Hagen Rodrian (Eds.)

Please write here your book title

Springer Internal Project number if known

January 12, 1999

Springer-Verlag

Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Preface

Here come the golden words

Place, Month 1999

First Name Hagen
First Name Rodrian

Contents

List of Contributors

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

Hans Bach
Tiergartenstr. 17
69121 Heidelberg
bach@springer.de

1. Perceptual Benchmarking for Multivariate Data Visualization

M. Ward, K. Theroux

1.1 Introduction

One of the prevalent behaviors one may observe among visualization researchers is the process of developing novel methods for visualization without rigorous testing, evaluation, and comparison to competing techniques. Claims to the effectiveness of a method are often supported with minimal examples, and little guidance is provided to users in terms of when, where, and for what purpose one would be advised to use the technique in question. This is not a phenomena unique to the field of visualization: most if not all areas within science and engineering go through phases of unconstrained innovation with minimal assessment. However, as each field matures it becomes time to examine the usefulness of the various methods and attempt to quantify or benchmark each method's performance. We believe it is time for this formalized assessment in the field of data visualization, and in this paper describe our efforts to develop perceptual benchmarks for multivariate visualization techniques.

We've identified three phases in the development of perceptual benchmarking:

- Define the goals of visualization, i.e. what features or characteristics of the data are being sought in the images, and to what degree of accuracy?
- Locate or create multivariate data sets with these features present in varying degrees.
- Perform user studies to determine the effectiveness and sensitivity of different visualization methods for identifying and measuring these features.

Once these steps have been taken, we can use the results not only for selecting an appropriate visualization method for our task, but also to identify weaknesses of different methods such that future enhancements may be well focussed. Additionally, such experiments may help identify new niches for data visualization research in the event that no technique tested is able to assist viewers in accurately and reliably identifying certain data phenomena.

Early work on perception of primitive graphical attributes such as line length and intensity was compiled in a seminal paper by Miller [1.1]. In this paper the author conjectures that humans have a rather limited capacity to accurately measure visual and other sensory stimulation, with experiments showing the limitations for many graphical attributes to be about six to ten levels. Other work has been done by Cleveland and McGill [1.2] in defining and categorizing the tasks that people use when they extract information from graphs. These "elementary perceptual tasks"[1.2] are then ordered by the subject's ability to accurately perceive what is shown.

1.2 The Goals of Visualization

The two dominant purposes of data visualization are the exploration of data sets and the presentation of information regarding the data. Both rely on effective and reliable communication of the patterns, trends, clusters, anomalies, and relations inherent in the data. More specifically, the goal of the visualization will fall into one or more of the following categories, which we derived based on the list found in [1.3]:

Identification: is there some interesting structure, feature or characteristic of the data visible within the display?
Classification: what is the nature or type of the structure?
Quantification: what is the size or extent of the structure? How many structures are there? What are their relative positions?
Understanding: are there correlations between different data dimensions, such as between temperature and velocity or value and direction?
Comparison: are datasets similar in terms of structure, subsets, range of values or other features?

1.3 Data Sets for Perceptual Benchmarking

Most experiments which have been designed to measure visual perception have used synthetic rather than real data sets. There are many good reasons for doing so:

- it is easy to adjust the characteristics of the data with synthetic sets.
- many real data sets can be interpreted in multiple ways, leading to ambiguous test results.
- real data sets often contain multiple structures, making it difficult to isolate the effectiveness of perceiving a single phenomena.
- real data can contain noisy or erroneous data points.
- many existing data sets are either univariate or do not contain clearly defined structure/features.

On the other hand, there are some disadvantages of using simulated data:

- the validity of an experiment with synthetic data may not readily apply to real-world situations.

- it is hard to create synthetic data which accurately models structures found in real data.

We are using both real and synthetic data in the development of our perceptual experiments. Real data sets have been acquired from repositories of statistical data found on the World Wide Web at StatLib[1]. Many of these data sets have been cited in the literature [1.4] and have been annotated to identify some of the structures, features, and anomalies of the data. Our experiments focus on the detection, classification, and measurement of these data characteristics. However, it is very difficult to modify existing data sets such that minor variations in the characteristics can be achieved, such as changing the degree of "outlyingness" of a point or varying the relative positions of two clusters to ascertain user sensitivity to these features. For this we have created synthetic data sets which allow us to control the number, distribution, and size of characteristics such as trends, clusters, and anomalies.

