TUCS

Dorina Marghescu

# Multidimensional Data Visualization Techniques for Financial Performance Data: A Review

Turku Centre *for* Computer Science

TUCS Technical Report
No 810, February 2007

# Multidimensional Data Visualization Techniques for Financial Performance Data: A Review

Dorina Marghescu

Turku Centre for Computer Science,
Åbo Akademi University, Department of Information Technologies,
Institute for Advanced Management Systems Research
Joukahaisenkatu 3-5 B, 20520 Turku, Finland
dorina.marghescu@abo.fi

TUCS Technical Report
No 810, February 2007

# Abstract

In this paper, we review 11 visualization techniques that can be used for visual exploration of multidimensional financial data. We illustrate the use of these techniques in studying the financial performance of companies from the pulp and paper industry. We also illustrate the use of visualization techniques for detecting outliers, and other patterns in financial performance data in the form of clusters, relationships, and trends. We highlight the potential benefits of using multiple visualization techniques to get insight into financial performance data.

**Keywords**: multidimensional data visualization techniques; visual data mining; financial performance; financial data visualization; multiple visualization techniques

**TUCS Laboratory**
Data Mining and Knowledge Management Laboratory

## 1. Introduction

The large amounts of high-dimensional business data that exist in organisations' databases require automated tools for data pre-processing and analysis. Data mining algorithms and techniques represent one class of such tools and they are used to automatically extract knowledge from the data. The discovered knowledge is referred to as patterns found in data, patterns that must be interesting (novel, valid, potentially useful and understandable) to the user [1]. These patterns are typically represented in the form of clusters, classes, trends, relationships, and summaries of the original data. However, in most of the cases, these data mining results are communicated to the business user in a format that is difficult to understand and/or interpret.

To overcome this problem of representing business data and data mining results in an accessible format, researchers investigate the possibilities and limits of using *information visualization* techniques for this purpose. Information visualization is defined as the use of computer-supported, interactive, visual representations of abstract (non-physical) data in order to amplify cognition [2]. Information visualization has different areas of application. One, which is the subject of our concern in this paper, is knowledge discovery. Card et al. refers to the techniques belonging to this category with the term *visual knowledge tools* [2]. Keim name these techniques as *visual data mining tools* [3]. These tools employ information visualization techniques in order to represent graphically large amounts of high-dimensional data and to involve the user in effectively and efficiently exploring data at different levels of detail. The users of these tools are capable of finding outliers and anomalies and detecting patterns and models (in the form of clusters, classes, trends, and relationships) in different categories of data (e.g., financial data, business information, document collections, etc.).

Visual data mining tools are used in two different situations. First, they serve as visual data exploration tools. This means that these tools represent the original data graphically so that the user can employ his perceptual abilities to discover interesting patterns in the data. Second, visual data mining tools represent graphically the information obtained after applying an automated data mining algorithm or technique. This information is represented by patterns or models obtained after performing a data mining task such as clustering, classification, estimation, dependency analysis, etc.

Recently, novel and interesting visualization techniques have flourished in the field of information visualization and visual data mining. However, the research literature concerning the use of information visualization and visual data mining to get insight into *financial data* is relatively sparse, despite the fact that this technological approach is suitable for both financial data and business users. Financial data is very complex due to its high dimensionality, large volume and diversity of data types. Business users are demanding straightforward visualizations and task-relevant outputs, due to the time and performance constraints under which they work [4].

Most research focuses on developing new information visualization techniques, and exploring their capabilities on financial datasets. Wright illustrated the application of animated visualization techniques to the securities industry (equity trading analytics, fixed-income risk management) [5]. Feiner and Beshers developed a system, called n-Vision, which uses the interaction metaphor "worlds within worlds" especially developed for the system, in order to support the exploration of high-dimensional data [6]. They illustrated the capabilities of their system on a financial dataset representing European options whose values are determined by functions of six variables. Ankerst et al. [7] and Keim [8] introduced and developed pixel-oriented techniques for exploring huge amounts of data, and they illustrated the performances of these techniques on financial data representing stock prices.

The most popular commercial tool used by business users for visualizing financial data is Microsoft Excel, which provides the standard visualization tools (bar, column, line, scatter plot, pie, stacked bars, etc.). However, Excel lacks the capabilities of interacting with and linking different visualizations and does not implement some of the recently developed visualization techniques for *multidimensional* data. To overcome these limitations, one can use the available add-in applications or linked applications to Excel that provide more sophisticated visualization tools (e.g., Excel Dashboards, Crystal Xcelsius, etc.). A new approach to financial visualization is realized by SmartMoney, which provides web-supported visualizations to explore high-dimensional financial data (e.g., Map of the market, Map of the portfolio – tools based on *treemap* concept [9]).

In this paper, we describe 11 visualization techniques suitable for visualizing multidimensional data and illustrate them on financial performance data. We focus on those visualization techniques that are suitable for representing graphically *table data*, that is, datasets that are expressed as data tables in which the rows represent cases or records and the columns represent attributes, dimensions of data. Hoffman and Grinstein use the term "table visualizations" to refer to this class of visualization techniques [10]. Moreover, we focus on table visualization techniques that are capable of displaying multidimensional or multivariate data. These are referred to in the literature as *multidimensional data visualizations* [10] or *multidimensional visualizations* [11].

The *research question* that has driven our work was to determine to which extent the use of *multiple* visualization techniques has advantages over the use of a single technique in exploring and extracting knowledge from data. To answer this question, we apply several visualization techniques to studying the problem of financial benchmarking. The financial benchmarking problem focuses on comparing competing companies with respect to their financial performance. The problem is complex because many variables are involved. First, we formulate the problem in terms of business questions and associated data mining tasks. Then we investigate the capabilities of each visualization technique in helping the user to solve the derived data mining tasks and identify interesting patterns in data.

Moreover we analyse the visualization techniques from different perspectives such as the capability to visualize data items or data models, and the type of data processed (i.e., original data or normalized data).

The results of these analyses show that the use of multiple visualization techniques for understanding multidimensional financial data and solving a business problem may have many benefits.

The problem of integrating or combining multiple methods for visualizing multidimensional data is also explored by Ward [12] and Kreuseler et al. [13] but from a system design perspective. In this paper, we do not address the interactive capabilities that each technique possesses or it is desirable to possess, but its capabilities in uncovering interesting patterns in data.

The paper is organised as follows. In the next section, we outline the problem of financial benchmarking and describe the dataset used for the illustration of the visualization techniques. In section 3, we describe different visualization techniques and present how various visualization techniques help in exploratory data analysis for our problem. Section 4 summarizes and discusses the various techniques used. We conclude with final remarks and future work ideas.

## 2. The problem of financial benchmarking

One of the problems that business intelligence people are confronted with nowadays is performing comparisons of companies' financial performance. The scope of the comparisons can be restricted depending on the objective of the study. For example, analysts might be interested in comparing companies of a particular profile, which operate in one country. Others might broaden the analysis to include the companies of the same industry from particular countries, one continent, or all continents. This problem of comparing financial performance of companies is known as *financial competitor benchmarking* [14]. The problem is non-trivial since many variables (financial ratios) must be considered. One part of the problem is choosing the ratios to be used when describing the financial performance of a company. Eklund proposed a model for financial competitor benchmarking, in the pulp and paper industry, with seven financial ratios as a basis for companies' performance comparison and the Self-Organizing Maps (SOM) as the method for data analysis [14, 15]. In this paper, we build on the mentioned research to explore the use of multiple visualization techniques for getting insight into financial data.

### 2.1.Illustrative dataset

The dataset analysed in this paper is a subset of a larger dataset whose collection process including variable and company selection, together with the use of SOMs for financial benchmarking are described by Eklund [14]. The data values are entirely based on the information obtained from companies' financial reports available on the Internet.

The data in this study refer to 80 companies that function in the pulp and paper industry worldwide, observed during 1997 and 1998. A total of 160 observations are analysed. The dataset contains seven numerical variables, namely seven ratios that characterize the financial performance of companies in the pulp and paper industry. The ratios are grouped in four categories: *profitability* (**O**perating **M**argin, **R**eturn **o**n **E**quity, and **R**eturn **o**n **T**otal **A**ssets), *solvency* (**I**nterest **C**overage, **E**quity to **C**apital), *liquidity* (**Q**uick **R**atio), and *efficiency* (**R**eceivables **T**urnover). In the remainder of the paper, we use acronyms when referring to any of the financial ratios (that is, OM, ROE, ROTA, IC, EC, QR, and RT respectively). Besides the numerical variables, the dataset contains three categorical variables: companies' names, regions (Europe, Northern Europe, USA, Canada and Japan), and year (1997 or 1998). The choice of this particular dataset was driven by two factors. First, there was the availability of the dataset, and second, its suitability for data mining (e.g., cluster detection, cluster characterization, class characterization, outlier detection, and dependency analysis) in order to study financial performance of companies. The size of the dataset is relatively small (160 observations, 7 numerical variables, 3 categorical variables), but appropriate for our illustrative purposes.

## 2.2. Business questions and data mining tasks

In order to use information visualization in solving a business problem, this should be translated in terms of business questions and further in visualization or data mining tasks [11]. For the problem of financial benchmarking we have derived the business questions and data mining tasks as follows:

a) Outlier detection: Does the data present outliers or anomalies? Are there any companies that present unusual values of financial ratios?
b) Dependency analysis: Are there any relationships between variables?
c) Data clustering: Are there clusters (groups of companies with similar financial performance) in the data? How many clusters do exist?
d) Cluster description: What are the characteristics of each cluster?
e) Class description: Are there any relationships (common features) among companies located in one region or another? What are these common features?
f) Comparison of data items: Compare two or more companies with respect to their financial performance.

