NBER WORKING PAPER SERIES

GRAPHICS FOR DATA ANALYSIS

Roy E. Welsch*

Working Paper No. 43

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

June 1974

Preliminary: not for quotation

*NBER Computer Research Center and Massachusetts Institute of Technology, Sloan School of Management. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

Abstract

In recent years, graphics has become an essential part of modern data analysis. It is particularly useful for interactive data analysis. This paper describes a system called CLOUDS which is designed to make available on inexpensive storage tube terminals a wide range of graphic tools related to data analysis, economics, and management science. The system can be accessed nationwide by nonprofit organizations via the National Bureau of Economic Research computer network.

Contents

Exhibits

Introduction

Data analysis might be described as the art (science?) of communicating with ourselves and others by means of summaries of data. If there is too much detail in the summaries the main points are obscured, if too little, valuable information may be lost. Our first attempt to find a suitable summary (model) will probably prove to be unsatisfactory and we shall need to interact with our data in order to proceed to a revised summary. We often reread a portion of text several times in order to provide ourselves with an effective mental summary; there is no reason why we should not find ways to do this with data.

One of the most effective ways to summarize and communicate is with visual summaries, such as plots, graphs, and pictures. Recently there has been a surge of activity on the part of data analysts to discover useful visual summaries, e.g., Tukey [1971,1972]. This is in part due to the fact that hardware and software developments have made it much easier to produce usable visual summaries at reasonable cost, and also because of new trends in statistics toward exploratory and diagnostic data analysis.

The need for a data analyst to interact with his data was satisfied to some extent in the early days because he did his computing on a desk calculator. Computers have removed the computational drudgery, but large batch oriented statistical packages make it difficult to interact with the data. Time-sharing statistical packages have lowered this barrier, but typewriter output actually makes it more difficult to get many visual summaries at reasonable cost. The advent of graphic terminals that rent for about the same cost as a typewriter terminal makes it feasible to fully explore the possibilities of interactive visual data analysis. In what follows we describe the progress we have made in this area.

The TROLL System

During the period 1966-1971 an interactive computer system for quantitative research in economics and related areas called TROLL (Time-Shared Reaction On-Line Laboratory) was developed at MIT. TROLL was written in AED and now runds on an IBM 360/67 at Yale University. Since 1971, TROLL has served as a point of departure for new quantitative research systems being developed by the National Bureau of Economic Research at its Computer Research Center for Economics and Management Science in Cambridge, Mass. One phase of this project involves data analysis and, in particular, graphic data analysis. The initial systems are being developed and tested within TROLL. TROLL and the new software under development will be incorporated into a comprehensive new system within the next two years.

The NBER Computer Research Center (CRC) is viewed as a national research center and therefore the accessibility of the computer systems to collaborative researchers elsewhere in the country is an important factor. This has been accomplished by making the CRC a part of the nationwide TYMNET network.

In order to make graphics accessible to this research community we decided to develop a system based on graphic terminals comparable in cost to a standard dial-up typewriter terminal. Then graphics could be used as a part of large projects that involve many of the applications software systems at the Center with little additional cost. The hardware in use at the Center consists of two Tektronix 4010-1 storage tube terminals running at 300 baud and a Tektronix 4610 hard copy unit.

The graphics project has been supervised by Roy Welsch with the assistance of Paul Holland and David Hoaglin. Helge Bjaaland did most of the system programming, which has now been taken over by David Rice.

Standard TROLL Graphic Capabilities

The standard TROLL system, completed in 1971, contains the following basic graphic facilities, many of which are used in the experimental system to be described in the next section. Three types of plots are available: histograms, overlayed time series plots, and scatter plots.

The OUTOPT command controls the appearance of the plots, e.g. plot length and width, contents of the legend, use of plot grids, and whether or not points are marked or connected by lines. The command is also used to specify an online output device (e.g., typewriter terminal or scope) and an offline device (e.g., printer or Calcomp plotter).

In many cases the offline plots are produced when the offline version of a plotting command is given. If the specified offline device is a tape driven plotter the offline plotting commands produce a temporary file (called a GRAPHIC file) describing the requested plot. A command called DRAW translates this file into the particular set of plotter instructions needed to produce the plot and writes these instructions on tape for use on the offline plotter.

Each plotting command contains provisions for individual scaling of data sets and axes. A separate command is used to put a title on the graphs. For typewriter terminals the TABSET command is used to set tabs. This greatly speeds up typewriter plotting.

TROLL commands may be combined to form macros. These can be stored in a user's macro file and are very useful for creating special purpose plots. There are some system macros, such as NORMPLOT which produces a Gaussian probability plot of a data file. Further examples of plotting macros will be given below.

Experimental TROLL Graphics

The basic TROLL graphic facility along with the related system macros represented for most users a real step forward when compared to existing batch statistics packages and most time-sharing systems. There was an increasing demand for more graphics to aid the data analyst, particularly in the display of multivariate data. Therefore the primary motivation for expanding the basic graphic capability was to find ways to use a two-dimensional plotting device to summarize and explore high dimensional data sets.

