend analysis.

4.5.3 Discussion

In the second case study, we demonstrate that parallel coordinates, a visualization technique designed specifically for multivariate information, can be used to confirm and clarify the results of stepwise regression for a tropical cyclone climate study. While the regression analysis gives us an ordering of the most important environmental variables, visual analysis using parallel coordinates facilitates a deeper understanding of the environmental causes for above average, normal, and below average seasonal statistics. Using the interactive visual analysis features of the MDX interface, the viewer can intuitively and rapidly explore these relationships.



57 of 57 lines selected (100.00%) -- Significant correlation coefficient |r| >= 0.5

Figure 4.42

Correlation analysis between the AMM variables and *June–July Niño 3* (1).

In the third case study, we show that interactive parallel coordinates can be used in conjunction with automated decision support algorithms to discover and confirm hypotheses about climate data. The expanded version of MDX used in this case study provides the same promising capabilities shown in the first evaluation, as well as new analytic components that guide the exploration. The expanded system effectively blends the analytic spotlight of statistical analytics and the inferential floodlight of visual exploration to facilitate high-dimensional geovisual analytics [156]. Using traditional analysis alone in the third case study would require the examination of 136 scatterplots to observe the same associations that are efficiently captured in a single frame of the interactive visualization system presented the second case study.

The fourth case study used the enhanced MDX capabilities to demonstrate the importance of the AMM data set in relation to the CSU predictors. The AMM data set has recently been shown in weather science literature to have strong associations with tropical cyclone activity. In addition to further corroborating the effectiveness of the MDX system, the simultaneous statistical analysis of both the CSU and the AMM data sets is a significant contribution from a weather science perspective. Furthermore, this case study demonstrated that, when combined with the CSU predictor data set, the AMM variables significantly boost the variance explained for each of the three dependent variables.

4.6 Lessons Learned from Climate Analysis with Geovisual Analytics

The development and evaluation of MDX has highlighted several strengths and weaknesses related to the application of geovisual analytics in climate analysis. In the remainder of this section, some of the more salient observations are discussed.

Traditionally, climate studies are performed using a collection of separate visualization and data analysis tools. Furthermore, the data visualization is often facilitated using only static plots. As a result, the scientists have limited interactivity with the data, which hinders the discovery of new hypotheses. The integrated statistical and visual analysis capabilities in MDX allow the scientist to connect the statistical and visualization processes on-the-fly for more rapid and creative knowledge discovery. The benefits of having an all-in-one environment became evident when we used the expanded version of MDX. In the second case study, the statistical analysis was executed in an external statistics package and the results were visualized using MDX parallel coordinates plots. That is, a weather scientist familiar with the statistics package generated tabular outputs from the statistical analysis tools, which were then fed to a computer scientist who generated parallel coordinates plots using MDX. This disjoint process required several days to reach conclusions. But with the expanded version of MDX, the statistical analysis and visualization capabilities are integrated a single environment, which makes independent analysis possible for each collaborator. The integrated environment represents perhaps the most significant improvement in the expanded version of MDX. In addition, the ability to directly interact with the data via dynamic visual queries is a vast improvement over the traditional static plots. Perhaps the greatest evidence of the promise of this approach came from the hurricane expert in our climate studies, Dr. Fitzpatrick, who indicated that the MDX system made it possible to validate the AMM data set more quickly and more comprehensively than what would have been possible with conventional analysis techniques.

Of the statistical indicators added in the expanded version of MDX, the new correlation indicators that are shown beneath each predictor axis provide the most significant benefit. The value of these indicators is apparent when we consider the task of identifying and reducing multicollinearity among the independent variables. In the second case study, we reduced multicollinearity by conducting correlation analysis with a tabular correlation matrix. The correlation matrix is a square matrix where each i, j element is equal to the correlation coefficient between the i and j variable. In order to perform this analysis, the scientist first located the coefficients that were higher than the predefined correlation threshold. Then, the row and column index elements were used to identify the variables and the correlation coefficient of each variable with the dependent variable was found from the table. Of the two predictors, the one with the weaker correlation with the dependent variable is removed from the display. In the expanded version of MDX, the interface is designed to streamline the correlation analysis by positioning the rows of the correlation matrix beneath each predictor axis in a set of color-coded blocks. As shown in Figure 4.20, the saturation of the box color indicates the strength of the correlation which effectively highlights the strongest correlations. Furthermore, the additional scatterplot displays, parallel coordinates plot, and numerical correlation coefficient display further enhance correlation analysis with multiple linked views of the associations at varying levels of detail.

Without the enhanced correlation indicators, the scientists may have to examine each pair of variables separately, a task that could require hundreds of scatterplots.

In addition to providing enhanced correlation analysis capabilities, the graphical statistical indicators provide a visual uncertainty metric which supplements knowledge discovery. For example, a tight clustering of lines in parallel coordinates might indicate a relatively stable predictor that may yield better results than another predictor with more dispersed lines. The statistical indicators can also directly guide the analysis as illustrated in our use of the central tendency and variability indicators to determine active, inactive, and normal hurricane seasons in the two case studies.

The utilization of multiple linked views in the expanded version of MDX helps the scientist by facilitating more creative exploratory analysis and offering additional views of the data. 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 parallel coordinates are only restricted by the horizontal screen resolution while a scatterplot is generally restricted to two or three dimensions. Moreover, it is difficult to decipher the correlation between axes in all but the extreme cases in parallel coordinates, but the scatterplot is more useful in less extreme cases that we encounter more often in real-world data. Having both the parallel coordinate plot and scatterplot in MDX gives the scientist access to both views in a complementary fashion, which offsets said deficiencies. Furthermore, the inclusion of the enhanced parallel coordinates encourages the weather scientist to consider the associations in new ways, which may lead to fresh insight.

We also noticed that using animation with the continuous aerial perspective line shading is helpful for identifying varying behavior of a particular axis over multiple dimensions (see Figure 4.19). The proximity-based parallel coordinates line shading is also applied to the data points in the small scatterplots beneath each axis. With this capability, we can move the mouse cursor from the top to the bottom of the *Year* axis to animate the temporal variability in the predictors. Capturing the animation effect in a movie or sequence of images is an effective means for visually communicating the behavior of a variable. The investigation of new automated animation capabilities is a promising direction for future research using the parallel coordinates technique. For instance, animating the axis arrangements and predictor relationships might further amplify cognition and improve the communication of results in climate studies.

Although the assortment of statistical indicators and multiple plots demonstrates great promise for enhanced knowledge discovery, the display can quickly become cluttered. The parallel coordinates plot alone is infamous for its tendency to easily become incomprehensible with large data sets. So, we are naturally aware of the potential for overwhelming clutter in the MDX interface which combines parallel coordinates and several graphical indicators. To control clutter, MDX provides access to display parameters in a settings panel. From this panel, the scientist can toggle options and change settings. The incorporation of predefined display settings based on skill levels or specific tasks might also help reduce the user's cognitive load. For example, we can predefine display settings for specific tasks (e.g. correlation analysis, regression analysis, statistical analysis) or based on the skill level (e.g. novice, intermediate, expert).

The parallel coordinates plot, although it has been used often in visualization literature, is still relatively unknown to most domain specific scientists (in this case weather scientists). Consequently, conveying results that are discovered in parallel coordinates (or any unconventional visualization technique for that matter) is difficult and requires some degree of training to teach the audience how to read the results. So, communicating results to the domain experts using MDX plots can be challenging, at least initially. To alleviate this issue, we might reproject the results from the parallel coordinates plots into more common displays, like scatterplots, scatterplot matrices, or line graphs. Having these export capabilities built into MDX would allow the scientist to explore the data set using more advanced visual analysis methods, but use more common plots to share insight.

We devoted a significant amount of time to finding an optimal color scheme for the MDX interface. The color scheme and layout was formulated by drawing on color design principles from fine art and graphic design, as well as empirical perceptual studies [153] discussed in the MSU Information Visualization course and practical evaluation in the software. For example, we use muted colors in most 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 are a good illustration of the significance of a well-planned color design. 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.

CHAPTER 5

CONCLUSION

This research has demonstrated that interactive parallel coordinates, a visualization technique designed specifically for complex multivariate information, can be used in conjunction with advanced statistical analysis to discover and confirm hypotheses in environmental data. While automated statistical analysis techniques yield an ordering of the most significant associations among a set of inter-related parameters, the dynamic visual analysis capabilities facilitate a deeper understanding of the relationships. The capabilities have been fused into a powerful geovisual analytics system, called MDX, and evaluated via a pedagogical case study as well as three real-world tropical cyclone climate studies using a systematic workflow. This dissertation has validated the following hypotheses:

1. The development of an advanced geovisual analytics approach using parallel coordinates and statistical techniques reveals a deeper level of understanding than traditional methods when applied to the task of finding complex multivariate trends in environmental data sets. With new ways to creatively explore the data, the approach offers a more effective visual interface to glean new insight about the data behind the visualization.

2. The effectiveness of the geovisual analytics approach is necessarily explored in the context of practical environmental studies, which are grounded in real-world data sets instead of invented or abstract data sets, in close collaboration with domain experts. The discovery of new associations and the confirmation of known patterns by domain experts will validate the promise of this new approach in environmental data analysis.

In the remainder of this chapter, several specific impacts of this research and future work are discussed.

