Parallel Coordinates as Tool of Exploratory Data Analysis

S.Savoska, S.Loskovska

Abstract — **The huge amount of multidimensional data everyday becomes bigger and more complex. For these reasons, data analysis of multivariate data becomes very difficult. In this paper, we present a visualization technique for multidimensional data sets named Parallel Coordinates as a technique for exploratory data analysis. The paper describes the technique and its applications. We present several examples of technique successful application as a tool for visual data analysis of the multivariate data.**

Keywords — **Parallel Coordinates, exploratory data analysis, multivariate visualization**

I. Introduction

Researcers require more and more efficient ways to analyze and interpret large amount of information, as large multivariate datasets become increasingly common. The visualization methods are very efficient when displaying multidimensional and multivariate datasets. The multivariate dataset is an *N*-dimensional set **E** with elements described by $e_i = (x_{i1}, x_{i2},..., x_{in})$. Each observation x_{ij} is usually independent of the other observations and, observations, in nature, may be discrete or continuous, or may take nominal values.

Several techniques have been proposed for multivariate data representation. They include axes reconfiguration techniques (such as Parallel Coordinates and glyphs); dimensional embedding techniques (such as dimensional stacking and worlds within worlds); dimensional sub-setting (such as scatter-plots) and dimensional reduction techniques (such as multidimensional scaling, self-organizing maps and principal component analysis).

In this paper we describe the multidimensional visualization technique called Parallel Coordinates. The idea for this technique comes from multi-dimensional geometry, which frustrates researchers by the absence of visualization. The question was "How to see geometry without the benefit of the picture?" The 2D or 3D Descartes coordinate systems couldn't solve the multidimensional problems. Inserberg proposed the use of a multidimensional coordinate system based on Parallel Coordinates. In the Euclidian plane (R^2), N-copies of the real line R (labeled $\overline{x_1}$, x_2, ... x_n) are placed equidistant and perpendicular to the X-axis (Figure 1). They correspond to the axes of the Parallel Coordinates system that represents the Euclidian N-dimensional space R^n [1]. All axes have the same positive orientation as the y-axis. The complete polygonal line $\overline{C} = \{C_1, C_2, ...C_n\}$ is represented by the segments between the axes [1]. In this way, a 1-1 correspondence is established between points in R^n and planar polygonal lines in Parallel Coordinates systems (Figure 1). This is the efficient way to place a large number of axes and to visualize the multivariate relations or multidimensional objects. This technique is introduced by Inserberg and Dimsdale during the 1977-1980's. In the early 90's, Parallel Coordinates was used as a two-dimensional technique for multidimensional data sets representation [2]. The technique has been enhanced during the next few years. Researchers have been working on improving this technique for better data investigation and easier user-friendly interaction by adding data clustering [3], brushing [1; 8], etc. With these improvements, Parallel Coordinates becomes very efficient technique for visualization relationships between designated neighboring dimensions.

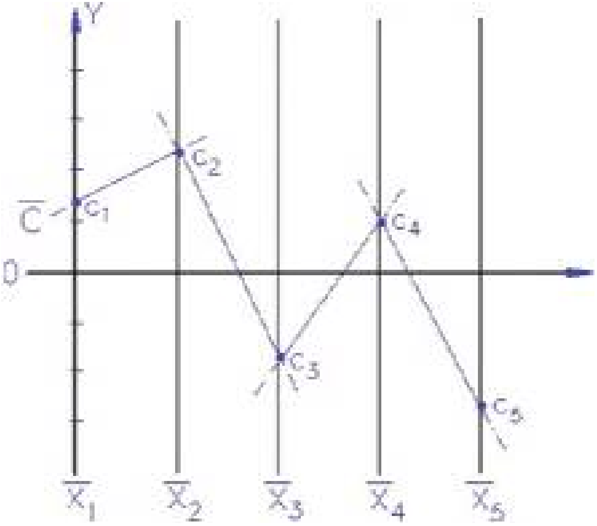

Figure 1 Representation of a point c in a Parallel Coordinate system $\overline{c} = \{c_1, c_2, c_3, c_4, c_5\}$

II. Parallel Coordinates as Visualization Tool

Parallel Coordinates technique is one of the most

S. Savoska, FAMIS – UKLO Bitola, Partizanska bb, 7000, Bitola, Makedonija (tel: 389-71-242 193; e-mail: savoskasnezana@yahoo.com).