For the task f), we have chosen three companies to be compared as to their financial performance in 1998: Reno de Medici, Buckeye Technologies, and Donohue. For Reno de Medici we look also at its evolution from 1997 to 1998. These companies are identified on the graphs using the letters A, B, C, and D, respectively. Table 1 presents the financial ratios of these companies.

**Table 1. Financial ratios of the companies chosen for comparison**

| ***Company*** | ***Reno de Medici 1997*** | ***Reno de Medici 1998*** | ***Buckeye technologies 1998*** | ***Donohue 1998*** |
|---|---|---|---|---|
| *Id.* | A | B | C | D |
| *Year* | 1997 | 1998 | 1998 | 1998 |
| *Region* | Europe | Europe | USA | Canada |
| *OM* | 4.02 | 6.7 | 19.42 | 21.24 |
| *ROE* | -15.38 | 5.34 | 38.96 | 17.96 |
| *ROTA* | 0.64 | 5.27 | 16.21 | 15.92 |
| *EC* | 27.94 | 28.19 | 20.91 | 46.35 |
| *QR* | 1.29 | 1.03 | 1.36 | 0.91 |
| *IC* | 0.15 | 1.68 | 3.28 | 5.15 |
| *RT* | 3.3 | 2.63 | 7.79 | 7.96 |

## 3. Visualization techniques

We have selected the visualization techniques based on Hoffman and Grinstein's survey [10]. The presentation of the techniques follows the classification of visualization techniques proposed in [3]:

- Standard 2D/3D visualizations,
- Geometrically transformed displays,
- Icon-based displays, and
- Stacked displays.

Since our dataset is relatively small, we did not use the dense-pixel display techniques, despite their effectiveness for visualizing huge amounts of data. We have considered separately the *projection techniques*, due to their capability to map data from original high-dimensional space to a transformed low-dimensional space. In addition, we have considered, also separately, one *clustering technique*, that is hierarchical clustering, whose output can be easily displayed graphically using the dendrogram technique.

Table 2 presents the visualizations techniques illustrated in our paper and the tools used for creating the visualizations.

**Table 2 Visualization techniques and tools used**

| *Visualization classes* | *Visualization techniques* | *Tools* |
|---|---|---|
| *Standard 2D/3D and variations* | Multiple line graphs | Matlab [16] |
| | Permutation matrix | Visulab [17] |
| | Survey plot | Orange [18] |
| *Geometrically transformed displays* | Scatter plot matrix | Visulab [17] |
| | Parallel coordinates | Visulab [17] |
| *Icon-based displays* | Star-glyphs | XmdvTool [12] |
| *Stacked displays* | Treemap | Treemap 4.1 [19] |
| *Projection techniques* | Principal Components Analysis | Statistics Toolbox for Matlab [20] |
| | Sammon's Mapping | SOM Toolbox for Matlab [21] |
| | Self-Organizing Map | SOM Toolbox for Matlab [21] and Nenet [22] |
| *Clustering techniques* | Dendrogram | Statistics Toolbox for Matlab [20] |

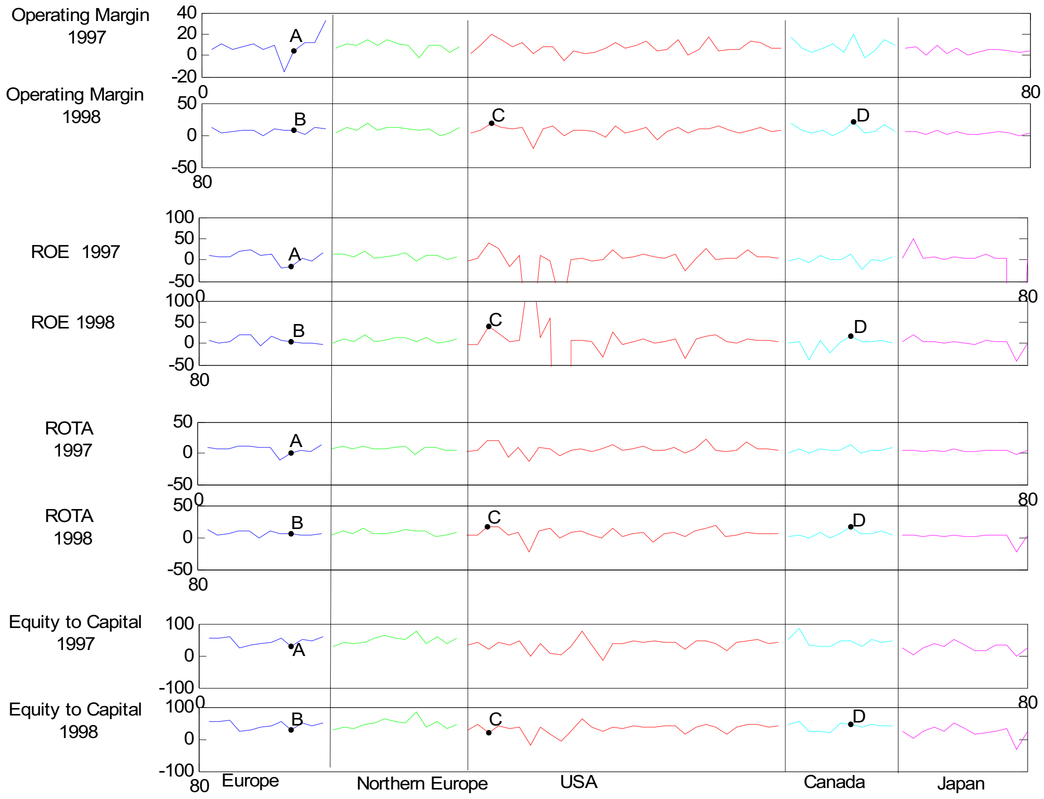
### 3.1. Standard visualizations

This class of visualization techniques comprises the most popular techniques, which are very effective in presenting two- or three-dimensional data. Among the techniques that belong to this category are x-y (-z) plots, line graphs, bar and column charts, area, stacked bar and column graphs, histograms, pie charts, doughnut charts, box plots, radar graphs, Pareto graphs, etc. However, in this section and in the next one, we will focus on those techniques that are suitable for displaying multidimensional data.

## *Multiple line graphs*

Line graphs are used for one dimensional data. On the horizontal axis (Ox) the values are not repeated (e.g., time or the ordering of the table). The vertical axis (Oy) shows the values of the variable of interest. *Multiple line graphs* can be used to show more than two variables or dimensions (x, y1, y2, y3, etc.).

Figure 1 shows line graphs for four ratios (OM, ROE, ROTA, and EC), observed in 1997 and 1998. The companies are mapped to the horizontal axis, in the order of appearance in the data table. The graph presents companies from different regions (Europe, Northern Europe, USA, Canada and Japan) with different colours, this facilitating the characterization of companies from one region or another. By positioning the two years data one under the other, the user can follow the evolution of some company's financial ratios, and make comparisons between companies' financial states.



A – Reno de Medici 1997
B – Reno de Medici 1998
C – Buckeye Technologies 1998
D – Donohue 19998

**Figure 1 Multiple line graphs created with Matlab**

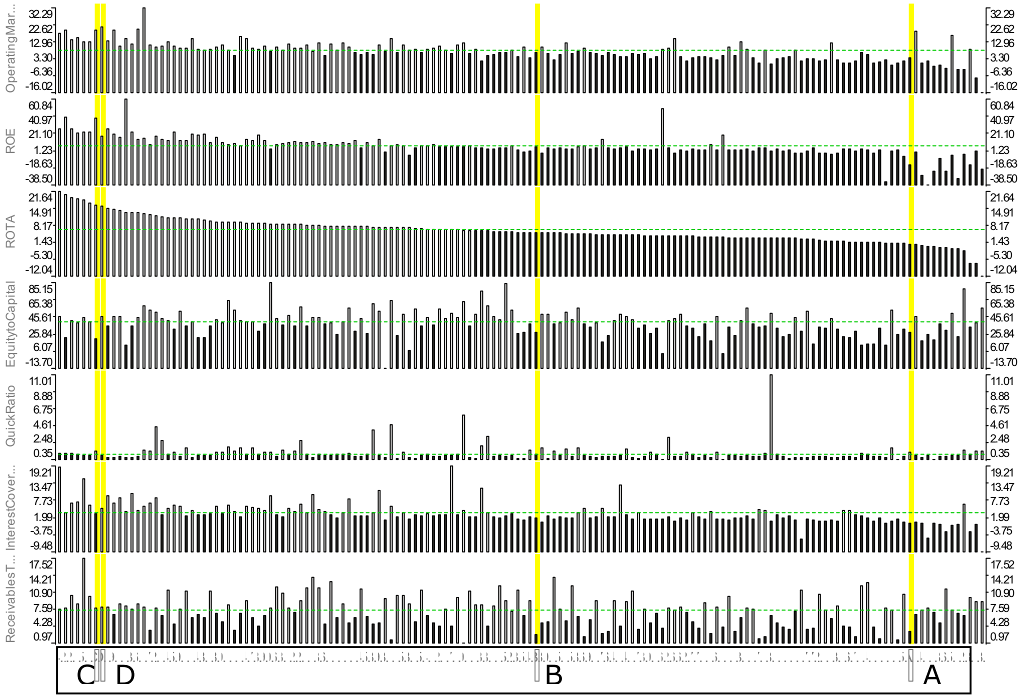
Another use of this type of graph is in showing outliers or anomalies in the data, for example, the very low and very high values of ROE for three of the companies (Crown Vantage 1997: -229.14 and 1998: +202; Gaylord Container 1997: -109.99 and 1998: -1279.06; and Settsu 1997: -4932.18). These companies were removed from the data set. By highlighting the companies to be compared, one can easily see the differences and similarities among them. Company Buckeye Technologies (C) is, in 1998, among the most profitable in USA and in the world. The same thing is illustrated in the case of Donohue 1998 (D), this company being among the most profitable in Canada and in the world. It is also possible to see the increase of ROE and ROTA in the case of Reno de Medici in 1998 (B) compared with 1997 (A).