5.1 Impact

This dissertation makes several contributions to the state of the art of environmental data analysis using geovisual analytics. First of all, we have fused together variants of several previously introduced interactive capabilities into one of the most advanced parallel coordinates based geovisual analytics environment. That is, although many of the techniques of our our interface were first introduced in earlier publications, no one has combined these techniques into a single interface.

One way that we have extended the functionality of the parallel coordinates display is by applying dynamic visual query techniques to provide enhanced access to the data behind the visualization. Although many of these interactions are considered fundamental parallel coordinates features, we have also incorporated several variants of more innovative extensions such as the dynamic axis scaling and the continuous aerial perspective shading. Our axis scaling approach is unique in the way it is controlled by the mouse wheel movement. Also, the aerial perspective shading utilizes an innovative line shading scheme based on a non-linear fall-off function that effectively shades the line closest to the mouse cursor in a more prominent manner.

Although other researchers have integrated statistical analysis algorithms with information visualization techniques, we found no prior works that linked automated regression and correlation analysis techniques into a seamless parallel coordinates interface like our application. This unique approach yields an effective means to quantify individual associations between parameters and it also helps quickly identify and quantify the most significant associations.

These advances in the parallel coordinates representation, dynamic visual query methods, and automated statistical analytics have also been combined into an a unique geovisual analytics application that is specifically designed to help one conduct environmental data analysis by correlation and regression processes.

Over the course of this research, we formulated a systematic workflow for using geovisual analytics to conduct exploratory environmental data analysis. This workflow is a formalization of the traditional climate study approach using correlation and regression analysis but utilizing a visual analysis environment. Using the MDX system and this workflow, we evaluated the effectiveness of the new geovisual analytics approach in three tropical cyclone climate studies. These case studies corroborate the claims that geovisual analytics hold much potential for improving the investigation of climate data sets. The case studies also provided an excellent opportunity to identify the strengths and weaknesses of this approach over more traditional climate analysis systems and they highlighted several significant associations for understanding tropical cyclone activity. These case studies are the first application of geovisual analytics to the study of tropical cyclone trends.

In the three tropical cyclone climate studies used to evaluate the effectiveness of the geovisual analytics approach, several significant discoveries were brought to light. In addition to validating the results of regression analysis using our representation techniques, interesting physical relationships in the CSU tropical cyclone predictor data set were highlighted. Furthermore, we isolated several global signals by combined analysis of the CSU and the AMM data sets—a significant investigation of twenty-nine variables. These discoveries validate the promise of our new approach in climate analysis.

The researchers involved in this work provide a unique blend of expertise in both the visualization and earth sciences realms. In addition to my earth sciences background, this team consisted of three visualization and computer graphics professors (Drs. Jankun-Kelly, Moorhead, and Swan), a hurricane expert (Dr. Fitzpatrick), and a software engineering professor (Dr. Allen) with extensive experience in software evaluations. Furthermore, this research effort addresses the NIH/NSF Visualization Challenges Report recommendation that visualization researcher "collaborate closely with domain experts who have driving tasks in data-rich fields to produce tools and techniques that solve clear real-world needs [92]" by the inclusion of Dr. Fitzpatrick throughout the design and evaluation of the system.

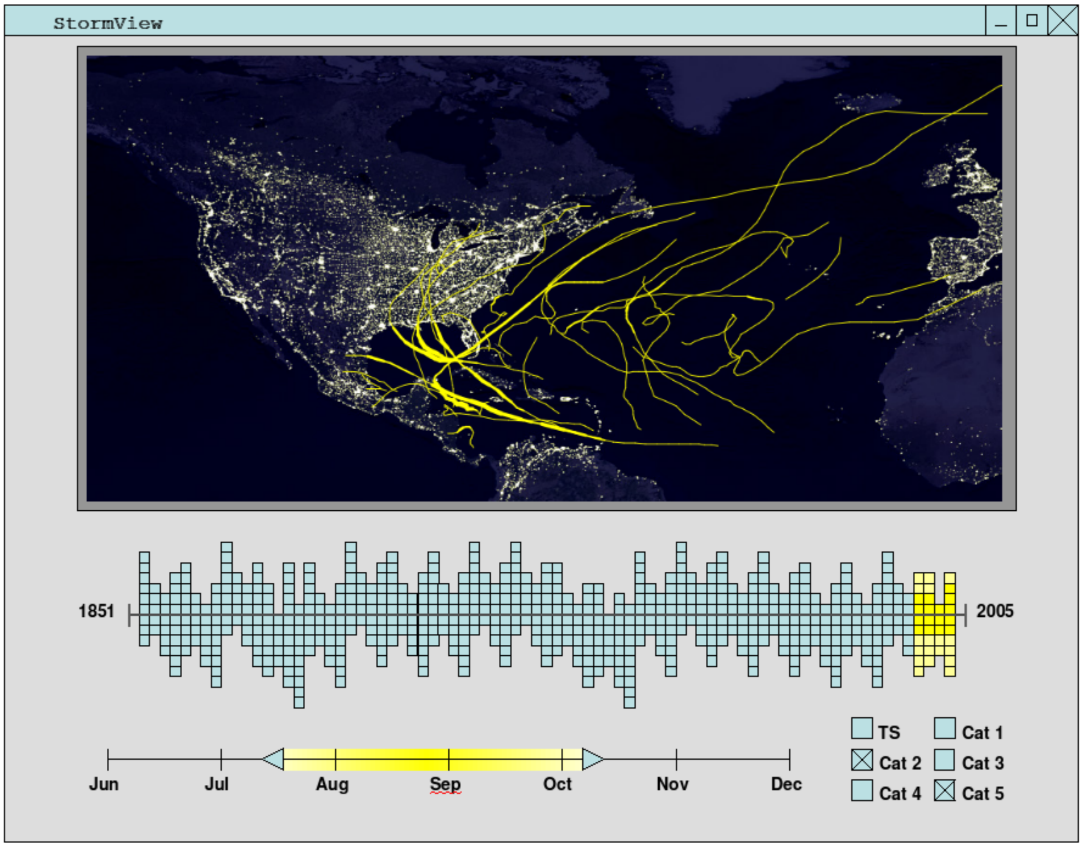
5.2 Future Work

Over the course of this project, several ideas for future research that extend the concepts presented in this dissertation have been identified. In this section, these topics are described.

Although removing the geographical map from the MDX interface helps locate non-geospatial parameter associations, geospatial associations are most evident in the context of a geographical map. Adding an additional linked map view to MDX would help the scientist identify these patterns. However, traditional map views are restricted to representing three or four variable fields simultaneously. Some visualization researchers have begun to explore the application of illustrative rendering techniques from the overarching field of computer graphics to the representation of environmental data in spatial maps. For example, Kirby et al. [98] successfully applied perceptual techniques to represent six velocity-derived quantities at each spatial location in a single flow visualization image. Prior to this work, the deformation of fluid elements was represented in qualitative sketches (in this case six) that were difficult to directly connect to the rest of the flow field. Later, Healey et al. [64] introduced new visualization methods for superior pattern recognition in multidimensional weather data. Future work is planned to expand these techniques to provide additional spatial representations of the climate data that complement the non-geographical views.

Additional advanced interaction techniques can be investigated and applied to the visualization interface. Such efforts will naturally extend the work by Shneiderman [133] and Tweedie [147] on dynamic visual queries and advanced interaction widgets, respectively. Incorporation of these concepts into the previously mentioned visualization systems should yield more effective exploratory analysis. In preparation for this work, we formulated a “mock-up” of an interaction widget, which is shown in Figure 5.1.

It is envisioned that the user can interact with this system via sliders and a unique widget that displays each storm in the data set as a color-coded square. With this interface, the storm squares are aligned along the horizontal axis according to the storm years, and they are ordered chronologically along the vertical axis. This unique widget will drive the



□ TS □ Cat 1
☒ Cat 2 □ Cat 3
□ Cat 4 ☒ Cat 5

Figure 5.1

NOAA Best Track data set analysis user interface concept.

information displayed in the map context as well as any non-geographic representations, such as parallel coordinates.

Another promising project is to explore the application of additional multivariate information visualization techniques to environmental data analysis. That is, additional multivariate visualization techniques—such as star coordinates [94], VisDB [96], and Table Lens [127]—can be used to explore the climate data used in the tropical cyclone climate studies. In addition, to expanding the techniques and formulating new approaches, this project would provide an excellent opportunity to compare and contrast the various approaches in their ability to improve the knowledge discovery and decision making processes in real-world scenarios.

Some of the weather information that has been gathered has included categorical data. While the continuous, numerical dimensions are well understood in multivariate visualization, categorized data is not sufficiently addressed. Examples of this type of data include: product or customer categories, location of hurricane landfall, and bank account types. These type of data generally have a small number of different values that usually have special meanings. Furthermore, the categories usually have no inherent order [103].