Previous test data set generation described by Keim, Bergeron and Pickett [1.5] provided us insight for creating data sets that would mimic some of the characteristics of real data. The goal in creating the test data is to produce sets that contain only a few such characteristics. This allows us to more precisely control the experiments by varying the simulated data characteristics.

1.4 Designing Perceptual Experiments

It is critical to take care in the design of perceptual experiments. The hypotheses to be tested by the experiment must be made explicit, and the experiments must be verified as to whether they will provide results adequate to validate or refute the hypotheses. Subjects must be carefully selected and trained such that they reflect to a reasonable degree the body of users for whom the evaluated system will be servicing. In our case, subjects needed to be trained both in understanding the data characteristics for which they will be searching and in how to interpret the different visualizations to be tested.

We examined some previous experiments with graphical perception from Cleveland and McGill[1.2] and Duncan[1.6]. Cleveland and McGill defined perceptual experiments that tested a set of elementary perceptual tasks that people use when they extract information from graphs. Their hypothesis on the accuracy of the perceptual tasks are (from most accurate to least accurate): position along a common scale; positions along nonaligned scales; length, direction, angle; area; volume, curvature; and shading, color, and saturation. Duncan tested the effectiveness of the subject's ability to distinguish between the shape and position of an object when there is more than one

[1] http://lib.stat.cmu.edu/datasets/

target object. We have included variations on some of these tests in our perceptual experiments.

It is also important to validate the data sets to be used to verify that they indeed contain the characteristics/features for which the subjects will be searching. For our experiments, we implemented quantitative algorithms to identify the features in question. Thus we can be assured that the subjects will not be spending time looking for non-existent features.

1.5 Case Studies

To validate our framework for perceptual benchmarking, we designed and performed a number of experiments in the evaluation of multidimensional data visualization techniques. We focussed on two tasks (detection and measurement), two data characteristics (outliers and clusters), and three visualization methods (scatterplot matrices, parallel coordinates, and star glyph plots based on principal components analysis). These techniques and the specifics of the experiments are described in the following subsections.

1.5.1 Visualization Methods Tested

A wide range of visualization techniques have been developed for the display of multivariate data. Each has strengths and weaknesses in terms of the number of dimensions or variables which can be effectively supported, the number of data points which can be effectively perceived, and the specific task being pursued using the visualization. The techniques we used in our experiments, which have been integrated into a multivariate data visualization tool developed at WPI called XmdvTool [1.7], are outlined below.

Scatterplot matrices: Perhaps the most popular form of multivariate data visualization technique, this consists of a grid of pairwise 2-D or 3-D scatterplots arranged in such a way that each row or column fixes one dimension while plotting it against each of the other dimensions in succession [1.8] (see Fig. 1.1).

Parallel coordinates: Pioneered by Wegman [1.9] and Inselberg [1.10], this method aligns the N axes parallel to each other with a vertical or horizontal orientation. Each data point is depicted as a polyline crossing each axis at a location proportional to the value of the data point for that dimension (see Fig. 1.2).

Principal components and glyphs: Another approach to the multivariate data visualization problem is to attempt to reduce the number of data dimensions to two or three and plot the points using traditional graphics methods. One such technique is Principal Components Analysis [1.11], which locates orthogonal axes through the data space which best capture the variations in the data. We use PCA to control the positioning of star