## 3.2. Variations of standard visualizations

### *Permutation matrix*

The permutation matrix is a special type of bar graphs described in [23]. One bar graph can represent one dimensional data so that the heights of the bars represent the data values. In a permutation matrix, for each data dimension there is built a bar graph. The horizontal axes of all bar graphs have the same information (e.g., the time or ordering of the data table). In a permutation matrix all data values below average are coloured black, and all data values above average are coloured white. A green dashed line plotted over the data represents the average value of each dimension. Implementations of permutation matrix allow to interactively changing the order of the observations for observing interesting patterns.

Figure 2 displays a permutation matrix created with Visulab [17], which has an automatic permutation mode for identifying patterns. The representation in Figure 2 is based on the sorting of the data in descending order according to the variable ROTA. The companies of interest are highlighted.



A – Reno de Medici 1997
B – Reno de Medici 1998
C – Buckeye Technologies 1998
D – Donohue 19998

**Figure 2 Permutation matrix created with Visulab**

One can easily see patterns in terms of relationships between ratios (companies with above average values of ROTA have high values for all other profitability ratios, and in most of the cases, good liquidity, solvency, and efficiency). The bar graphs also reveal “anomalies” or extreme values in the data. The user can easily compare two or more companies with respect to their financial performance. The graph displays low ratios for Reno de Medici, in 1997 (A). In 1998, the profitability of the company improved considerably (B), but it is still under the average. Buckeye Technologies (C) and Donohue (D) are in 1998 among the most profitable companies.

## *Survey plot*

The survey plot is a variation of the permutation matrix. The values of each data dimension are represented as horizontal bars. The width of the bars is proportional with the data values. The bars are centred and there are no spaces separating the bars. One can use colours to distinguish between different classes in the data (if a class variable is present).

Figure 3 displays a survey plot, in which the data is sorted along the ROTA dimension. This facilitates the detection of relationships between ROTA and other ratios, for example OM, ROE and IC.

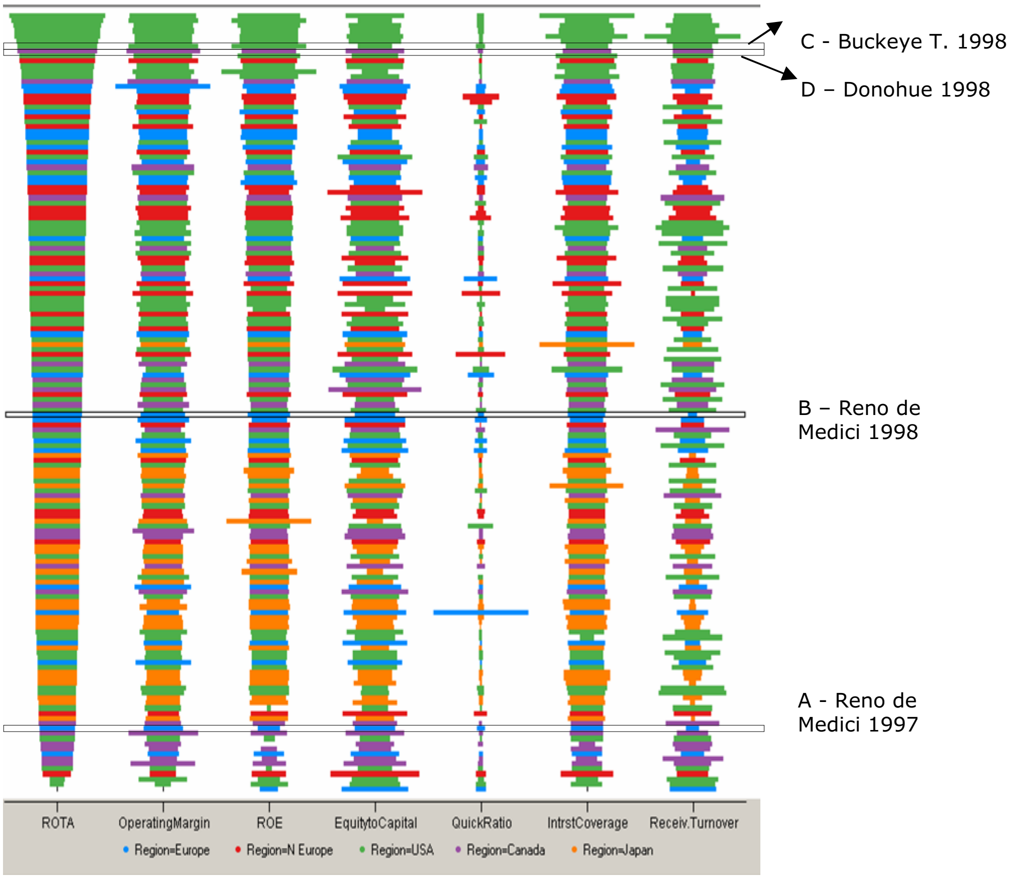


**Figure 3 Survey plot created with Orange**

Companies from different regions are displayed with different colours. The graph shows that the Japanese companies are not among the most profitable ones, while the American and European companies display the highest profitability. The technique facilitates the detection of outliers and comparison between two or more companies.

## 3.3. Geometrically transformed displays

According to Keim, these techniques aim at finding "interesting" transformations of multidimensional data sets [3]. The class of geometric display techniques includes techniques from exploratory statistics, such as scatterplots matrices, projection techniques and clustering techniques. Parallel Coordinates technique also belongs to this class. We illustrate three projection techniques separately in Section 3.6 and one clustering technique in Section 3.7.

## *Scatter-plot matrix*

A scatter plot is used to plot two dimensional data so that the horizontal axis shows the values of one variable and the vertical axis shows the values of another variable. The scatter-plot matrix is useful for looking at all possible pairs of variables in the dataset.

Figure 4 displays a scatter plot matrix for the financial ratios of the companies. The plots clearly reveal relationships between the profitability ratios ROE and ROTA, OM and ROTA, ROE and OM and between IC and ROE, ROTA, and OM.

Figure 4 reveals also outliers or extreme values, and makes possible comparison between companies. We have highlighted the companies of interest with different colours. It is clear the position of Buckeye Technologies 1998 (C) and Donohue 1998 (D) among the most profitable companies and the positive evolution of Reno de Medici in 1998 (B).

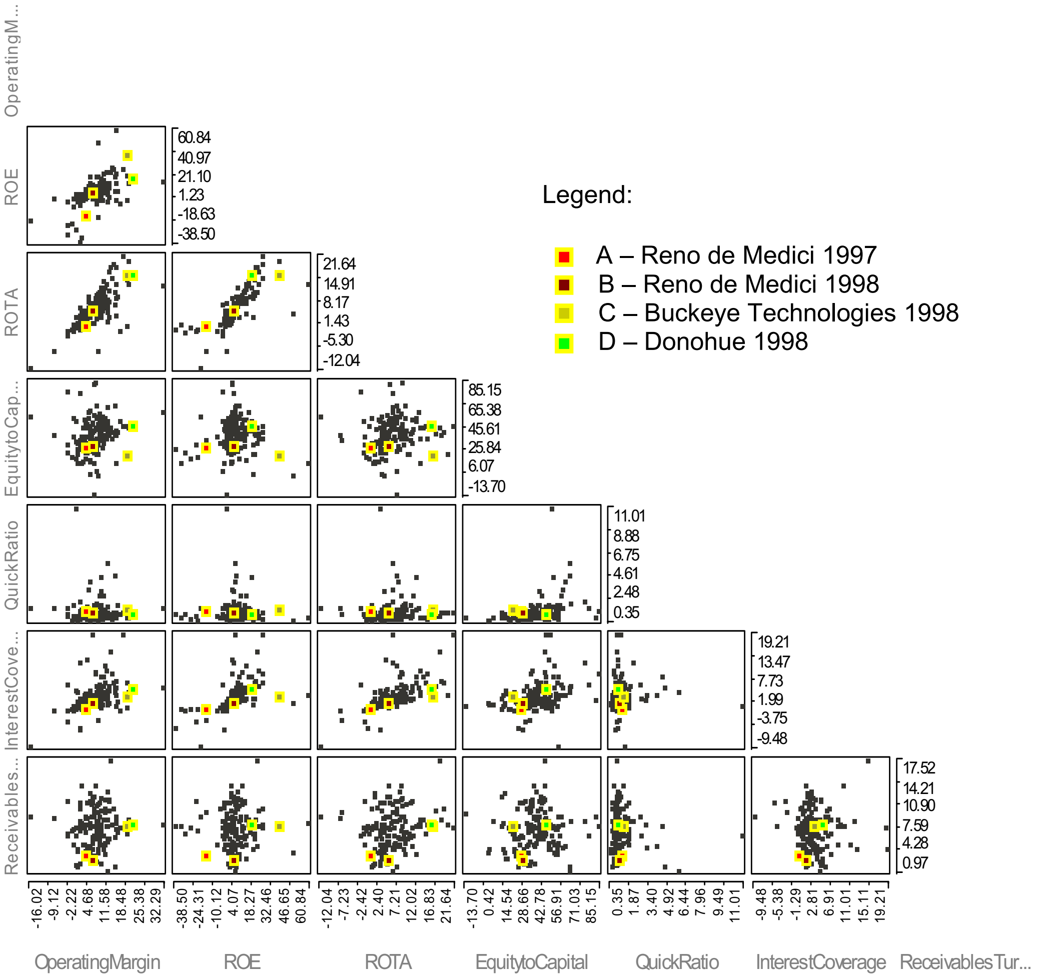


**Figure 4 Scatter-plot matrix created with Visulab**

## *Parallel coordinates*

Introduced by Inselberg [24], parallel coordinates represent multidimensional data using lines. The data dimensions are represented as parallel axes (coordinates). The maximum and minimum values of each dimension are scaled to the upper and lower points on a vertical axis. An n-dimensional data point is displayed as a *polyline* that crosses each axis at a position proportional to its value for that dimension.