Recently, Kosara et al. [103] presented a new approach to the visualization of categorical data that is called *Parallel Sets*. In this system, traditional Venn diagrams and parallel coordinates representation techniques are combined [103]. As shown in Figure 5.2, the authors addressed the simultaneous display of continuous and categorical data in the Parallel Sets technique. However, there is significant room for improving on the integration of continuous dimensions using this technique. To improve the capabilities of MDX, the

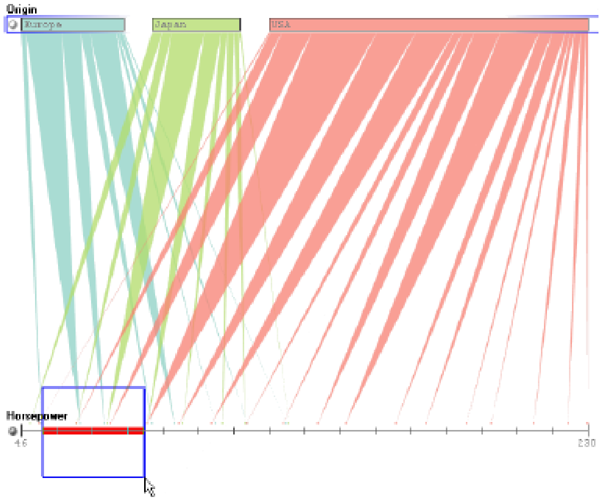


Figure 5.2

The Parallel Sets approach by Kosara et al. [103].

Parallel Sets concept can be extended to provide a solution that works well with data sets that contain both continuous and categorical data.

The evaluation of our geovisual analytics approach to climate study can also be extended to include more case studies with different data sets. In addition, comparing traditional techniques with the geovisual analytics method via a human subject user study would provide valuable insight on the potential of our new approach.

5.3 Related Publications

Publications that are related from this dissertation research are described in the following list:

- C. A. Steed, P. J. Fitzpatrick, T. J. Jankun-Kelly, A. N. Yancey, and J. E. Swan II, "Practical Application of Parallel Coordinates to Hurricane Trend Analysis," *Posters Compendium: IEEE Visualization 2007*, Sacramento, California, October 2007, pp. 4–5, IEEE Computer Society.

- C. A. Steed, P. J. Fitzpatrick, T. J. Jankun-Kelly, J. E. Swan II, and A. N. Yancey, *An Interactive Parallel Coordinates Techniques Applied to a Tropical Cyclone Climate Analysis*, Tech. Rep. NRL/MR/7440–08-9126, Naval Research Laboratory, Stennis Space Center, MS 39529, June 2008.

- C. A. Steed, P. J. Fitzpatrick, T. J. Jankun-Kelly, and J. E. Swan II, *Visual Analysis of North Atlantic Hurricane Trends using Parallel Coordinates and Statistical Techniques*, Tech. Rep. NRL/MR/7440–08-9130, Naval Research Laboratory, Stennis Space Center, MS 39529, July 2008.

- C. A. Steed, P. J. Fitzpatrick, T. J. Jankun-Kelly, and J. E. Swan II, "North Atlantic Hurricane Trend Analysis using Parallel Coordinates and Statistical Techniques," *Geospatial Visual Analytics Workshop*. International Cartographic Association, September 2008. http://geoanalytics.net/GeoVisualAnalytics08/ (current 9 Oct. 2008).

- C. A. Steed, P. J. Fitzpatrick, T. J. Jankun-Kelly, A. N. Yancey, and J. E. Swan II, "An Interactive Parallel Coordinates Techniques Applied to a Tropical Cyclone Climate Analysis," *Computers & Geosciences*, 2008, Under Review.

REFERENCES

[1] "About FedEx," http://www.fedex.com/us/about/ (current 15 Oct. 2008).

[2] "AT&T – Investor Relations," http://www.att.com/gen/investor-relations?pid=5711 (current 15 Oct. 2008).

[3] "Image Gallery - Walker Circulation," National Oceanic & Atmospheric Administration Geophysical Fluid Dynamics Laboratory, http://www.gfdl.noaa.gov/research/climate/highlights/GFDL_V1N3_gallery.html (current 15 Oct. 2008).

[4] "Engineering Statistics Handbook," National Institute for Standards and Technology, 2008, http://www.itl.nist.gov/div898/handbook/index.htm (current 9 Sept. 2008).

[5] "Monthly Climate Timeseries: Atlantic Meridional Mode SST Index," National Oceanic & Atmospheric Administration, 2008, http://www.cdc.noaa.gov/Timeseries/Monthly/AMM/ (current 19 Sept. 2008).

[6] "NHC Website," National Hurricane Center, 2008, http://www.nhc.noaa.gov (current 9 Sept. 2008).

[7] P. I. A.Godinho, B. S. Meiguins, A. S. Meiguins, R. M. do Carmo, M. d. Garcia, L. H. Almeida, and R. Lourenco, "PRISMA – A Multidimensional Information Visualization Tool Using Multiple Coordinated Views," *International Conference on Information Visualization*, Zürich, Switzerland, Jul. 2007, IEEE Computer Society, pp. 23–32.

[8] C. Ahlberg and B. Shneiderman, "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays," *Proceedings of Human Factors in Computing Systems*, Boston, Massachusetts, Apr. 1994, ACM, pp. 313–317, 479–480.

[9] H. Albazza and X. Z. Wang, "Historical Data Analysis Based on Plots of Independent and Parallel Coordinates and Statistical Control Limits," *Journal of Process Control Limits*, vol. 16, no. 2, Feb. 2006, pp. 103–114.

[10] G. Andrienko and N. Andrienko, “Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates,” *Computers, Environment and Urban Systems*, vol. 25, no. 1, Jan. 2001, pp. 5–15.

[11] G. Andrienko and N. Andrienko, “Parallel Coordinates for Exploring Properties of Subsets,” *Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization*, London, England, Jul. 2004, IEEE Computer Society, pp. 93–104.

[12] N. Andrienko, G. Andrienko, and P. Galas, “Exploratory Spatio-Temporal Visualization: An Analytical Review,” *Journal of Visual Languages & Computing*, vol. 14, no. 6, Dec. 2003, pp. 503–541.

[13] M. Ankerst, S. Berchtold, and D. A. Keim, “Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data,” *Proceedings of the IEEE Symposium on Information Visualization*, Research Triangle, California, Oct. 1998, IEEE Computer Society, pp. 52–60, 153.

[14] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz, “Uncovering Clusters in Crowded Parallel Coordinates Visualization,” *IEEE Symposium on Information Visualization*, Austin, Texas, Oct. 2004, IEEE Computer Society, pp. 81–88.

[15] A. S. Bair, D. H. House, and C. Ware, “Texturing of Layered Surfaces for Optimal Viewing,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, Sept. 2006, pp. 1125–1132.

[16] N. Barlow and L. J. Stuart, “Animator: A Tool for the Animation of Parallel Coordinates,” *Proceedings of the International Conference on Information Visualization*, London, England, Jul. 2004, IEEE Computer Society, pp. 725–730.

[17] R. G. Barry and R. J. Chorley, *Atmosphere, Weather and Climate*, 6th edition, Routledge, New York, New York, 1992.

[18] G. D. Bell, E. Blake, C. W. Landsea, M. Chelliah, R. Pasch, K. C. Mo, and S. B. Goldenberg, “The Tropics – Atlantic Basin,” *State of the Climate in 2006*, A. Arguez, ed., vol. 88, Bulletin of the American Meteorological Society, 2007, pp. S48–S51.

[19] F. Bendix, R. Kosara, and H. Hauser, “Parallel Sets: Visual Analysis of Categorical Data,” *Proceedings of the IEEE Symposium on Information Visualization*, Minneapolis, Minnesota, Oct. 2005, IEEE Computer Society, pp. 133–140.

[20] M. R. Berthold and L. O. Hall, “Visualizing Fuzzy Points in Parallel Coordinates,” *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 3, Jun. 2003, pp. 369–374.

[21] E. Bertini, L. D. Aquila, and G. Santucci, “SpringView: Cooperation of Radviz and Parallel Coordinates for View Optimization and Clutter Reduction,” *Proceedings*

of the International Conference on Coordinated & Multiple Views in Exploratory Visualization, London, England, Jul. 2005, IEEE Computer Society, pp. 22–29.

[22] D. Brodbeck and L. Girardin, “Visualization of Large-Scale Customer Satisfaction Surveys Using a Parallel Coordinate Tree,” *IEEE Symposium on Information Visualization*, Seattle, Washington, Oct. 2003, IEEE Computer Society, pp. 197–202.

[23] M. Caat, N. M. Maurits, and J. B. Roerdink, “Design and Evaluation of Tiled Parallel Coordinate Visualization of Multichannel EEG Data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 1, Jan. 2007, pp. 70–79.

[24] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, San Francisco, California, 1999.

[25] B. J. Chalmer, *Understanding Statistics*, Marcel Dekker, Inc., New York, New York, 1987.

[26] R. Chang, G. Wessel, R. Kosara, E. Sauda, and W. Ribarsky, “Legible Cities: Focus-Dependent Multi-Resolution Visualization of Urban Relationships,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, Nov. 2007, pp. 1169–1175.

[27] J. X. Chen and S. Wang, “Data Visualization: Parallel Coordinates and Dimension Reduction,” *Computing in Science & Engineering*, vol. 3, no. 5, Sept. 2001, pp. 110–113.

[28] S.-Y. Chou, C.-W. Lin, and C.-S. Yeh, “Cluster Identification with Parallel Coordinates,” *Pattern Recognition Letters*, vol. 20, Jun. 1999, pp. 565–572.

[29] P.-S. Chu, “ENSO and Tropical Cyclone Activity,” *Hurricanes and Typhoons: Past, Present, and Future*, R. J. Murnane and K.-B. Liu, eds., Columbia University Press, 2004, pp. 297–332.