S. Loskovska, FEIT – UKIM - Skopje, 1000 Skopje, Makedonija, (e-mail: suze@feit.ukim.edu.mk).

efficient techniques for visualization of the large scale data. It can be used to show multidimensional points that reside at the same poly-line which intersect at specific points between the vertical axes. It is very useful in preventing collisions such those in the air traffic and anytime where positive or negative correlations can be assumed. Crossing axis in opposite order shows negative correlation.

The first and more widespread application of Parallel Coordinates is exploratory data analysis (EDA) for discovering of data subset relations. If the dataset have M items, the subsets may be one of the 2^M. Searching a dataset with M items for interesting properties is very hard. But, our eyes can discriminate in a good data representation and navigate the discovery process. Good representation of datasets with M variables should preserve information and give good results for M (any) number of variables.

Parallel Coordinates transform multivariate relations into 2D patterns. These patterns are suitable for analysis and data exploration – searching for a clue in many dimensions. Even there are usually developed specialized queries to find patterns, they still could not handle all encountered situations. The requirement for successful exploratory data analysis needs to have an informative representation without the loss of data, good choice of queries and skillful interaction with the display.

III. Advantages and Disadventages of the Parallel Coordinates Technique

The main advantage of Parallel Coordinates technique is that the number of data dimensions is restricted only by the horizontal resolution of the screen. But, as the axes get closer it may become more difficult to perceive structures or data relations. Another advantage is that the correlations between variables in the dataset can be spotted easily.

A reduction of the amount of useful information when the level of clutter is present in the visualization is one disadvantage of the technique. There are other limitations of this technique. First of all, readability and efficiency of plots suffer when it is necessary to display a large data-set. In this case, there are problems with the high density representations. These problems cause the loss of speed and interactions, elements overlapping and reduced readability of visual representation of aggregated data because the objects are drown one over another.

IV. What really represent the parallel coordinates technique?

The Parallel Coordinates technique can be considered as a generalization of two-dimensional Cartesian technique. The axes in Parallel Coordinates are drawn parallel to each other. We can draw as many axes as we want, so we can represent the points of dimensionality larger than three. Instead of using a “dot” to represent the location, a “line” is used to connect the coordinates of the point on the axes. In this way the points become lines. In Parallel Coordinates plots, the dual points are lines and the dual lines are points. A point in the Cartesian coordinates becomes a line in Parallel Coordinates (a poly line if we have more than two dimensions). Also, an ellipse in Cartesian coordinate maps into a line hyperbola in Parallel Coordinates (Figure 2 (a, b)). In general the point conic in Cartesian coordinates becomes a line conic in Parallel Coordinates. Also, the rotation in the Cartesian becomes translation in Parallel Coordinates.

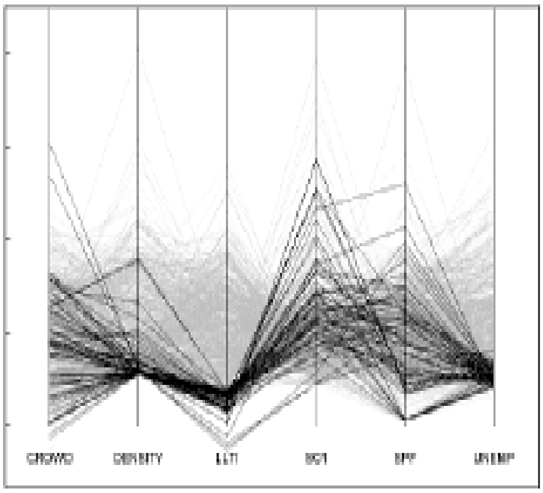

Figure 2 a) Parallel Coordinates with lines [1]

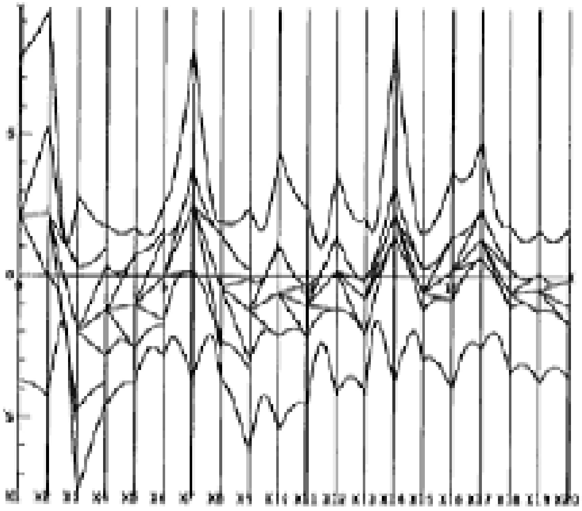

Figure 2 b) Hyperbola in Parallel Coordinates – envelope of its points [1]