Figure 5 represents the financial ratios as parallel axes and each company as a polyline that crosses each axis at a point proportional with the value of the company for that ratio. The companies of interest are highlighted with different colours.

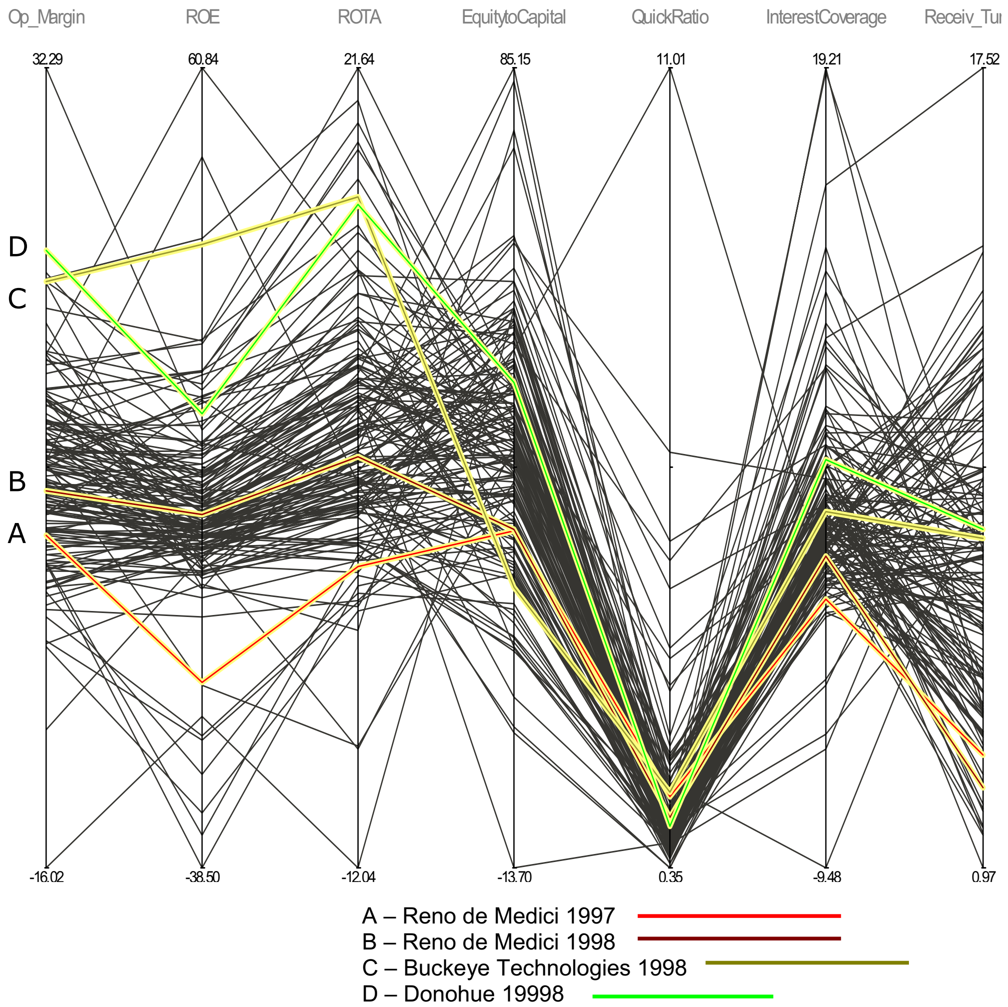


**Figure 5 Parallel coordinates created with Visulab**

The display facilitates detection and characterization of outliers, as well as comparison between companies and detection of relationships among variables. The user can observe the increase in the profitability ratios in the case of the company Reno de Medici in 1998. Moreover, one can analyse the differences and similarities between Buckeye Technologies 1998 (C) and Donohue 1998 (D). The relationships between two or more variables can be detected if the correlated variables are arranged consecutively (for example, ROE and ROTA). Therefore, a challenge to designers is to choose the optimum arrangement of the axes or to allow the users to change the order of dimensions interactively.

## 3.4.Icon-based displays

The idea of icon-based techniques is to map the attribute values of a multidimensional data item to the features (colour, shape, etc.) of an icon. Keim [3] distinguishes among different types of icons used to visualize high-dimensional data items: little faces (Chernoff faces), needle icons, star-icons or star-glyphs, stick figure icons, colour icons, and Tilebars. These displays are especially effective when the data items are relatively dense with respect to the two display dimensions and consequently, the resulting visualization presents texture patterns that vary according to the characteristics of the data.

### *Star-Glyphs*

In a star-glyph, a data item is represented by a glyph (symbol or icon) consisting of n lines emanating from a point at uniformly separated angles [12]. The number of lines, n, is equal to the number of data dimensions. The lengths of the lines are proportional to the values of each data dimension. The end-points of the lines for a given glyph are connected to form a polygon in order to reduce misinterpretation when glyphs overlap. The problem in using the star glyphs is to organise all the star glyphs on the screen in a meaningful manner. One can use two of the dimensions of the data to control their position, or the glyphs can be ordered according to one dimension, based on user's decision.

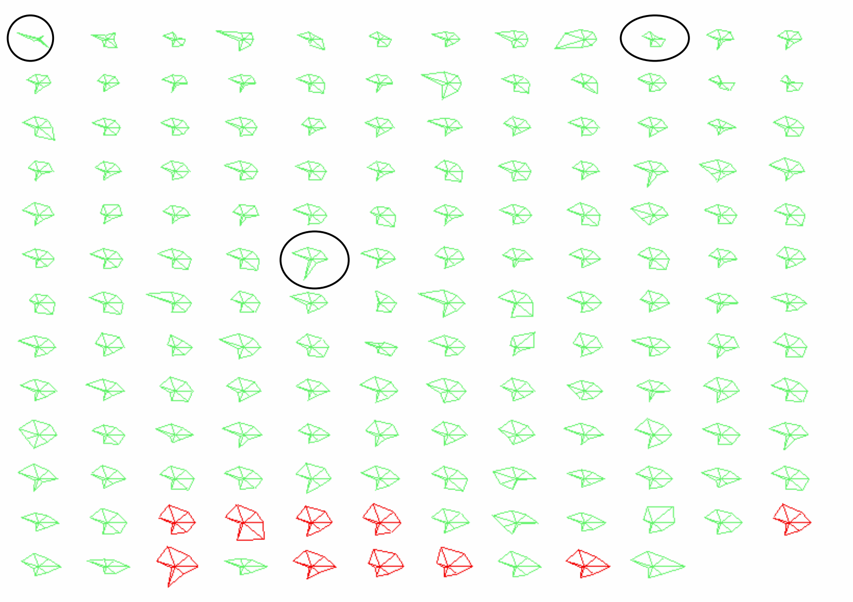


**Figure 6 Star-glyphs created with XmdvTool**

Figure 6 displays the data sorted according to OM and highlights with red the companies with ROTA higher than 14. The visual presentation can reveal relationships between financial ratios, for example between OM and ROTA (the companies with the highest ROTA are among the companies with high OM). The graph can also reveal multivariate outliers, represented by star-glyphs whose shape does not resemble the shapes of its neighbours. We have circled on the graph some of the outliers we have identified. If labels are attached to the glyphs, comparisons between companies are possible.

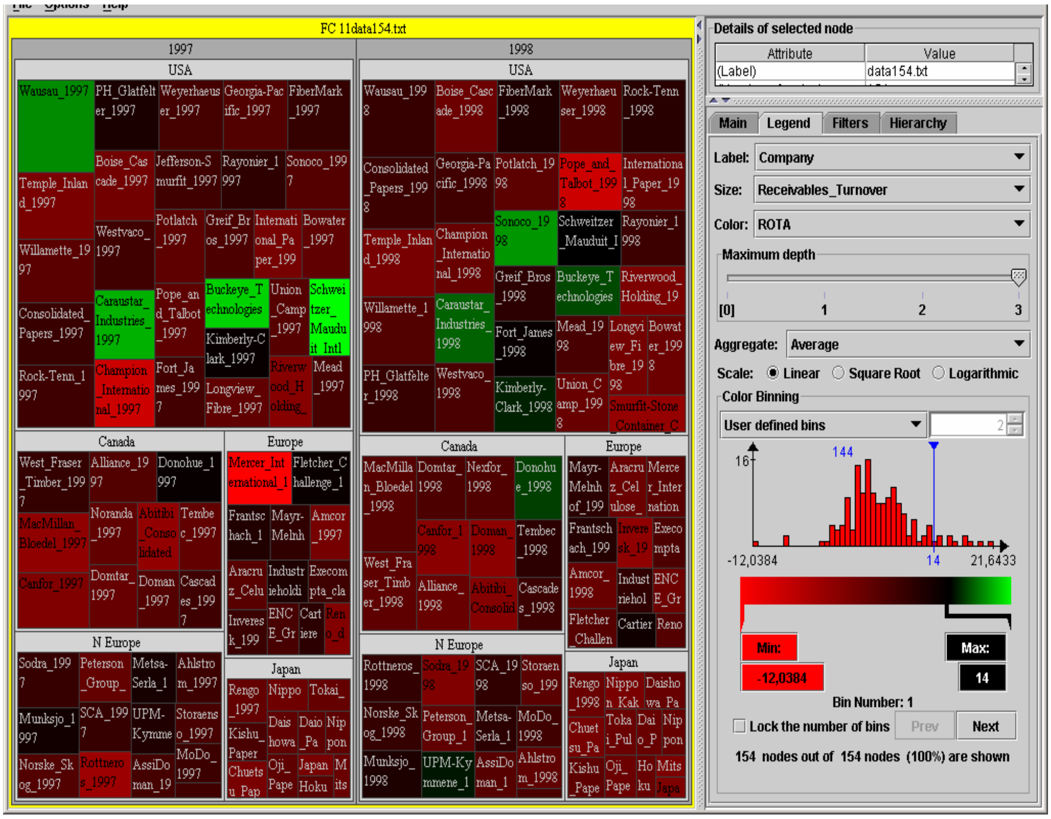
## 3.5.Stacked displays

These techniques are suitable for partitioning the data in a hierarchical fashion [3]. The partitioning is done based on the information contained in the data dimensions. The techniques are especially effective for presenting hierarchical data. However, interesting patterns can be found with non-hierarchical data as well.