[30] Computer Science and Telecommunications Board, National Research Council, *IT Roadmap to a Geospatial Future*, National Academic Press, Washington, D. C., 2003.

[31] M. Crider, S. Bergner, T. N. Smyth, T. Möller, M. Tory, A. E. Kirkpatrick, and D. Weiskopf, “A Mixing Board Interface for Graphics and Visualization Applications,” *Proceedings of the Graphics Interface Conference*, Montréal, Canada, May 2007, ACM, pp. 87–94.

[32] J. Depner, B. Reed, S. Byrne, J. Parker, M. Paton, L. Gee, L. A. Mayer, and C. Ware, “Dealing with Increasing Data Volumes and Decreasing Resources,” *Proceedings of Oceans 2002*. Oct. 2002, vol. 2, pp. 1212–1222, IEEE.

[33] T. Dwyer, S.-H. Hong, D. Koschützki, F. Schreiber, and K. Xu, "Visual Analysis of Network Centralities," *Proceedings of Asia-Pacific Symposium on Information Visualization*, Tokyo, Japan, Feb. 2006, Australian Computer Society, pp. 189–197, ACM.

[34] J. A. Dykes and D. M. Mountain, "Seeking Structure in Records of Spatio-Temporal Behavior: Visualization Issues, Efforts and Applications," *Computational Statistics and Data Analysis*, vol. 43, no. 4, Aug. 2003, pp. 581–603.

[35] R. M. Edsall, "The Parallel Coordinate Plot in Action: Design and Use for Geographic Visualization," *Computational Statistics and Data Analysis*, vol. 43, no. 4, Aug. 2003, pp. 605–619.

[36] G. Ellis and A. Dix, "Enabling Automatic Clutter Reduction in Parallel Coordinate Plots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, Sept. 2006, pp. 717–723.

[37] N. Elmqvist, J. Stasko, and P. Tsigas, "DataMeadow: A Visual Canvas for Analysis of Large-Scale Multivariate Data," *IEEE Symposium on Visual Analytics Science and Technology*, W. Ribarsky and J. Dill, eds., Sacramento, California, Oct. 2007, IEEE Computer Society, pp. 187–194.

[38] D. Ericson, J. Johansson, and M. Cooper, "Visual Data Analysis using Tracked Statistical Measures within Parallel Coordinate Representations," *Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization*, London, England, Jul. 2005, IEEE Computer Society, pp. 42–53.

[39] G. Falkman, "Information Visualization in Clinical Odontology: Multidimensional Analysis and Interactive Data Exploration," *Artificial Intelligence in Medicine*, vol. 22, no. 2, May 2001, pp. 133–158.

[40] E. Fanea, S. Carpendale, and T. Isenberg, "An Interactive 3D Integration of Parallel Coordinates and Star Glyphs," *IEEE Symposium on Information Visualization*, Minneapolis, Minnesota, Oct. 2005, pp. 149–156.

[41] N. Feldt, H. Pettersson, J. Johansson, and M. Jern, "Tailor-made Exploratory Visualization for Statistics Sweden," *Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization*, London, England, Jul. 2005, IEEE Computer Society, pp. 133–142.

[42] S. Few, "Bridging the Chasm Between Infovis and the World Out There," *IEEE Visualization 2007 Tutorial*, Sacramento, California, Oct. 2007, IEEE Computer Society, http://vis.computer.org/vis2007/session/tutorials.html (current 9 Sept. 2008).

[43] P. Fiorini and A. Inselberg, "Configuration Space Representation in Parallel Coordinates," *Proceedings of the International Conference on Robotics and Automation*, Scottsdale, Arizona, May 1989, IEEE, pp. 1215–1220.

[44] P. J. Fitzpatrick, *Understanding and Forecasting Tropical Cyclone Intensity Change*, Doctoral Dissertation, Department of Atmospheric Sciences, Colorado State University, Fort Collins, Colorado, 1996.

[45] P. J. Fitzpatrick, "Understanding and Forecasting Tropical Cyclone Intensity Change with the Typhoon Intensity Prediction Scheme (TIPS)," *Weather and Forecasting*, vol. 12, no. 4, 1997, pp. 826–846.

[46] P. J. Fitzpatrick, *Hurricanes: A Reference Handbook*, 2nd edition, Contemporary World Issues. ABC–CLIO, Santa Barbara, California, 2006.

[47] C. Forsell and J. Johansson, "Tasked-Based Evaluation of Multi-Relational 3D and Standard 2D Parallel Coordinates," *Visualization and Data Analysis*, R. F. Erbacher, J. C. Roberts, M. T. Gröhn, and K. Börner, eds., San Jose, California, Jan. 2007, SPIE, vol. 6495, pp. 64950C–1–12.

[48] M. Friendly and E. Kwan, "Effect Ordering for Data Displays," *Computational Statistics and Data Analysis*, vol. 43, no. 4, Aug. 2003, pp. 509–539.

[49] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Hierarchical Parallel Coordinates for Exploration of Large Datasets," *Proceedings of IEEE Visualization*, San Francisco, California, Oct. 1999, IEEE Computer Society, pp. 43–50.

[50] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Structure-Based Brushes: A Mechanism for Navigating Hierarchically Organized Data and Information Spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 2, Apr. 2000, pp. 150–159.

[51] P. Gatalsky, N. Andrienko, and G. Andrienko, "Interactive Analysis of Event Data Using Space-Time Cube," *Proceedings of the International Conference on Information Visualisation*, London, England, Jul. 2004, IEEE Computer Society, pp. 145–152.

[52] E. W. Gilbert, "Pioneer Maps of Health and Disease in England," *Geographical Journal*, vol. 124, 1958, pp. 172–183.

[53] A. Goel, C. Baker, C. A. Shaffer, B. Grossman, R. T. Haftka, W. H. Mason, and L. T. Watson, "VizCraft: A Multidimensional Visualization Tool for Aircraft Configuration Design," *Proceedings of the IEEE Visualization Conference*, San Francisco, California, Oct. 1999, IEEE Computer Society, pp. 425–128, 555.

[54] S. B. Goldenberg, C. W. Landsea, A. M. M.-N. nez, and W. M. Gray, "The Recent Increase in Atlantic Hurricane Activity: Causes and Implications," *Science*, vol. 293, Jul. 2001, pp. 474–479.

[55] M. Graham and J. Kennedy, "Using Curves to Enhance Parallel Coordinate Visualizations," *Proceedings of the International Conference on Information Visualization*, London, England, Jul. 2003, IEEE Computer Society, pp. 10–16.

[56] E. Gröller, H. Löffelmann, and R. Wegenkittl, "Visualization of Analytically Defined Dynamical Systems," *Proceedings of the Conference on Scientific Visualization*, Dagstuhl, Germany, Jun. 1997, IEEE Computer Society, pp. 71–84.

[57] K. Grünfeld, "Integrating Spatio-temporal Information in Environmental Monitoring Data – A Visualization Approach Applied to Moss Data," *Science of the Total Environment*, vol. 347, no. 1–3, Jul. 2005, pp. 1–20.

[58] D. Guo, J. Chen, A. M. MacEachren, and K. Liao, "A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP)," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, Nov. 2006, pp. 1461–1474.

[59] L. O. Hall and M. R. Berthold, "Fuzzy Parallel Coordinates," *Proceedings of the International Conference of the North American Fuzzy Information Processing Society*, Atlanta, Georgia, Jul. 2000, IEEE, pp. 74–78.

[60] R. W. Hamming, *Numerical Analysis for Scientists and Engineers*, McGraw-Hill, New York, New York, 1973.

[61] M. C. Hao, U. Dayal, D. A. Keim, D. Morent, and J. Schneidewind, "Intelligent Visual Analytics Queries," *IEEE Symposium on Visual Analytics Science and Technology*, W. Ribarsky and J. Dill, eds., Sacramento, California, Oct. 2007, IEEE Computer Society, pp. 91–98.

[62] S. Haroz, K. Ma, and K. Heitmann, "Multiple Uncertainties in Time-Variant Cosmological Particle Data," *IEEE Pacific Visualization Symposium*, Kyoto, Japan, Mar. 2008, IEEE Computer Society, pp. 207–214.

[63] H. Hauser, F. Ledermann, and H. Doleisch, "Angular Brushing of Extended Parallel Coordinates," *Proceedings of the IEEE Symposium on Information Visualization*, Boston, Massachusetts, Oct. 2002, IEEE Computer Society, pp. 127–130.

[64] C. G. Healey, L. Tateosian, J. T. Enns, and M. Remple, "Perceptually-Based Brush Strokes for Nonphotorealistic Visualization," *ACM Transactions on Graphics*, vol. 23, no. 1, 2004, pp. 64–96.

[65] G. Heinz, L. J. Peterson, R. W. Johnson, and C. J. Kerk, "Exploring Relationships in Body Dimensions," *Journal of Statistics Education*, vol. 11, no. 2, Jul. 2003, http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html (current 9 Sept. 2008).

[66] M. Henley, M. Hagen, and R. D. Bergeron, "Evaluating Two Visualization Techniques for Genome Comparison," *International Conference on Information Visualization*, Zürich, Switzerland, Jul. 2007, IEEE Computer Society, pp. 551–558.