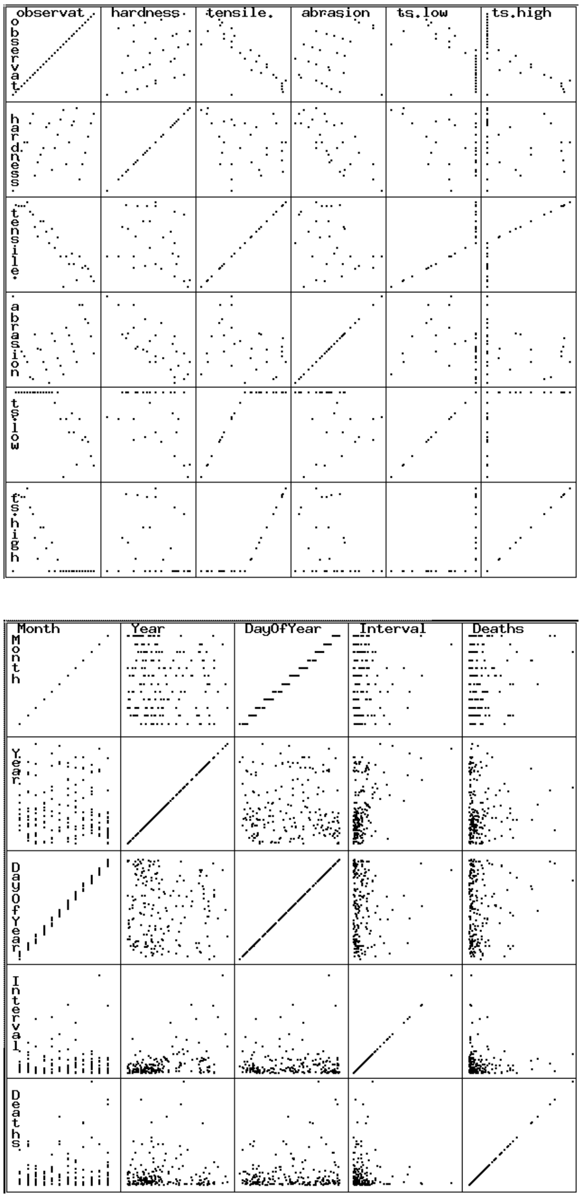


Fig. 1.1. Scatterplot matrix showing Rubber parameters **(a)** and Coal Mining Disasters **(b)** data sets

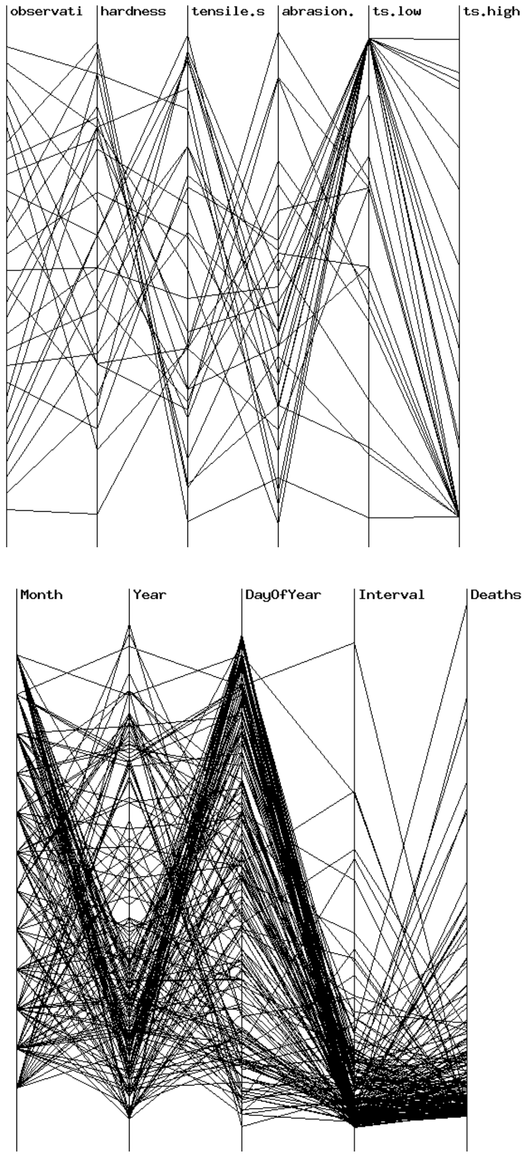


Fig. 1.2. Parallel Coordinates display of Rubber parameters **(a)** and Coal Mining Disasters **(b)** data sets

glyphs [1.12] [1.13]. Each glyph is formed by emanating N rays (where N is the number of dimensions) from a point, with the length of each ray proportional to the value for that dimension. The end points of the rays are then connected to form a polygon (see Fig. 1.3).

1.5.2 Outlier Detection and Measurement Experiments

Stage 1: We first developed a quantifiable definition for an outlier and created an algorithm capable of labeling data points appropriately. The first part of the algorithm calculates a one-dimensional projection regression that results in a measure of fit for each point within each dimension [1.14]. The standard deviations of the measures of fit for each dimension are then calculated and compared against a threshold standard deviation to determine if the point is an outlier. If the standard deviation for a point is greater than the threshold value, the point is labeled an outlier.