### *Treemaps*

The treemaps [9] are hierarchical visualizations of multidimensional data. Data dimensions are mapped to the size, position, colour, and label of nested rectangles.

Figure 7 displays the dataset with treemaps technique. The figure was created with Treemap 4.1 tool [19]. Each company is represented by a rectangle. The size of the rectangle indicates the value of Receivables Turnover ratio. The colour of the rectangle indicates the value of ROTA ratio as follows: light green shows high values of ROTA; light red shows small values of ROTA, dark red and dark green shows values of ROTA close to 14 (see "colour binning" panel in the visualization below). The dataset is organised in this visualization in categories such as year and region.



Details of selected node

| Attribute | Value |
|---|---|
| (Label) | data154.txt |

Main | Legend | Filters | Hierarchy

Label: Company

Size: Receivables_Turnover

Color: ROTA

Maximum depth [0] 1 2 3

Aggregate: Average

Scale: ● Linear ○ Square Root ○ Logarithmic

Color Binning

User defined bins 2

Min: -12,0384

Max: 14

Bin Number: 1

□ Lock the number of bins Prev Next

154 nodes out of 154 nodes (100%) are shown

**Figure 7 Treemap created with Treemap 4.1**

This treemap representation allows the user to see where the most profitable companies in terms of ROTA are being located, and how the companies of interest have evolved in time. In addition, one can identify common features or patterns of the industry, for example, that Japanese companies have the lowest values of efficiency ratio. The user can also compare the financial performance of different companies. The companies Buckeye Technologies 1998 and Donohue 1998 are mapped to green coloured rectangles that show high ROTA, in contrast with Reno de Medici 1998, which shows low value of ROTA. Moreover, the later has lower efficiency than its competitors.

## 3.6.Projection techniques

Projection techniques are used to reduce the variable space. The dimensionality reduction is obtained by combining the original variables into a smaller number of new variables, in a linear or nonlinear manner. The projection methods are particularly useful because they lend themselves to visual representations of data, when the number of new dimensions is one, two or three.

### *Principal component analysis (PCA)*

PCA is a dimensionality-reducing technique employing *linear transformation of data* [25]. The projection of high-dimensional space onto a lower-dimensional space tries to preserve the variance of the original data as well as possible. The PCA technique creates new variables (called principal components), which are linear composites of the original variables and are uncorrelated among them. The maximum number of new variables that can be formed is equal to the number of original variables. The PCA output is judged in terms of how well the new variables represent the information contained in data, or, geometrically, how well the new dimensions can capture the original configuration of the data.

Figure 8 shows PCA plot that was constructed from the standardized dataset. The red dot shows the observation closest to the centre of the dataset. The companies of interest are marked with a yellow star and labelled on the graph.

One can interpret the principal components by inspecting the loadings of each original variable to the PCs. The higher the loading of a variable, the more influence it has in forming the PC score and vice versa. In our case, the first PC (horizontal axis) is highly correlated with the profitability ratios and the IC ratio. Therefore, companies placed towards the right of the horizontal axis, have high values for the profitability and IC. The second PC (vertical axis) is highly correlated with QR and EC. Companies located on the upper part of the graph have high liquidity and high solvency with respect to EC. The amount of variation explained by the two PCs is 40.926% + 19.455%= 60.38% of the total variance. While this amount of variance accounts for the variation of six of the ratios, it does not consider the variation of efficiency (RT) among companies.

If we want to minimize the loss of information resulted from dimensionality reduction, we have to include the third principal component. This is highly correlated with RT (0.933), and it explains 14.783 % of the total variation. The three PCs will represent the dataset in a 3D-space with a minimum loss of information generated by the data reduction.

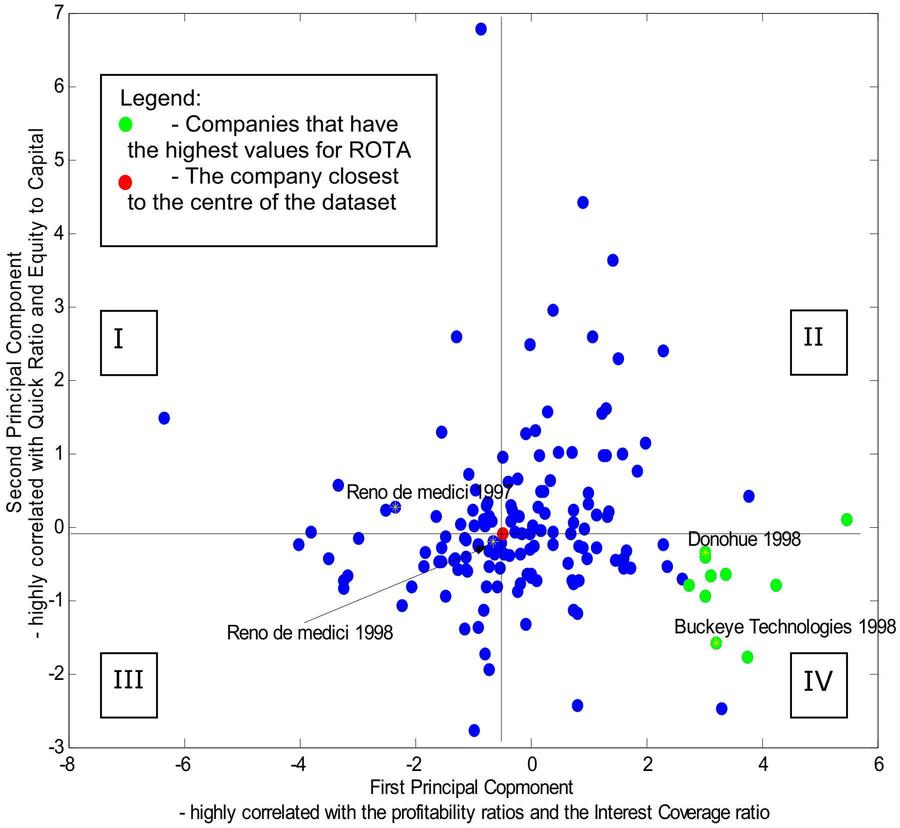


**Figure 8 Data projected on the first two PCs created with Statistics Toolbox for Matlab. In area I: medium-high liquidity, low-medium profitability; II: medium-high liquidity, solvency and profitability; III: low-medium liquidity, solvency and profitability; IV: low-medium liquidity, medium-high profitability**

Besides its usefulness as a data reduction method, the PCA is also useful in finding numerous patterns in data (Figure 8). The graph shows the high profitability of Buckeye Technologies 1998 and Donohue 1998, and the increase in profitability for Reno and Medici in 1998. The green dots highlight the companies with highest ROTA. The high correlation of the first PC with all profitability ratios and with IC ratio indicates that there exist also relationships between profitability ratios and IC. Similarly, the high correlation of the second PC with EC and QR indicates that EC and QR are also correlated.

By splitting the visual representation in four areas by two orthogonal lines that intersect in the centre of the dataset, one can divide the dataset into four groups of similar observations as shown and described in Figure 8. Based on the meaning of the first two PCs, one can evaluate that in area I there are situated companies with medium-high liquidity and low-medium profitability; in area II, companies with medium-high liquidity, solvency and profitability; in area III, companies with low-medium liquidity, solvency and profitability; and in area IV, companies with low-medium liquidity but medium-high profitability. Based on this evaluation, one can compare the financial performance of the companies of interest.

### *Sammon's mapping*

The Sammon's mapping is a *nonlinear projection* of the multidimensional data down to two dimensions so that the distances between data points are preserved [26]. It belongs to multidimensional scaling techniques.

Figure 9 illustrates the Sammon's mapping applied to our financial dataset. The data values were normalized using discrete histogram equalization method. The normalization method works in two steps: first, the data values of each variable are replaced by the order index, and then these values are normalized to be in the range [0, 1], by applying a linear transformation. Companies from different regions are displayed with different colours. The companies of interest are marked with yellow stars and labelled on the graph.