[67] Y. V. Heyden, V. Pravdova, F. Questier, L. Tallieu, A. Scott, and D. L. Massart, "Parallel Co-ordinate geometry and principal component analysis for the interpretation of large multi-response experimental designs," *Analytica Chimica Acta*, vol. 458, no. 2, May 2002, pp. 397–415.

[68] P. Hofman, G. Grinstein, and D. Pinkney, "Dimensional Anchors: A Graphic Primitive for Multidimensional Multivariate Information Visualizations," *Workshop on New Paradigms in Information Visualization and Manipulation*, Kansas City, Mosisurri, Nov. 1999, ACM, pp. 9–16.

[69] D. C. Howell, *Statistical Methods for Psychology*, 2nd edition, Duxbury Press, 1987.

[70] C.-K. Hung and A. Inselberg, "Visualizing Multidimensional Relations with Parallel Coordinates," *International Conference on Information Technology: Research and Education*, Tel Aviv, Israel, Oct. 2006, IEEE Communications Society, pp. 261–265.

[71] A. Inselberg, "Parallel Coordinates . . . Serious and Humor: Homepage Alfred Inselberg," http://www.math.tau.ac.il/ aiisreal/ (current 9 Sept. 2008).

[72] A. Inselberg, *N-Dimensional Graphics Part I: Lines and Hyperplanes*, Tech. Rep. G320-2711, IBM Los Angeles Scientific Center, Los Angeles, California, 1981.

[73] A. Inselberg, "The Plane with Parallel Coordinates," *The Visual Computer*, vol. 1, no. 4, Dec. 1985, pp. 69–91, http://www.springerlink.com/content/x3p504736mu14661/ (current 9 Sept. 2008).

[74] A. Inselberg, "Multidimensional Detective," *Proceedings of the IEEE Symposium on Information Visualization*, Phoenix, Arizona, Oct. 1997, IEEE Computer Society, pp. 100–107.

[75] A. Inselberg, "Conflict Detection and Planar Resolution for Air Traffic Control," *IEEE Intelligent Transportation Systems Conference Proceedings*, Oakland, California, Aug. 2001, IEEE Computer Society, pp. 1200–1205.

[76] A. Inselberg, "Visualization and Knowledge Discovery for High Dimensional Data," *International Workshop on User Interfaces to Data Intensive Systems*, Zurich, Switzerland, May 2001, IEEE Computer Society, pp. 5–24.

[77] A. Inselberg and T. Avidan, "The Automated Multidimensional Detective," *Proceedings of the IEEE Symposium on Information Visualization*, San Francisco, California, Oct. 1999, IEEE Computer Society, pp. 112–119, 151.

[78] A. Inselberg and T. Avidan, "Classification and Visualization for High-Dimensional Data," *Proceedings of the Conference on Knowledge Discovery in Data*, Boston, Massachusetts, Aug. 2000, ACM, pp. 370–374.

[79] A. Inselberg and T. Chomut, “Convexity Algorithms in Parallel Coordinates,” *Journal of the Association for Computing Machinery*, vol. 34, no. 4, Oct. 1987, pp. 765–801.

[80] A. Inselberg and B. Dimsdale, “Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry,” *Proceedings of IEEE Visualization*, San Francisco, California, Oct. 1990, IEEE Computer Society, pp. 361–378.

[81] A. Inselberg and B. Dimsdale, “Multidimensional Lines I: Representation,” *SIAM Journal on Applied Mathematics*, vol. 54, Apr. 1994, pp. 559–577.

[82] A. Inselberg and B. Dimsdale, “Multidimensional Lines II: Proximity and Applications,” *SIAM Journal on Applied Mathematics*, vol. 54, Apr. 1994, pp. 578–596.

[83] M. Jern, “An Information Visualization Approach to Retail Space Management (VisMT),” *Proceedings of the International Conference on Information Visualization*, Zürich, Switzerland, Jul. 2007, IEEE Computer Society, pp. 109–116.

[84] M. Jern and J. Franzén, ““GeoAnalytics” – Exploring Spatio-temporal and Multivariate Data,” *Proceedings of the Information Visualization*, London, United Kingdom, Jul. 2006, IEEE Computer Society, pp. 25–31.

[85] M. Jern, S. Johansson, J. Johansson, and J. Franzén, “The GAV Toolkit for Multiple Linked Views,” *Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization*, Zürich, Switzerland, Jul. 2007, IEEE Computer Society, pp. 85–97.

[86] J. Johansson, M. Cooper, and M. Jern, “3-Dimensional Display for Clustered Multi-Relational Parallel Coordinates,” *Proceedings of the International Conference on Information Visualization*, London, England, Jul. 2005, IEEE Computer Society, pp. 188–193.

[87] J. Johansson, C. Forsell, M. Lind, and M. Cooper, “Perceiving Patterns in Parallel Coordinates: Determining Thresholds for Identification of Relationships,” *Information Visualization*, vol. 7, no. 2, May 2008, pp. 152–162, http://staffwww.itn.liu.se/ jimjo/ (current 9 Sept. 2008).

[88] J. Johansson, P. Ljung, and M. Cooper, “Depth Cues and Density in Temporal Parallel Coordiantes,” *Joint EUROGRAPHICS – IEEE VGTC Symposium on Visualization*, K. Museth, T. Möller, and A. Ymeman, eds., Prague, Czech Republic, Sept. 2007, IEEE Computer Society, pp. 35–42.

[89] J. Johansson, P. Ljung, M. Jern, and M. Cooper, “Revealing Structure within Clustered Parallel Coordinates Displays,” *IEEE Symposium on Information Visualization*, Minneapolis, Minnesota, Oct. 2005, IEEE Computer Society, pp. 125–132.

[90] J. Johansson, P. Ljung, M. Jern, and M. Cooper, “Revealing Structure in Visualizations of Dense 2D and 3D Parallel Coordinates,” *Information Visualization*, vol. 5, no. 2, Jun. 2006, pp. 125–136.

[91] J. Johansson, R. Treloar, and M. Jern, “Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data,” *Proceedings of the International Conference on Information Visualization*, London, England, Jul. 2004, IEEE Computer Society, pp. 52–57.

[92] C. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, eds., *NIH/NSF Visualization Reserach Challenges*, 1st edition, IEEE Press, Los Angeles, California, 2006, http://tab.computer.org/vgtc/vrc/index.html (current 9 Sept. 2007).

[93] C. Jones, K. Ma, S. Ethier, and W. Lee, “An Integrated Exploration Approach to Visualizing Multivariate Particle Data,” *Computing in Science & Engineering*, vol. 10, no. 4, Jul. 2008, pp. 20–29.

[94] E. Kandogan, “Visualizing Multi-Dimensional Clusters, Trends, and Outliers Using Star Coordinates,” *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, Aug. 2001, ACM, pp. 107–116.

[95] B. B. Karki and R. Chennamsetty, “A Visualization System for Mineral Elasticity,” *Visual Geosciences*, vol. 9, no. 1, Jan. 2006, pp. 49–67.

[96] D. A. Keim and H.-P. Kriegel, “VisDB: Database Exploration Using Multidimensional Visualization,” *IEEE Computer Graphics and Applications*, vol. 14, no. 5, Sept. 1994, pp. 40–49.

[97] K. R. King and T. R. Harris, “Parallel-Coordinates Visualization of Capillary Transport Model Analysis,” *Proceedings of the First Joint BMES/EMBS Conference Serving Humanity, Advancing Technology*, Atlanta, Georgia, Oct. 1999, IEEE, p. 1193.

[98] M. Kirby, H. Marmanis, and D. H. Laidlaw, “Visualizing Multivalued Data from 2D Incompressible Flows Using Concepts from Painting,” San Francisco, California, Oct. 1999, IEEE Computer Society, pp. 333–340, 540.

[99] P. J. Klotzbach, “Re: El Nino / La Nina Data,” personal communication, Jan. 2007.

[100] P. J. Klotzbach and W. M. Gray, *Summary of 2006 Atlantic Tropical Cyclone Activity and Verification of Author’s Seasonal and Monthly Forecasts*, Tech. Rep., Colorado State University, Nov. 2006, http://hurricane.atmos.colostate.edu/Forecasts/2006/nov2006/ (current 9 Sept. 2008).

[101] P. J. Klotzbach, W. M. Gray, and W. Thorson, *Extended Range Forecast of Atlantic Seasonal Hurricane Activity and U.S. Landfall Strike Probability for 2007*, Tech. Rep., Colorado State University, 2006, http://tropical.atmos.colostate.edu/Forecasts/2006/dec2006/ (current 9 Sept. 2008).

[102] Z. Konyha, K. Matković, D. Gracanin, M. Jelović, and H. Hauser, "Interactive Visual Analysis of Families of Function Graphs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, Nov. 2006, pp. 1373–1385.

[103] R. Kosara, F. Bendix, and H. Hauser, "Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, Jul. 2006, pp. 558–568.

[104] J. P. Kossin and D. J. Vimont, "A More General Framework for Understanding Atlantic Hurricane Variability and Trends," *Bulletin of the American Meteorological Society*, vol. 88, no. 11, Nov. 2007, pp. 1767–1781.

[105] M.-J. Kraak, "Playing with maps: Explore, discover, learn, categorize, model, analyze, explain, present geographic and non-geographic data," *Proceedings of the Information Visualization*, London, England, Jul. 2006, IEEE Computer Society, pp. 291–296.

