The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 465-474

Principles of Multivariate Data Visualization[1]

Moon Yul Huh[2] and Woon Ock Cha[3]

Abstract

Data visualization is the automation process and the discovery process to data sets in an effort to discover underlying information from the data. It provides rich visual depictions of the data. It has distinct advantages over traditional data analysis techniques such as exploring the structure of large scale data set both in the sense of number of observations and the number of variables by allowing great interaction with the data and end-user. We discuss the principles of data visualization and evaluate the characteristics of various tools of visualization according to these principles.

Keywords : visualization, MDL principle, line mosaic plot.

1. Introduction

Data visualization takes advantage of rich computer graphics, and is an exciting area of current research by statisticians, engineers and those involved in data mining (McLeod & Provost, 2001). It can be defined as the process of applying automation technology and a discovery process to data sets in an effort to derive underlying information from the data (Nicholas, 1999). It provides distinct advantages over traditional data analysis techniques such as analyzing large data sets and multivariate data both of which are very difficult to analyze, allowing great interaction with the data and end-user, and providing rich visual depictions of the data for easier understanding and analysis.

In today's high speed computing with great power of network and data oriented world, the demand and the use of data visualization technology will rapidly grow. The focus of this paper is to provide the principles of multivariate data visualization which would be indispensable for designing efficient visualization tools.

In this paper, we first state Minimum Description Length (MDL) principle and discuss the

[1] This research was financially supported by Hansung University in the year of 2004

[2] Professor, Department of Statistics, SungKyunKwan University, Seoul, 110-745, Korea
E-mail : myhuh@skku.ac.kr

[3] Professor, Division of Computer Engineering, Hansung University, Seoul, 136-792, Korea
E-mail : wcha@hansung.ac.kr

principles of data visualization from the viewpoint of psychophysics. We then evaluate various existing visualization tools to see if they are suitably designed according to the principles. We finally present parallel coordinates and line mosaic plots as the prototypes for data visualization using hDAVIS (extension of DAVIS, Huh and Song, 2002) which is developed for visualization of multidimensional complex data.

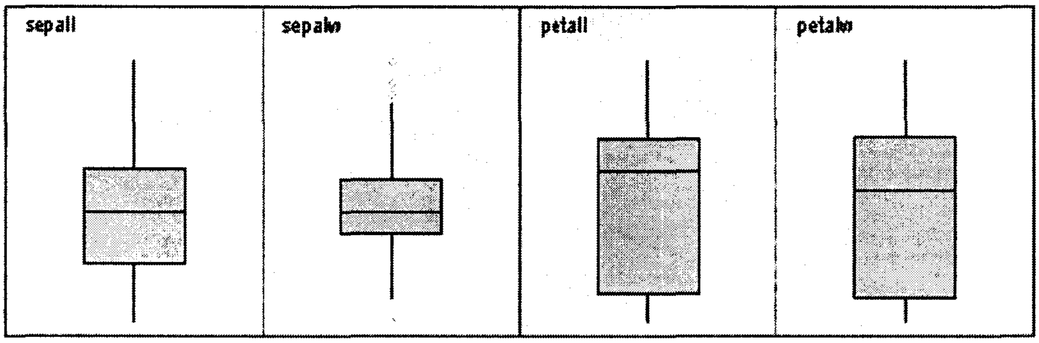
2 MDL principle

Consider we are interested in the exploration of the structure of famous iris data (Blake and Merz, 1998). First approach would be to obtain descriptive statistics of the 5 variables of the data as follows. The statistics give mean, standard deviation, and 5 numbers for each continuous variables. Iris data consists of 3 species, and there are 50 observations for each species.

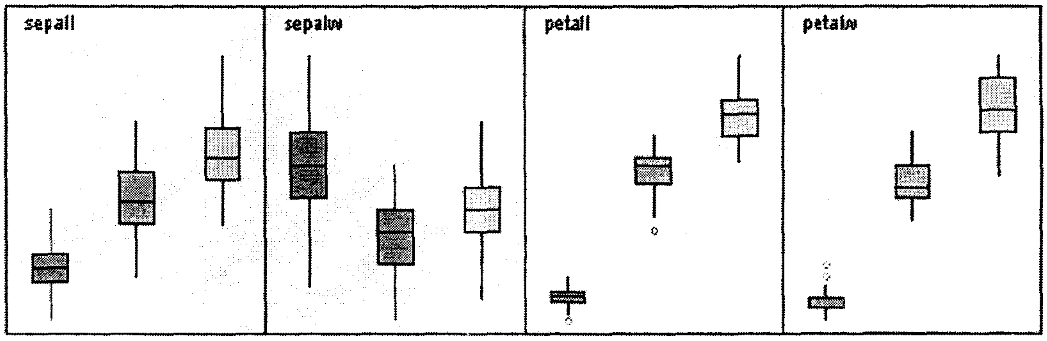
<Table 1> Descriptive statistics of Iris data