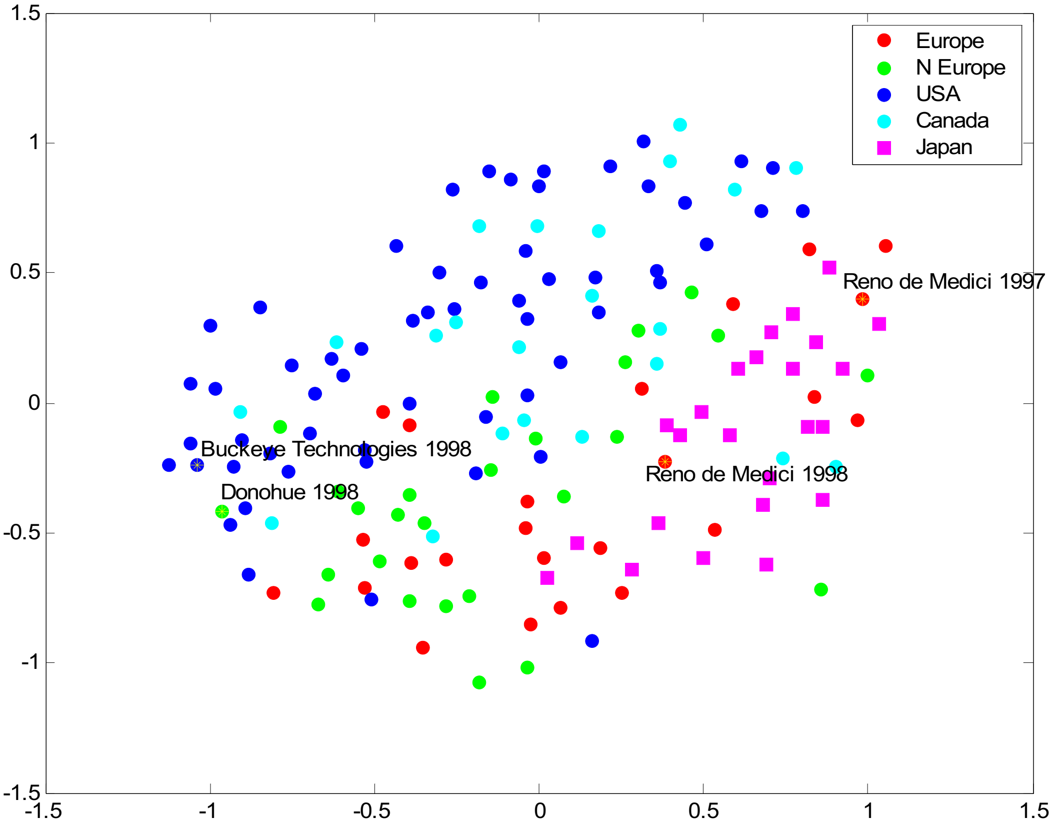


**Figure 9 Sammon's mapping created with SOM Toolbox for Matlab**

The technique is useful in visualizing class distributions, especially the degree of their overlap. One can see that companies from Canada and USA overlap and map to the same area of the graph, whereas Japan, Europe and Northern Europe form three different groups. However, the degree of overlapping between all these classes is quite high; especially Europe and Northern Europe do not separate well from the other groups. The differences and similarities between the companies are easy to distinguish, but not easy to interpret.

### *Self-Organizing Maps (SOM)*

The SOM technique, developed by Kohonen [26] is a special type of neural network based on unsupervised learning. The SOM algorithm is similar to K-Means clustering algorithm, but the output of a SOM is topological and neighbouring clusters are similar. As a projection technique of multidimensional data onto a two-dimensional grid, the SOM method is similar to multidimensional scaling techniques, such as Sammon's mapping. The grid consists of units that have assigned reference vectors with the same dimensionality as the original data. After learning is complete, the reference vectors are updated such that they resemble most of the data items, as much as possible. Each data item is then mapped to the unit where the highest similarity between the reference vector and the data item is calculated. Multiple data items mapped onto the same unit are similar and form a cluster. There are many ways to represent the SOM output.

One way to represent the data is to use the scatter plot technique (usually with jittering), in which the horizontal and vertical axes are produced by the Kohonen network (i.e., the map size). Figure 10 is a scatter plot of the dataset based on the SOM coordinates.

Figure 10 shows the companies as they are mapped to the units of the SOM grid. The size of the map is given by multiplying the 6 units on the horizontal with the 5 units on the vertical axis, in total 30 units. Companies from different regions are highlighted by using different colours. The data values were normalized using discrete histogram equalization method. The companies of interest are highlighted with yellow stars. The technique of jittering was used in order to change with a small value the position of each company; otherwise the companies mapped to the same unit would have overlapped.

Figure 10 allows the user to identify many clusters in the data (if more companies are mapped to the same map unit, they may be interpreted as forming a cluster). However, the interpretability of this map is not easy. One can distinguish among the companies belonging to the same region, or identify the placement of these companies on the map but cannot interpret these classes or the clusters formed.

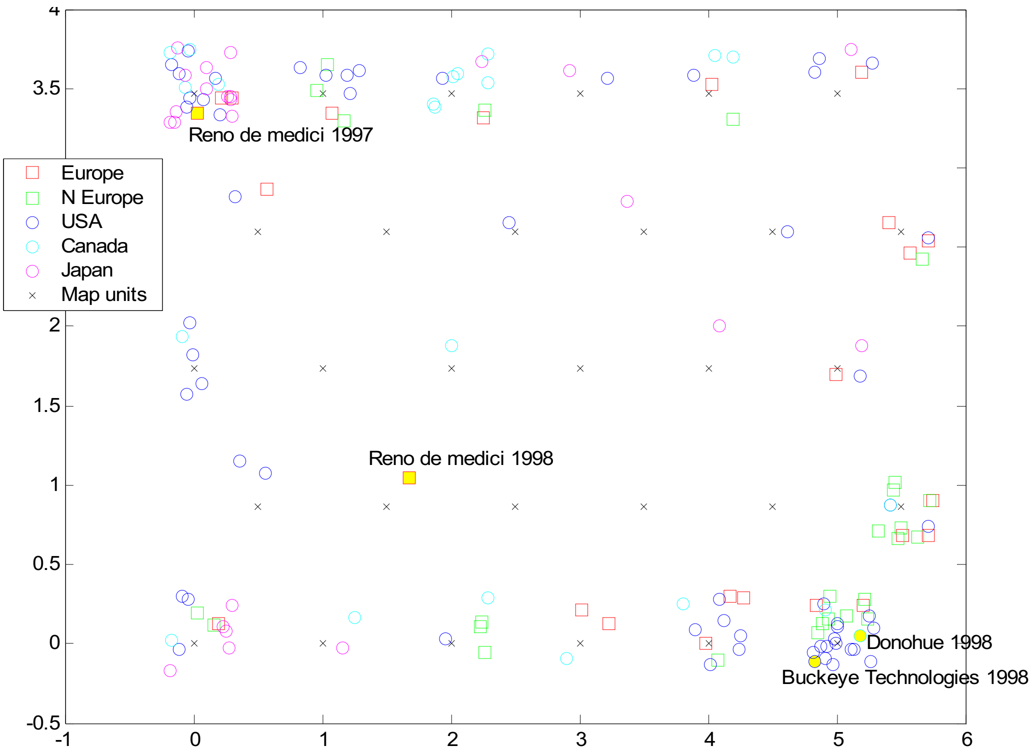
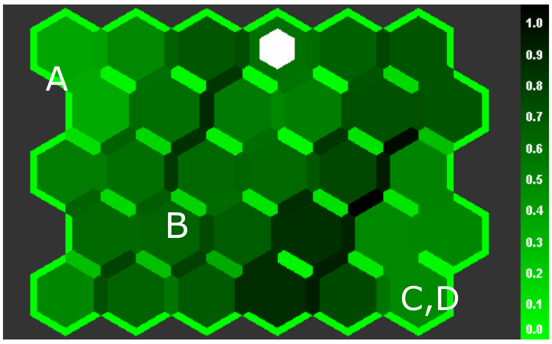


**Figure 10 Self-Organizing Map – scatter plot view created with SOM Toolbox for Matlab**

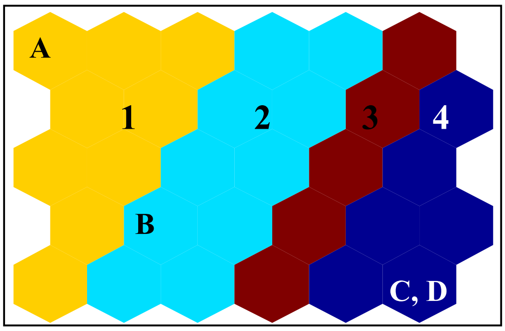
Ultsch and Siemon [27] developed the U-matrix graphic display to illustrate the clustering of the reference vectors, by representing graphically the distances between map units. In this visual representation, each map unit is represented typically by a hexagon. The line or border between two neighbouring map-units (hexagons) has a distinguishable colour that signifies the distance of the two corresponding reference vectors. Dark green accounts for large distances, and light green signifies similarities between the vectors, as indicated by the colour bar (Figure 11).



A – Reno de Medici 1997
B – Reno de Medici 1998
C – Buckeye Technologies 1998
D – Donohue 19998

**Figure 11 Self-Organizing Map - U-matrix view created with Nenet**

By looking at the borders' colours in Figure 11, the user can distinguish the main clusters that exist in the data. A clustering algorithm (e.g., K-means) can be used to automatically partition the map in similar clusters (Figure 12). The dataset appears to contain four clusters.

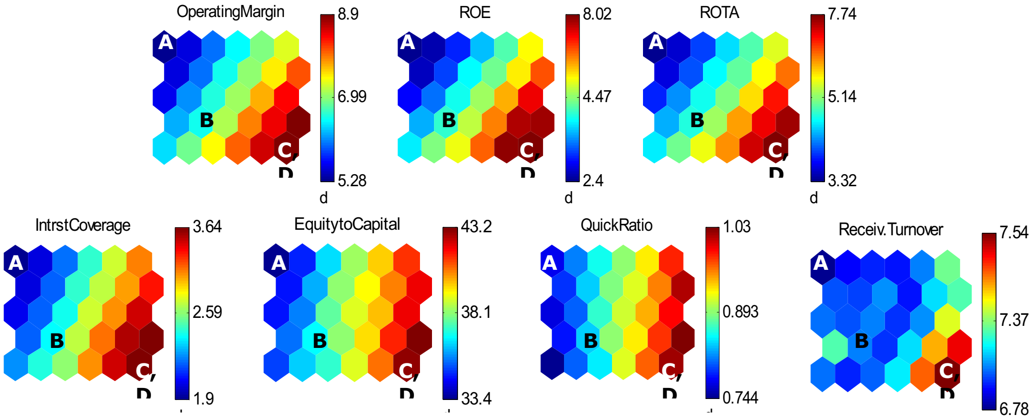


A – Reno de Medici 1997
B – Reno de Medici 1998
C – Buckeye Technologies 1998
D – Donohue 19998

**Figure 12 Self-Organizing Maps - Clustering of SOM view created with SOM Toolbox for Matlab**

Based on Figure 10, Figure 11 and Figure 12, one can compare the companies of interest with respect to their membership to the identified clusters.