Stage 2: We then acquired data sets, both real and simulated, which contained outliers according to the definition in Stage 1. For some data sets, we varied the degree of separation between the outliers and the main bodies of data points to test subject sensitivity (see Figs. 1.4 & 1.5). A total of six outlier experiments were designed, and each was repeated three times for the three visualization techniques being tested.

Stage 3: Once the subjects (19 Computer Science graduate students with minimal exposure to data visualization) were trained to interpret the visualization technique(s) to be assessed and given examples of data sets with identified outliers (approximately one hour of training), they were then shown a set of 18 images of data sets containing between 0 and 6 outliers. The subjects were asked to not spend more than one minute per image. The tasks given to them were:

1. Determine if an image contains one or more outliers.
2. Identify the points believed to be outliers.
3. Estimate the degree of separation of each outlier on a 5-point scale (marginal to extreme).

Stage 4: Given the subject responses, we can now assess the usefulness of each visualization method tested in terms of outlier detection and measurement across data sets with different characteristics. The percentage of correctly and incorrectly detected outliers were tallied, as well as the average error in estimating the degree of separation.

1.5.3 Cluster Detection and Measurement Experiments

Stage 1: We developed a quantifiable definition for a cluster and adapted an algorithm capable of labeling data points as to which cluster they belong to. The algorithm is called *fanny* and performs a fuzzy clustering on the

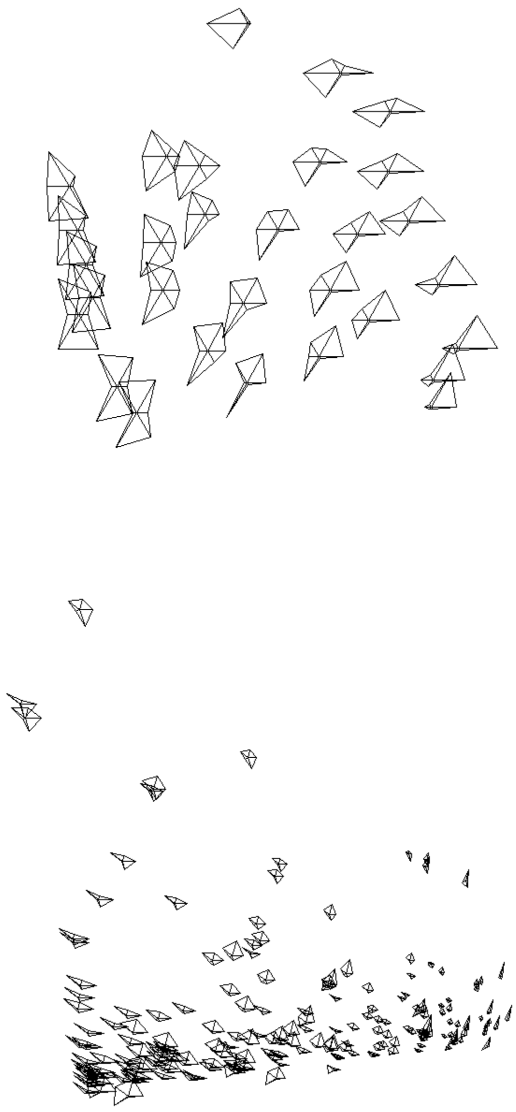


Fig. 1.3. Glyph display of Rubber parameters **(a)** and Coal Mining Disasters **(b)** data sets, with positions based on principal components analysis