	Mean	Std.Dev	Min	0.25Q	Median	0.75Q	Max
sepall	5.843	0.828	4.300	0.100	5.800	6.400	7.900
sepalw	3.054	0.434	2.000	2.800	3.000	3.300	4.400
petall	3.759	1.764	1.000	1.600	4.400	5.100	6.900
petalw	1.199	0.763	0.100	0.300	1.300	1.800	2.500

class	Iris-setosa	Iris-versicolor	Iris-virginica
	50	50	50



<Figure 1> Parallel box plots for each variables of Iris data.



<Figure 2> Parallel box plots for each variables of Iris data when observations are divided into 3 categories (Iris-setosa, Iris-virginica, Iris-versicolor) according to the class variable.

Compare the statistics in Table 1 with the parallel box plots shown in Figure 1 and 2. It is apparent that Figure 1 needs less attention to understand the basic characteristics of the 4 variables of the Iris data than Table 1. Also, Figure 2 gives comprehensive information regarding the structure of the 4 variables of Iris data for each categories of the Iris class.

Humans tend to perceive the underlying structure of the complex data better by figures than by numbers. For this proposition, we consider a question: to perceive a human's face, how much memory is required? To answer this question, we borrow a concept of minimum description length (MDL) principle from information theory.

Let DL be the length needed to describe the person's face. Then, MDL principle gives the following formula.

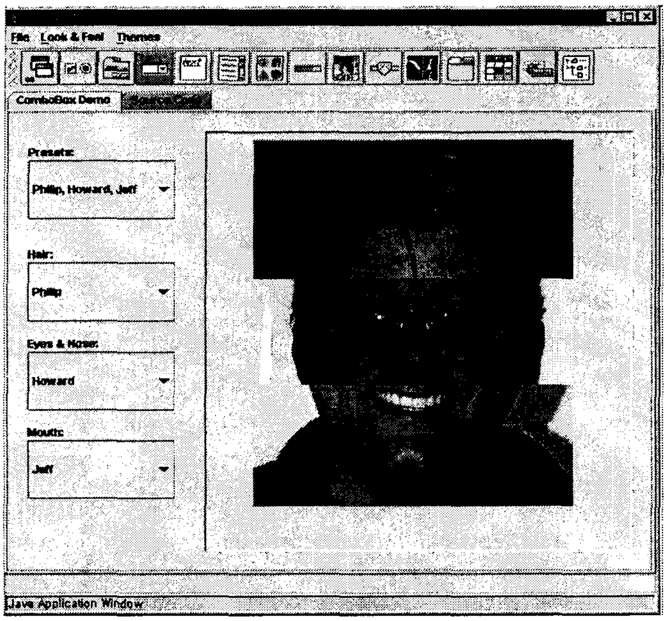
$$DL[E]= DL[T] + DL[E \mid T]$$

where, T denotes the theory and E denotes an example. In words, this is as follows.

description length(example) = length(theory) + length(example | theory)

Here, length(example | theory) equals description length of a particular example that cannot be explained by the theory. We are concerned about this quantity only since the theory is well described, and we do not need to pass this information to understand an example.

To describe the image of a person's face, how much information is needed? To explain this, we consider an example of Swingset demo (http://java.sun.com/products/javawebstart/ apps/remoteApps.html, Figure 3). We need 3 parts of face to consider: upper part; middle part; lower part. Upper part consists of hair, forehead; middle part consists of two eyes with or without glasses, and nose. Lower part consists of mouth and a chin, and etc. This is a simplified theory for the perception of a human face, and we are trained with this theory from our everyday life. In the Swingset demo, each part has 10 different prototypes which needs log2(10) = 3.3 bits. Hence 3 X 3.3 = 9.9 bits total are needed to describe a specific person's face.