Parallel Coordinates representations can provide statistical data interpretations. In the statistical setting, the following interpretations can be made: For highly negative correlated pairs, the dual line segments in Parallel Coordinates tend to cross near a single point between the two Parallel Coordinates axes. Parallel or almost parallel lines between axes indicate positive correlation between variables. In this way, most common objectives to this technique are to represent the dependency on the order to the axes to identify the relations between variables.

But, when using the Parallel Coordinates as tools for EDA growth, researchers use the brushing mechanism and the grand tour which allow fast calculation and application of colors and saturation. Fua [4] proposed hierarchical Parallel Coordinates based on hierarchical clustering to create a multi-revolutionary view of data enabling data exploration at varying levels of details. Also, it "uses data aggregation technique to collapse data into clusters". Novotny [14] uses a binning algorithm based on a k-means clustering approach for creating aggregate Parallel Coordinates visualization [2]. All these approaches give good results for presenting static data, but not for time-varying data. Some improvements of these techniques were made with interaction techniques [1].

Basic approaches for efficient working with Parallel Coordinates are the possibility for supporting logical operations. Parallel Coordinates explorer [2] allows some aggregated poly-lines on the axes by averaging the values of all axes. The example for conventional brushing of cars data by the origin country is shown on Figure 3 and the reduction to one average car for each country manufacturer is shown on the Figure 4.

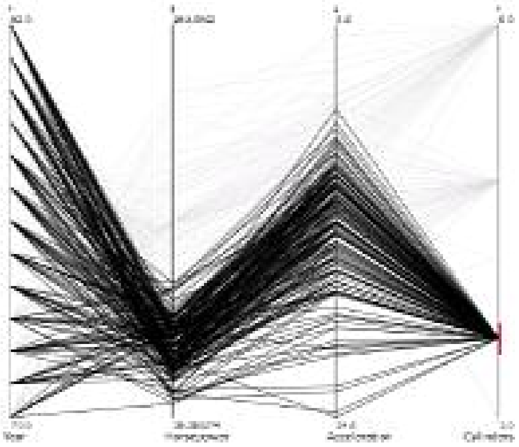

Figure 3 Brushing Parallel Coordinates, all cars with four cylinders are marked and emphases

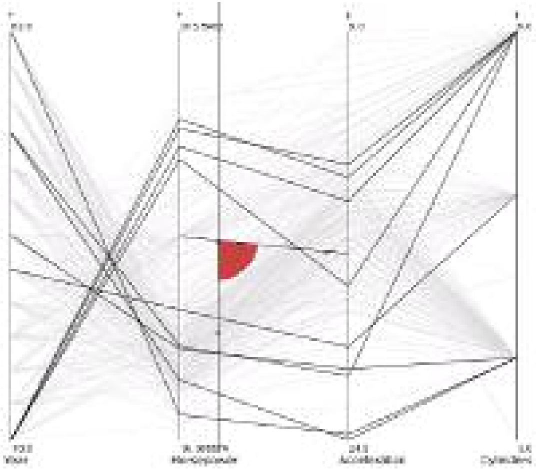

Figure 4 Reading between the lines: visualizing a positive correlation [13]

V. 3D Parallel coordinates

The Parallel Coordinates technique provides the possibility to display and analyze many data dimensions in a Parallel Coordinates system in 2-D plane. To provide spatial information, the parallel axes can be extended into the third dimension. In this case, data lines are ordered according to their real physical position, depending by the problem which is visualized. Data selection can be done in Parallel Coordinates by defining data ranges in different data dimensions using sliders displayed on the axes. Selected data parts are displayed using line color, color groups, or min/max or average lines defined by the according data selection. Some examples of 3D Parallel Coordinates are shown on the Figure 5.