[106] M.-J. Kraak, "Visualization Viewpoints: Beyond Geovisualization," *IEEE Computer Graphics and Applications*, vol. 26, no. 4, Jul. 2006, pp. 6–9.

[107] S. Krasser, G. Conti, J. Grizzard, J. Gribschaw, and H. Owen, "Real-Time and Forensic Network Data Analysis Using Animated and Coordinated Visualization," *Proceedings of the IEEE Workshop on Information Assurance and Security*, West Point, New York, Jun. 2005, IEEE Computer Society, pp. 42–49.

[108] N. Kumasaka and R. Shibata, "High-dimensional data visualization: The textile plot," *Computational Statistics and Data Analysis*, vol. 52, Mar. 2008, pp. 3616–3644.

[109] C. W. Landsea, "Hurricanes and global warming," *EOS*, vol. 438, 2005, pp. E11–E13.

[110] E. Lanthier, "Correlation," *Psychology Research Methods*, Northern Virginia Community College Website, 2002, http://www.nvcc.edu/home/elanthier/methods/correlation.htm (current 9 Sept. 2008).

[111] M. Lanzenberger, S. Miksch, and M. Pohl, "Exploring Highly Structured Data A Comparative Study of Stardinates and Parallel Coordinates," *Proceedings of the International Conference on Information Visualization*, London, England, Jul. 2005, IEEE Computer Society, pp. 312–320.

[112] M. Lawrence, E.-K. Lee, D. Cook, H. Hofmann, and E. Wurtele, “exploRase: Exploratory Data Analysis of Systems Biology Data,” *Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization*, London, England, Jul. 2006, IEEE Computer Society, pp. 14–20.

[113] H.-Y. Lee and H.-L. Ong, “Visualization Support for Data Mining,” *IEEE Expert*, vol. 11, no. 5, Oct. 1996, pp. 96–75.

[114] H.-Y. Lee, H.-L. Ong, E.-W. Toh, and S.-K. Chan, “A Multi-Dimensional Data Visualization Tool for Knowledge Discovery in Databases,” *Proceedings of the International Computer Software and Applications Conference*, Dallas, Texas, Aug. 1995, IEEE, pp. 26–31.

[115] J. Lee, M. Podlaseck, E. Schonberg, and R. Hoch, “Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising,” *Data Mining and Knowledge Discovery*, vol. 5, no. 1–2, Jan.–Apr. 2001, pp. 59–84.

[116] P. Lyman and H. R. Varian, “How Much Information 2003?,” University of California at Berkeley, 2003, http://www.sims.berkeley.edu/how-much-info-2003 (current 9 Sept. 2008).

[117] A. M. MacEachren, M. Gahegan, W. Pike, I. Brewer, G. Cai, E. Lengerich, and F. Hardisty, “Visualization Viewpoints: Geovisualization for Knowledge Construction and Decision Support,” *IEEE Computer Graphics and Applications*, vol. 24, no. 1, Jan. 2004, pp. 13–17.

[118] A. M. MacEachren and M.-J. Kraak, “Research Challenges in Geovisualization,” *Cartography and Geographic Information Science*, vol. 28, no. 1, 2001, pp. 3–12.

[119] A. R. Martin and M. O. Ward, “High Dimensional Brushing for Interactive Exploration of Multivariate Data,” *Proceedings of 6th IEEE Visualization 1995 Conference*, Atlanta, Georgia, Oct. 1995, IEEE Computer Society, pp. 271–278.

[120] K. Matkovic, M. Jelović, J. Jurić, Z. Konyha, and D. Gracanin, “Interactive Visual Analysis and Exploration of Injection Systems Simulations,” *Proceedings of IEEE Visualization*, Minneapolis, Minnesota, Oct. 2005, IEEE Computer Society, pp. 391–398.

[121] T. Munzner, C. Johnson, R. Moorhead, H. Pfister, P. Rheingans, and T. S. Yoo, “Visualization Viewpoints: NIH-NSF Visualization Research Challenges Report Summary,” *IEEE Computer Graphics and Applications*, vol. 26, no. 2, Mar. 2006, pp. 20–24.

[122] H. Notsu, Y. Okada, M. Akaishi, and K. Niijima, “Time-tunnel: Visual Analysis Tool for Time-series Numerical Data and Its Extension toward Parallel Coordinates,” *Proceedings of the Computer Graphics, Imaging and Vision: New Trends*, Beijing, China, Jul. 2005, pp. 167–172.

[123] M. Novotný and H. Hauser, "Outlier-preserving Focus+Context Visualization in Parallel Coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, Sept. 2006, pp. 893–900.

[124] S. Park and A. Martin, "A Novel Assessment Tool for Reusability of Wastes," *Journal of Hazardous Materials*, vol. 139, no. 3, Jan. 2007, pp. 575–583.

[125] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering," *IEEE Symposium on Information Visualization*, Austin, Texas, Oct. 2004, IEEE Computer Society, pp. 89–96.

[126] R. M. Pillat and C. M. Freitas, "Coordinating Views in the InfoVis Toolkit," *Proceedings of Advanced Visual Interfaces*, Venezia, Italy, May 2006, ACM, pp. 496–499.

[127] P. Pirolli and R. Rao, "Table Lens as a Tool for Making Sense of Data," *Proceedings of the Workshop on Advanced Visual Interfaces*, Gubbio, Italy, May 1996, ACM, pp. 67–80.

[128] S. Potts, M. Tory, and T. M. öller, "A Parallel Coordinates Interface for Exploratory Volume Visualization," *Posters Compendium: IEEE Visualization 2003*, Seattle, Washington, Oct. 2003, IEEE Computer Society, pp. 102–103.

[129] 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*, vol. 13, no. 6, Nov. 2007, pp. 1408–1415.

[130] R. A. Rensink, "Change Detection," *Annual Review of Psychology*, vol. 53, 2002, pp. 245–577.

[131] A. Schneidewind, P. Neumann, and I. Schmitt, "An Approach to Visualize Image Retrieval Results," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.*, Washington, D.C., Jun. 2004, IEEE Computer Society.

[132] J. Shearer, M. Ogawa, K. Ma, and T. Kohlenberg, "Pixelplexing: Gaining Display Resolution Through Time," *IEEE Pacific Visualization Symposium*, Kyoto, Japan, Mar. 2008, IEEE Computer Society, pp. 159–166.

[133] B. Shneiderman, "Dynamic Queries for Visual Information Seeking," *IEEE Software*, vol. 11, no. 6, 1994, pp. 70–77.

[134] M. Sifer, "User Interfaces for the Exploration of Hierarchical Multi-dimensional Data," *IEEE Symposium on Visual Analytics Science and Technology*, Baltimore, Maryland, Oct. 2006, IEEE Computer Society, pp. 175–182.

[135] H. Siirtola, "Direct Manipulation of Parallel Coordinates," *Proceedings of the International Conference on Information Visualisation*, London, England, Jul. 2000, IEEE Computer Society, pp. 373–378.

[136] H. Siirtola, "Combining Parallel Coordinates with the Reorderable Matrix," *Proceedings of the Coordinated and Multiple Views in Exploratory Visualization Conference*, London, England, Jul. 2003, IEEE Computer Society, pp. 63–74.

[137] H. Siirtola and K.-J. Räihä, "Interacting with Parallel Coordinates," *Interacting with Computers*, vol. 18, no. 6, Dec. 2006, pp. 1278–1309.

[138] Y. Singer, O. Greenshpan, and A. Inselberg, "Multidimensional Visualization of Communication Networks," *IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Nov. 2006, IEEE, pp. 371–375.

[139] S. L. Spraragen and M. Podlaseck, "A High-Density Catalog for Online Browsing," *Proceedings of the Hawaii International Conference on System Sciences*, Maui, Hawaii, Jan. 2001, IEEE, pp. 1–9.

[140] J. J. Thomas, "Visual Analytics: Why Now?," *Information Visualization*, vol. 6, 2007, pp. 104–106.

[141] J. J. Thomas and K. A. Cook, eds., *Illuminating the Path: The Reserach and Development Agenda for Visual Analytics*, IEEE Computer Society, Los Alamitos, California, 2005, http://nvac.pnl.gov/agenda.stm (current 9 Sept. 2008).

[142] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D Information Visualization for Time Dependent Data on Maps," *Proceedings of the Ninth International Conference on Information Visualization*, London, England, Jul. 2005, IEEE Computer Society, pp. 175–181.

[143] M. Tory, S. Potts, and T. Möller, "A Parallel Coordinates Interface for Exploratory Volume Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, Jan. 2005, pp. 71–80.

[144] E. R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1983.

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

[146] L. Tweedie, R. Spence, H. Dawkes, and H. Su, "Externalising Abstract Mathematical Models," *Proceedings of the Conference on Human Factors in Computing Systems*, Vancouver, British Columbia, Canada, Apr. 1996, ACM, pp. 406–412.

[147] L. Tweedie, R. Spence, H. Dawkes, and H. Su, "Externalising Abstract Mathematical Models," *Proceedings of the Conference on Human Factors in Computing Systems*, Vancouver, British Columbia, Canada, Apr. 1996, ACM, pp. 406–412.