Presets: Philip, Howard, Jeff

Hair: Philip

Eyes & Nose: Howard

Mouth: Jeff

<Figure 3> Swingset demo

In case of a person's name of a Korean, we all know that the name requires 2-4 Korean characters. 1 Korean character = 2 bytes = 16 bits. This is a theory for the Korean names. To describe a specific person's name, we need 32~64 bits. Therefore, small amount of memory bits are needed to describe a face than a name. Hence, to understand, or perceive a person's face need far less information than a person's name. This is a partial explanation that figures are more efficient tool than numbers or symbols to perceive the structure of an object.

3 Principles of data visualization

Data visualization means presenting data in pictorial form and using human recognition capabilities to detect patterns. Data visualization techniques focus on the ability to explore complex multidimensional data and provide the user an interactive interface.

Principles of visualization are essential for designing efficient visualization tools. We give principles of data visualization and evaluate the well known statistical plots according to these principles.

3.1 Principles of relative change (Weber's Law)

Weber's Law (Grinstein and Ward, 2002) states that the increase in a change which is just noticeably different is a constant proportion of the change.

Let x be the magnitude of a physical attribute of an object, and let $x+w_p(x)$ be the magnitude of a second object. Let the subscript p be the probability that a person perceives the second object to have the larger magnitude.

Weber's law is as follows.

$$w_p(x)=k_p x \text{ , where } k_p \text{ is a linearization constant.}$$

The function is linear in x, and for example, this means that person apparently detect the presence of a change from 100 centimeters to 101 centimeters with the same probability as detecting the presence of change from 1 to 1.01 centimeters. The absolute error is much smaller in the second case, so comparison against higher resolution scales is advantageous.

Therefore, the likelihood of detection is proportional to the relative change, not the absolute change, of graphical attribute. This means that regardless of the absolute size of graphical entity, person's ability to detect changes is based on relative changes and relative changes is easier to detect. Parallel box plot (Figure 2) is a good example demonstrating this principle.

3.2 Principles of dimensionality (Steven's Power Law)

Dimensionality playes an important role in visualization. Steven's Power Law (Steven, 1976, Grinstein and Ward, 2002) states that the perceived scale in absolute measurements is the actual scale raised to a power. For linear feature, power is .9 ~1.1; for area feature, .6 ~.9; for volume, .5 ~ .8.

Stevens's law can be represented as

$$p(x)=kx \wedge b$$

where x is the actual magnitude the attribute, p(x) is the perceived magnitude of the x, b is the power depending on the dimensionality of x, and k is some constant. For length, b is roughly 1 ^ (-.5) = 1; for area, b is roughly 2 ^ (-.5) ≈ .71; and for volume, b is roughly 3 ^ (-.5) ≈ .58.

Here is an area example. When we have two objects (x and y) with the actual magnitude of 4 and 1, the perceived magnitude of these two objects are as follows.

$$p(x)=1 \wedge (.71)=1$$
$$p(y)=4 \wedge (.71)=2.68$$

Hence, the larger area appears 2.7 times larger, not 4 times larger. People underestimate the larger areas. For example, the relative lengths of road segments will be correctly perceived, but a lake represented on a map with an area graphically 10 times larger than another will be perceived as only 5 times larger (Catarci et al., 2004). This principle says that mapping single values to attribute such as area and volume is more likely to result in errors in judgment than mapping to simpler attributes such as length.

3.3 Cleveland's error in perception from the statistical graphics

The above 2 principles suggest us to use lines for the elements of statistical graphics, and to consider relative changes instead of placing a graphics component in an absolute position. Cleveland and McGill (1985) suggests a comprehensive principles for statistical graphics. They listed "errors in perception" from the statistical graphics in the following order of increasing error.