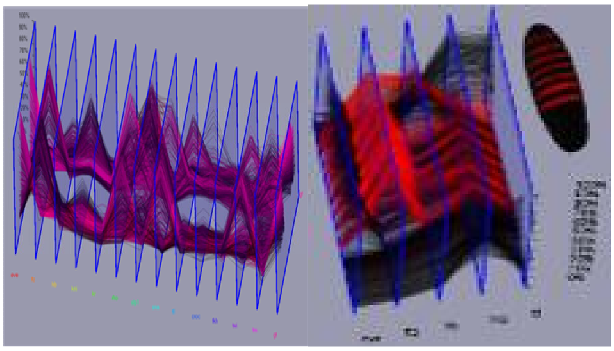

Figure 5 3D Parallel Coordinates visualization examples – Interactive 3D visualization technique [16]

VI. Parallel coordinates as EDA tool

The Parallel Coordinates technique is implemented in many applications, like XmdvTools [8], Matlab [9], GGobi [10], Parallel Coordinate Explorer [11] and many others [16]. These applications enable Cooridinates to be used as a tool for EDA.

Parallel Coordinates as a tool for exploratory data analysis commonly is used for satellite data analysis. For example, the analysist uses query to find patterns for water's edge. Figure 6 shows the visualization with Parallel Coordinates, where the water region due to density differences can be seen. Figure 7 shows the process of finding regions with vegetation [1].

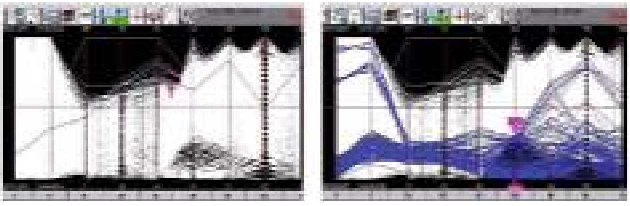

Figure 6 Using Parallel Coordinates with helping queries that select data. On the right figure, the found water regions are shown.

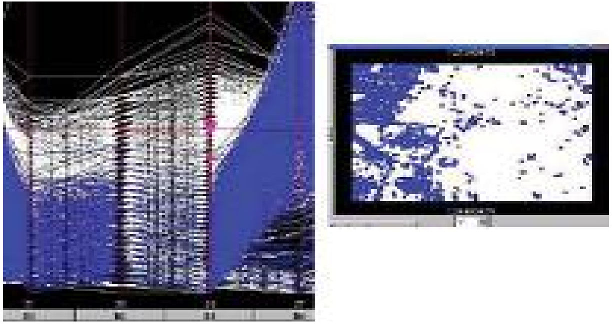

Figure 7 Parallel Coordinates –for finding regions with vegetation

Parallel Coordinates as a tool for exploratory data analysis are also used for analysis of financial data obtained with compound queries. In Figure 8, the gold prices are compared with quotes of other money value (Sterling, Dmark, Yen) by week on Mondays, month and year. Figure 9 represents extraction of the gold prices in 1986, where gold price jumped in the second week of august. The correlation between low Dmark and gold price range is obvious.

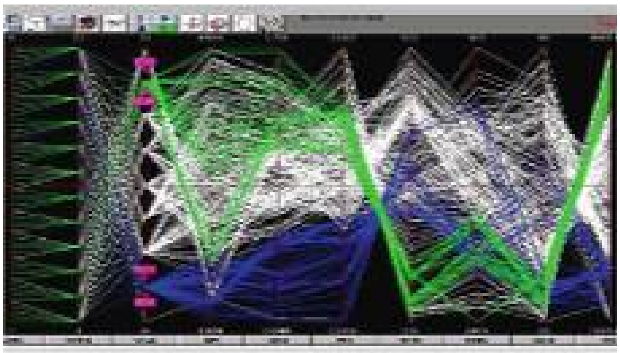

Figure 8 Financial data presented in Parallel Coordinates. Quotes by Week-on Mondays, Month, Year – the first 3 axis fix the date [1]

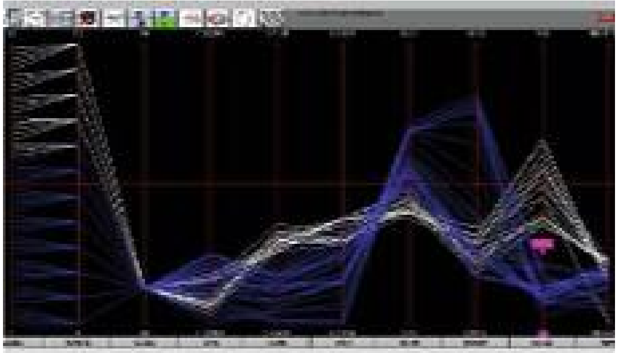

Figure 9 Gold prices in 1986. The correlation between the low Dmark and low Gold price range