[148] A. Unwin, C. Volinsky, and S. Winkler, "Parallel Coordinates for Exploratory Modelling Analysis," *Computational Statistics and Data Analysis*, vol. 43, no. 4, Aug. 2003, pp. 553–564.

[149] D. J. Vimont and J. P. Kossin, "The Atlantic Meridional Mode and hurricane activity," *Geophysical Research Letters*, vol. 34, 2007, pp. 1–5.

[150] F. Vitart, "Dynamical Seasonal Forecasts of Tropical Storm Statistics," *Hurricanes and Typhoons: Past, Present, and Future*, R. J. Murnane and K.-B. Liu, eds., Columbia University Press, 2004, pp. 354–392.

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

[152] H.-B. Wang, C.-B. Wang, K. Kiu, B. Meng, and D.-R. Zhou, "VISDM-PC: A Visual Data Mining Tool Based on Parallel Coordinate," *Proceedings of the International Conference on Machine Learning and Cybernetics*, Shanghai, China, Aug. 2004, IEEE Computer Society, pp. 1244–1248.

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

[154] V. Wegenkittl, H. Löffelmann, and E. Gröller, "Visualizing the Behavior of Higher Dimensional Dynamical Systems," *Proceedings of 8th IEEE Visualization '97 Conference*, Phoenix, Arizona, Oct. 1997, IEEE Computer Society, pp. 119–125.

[155] E. J. Wegman, "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of the American Statistical Association*, vol. 85, no. 411, Sept. 1990, pp. 664–675.

[156] 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*, vol. 12, no. 6, Nov. 2006, pp. 1363–1372.

[157] P. C. Wong and R. D. Bergeron, "30 Years of Multidimensional Multivariate Visualization," *Scientific Visualization - Overviews, Methodologies, and Techniques*, G. M. Nielson, H. Hagan, and H. Muller, eds., IEEE Computer Society Press, Los Alamitos, California, 1997, pp. 3–33.

[158] Y. Xu, W. Hong, X. Li, and J. Song, "Parallel Dual Visualization of Multidimensional Multivariate Data," *Proceedings of the IEEE International Conference on Integration Technology*, Shenzhen, China, Mar. 2007, IEEE Computer Society, pp. 263–268.

[159] Y. Xu, W. Hong, X. Li, and J. Song, "Visual Pattern Recognition Method Based on Optimized Parallel Coordinates," *Proceedings of the IEEE International Conference on Integration Technology*, Shenzhen, China, Mar. 2007, IEEE Computer Society, pp. 127–132.

[160] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner, "Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets," *IEEE Symposium on Information Visualization*, Seattle, Washington, Oct. 2003, IEEE Computer Society, pp. 105–112.

[161] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang, "Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets," *Joint EUROGRAPHICS – IEEE VGTC Symposium on Visualization*, Grenoble, France, May 2003, Eurographics Association, pp. 19–28, 282.

[162] L. Yang, "Pruning and Visualizing Generalized Association Rules in Parallel Coordinates," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 1, Jan. 2005, pp. 60–70.

[163] H. Ye and Z. Lin, "Speed-up Simulated Annealing by Parallel Coordinates," *European Journal of Operational Research*, vol. 173, no. 1, Aug. 2006, pp. 59–71.

[164] K. Zhao, B. Liu, T. M. Tirpak, and A. Schaller, "Detecting Patterns of Change Using Enhanced Parallel Coordinates Visualization," *Proceedings of the IEEE International Conference on Data Mining*, Melbourne, Florida, Nov. 2003, IEEE, pp. 747–750.

[165] K. Zhao, B. Liu, T. M. Tirpak, and A. Schaller, "V-Miner: Using Enhanced Parallel Coordinates to Mine Product Design and Test Data," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, Washington, Aug. 2004, ACM, pp. 494–502.

APPENDIX A

ANALYSIS PARALLEL COORDINATES LITERATURE REVIEW

Table A.1: Comparative summary of parallel coordinates literature review

When	Who	Objective	Case Study	User Study
1981	Inselberg [72]	N-dimensional geometry, PCP theory	-	-
1985	Inselberg [73]	N-dimensional geometry, PCP theory	X	-
1987	Inselberg & Chomut [79]	N-dimensional geometry, PCP theory	-	-
1989	Fiorini & Inselberg [43]	PCP theory, control robotic arm	-	-
1990	Inselberg & Dimsdale [80]	N-dimensional geometry, PCP theory, air traffic control	X	-
1990	Wegman [155]	multidimensional data analysis, PCP theory	X	-
1994	Inselberg & Dimsdale [81]	N-dimensional geometry, PCP theory	-	-
1994	Inselberg & Dimsdale [82]	N-dimensional geometry, PCP theory, air traffic control	X	-
1995	Lee et al. [114]	grouping lines, statistical analysis, WinVis	-	-
1995	Martin & Ward [119]	multidimensional brushing, XmdvTool	-	X
1996	Lee & Ong [113]	clustering, grouping, statistical analysis, WinVis	X	-
1997	Gröller et al. [56]	3D PCP, extruded PCP	-	-
1997	Inselberg [74]	Analysis strategies with PCP	X	-
1997	Wegenkittl et al. [154]	3D PCP, extruded PCP	X	-
1998	Ankerst et al. [13]	similarity-based axis arrangements, clustering, stock market data	X	-
1999	Chou et al. [28]	PCP theory, clustering	-	-
1999	King & Harris [97]	pulmonary capillary exchange data	X	-
1999	Inselberg & Avidan [77]	PCP theory, automated classification	X	-
1999	Goel et al. [53]	PCP for aircraft design	X	-
1999	Fua et al. [49]	brushing, large data sets, clustering, tree-map, proximity-based shading, distortion	X	-
1999	Hoffman et al. [68]	visualization framework, dimensional anchors	X	-
2000	Fua et al. [50]	brushing, large data sets, clustering, tree-map, proximity-based shading, distortion, remote sensing data	X	-
2000	Hall & Berthold [59]	fuzzy data, clustering	X	-

Continued on next page

Table A.1 – continued from previous page

When	Who	Objective	Case Study	User Study
2000	Siirtola [135]	polyline averaging, correlation indicators, statistical analysis	X	-
2000	Inselberg & Avidan [78]	PCP theory, classification	X	-
2001	Lee et al. [115]	analysis of clickstream data for online sales	X	X
2001	Spraragen & Podlaseck [139]	visualization of musical works	-	X
2001	Andrienko & Andrienko [10]	geovisualization, linked views, comparable attributes, statistical indicators	X	-
2001	Falkman [39]	daily clinicians diagnostic work, 3D PCP, linked displays, clustering, focus+context	X	-
2001	Inselberg [76]	PCP theory, automatic classification	X	-
2001	Inselberg [75]	air traffic control	X	-
2001	Chen & Wang [27]	brushing, dimensional reduction, overcrowded PCP	-	-
2002	Heyden et al. [67]	principal component analysis, multivariate data analysis, multi-response experimental design data	X	-
2002	Hauser et al. [63]	angular brushing, linked displays, visual query, computational fluid dynamics data	-	-
2003	Yang et al. [161]	visual hierarchical dimension reduction	X	-
2003	Berthold & Hall [20] & Hall	fuzzy data, clustering, ocean satellite image data	X	-
2003	Siirtola [136]	linked views, comparison of PCP and the Reorderable Matrix view	-	X
2003	Graham & Kennedy [55]	line clutter, smooth curved polylines, focus+context	-	-
2003	Dykes & Mountain [34]	linked views, geovisualization, statistical analysis, geocentric PCP	X	-
2003	Unwin et al. [148]	comparing models, medical trial data analysis	X	-
2003	Friendly & Kwan [48]	information ordering, crime statistics data	X	-
2003	Edsall [35]	spatio-temporal, brushing, linked views, climate modeling & analysis, epidemiology	X	-
2003	Yang et al. [160]	similarity-based dimensional ordering, dimension spacing, dimension filtering	-	-

Continued on next page

Table A.1 – continued from previous page

When	Who	Objective	Case Study	User Study
2003	Brodbeck & Girardin [22]	combines tree layout with PCP, bifocal lens distortion, focus+context	-	-
2003	Potts et al. [128]	visualization of parameter space used in volumetric rendering	-	-
2003	Zhao et al. [164]	change patterns, edit-distance based axis arrangement, product test and design data	X	-
2004	Schneidewind et al. [131]	image retrieval analysis, linked views	X	-
2004	Johansson et al. [91]	classification, self-organizing map, clustering, zooming, large data sets, linked views, molecular data	-	-
2004	Barlow & Stuart [16]	temporal PCP animation, neurophysiological data visualization	X	-
2004	Andrienko & Andrienko [11]	clustering, statistical-based axis scaling, classes in PCP, demographic data	X	-
2004	Zhao et al. [165]	edit-distance based axis arrangement, product test and design data	X	-
2004	Wang et al. [152]	correlation analysis, zooming, classification, similarity analysis, test score data set analysis	X	-
2004	Artero et al. [14]	large data sets, overcrowding, clustering	X	-
2005	Yang [162]	smooth polylines, visualizing association rules, network traffic data analysis	X	-
2005	Tory et al. [143]	visualization of parameter space used in volumetric rendering, medical visualization	-	X
2005	Lanzenberger et al. [111]	comparison of stardinates and parallel coordinates, psychotherapeutic data	-	X
2005	Notsu et al. [122]	new PCP display technique, 3D PCP, time series data	X	-
2005	Johansson et al. [86]	3D PCP, clustering, axis arrangement, pollution data, meteorological data	-	-
2005	Grünfeld [57]	comparing PCP to other visualization techniques, environmental data, spatio-temporal data	X	-
2005	Feldt et al. [41]	linked views, statistical database for Sweden	X	-