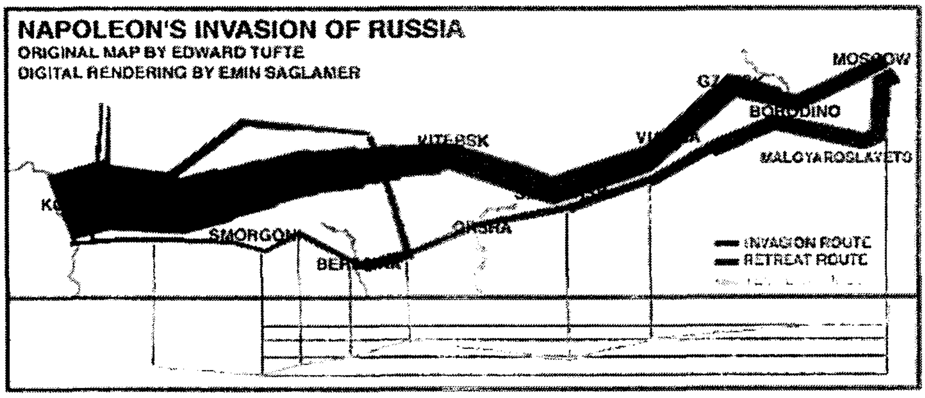
- Position along a common scale
- Position along identical nonaligned scales
- Length
- Angle/Slope (though error depends greatly on orientation and type)
- Area
- Volume
- Color Hue, Saturation, Density (only informal testing)

Hence, for statistical graphics, it is suggested to use lines instead of areas or volumes, and draw the components of a plot along identical aligned axis. Examples following this principle are Tukey's hanging bar and residual plot of regression analysis.

4. Tools for multidimensional data visualization

Historically, many excellent plots have been devised. Just for reference, two examples are given here. The first one is Charles Minard's Napoleon map and the second one is Nightingale's coxcomb. Minard's Napoleon map (Figure 4) gives several information: number of soldiers and temperatures at different battle fields in time sequence on the map. The plot uses lines and their widths along a common scale to represent the number of soldiers at battlefields. Full description of the plot is available at the website. This plot is regarded as one of the best plots ever drawn.

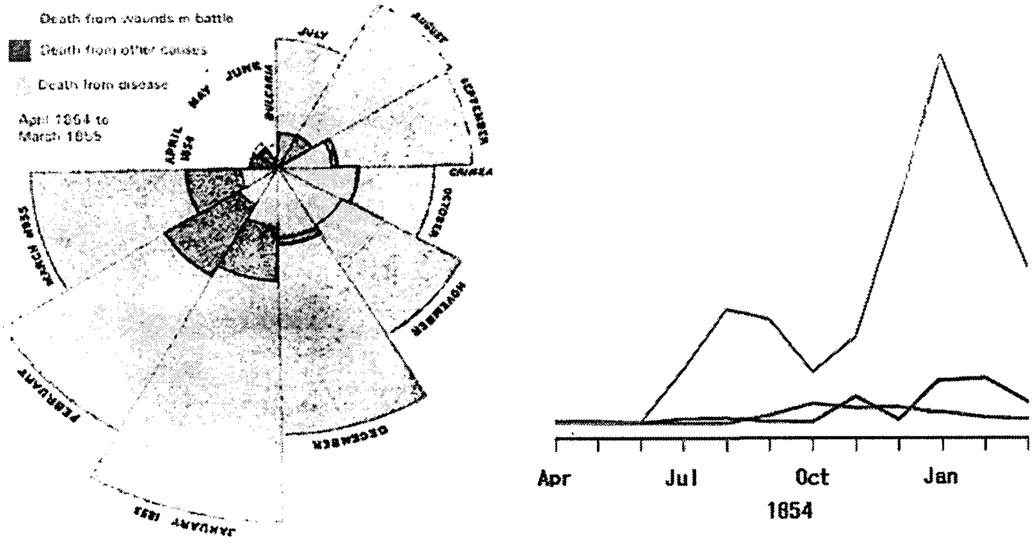
Coxcomb (Figure 5) is a polar-area diagram that shows how people had died in Crimean war from April, 1854 to March, 1855. Three different causes of death in time sequence are depicted. To represent Nightingale's coxcomb using pie chart, we need 3 separate pie charts for each cause of deaths. Hence, it may seem at first that coxcomb is more efficient than pie


NAPOLEON'S INVASION OF RUSSIA
ORIGINAL MAP BY EDWARD TUFTE
DIGITAL RENDERING BY EMIN SAGLAMER
MOSCOW
BORODINO
MALOYAROSLAVETS
VITEBSK
SMORGONI
ORSHA
BE
INVASION ROUTE
RETREAT ROUTE