It is also possible to visualize each data dimension via the feature planes. The features planes show for each variable what the level of the values in each map unit is. Colour red signifies high values of the variables, and blue and black correspond to low values of the variables (as indicated by the colour bars, Figure 13).



A – Reno de Medici 1997
B – Reno de Medici 1998
C – Buckeye Technologies 1998
D – Donohue 19998

**Figure 13 Feature planes created with SOM Toolbox for Matlab**

The feature planes help the user in describing the data clusters and the companies of interest. The feature planes also help the user to identify rapidly any relationships between variables (e.g., profitability ratios are well correlated among themselves and with solvency and liquidity ratios).

By examining the features planes in parallel with the clustering of the SOM, the user obtains the description of the four clusters identified previously as follows. Cluster 1 shows very low profitability, liquidity, solvency and efficiency. It contains the companies with the worst financial performance. Reno de Medici 1997 is situated in this cluster (A). Cluster 2 shows medium profitability, solvency, and liquidity, but low efficiency. Reno de Medici 1998 belongs to this cluster (B). Cluster 3 shows good profitability, liquidity and solvency. Efficiency is medium to low. Cluster 4 shows very high profitability, solvency, liquidity and efficiency. It contains the companies with the best financial performance, among which Buckeye Technologies 1998 and Donohue 1998 (C and D) are situated.

By looking at both Figure 10 and Figure 12, one can see the composition of each cluster with respect to variable Region (e.g., Cluster 4 contains mostly American, Northern European and European companies). Because the SOM maps to each unit a number of companies, one can easily compare different companies based on their position on the map (or cluster). For example, Buckeye Technologies 1998 and Donohue 1998 belong to Cluster 4, among the best performing companies. Reno de Medici improved its performance in 1998, and moved from Cluster 1 to Cluster 2.

## 3.7. Clustering techniques

Clustering techniques aim at partitioning the dataset in distinct groups so that the observations in each group are similar to each other, while observations from different groups are dissimilar. The similarity and dissimilarity between observations are typically calculated based on distance metrics. The SOM can also be described as a clustering method.

### *Dendrograms*

Dendrograms represent graphically the nested groupings of data items produced by a hierarchical clustering technique [25]. The dendrogram also represents the similarity levels at which groupings are formed. The dendrogram can be cut at different levels to reveal different clusterings of the data. The hierarchical clustering starts by defining a similarity measure between the data elements. The most similar data items are clustered first. Then the distance between the remaining items and the clusters formed is computed and the minimum distance is chosen to yield a new cluster. The procedure continues until a stopping criterion is met. Usually, data should be normalized so that all variables will have values that lie within similar ranges. In this way, the contribution of each variable in defining the distance between observations (and clusters) is equal.

Figure 14 represents the dendrogram using Ward hierarchical clustering technique and Euclidian distance as the similarity measure. The data dimensions were normalized using discrete histogram equalization method. In the figure, different colours represent different clusters obtained at the distance level 2.4, and the numbers in brackets are the companies' identifiers.

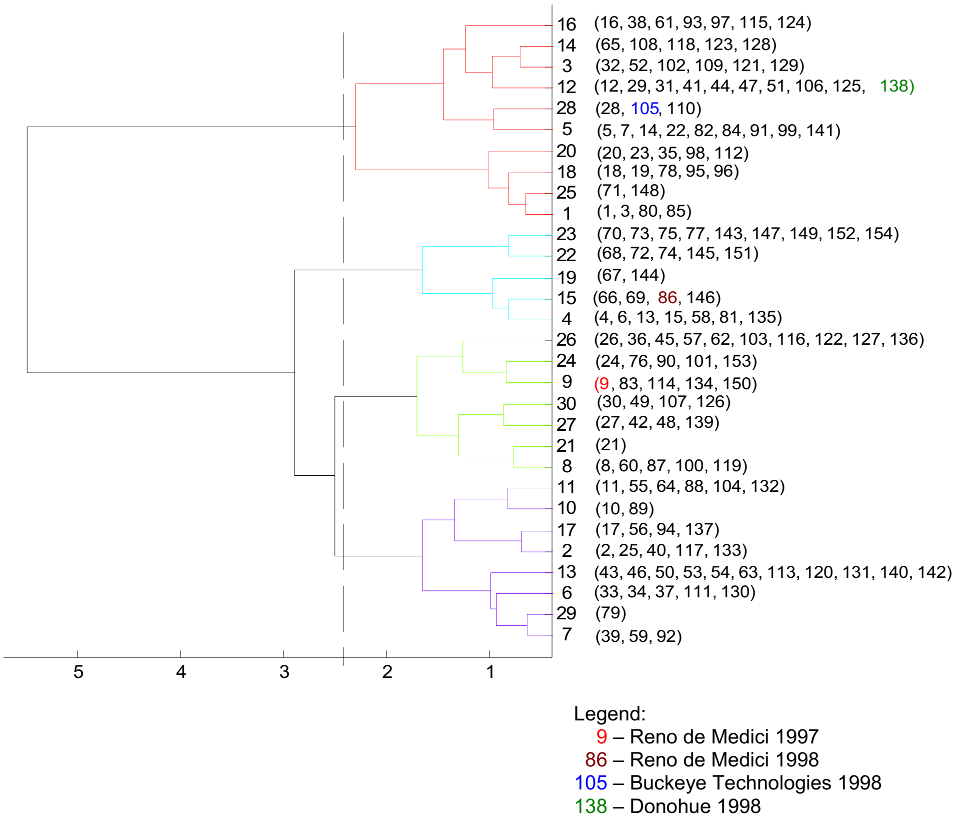


**Figure 14 Dendrogram created with Statistics Toolbox for Matlab**

From this representation, we get the idea of how many clusters exist in the data and what are their compositions. We have highlighted the companies of interest. Their membership to the clusters appears to be in conformity with previous findings. Companies Buckeye Technologies and Donohue belong to the same cluster in 1998, which seem to be one with the most profitable companies. Reno de Medici moved from one cluster, in 1997, into another, in 1998, which is in line with the increase in profitability and other changes that the company experienced. However, the dendrogram does not enable the user to interpret the characteristics of each cluster.

## 4. Summary and discussion

Based on our understanding of the visualization techniques and on our experience with applying the visualization techniques to the financial benchmarking problem, we subjectively assess the extent to which each technique is capable in solving the tasks defined in Section 2.1. We observed that all visualization techniques are capable of providing an overview of the dataset under analysis, and different techniques uncover different patterns in the data.

We summarize in Table 3 the capabilities of each technique to answer the questions and data mining tasks formulated in Section 2.1. The assessment in Table 3 concerns only the dataset and business problem presented in Section 2. We do not intend to generalize the results to other datasets, because for a different dataset (with different types of data, number of variables, number of observations, underlying structure) the results of the evaluation could be different.

**Table 3. The capabilities of the visualization techniques on the dataset under analysis**

| | *Task – See Section 2.1* | | | | | |
|---|---|---|---|---|---|---|
| *Visualization technique* | *Outliers detection* | *Dependency analysis* | *Clustering* | *Cluster description* | *Class description* | *Comparison* |
| *Line graphs* | ✓ | ✓ | | | ✓ | ✓ |
| *Permutation matrix* | ✓ | ✓ | | | | ✓ |
| *Survey plot* | ✓ | ✓ | | | ✓ | ✓ |
| *Scatter plot matrix* | ✓ | ✓ | | | | ✓ |
| *Parallel coordinates* | ✓ | ✓ | | | | ✓ |
| *Star glyphs* | ✓ | ✓ | | | | |
| *Treemaps* | ✓ | | | | ✓ | ✓ |
| *PCA* | ✓ | ✓ | ✓ | ✓ | | ✓ |
| *Sammon's mapping* | | | | | ✓* | |
| *Self organizing maps – all views combined* | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| *Dendrogram* | | | ✓ | | | |

* Sammon's mapping is capable of organizing the dataset so that different classes are distinguishable but does not provide a means to interpret and describe the classes.

From this summary, one can observe that the most effective techniques in uncovering patterns in this specific dataset are the SOM and PCA. However, only the SOM appears to answer all data mining tasks formulated in Section 2.1, but this is explained by the fact that all the four SOM-based views were analysed together. If we assess separately each SOM-based visualization technique, the results show that different SOM views show different patterns in the data (Table 4).

**Table 4. The capabilities of SOM-based visualization techniques on the dataset under analysis**

| | *Task – See Section 2.1* | | | | | |
|---|---|---|---|---|---|---|
| *Technique* | *Outliers detection* | *Dependency analysis* | *Clustering* | *Cluster description* | *Class description* | *Comparison* |
| *SOM – scatter plot* | ✓ | | ✓ | | ✓* | |
| *SOM – U-matrix* | | | ✓ | | | |
| *SOM - clustering* | | | ✓ | | | |
| *SOM – feature planes* | ✓ | ✓ | ✓ | ✓ | | ✓ |
| *Self organizing maps – all views combined* | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

* SOM – scatter plot view is capable of showing where the companies from different classes (regions) are mapped but does not provide a means to interpret and describe the classes.