Continued on next page

Table A.1 – continued from previous page

When	Who	Objective	Case Study	User Study
2005	Ericson et al. [38]	clustering, linked views, brushing, animation, statistical indicators	X	-
2005	Bertini et al. [21]	linked views, clutter, clustering	X	-
2005	Matković et al. [120]	injection system simulation data analysis, linked views	X	-
2005	Fanea et al. [40]	3D PCP, parallel glyphs, interaction, distortion, focus+context, color scales	X	-
2005	Bendix et al. [19]	categorical data, statistical indicators, parallel sets, sales, marketing, and demographic data	X	-
2005	Johansson et al. [89]	large data sets, clustering, interaction, transfer function, animation, statistical indicators	-	-
2006	Karki et al. [95]	linked views, mineral elasticity data	X	-
2006	Albazzaz & Wang [9]	dimension reduction, outlier detection, process control	X	-
2006	Dwyer et al. [33]	network structured data centrality analysis, biological and social network data	X	-
2006	Pillat & Freitas [126]	linked views	-	X
2006	Johansson et al. [90]	3D multi-relational PCP, clustering, transfer function, density map, feature animation	X	-
2006	Lawrence et al. [112]	linked views, systems biology data analysis, GGobi	X	-
2006	Kosara et al. [103]	categorical data, statistical indicators, parallel sets, sales, marketing, and demographic data	X	-
2006	Jern & Franzén [84]	geovisualization, visual analytics, spatio-temporal data, linked views, statistical analysis	-	-
2006	Kraak [105]	geovisualization, new roles for maps	-	-
2006	Ye & Lin [163]	optimization problems, simulated annealing, curved PCP polylines	X	-
2006	Novotný & Hauser [123]	focus+context, outlier detection, large data sets	-	-
2006	Ellis & Dix [36]	automatic clutter reduction, lens distortion, measuring occlusion in PCP displays	X	-

Continued on next page

Table A.1 – continued from previous page

When	Who	Objective	Case Study	User Study
2006	Bair et al. [15]	analysis of perception experiment data with PCP	X	-
2006	Hung & Inselberg [70]	PCP theory, multidimensional geometry	-	-
2006	Sifer [134]	extension of parallel sets to remove clutter, sales and network traffic data analysis	-	X
2006	Konyha et al. [102]	linked views, brushing, time series data, function graphs, optimization of fuel injection system	X	-
2006	Singer et al. [138]	analysis of communication network data	X	-
2006	Guo et al. [58]	geovisualization, spatio-temporal, self-organizing map, linked views, axis scaling, small multiples	-	-
2006	Siirtola & Räihä [137]	survey of PCP interaction techniques, PCP usability study	-	X
2007	Forsell & Johansson [47]	axis permutation, 3D PCP, 3D vs. 2D PCP, comparing 3 types of PCP	-	X
2007	Caat et al. [23]	visualization of time-varying multichannel electroencephalography (EEG) data	-	X
2007	Park & Martin [124]	analysis of waste reusability	X	-
2007	Xu et al. [158]	combined scatterplot, star diagram and PCP into parallel dual plot, overplotting, pattern recognition	X	-
2007	Xu et al. [159]	overplotting, pattern recognition, visual classification, machine learning, support vector machines	X	-
2007	Crider et al. [31]	using physical sliders on mixer board for interaction	-	X
2007	A.Godinho et al. [7]	linked views, comparing scatterplots, PCP, and treemap displays	-	X
2007	Henley et al. [66]	genomic study, PCP and scatterplots, disadvantage and advantages of PCP	-	X
2007	Jern [83]	retail space management	-	-
2007	Jern et al. [85]	geovisualization, visual analytics, linked views, visual query	-	-
2007	Johansson et al. [88]	temporal data, GPU, large data sets	X	-

Continued on next page

Table A.1 – continued from previous page

When	Who	Objective	Case Study	User Study
2007	Elmqvist et al. [37]	large data, PCP and starplot, dynamic queries, workflow, qualitative expert review	X	X
2007	Hao et al. [61]	interaction and automation techniques, correlation analysis, similarity measures, clustering	X	-
2007	Chang et al. [26]	urban information visualization, linked views	X	X
2007	Qu et al. [129]	weather data, curved PCP axis, axis arrangement, correlation analysis	X	-
2008	Haroz et al. [62]	cosmology, uncertainty visualization, large data sets, spatio-temporal data	X	-
2008	Shearer et al. [132]	animation, small displays, clutter reduction, scalability, geospatial data	X	-
2008	Kumasaka & Shibata [108]	numerical and categorical data, textile plot PCP extension	X	-
2008	Johansson et al. [87]	comparing 2D PCP and 3D PCP	-	X
2008	Jones et al. [93]	time-varying data, plasma physics simulations, particle data, linked views	X	-

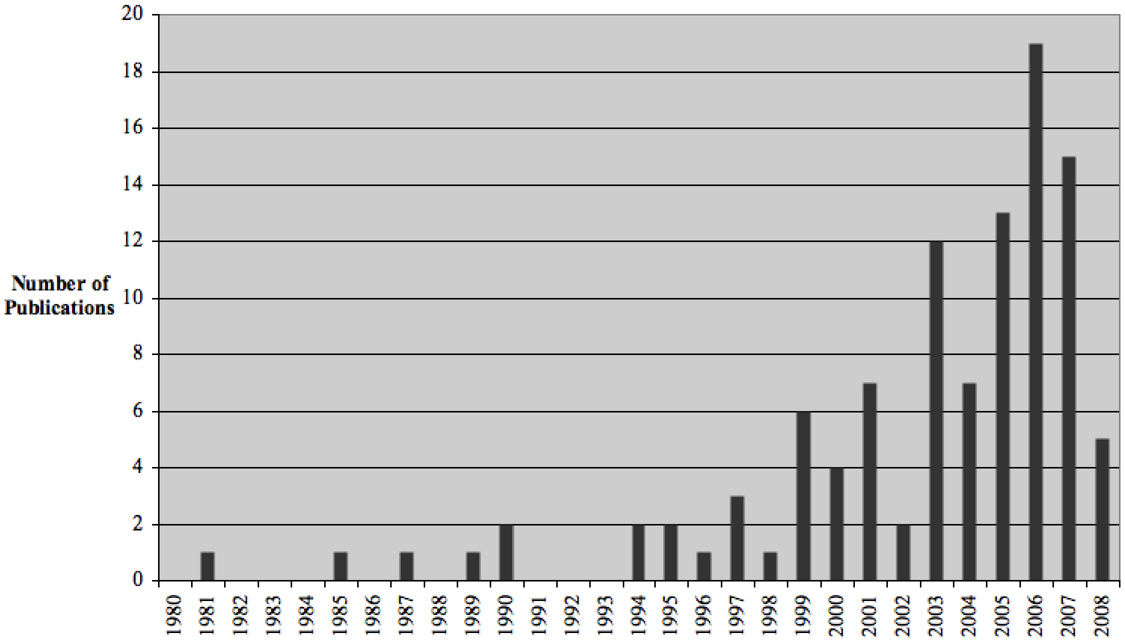


Figure A.1

Plot of parallel coordinates paper count per year.

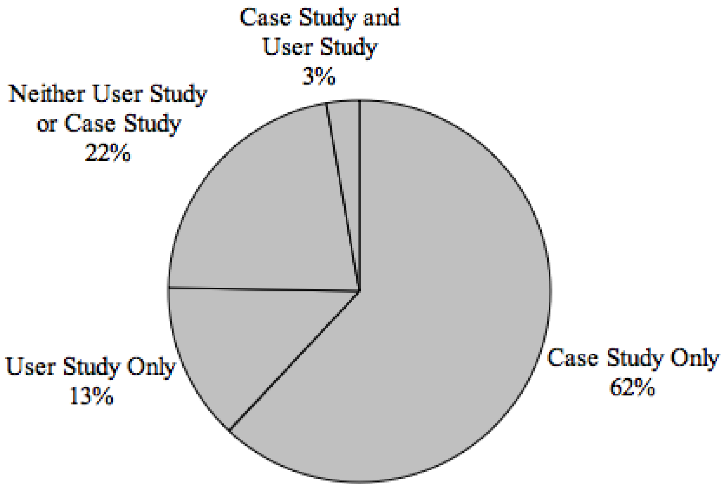


Figure A.2

Plot of parallel coordinates literature evaluations.

Table A.2

Number of parallel coordinates papers per year.

Year	Number of Publications
1981	1
1985	1
1987	1
1989	1
1990	2
1994	2
1995	2
1996	1
1997	3
1998	1
1999	6
2000	4
2001	7
2002	2
2003	12
2004	7
2005	13
2006	19
2007	15
2008	5

Table A.3

Analysis of evaluations in parallel coordinates literature.

Neither User Study or Case Study	23 (3%)
User Study Only	17 (13%)
Case Study Only	65 (62%)
Case Study and User Study	3 (3%)
Total Papers	105 (100%)