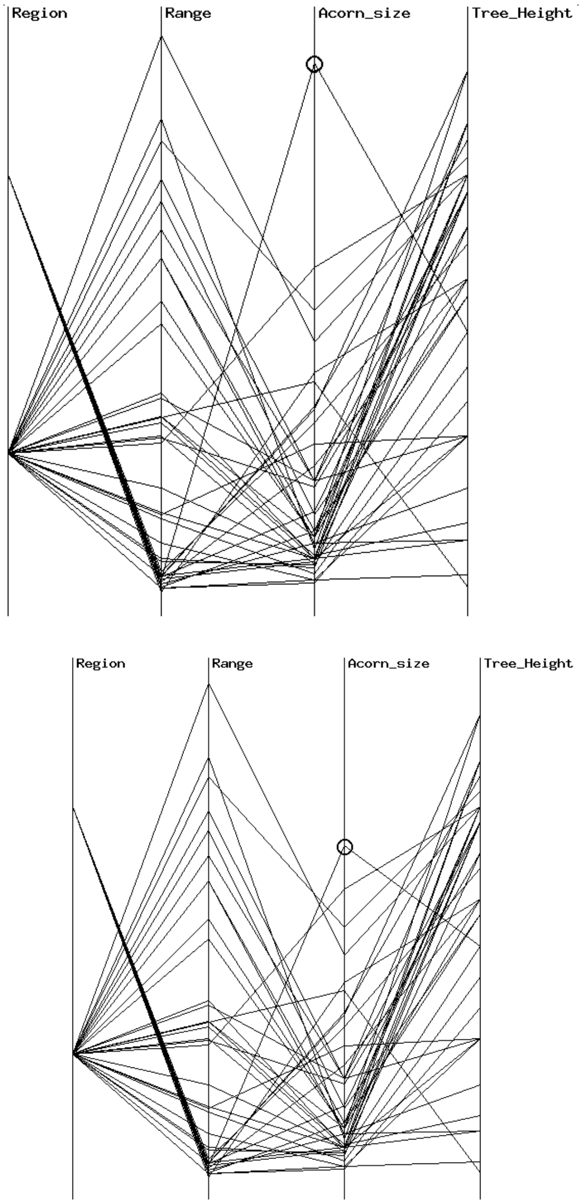
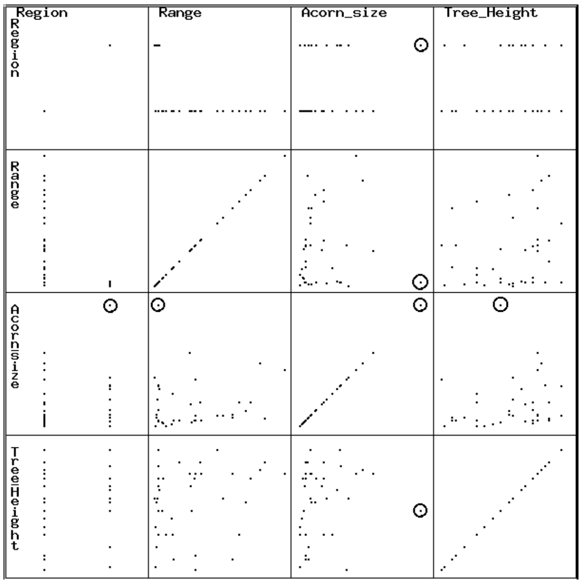


Fig. 1.4. Parallel Coordinates view of data set describing Acorn attributes, with a single outlier (circled, in the Acorn Size dimension) in its original position **(a)** and with the distance artificially shortened **(b)**



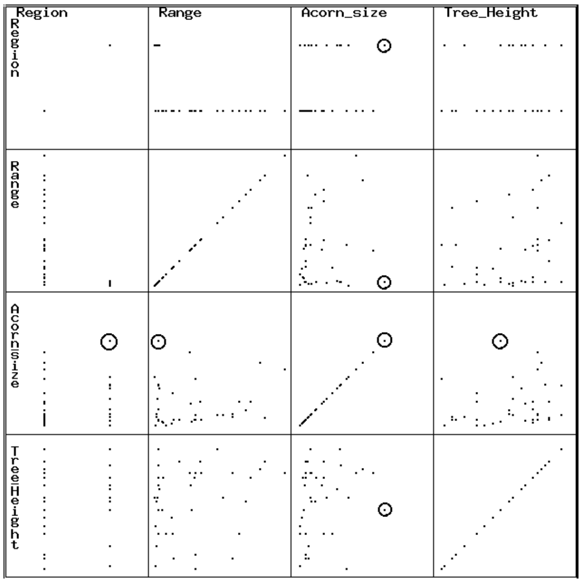


Fig. 1.5. Same data as previous figure using scatterplot matrices

data [1.15]. It is advantageous to use a fuzzy clustering algorithm for our experiments because many of our clusters are not well separated, resulting in data points that may be between clusters. The algorithm output shows the probabilities of each point belonging to a particular cluster.

Stage 2: We then acquired data sets, both real and simulated, which contained clusters according to the definition in Stage 1. For simulated data, we varied the number, size, and orientation of clusters and their relative positions. We also added random noise to some of the simulated data sets using an adaptation of Perlin's noise generation function [1.16] (see Fig. 1.6). 16 distinct experiments were designed, and each was presented using the three display techniques.