We use Matlab [9] for multidimensional data exploration. Figure 10 represents the Parallel Coordinate's technique visualization as exploratory data analysis technique for analyzing local government data [12]. The local governament managers analyse data for quantity of building space for living and working place, depending of urban zone. The conclusions help menagers in the process of decision making for creating infrastucture plan in the city. Data is taken from building data warehouse for three years from city in Macedonia.

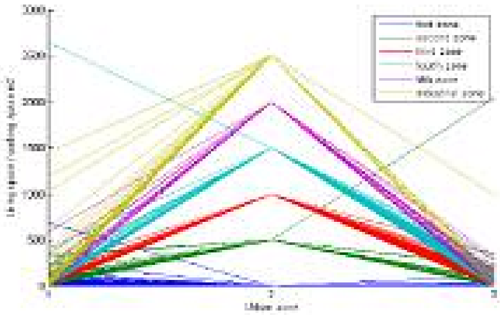

Figure 10 Parallel Coordinates visualization for Local government [12]

Figure 10 represents the Parallel Coordinate's technique visualization as exploratory data analysis technique for analyzing local government data [12]. The figure represents data for quantity of building space for living and working place, depending of urban zone. This representation helps menagers in the process of decision making for creating infrastucture plan in the city. Data is taken from building data warehouse for three years from city in Macedonia.

VII. Conclusion

The effective exploratory data analysis of great volume of multidimensional data demands some specific data visualization techniques for analyzing a complex and huge databases. Because of the possibility to pose many axes in 2-D surface and some interactive technique, Parallel Coordinate is a very useful technique for exploratory data analysis.

References

[1] Zudilova-Seinstra E., Adriaansen T., Van Liere R., Trends in interactive Visualization, Springer, 2009-09-12

[2] Brandst¨atter A., Visualization of Online Sales Databases, Wien, 02-2007

[3] Yang J., Ward M.O., Rundensteiner E.A, Interactive Hierarhical Display: A General Framework for Visualization and Exploration of large Multivariate Data Sets, Worcester, WPI, 15.02.2002

[4] Fua Y.H., Ward M.O., Rundensteiner E.A, Hierarhical parallel Coordinates for Exploration of large Data Sets, 2000 In Proc. of IEEE Visualization '99, pages 43–50.

[5] Rubel O., Prabhat, Wu K., High Performance Multivariate Visual Data Exploration for Extremly Large Data, LB National Laboratory, Berkeley, CA 94720, USA, 2009

[6] Yang D., Rundensteiner E.A., Ward M., Analysis Guided Exploration of Multidimensimal Data, NSF grant IIS-0414380, NSA, 2006

[7] Solka J.L., Marchette J.D., Adams M.L., Applications of Statistical Visualization to Computer security, NSWC, DTD, 2002

[8] Rundensteiner E.,Ward M., Xie Z., Cui Q, Wad C, Yang D, Huang S., XmdvTool: Quality-Aware Interactive Data Exploration, ACM 978-1-59593-686-8/07/0006, (http://davis.wpi.edu/~xmdv/)

[9] Parallel coordinate graphics using MATLAB - http://isomap.stanford.edu/IsomapR1.tar

[10] GGoby (http://www.ggobi.org)

[11] http://www.cs.uta.fi/~hs/pce/

[12] Savoska S., Loskovska S., Dimitrovski I., Information Visualization from the Public Utilities Databases of Local Municipality for Municipalities Managers, Proceedings, Cavtat, 2008

[13] Hauser H. , Ledermann F., Doleisch H., Angular Brushing of Extended Parallel Coordinates, http://www.VRVis.at/vis/

[14] M. Novotn´y, "Visually effective information visualization of large data," in Proceedings of Central European Seminar on Computer Graphics (CESCG), 2004

[15] http://vis.lbl.gov/Events/SC05/Drosophilia/index.html (Drosophila Gene Expression Data Exploration and Visualization)

[16] Parallel Coordinates Proposal, Yisheng Chen & Tianfang Xu;