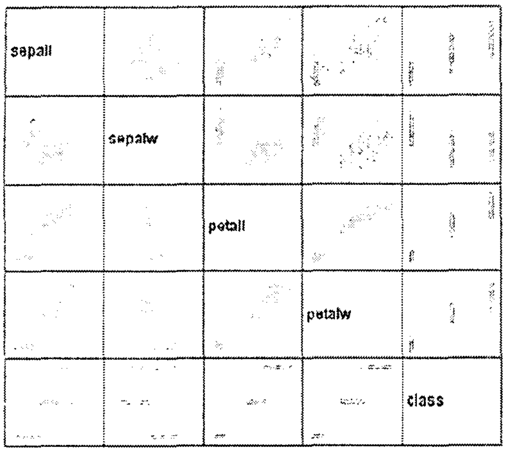
<Figure 4> Napoleon map (origin : Charles Joseph Minard's map) (http://www.ddg.com/LIS/InfoDesignF96/Emin/napoleon/index2.html)

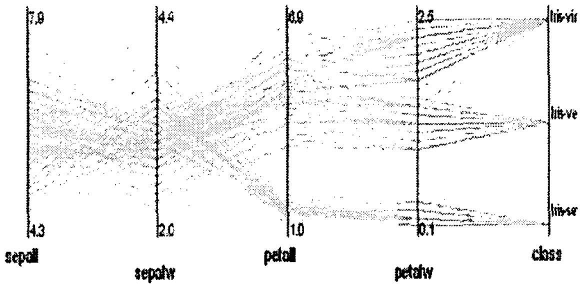
chart representation. We know that the area of an wedge of a circle is $(r^2 \times \Theta)/2$. Coxcomb fixes Θ while pie chart fixes r to represent an event. Hence, the perception of the relative importance of an event with coxcomb and piechart depends on r^2 (area) and Θ (angle) respectively. Steven's Law states that the perception of area tends to provide underestimation of the true magnitude.

This makes pie chart more readable than coxcomb. However, Cleveland's result states that angle is far more difficult to perceive than the height (position) along a common scale. This principle suggests bar chart or line chart is far more efficient than pie chart, and pie chart is not suggested if possible.



<Figure 5> Nightingale's coxcomb (http://www.uh.edu/engines/epi1712.htm), and the line plot for the same data.



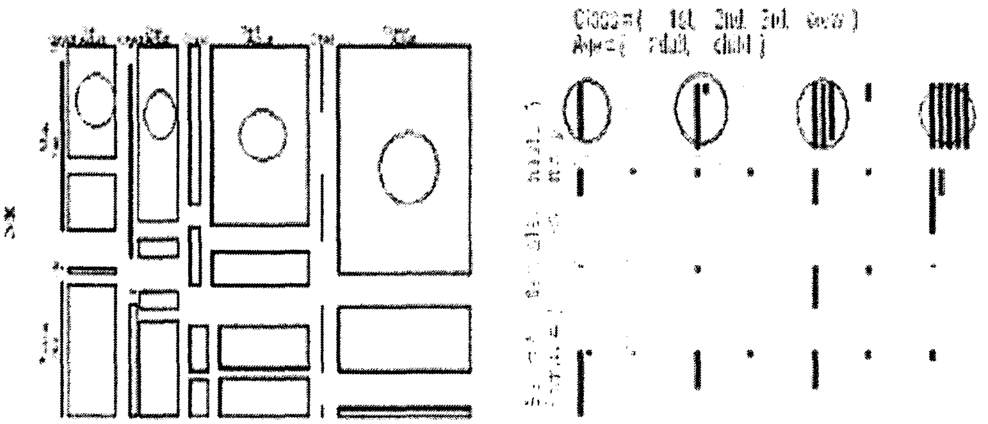


<Figure 6> Scatterplot matrix of Iris data.

<Figure 7> Parallel coordinates of Iris data.

We now discuss scatterplot matrix. Figure 6 gives the scatterplot matrix of the Iris data. The plot is very concise and is an efficient tool for data visualization because we can identify each data point, or we can select a subset of data points with this plot. This feature makes the plot a popular tool among statisticians. This is basically a two-dimensional plot, and is drawn along nonaligned axes. Another serious problem of this plot is that this is hard to perceive when data size gets larger. We need some methods for data reduction, either for feature selection and data size (number of observations) reduction. Cleveland's coPlot (1993) is another ingenious method for the multidimensional data representation. But this has the same problem as the scatterplot matrix has besides its complexity and limitations in applicable data dimensionality (maximum is 4). In this respect, parallel coordinates is becoming a more popular tool for data visualization. The plot is basically 1-dimensional and is aligned along a common axis, and this makes the plot easier to perceive. Figure 7 gives the parallel coordinates of the Iris data.