Stage 3: Once the subjects (19 Computer Science graduate students) were trained to interpret the visualization technique(s) to be assessed and given examples of data sets with known clusters (approximately 1 hour of training), they were given a packet containing 48 images of data sets which contained between 0 and 4 clusters. The subjects were asked to not spend more than one minute per image. The tasks given to them were:

1. Determine if an image contains one or more clusters.
2. Highlight the groups of points believed to lie in distinct clusters.
3. Estimate the size of each cluster on a 5-point scale (small to very large).

Stage 4: Given the subject responses, we can now assess the usefulness of each visualization method tested in terms of cluster detection and measurement across data sets with different characteristics. The percentage of correctly and incorrectly detected clusters have been tallied, as well as the average error in estimating the size.

1.6 Results

In all, over 4000 data points were gathered during our experiments, categorized as cluster identification, cluster size assessment, outlier identification, and outlier separation identification. The details of our results will be presented in a future paper; here we will just summarize our findings. Bear in mind that these experiments are sensitive to a large number of factors, including the specific data sets used, the background of the subjects, the duration of training, and the amount of time given to complete the experiments.

For both tasks, scatterplot matrices generally fared best, followed by star glyphs positioned by principal component analysis, and lastly by parallel coordinates. Each had measurable strengths and weaknesses, as outlined below.

- Scatterplot matrix
  - less effective for overlapping clusters
  - less effective in size assessment for large clusters
  - less effective when outliers fell between clusters

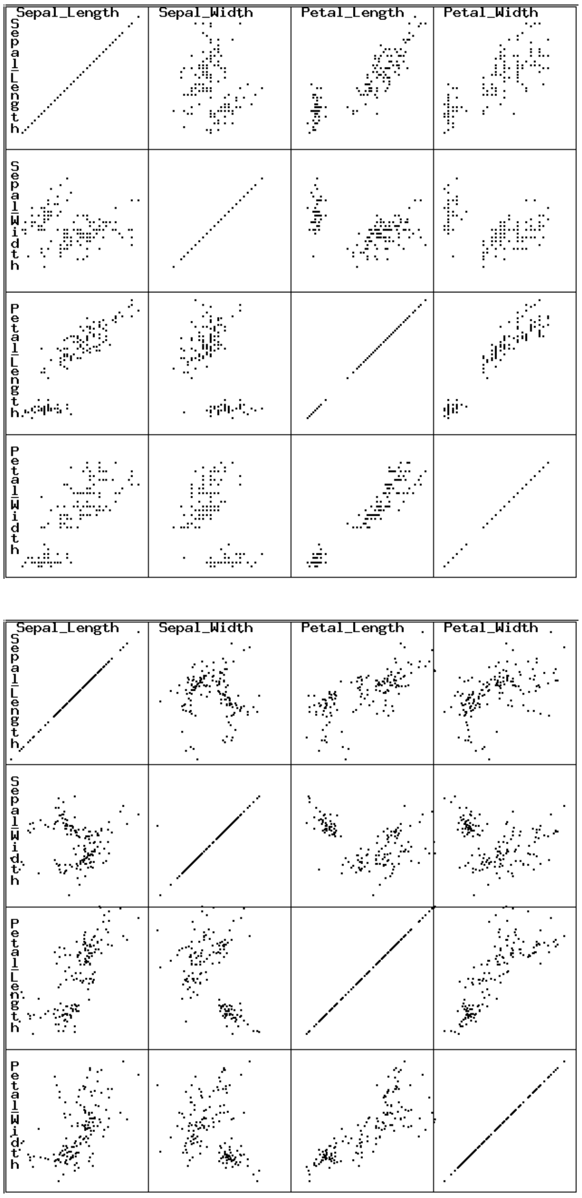


Fig. 1.6. Iris data showing different degrees of clustering. Image **(a)** is the original data and image **(b)** has a 40% added noise factor

- Glyphs
  - best for identifying internal outliers (between clusters)
  - poor for differentiating non-outliers
  - good for conveying outlier separation
  - good for overlapping clusters
  - good for measuring moderate sized clusters
- Parallel Coordinates
  - good for differentiating non-outliers

In some cases, the difference in performance by the different techniques was quite significant. For example, there were only 27 instances of incorrectly identified clusters (false positives), while for glyphs and parallel coordinates there were 126 and 159 occurrences, respectively. On the other hand, there were many circumstances where the performance difference was less striking. For example, we did not find a major difference in the number of accurately identified clusters amongst the different visualization methods (78% for scatterplots, 74% for glyphs, 71% for parallel coordinates).