Table 3 also shows that almost all visualization techniques can facilitate the comparison among companies, especially if the tools integrate interactive tools for selecting and highlighting the items of interest. Moreover, all techniques but dendrogram and Sammon's mapping are effective in finding outliers or anomalies in this dataset. Besides the tasks mentioned in Table 3, some visualization techniques give more insight into the data by highlighting trends such as evolution of financial ratios in time for selected companies (e.g., line graphs, permutation matrix, survey plot, scatter plot, parallel coordinates, treemap, PCA, and SOM).

Table 3 can therefore be used as a means to map the data mining tasks to different visualization techniques for this dataset. The table can serve also as a method to compare different visualization techniques with respect to their capabilities in solving certain data-mining tasks. This table can be therefore used in the process of selection of visualization techniques for representing and exploring the data in the financial benchmarking problem.

The visualization techniques can also be compared with respect to their capability in showing data items or data models. For example, some techniques display a data mining model (e.g., SOM, dendrogram) and other techniques display the data items (multiple line graphs, permutation matrix, survey plot, scatter plots, parallel coordinates, star-glyphs, treemap, PCA, Sammon's mapping). Regarding the data mining models or patterns visualized, some techniques are effective in showing clusters (SOM and dendrogram), correlations (scatter plots), or other patterns (e.g., treemap, Sammon's mapping). Treemap is especially effective in displaying hierarchical data, and in our example proved to be very effective in making comparisons between companies and highlighting the characteristics of companies from one region or another with respect to the value of financial ratios.

Moreover, the visualization techniques can be compared with respect to the type of data processed. For example, the following visualization techniques represent the original data: multiple line graphs, permutation matrix, survey plot, scatter plot, parallel coordinates, treemap, star glyphs, while others represent standardized or normalized data: PCA, Sammon's mapping, SOM, dendrogram. The visualizations obtained using standardized or normalized data are more difficult to interpret.

The mapping from data mining tasks to visualization techniques and the comparison of the techniques with respect to their capability to display the data items, data models, original data or normalized data show that the visualization techniques rather complement one to another than compete. Table 3 shows that there are data mining tasks for which more than one visualization techniques can be used. On the other hand, one data mining task may be addressed by different visualization techniques but with a different outcome (e.g., clustering solutions produced by the SOM, PCA and the dendrogram based on hierarchical clustering).

Observing this complementarity of the techniques analysed in this paper determines us to state that the use of multiple techniques may have advantages over the use of a single technique. One argument that supports this statement is given by the fact that when combining different visualizations that are based on the SOM, we obtain an almost complete understanding of the data, while if we consider only one view, for example the scatter-plot view, we understand very little about the data (Table 4).

The potential benefits of using multiple visualizations are identified as follows. One benefit is that the user has the possibility to see different facets of the data and problem under investigation by using different visualizations of the data that uncover distinct patterns. Another benefit is that the analyst has the possibility to confirm that the patterns or outliers highlighted by one visualization technique are indeed real, and not an artefact, gaining therefore more confidence in the findings. A third benefit is given by the descriptive power of some techniques over the others.

## 5. Conclusion

In this paper, we reviewed 11 multidimensional data visualization techniques for representing financial performance data. We illustrated how different visualization techniques can be used to analyze multidimensional financial datasets representing financial performance of companies. We investigated the capabilities of different visualization techniques in uncovering interesting patterns in financial data, patterns described in terms of outliers, clusters, classes, relationships and trends.

By deriving business questions and data mining tasks from the financial benchmarking problem as recommended in [11], and mapping these tasks to appropriate visualization techniques, we provided a means to subjectively compare and assess the capabilities of different visualization techniques to solve the financial benchmarking problem. This approach can serve visual data mining systems' developers in assessing the strength of various techniques in the early stage of system development and select accordingly the most appropriate techniques. The approach can be extended by involving users in further evaluation studies of selected visualization techniques.

We also highlighted the potential benefits of using multiple visualization techniques for solving a business problem such as the financial benchmarking problem and uncovering all interesting patterns in data. One benefit is that the user has the possibility to see different facets of the data and problem under investigation by using different visualizations of the data that uncover distinct patterns. Another benefit is that the analyst has the possibility to confirm that the patterns or outliers highlighted by one visualization technique are indeed real, and not an artefact. A third benefit is given by the descriptive power of some techniques over the others.

The study can also be extended by analysing other financial datasets and/or other multidimensional data visualization techniques.

We did not consider for this study the interaction techniques required by an efficient exploration of the data. In the context of multiple views of the same dataset, the implementation of interaction techniques such as linking and brushing techniques would enable the user to see the effects of his/her actions on more than one view and highlight the items of interest. This is a possible research direction for designers of visualization systems.

## Acknowledgements

I would like to thank Professor Barbro Back and Dr. Tomas Eklund for their contribution and for making available the dataset. I gratefully acknowledge also the financial support of Tekes (grant number 40435/05) and the Academy of Finland (grant number 104639).

## References

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P., (1996) "From Data Mining to Knowledge Discovery: An Overview", in *Advances in Knowledge Discovery and Data Mining*, Chapter 1, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., AAAI Press, Menlo Park, CA, and MIT Press, Cambridge, AM, p. 1-34.

[2] Card, S.K., Mackinlay, J.D., and Shneiderman, B., (1999) *Readings in Information Visualization - Using Vision to Think*, Morgan Kaufmann Publishers.

[3] Keim, D.A., (2002) "Information Visualization and Visual Data Mining", IEEE Transactions on Visualization and Computer Graphics, Vol. 8, No. 1, January-March.

[4] Kohavi, R., Rothleder, N. J., Simoudis, E., (2002) "Emerging Trends in Business Analytics", Communications of the ACM, 45(8), p. 45-48.

[5] Wright, W., (1995) "Research Report: Information Animation Applications in the Capital Markets", In Proceedings of the 1995 IEEE Symposium on Information Visualization (Atlanta, Georgia, October 30 - 31, 1995). INFOVIS. IEEE Computer Society, Washington, DC.

[6] Feiner, S., and Beshers C., (1990) "Worlds within Worlds: Metaphors for Exploring N-Dimensional Virtual Worlds. In Proceedings of the ACM Symposium on User Interface Software and Technology, p. 76-83.

[7] Ankerst, M., Keim, D. A., and Kriegel, H. -P., (1996) "Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets", Proc. Visualization '96.

[8] Keim, D. A., (1996) "Pixel-Oriented Visualization Techniques for Exploring Very Large Databases", Journal of Computational and Graphical Statistics, 5(1), p. 58-77.

[9] Johnson, B., and Shneiderman, B., (1991) "Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures", Proc. of IEEE Visualization '91 Conf., San Diego, p. 284–291.

[10] Hoffman, P. E., and Grinstein, G. G., (2002) "A Survey of Visualizations for High-Dimensional Data Mining", in Fayyad, U., Grinstein, G. G., and Wierse, A. (eds.), *Information Visualization for Data Mining and Knowledge Discovery*, San Francisco: Morgan Kaufmann.

[11] Soukup, T. and Davidson, J., (2002) *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, John Wiley and Sons.

[12] Ward, M. O., (1994) "XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data", Proc. Visualization '94, p. 326–336.

[13] Kreuseler, M., Lopez, N., and Schumann, H., (2000) "A Scalable Framework for Information Visualization", In Proceedings of the IEEE Symposium on Information Visualization 2000 (October 09 – 10).

[14] Eklund, T., (2004) *The Self-Organizing Map in Financial Benchmarking*, TUCS Doctoral Dissertation, Åbo Akademi University, Turku.

[15] Eklund, T., Back, B., Vanharanta, H., and Visa, A., (2003) "Using the Self-Organizing Map as a Visualization Tool in Financial Benchmarking", Information Visualization Journal, Vol. 2, No. 3, p. 161-171.

[16] The MathWorks, (2000) *MATLAB - The Language of Technical Computing, Using Matlab Graphics*, Version 6.

[17] Hinterberger, H. H., and Schmid, C., (1993) "Reducing the Influence of Biased Graphical Perception with Automatic Permutation Matrices", SoftStat '93, Proceedings of the Seventh Conference of the Scientific Use of Statistic-Software, March 14–18.

[18] Demsar J., Zupan B., Leban G., (2004) "Orange: From Experimental Machine Learning to Interactive Data Mining", White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

[19] Treemap 4.1, (2004) http://www.cs.umd.edu/hcil/treemap/

[20] The MathWorks, (2002) *Statistics Toolbox for Use with MATLAB - User's Guide*, Version 4.

[21] SOM Toolbox 2.0 for Matlab, (2005) http://www.cis.hut.fi/projects/somtoolbox/

[22] Nenet 1.1, (1999) http://koti.mbnet.fi/~phodju/nenet/Nenet/General.html

[23] Bertin, J., (1967/1983) *Semiology of Graphics: Diagrams, Networks, Maps*, (W.J. Berg, Trans.), Madison, WI: University of Wisconsin Press.

[24] Inselberg, A., (1985) "The Plane with Parallel Coordinates", Special Issue on Computational Geometry, The Visual Computer 1, p. 69–91.

[25] Sharma, S., (1995) *Applied Multivariate Techniques*, John Wiley & Sons.

[26] Kohonen, T., (2001) *Self-Organizing Maps*, Springer-Verlag.

[27] Ultsch, A., and Siemon, H., (1989) Technical Report 329, Univ. of Dortmund, Dortmund, Germany.

Turku
Centre *for*
Computer
Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

**University of Turku**
- Department of Information Technology
- Department of Mathematics

**Åbo Akademi University**
- Department of Computer Science
- Institute for Advanced Management Systems Research

**Turku School of Economics and Business Administration**
- Institute of Information Systems Sciences

ISBN 978-952-12-1860-6
ISSN 1239-1891