To visualize multidimensional categorical data, we often use mosaic plot. Figure 8 is the mosaic plot with Titanic data (Dawson, 1995). Titanic data consists of 2201 cases and 4 variables (Class, Gender, Age and Survived). The values of each variable are: Class = {1st, 2nd, 3rd, crew}; Gender = {male, female}; Age = {adult, child}; Survived = {yes, no}. Left plot is the conventional mosaic plot (Friendly, 1994, 1999), and line mosaic plot (Huh, 2004) is given in the right-hand-side. The 4 circled objects are adult males who did not survive, but whose class were in the 1st, 2nd, 3rd and crew respectively. The left plot (conventional mosaic plot) shows that as the class goes down (assuming the crew class is lower than the 3rd class), number of *not*survived gets larger. However, it is hard to estimate the relative magnitude from this plot. With line mosaic plot, however, we can estimate quite accurately the relative magnitude of the number belonging to one class to the number belonging to the other class. For example, the number of *not*survived in 3rd class is easily seen to be more than 3 times the number in 1st class.



<Figure 8> Conventional mosaic plot (left) for the Titanic data, and the line mosaic plot for the same data.

The conventional mosaic plot is trying to show the size of the combination of each cell of a contingency table through an area, is not drawn along a common scale, and is drawn along nonaligned axis. This makes the plot hard to perceive. With line mosaic plot, however, we can easily perceive the characteristics of the multidimensional categorical data since this plot is aligned along a common axis, and uses lines to express the size of a cell.

5 Conclusion

Data visualization techniques focus on the ability to analyze multidimensional data and provide the user with an interactive interface. In this paper, we discuss the principles of data visualization and discuss the well known tools for visualization. It is recommended to apply the principles of visualization when we develop the effective visualization tools. Line mosaic plot and parallel coordinates are good visualization tools which are suitably designed according to the principles of multidimensional data visualization. All the figures in this paper except Swingset demo in Figure 3, Napoleon map in Figure 4, coxcomb and line plot in Figure 5 and conventional mosaic plot in Figure 7, are drawn using hDAVIS which is publically available on the website: http://stat.skku.ac.kr/~myhuh/hDAVIS.html

References

[1] Blake, C .L. & Merz, C. J. (1998). UCI Repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA (http://www.ics.uci.edu/~mlearn/MLRepository.html)

[2] Catarci, T., D'Amore, F., Janecek, P., Spaccapietra, S., Interacting with GIS: from paper cartography to virtual environments Unesco Encyclopedia on man-machine Interfaces, Advanced Geographic Information Systems, Unesco Press (in Press). available as a pdf file at http://hci.epfl.ch/website/publications/2001/EOLSS -with_images.pdf.[3] Cleveland, W. S. and McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science*, 229, 828-833.

[4] Cleveland, W. S. (1993). Visualizing Data. AT&T Bell Lab, Murray Hill.

[5] Dawson, R. J. (1995). http://ssi.umh.ac.be/titanic.html

[6] Friendly, Michael (1994). Mosaic Displays for Multi-Way Contingency Tables, *Journal of the American Statistical Association*, 89, 190-200.

[7] Friendly, Michael (1999). Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data, *Journal of Computational and Graphical Statistics*, 8, 373-395.

[8] Grinstein, G. G. and Ward, M. (2002). Introduction to data Visualization in *Information Visualization in Data Mining and Knowledge Discovery* edited by Fayyad, U., Grinstein, G. G. and Wierse, A., Morgan Kaufmann publishers [9] Huh, M. Y. and Song, K. R. (2002). DAVIS: A Java-based data visualization system, *Computational Statistics*, 17(3), 411-423.

[10] Huh, M. Y. (2004). Line Mosaic Plot: Algorithm and Implementation, invited paper, *COMPSTAT* 4, Prague, Chech.

[11] McLeod, A. I. & Provost, S. B. (2001). Multivariate Data Visualization, *Encyclopedia of Environmetrics*, 1333-1344, Edited by Abdel El-shaarawi and Walter Piegorsch, New York, Wiley

[12] Nicholas, C. J. (1999). The Emergence of Data Visualization and Prospects for its Business Applications, Masters of Information systems Management Professional Seminar

[13] Steven, S. S. (1975). *Psychophysics : Introduction to its Perceptual, Neural, and Social Prospects.* New York: Wiley.

[Received July 2004, Accepted August 2004]