Summary

In this paper we have presented a framework for assessing the strengths and weaknesses of multivariate data visualization techniques through perceptual benchmarking and described some of our experiences in performing this procedure. We identified features of data sets which are often targets of visual data exploration and presentation, as well as tasks that viewers might wish to perform on these features (detect, classify, quantify). The next step was to create or obtain data sets containing the desired features and design experiments which test the abilities of human subjects at perceiving these features with the desired level of accuracy. The third stage of this overall process was to carry out the experiments, after which began an assessment and evaluation phase. Our experience was that although the procedure is time-consuming and sometimes tedious, the results provide us with tangible guidance in the use and development of visualization methods. One of the major problems encountered involved finding suitable subjects (and a sufficient number) to participate in the experiments. Clearly the use of graduate students, while relatively straightforward to organize, probably did not give us subjects with similar background to the typical users of visualization systems.

Future work will proceed along many dimensions, including experiments with other visualization techniques, data sets and attributes, and perceptual tasks. We also hope to experiment with different groups of test subjects and duration of training to study the variability dependent on these parameters. Finally, besides the visualization methods themselves it is critical to study the effect of interactivity on the various techniques; a method which performs poorly in a static evaluation may prove to be extremely effective in an interactive environment.

We believe it is important that developers of visualization techniques perform rigorous and quantifiable evaluation and comparison of the methods in terms of the tasks and data characteristics for which they are effective. By developing a suite of benchmarking tasks we hope to add structure to the evaluation process, allowing users to select visualization techniques most pertinent to their needs and assist researchers in the development of new and improved visualization tools.

References

1.1 Miller, G. (1956) The magic number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review **63**, 276–291

1.2 Cleveland, W. McGill, R. (1984) Graphical perception: theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association **79**, 531–554

1.3 Keller, P. R., Keller, M. M. (1993) Visual Cues. IEEE Press, Los Alamitos, CA

1.4 Cleveland, W. S. (1993) Visualizing Data. Hobart Press, Summit, NJ

1.5 Keim, D., Bergeron, R. D., Pickett, R. (1994) Test data sets for evaluating data visualization techniques. in Database Issues for Data Visualization, IEEE Visualization '93 Workshop, Springer-Verlag, Berlin

1.6 Duncan, J. (1993) Coordination of what and where in visual attention. Perception **22**, 1261–1270

1.7 Ward, M. (1994) XmdvTool: integrating multiple methods of visualizing multivariate data. Proceedings of Visualization '94, pp. 326–333, IEEE Press, Los Alamitos, CA

1.8 Fisherkeller, M., Friedman, J., Tukey, J. (1975) Prim-9, an interactive multidimensional data display and analysis system. in Dynamic Graphics for Statistics (edited by W. Cleveland and M. McGill), 91–109, Wadsworth & Brooks/Cole, Pacific Grove, CA

1.9 Wegman, E. (1987) A parallel coordinate approach to statistical graphics. Proc. National Computer Graphics Association Conference, 574–580

1.10 Inselberg, A., Dimsdale, B. (1990) Parallel coordinates: a tool for visualizing multidimensional geometry. Proceedings of Visualization '90, 361–378, IEEE Press, Los Alamitos, CA

1.11 Jackson, J. (1991) A User's Guide to Principal Components. John Wiley and Sons, New York

1.12 Siegel, J., Farrell, E., Goldwyn, R., Friedman, H. (1972) The Surgical implication of physiologic patterns in myocardial infraction shock. Surgery **72**, 126–141

1.13 Mezzich, J., Worthington, D. (1978) A comparison of graphical representations of multidimensional psychiatric diagnostic data. in Graphical Representations of Multivariate Data (edited by P. C. C. Wang), Academic Press, San Diego, CA

1.14 Judd, C., McClelland, G. (1989) Data Analysis: A Model-Comparison Approach. Harcourt Brace Jovanich, San Diego, CA

1.15 Kaufman, L., Rousseeuw, P. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York

1.16 Ward, G. (1991) A recursive implementation of the perlin noise function. in Graphics Gems II (edited by J. Arvo), 396–401, Academic Press, San Diego, CA