CLOUDS

A system called CLOUDS was designed to plot and manipulate p-dimensional point clouds. It consists of groups of commands which relate to defining the cloud, specifying the order in which the points are plotted, projection plotting, rotation, superimposition, working with individual points, masking, saving the position of the point cloud values, and utilities. A complete list is given in the appendix.

A point cloud is defined by p data files (variables) or an nxp matrix. Each data file (or column of the matrix) is considered to be a dimension. If any dimension has a missing element (NA), then this point is removed from the point cloud.

The ordering commands allow the user to change the plotting order of the points and randomly or selectively thin the number of points to reduce plotting time. Thus a plot with a small number of points can be examined before proceeding to the entire data set. An order derived externally (perhaps from a clustering algorithm) can be loaded as the plotting order.

The projection plotting commands make it possible to project the point cloud on planes defined by any two of the original p axes. Rotation is accomplished by specifying a rotation center, angle, and plane (e.g., if the rotation plane is defined by axes 2 and 1, the point cloud will be rotated from 2 toward 1 in that plane). The rotation can be specified in integer multiples of a prechosen rotation angle so that large and small rotations can be accomplished easily. In essence rotation makes it easy to explore certain linear combinations of the dimensions (variables). This often makes clusters apparent or brings to light anomalies in the data that merit further study.

The superimposition commands are used to overlay plots either with or without rotation and to place plots on different portions of the screen without erasing what is already there. This is especially important on a storage tube scope.

Individual points can be identified and labeled using the keyboard or cursor lines on the scope. Points can also be moved, added, or deleted in the same manner. These commands have proved to be very useful for identifying and working with outliers.

Portions of the plot can be masked to exclude a number of points either before or after rotation. The masks are determined via keyboard control or by placing the cursor lines in the appropriate position. A ZOOM command permits the user to mask and then fill the screen with what remains. Masks may be manipulated and combined using "and" and "or" operations.

Often when masks are being used a plot is obtained that the user wants to become a starting point for further work. In short, he no longer cares about the points that have been masked out. On the other hand, after many rotations, it is often desirable to be able to go back to the initial point cloud to start again. The group of commands for saving and restoring the position of the point cloud was designed for this purpose.

The current point cloud is obtained by applying the current rotation matrix to the initial point cloud and also accounting for masks and commands that affect the coordinates of points (APNT, DPNT, MPNT, RCENTER). The INIT command returns the system to the initial point cloud by making the rotation matrix an identity matrix. The SAVE command makes the current point cloud the initial point cloud. SVECTOR can be used to save the current point cloud as external files for later use. The rotation matrix and center can also be permanently filed.

A number of utilities are available that allow rescaling and normalizing. Special symbols can be used to mark points. OUTOPT options control plot size, position, and labeling. Corresponding offline plotting commands are being implemented.

Currently CLOUDS is being improved by the addition of commands for placing text anywhere on the screen. This will allow more extensive labeling than is now possible.

It has proved useful to be able to place many plots on the screen simultaneously. Conversely we plan to make it possible to use many single screen partial plots to construct one large plot. (This will involve clipping.) The utility of an associated hardcopy device is obvious in this application.

A difficulty with CLOUDS is that the user needs guidance about where to look for structure in a point cloud. For moderate p, the number of possible projections and rotations is very large. Recently, Friedman and Tukey [1973] have proposed pursuit projection algorithms as a way to help overcome this problem. These algorithms look for structure by trying to find a projection of the cloud into a low dimensional space which separates the projected points into clusters. Such a projection then becomes a useful starting point for future search.

We plan to try projection pursuit in CLOUDS. It is especially important for a moderate speed storage tube device because one cannot fly around the point cloud at a speed that makes it possible to examine a large number of projections visually.

The design for CLOUDS had a limited scope: to produce a set of commands to manipulate point clouds. As soon as it became operational, people found ways to combine the commands to make plots that appear to have little to do with point clouds. CLOUDS has therefore become a type of language for graphics or at least a dialect for data analysis applications. It naturally suffers from not having been designed for this purpose in the first place.

While CLOUDS was being developed, another project at the Computer Research Center was in the design stage. This involved the creation of a programming language for data analysis applicatioms called DASEL. It is designed to make operations on oriented n-way arrays of data simple and efficient, and has profited from experience with APL.

We are now planning to include in DASEL a set of graphic primitives that will provide, we hope, a useful language for graphic data analysis. A successor to CLOUDS will be programmed in this language.

The development of the CLOUDS system has benefited greatly from discussions with John W. Tukey about his work in 1972 on the PRIM-9 system at the Stanford Linear Accelerator Center. Complete documentation for CLOUDS is available from the Computer Research Center in Cambridge, Mass.

Examples Using CLOUDS

Marsaglia [1968] has shown that certain types of pseudo random number generators place their output on p-dimensional hyperplanes in k-space with $p < k$. These hyperplanes are, of course, sets of probability zero in k-space. Clearly one would like to have a generator that places its output on a large number of planes. One well-known linear congruential generator, RANDU, puts all its points on 15 planes in 3-space, a rather small number.

The theory [Hoaglin, 1973] is as follows. The multiplier, 65539, can be written as $(2^{16} + 3)$ and therefore the n+1 number generated on a 32 bit machine (one sign bit) is

$$X_{n+1} = (2^{16} + 3)X_n \qquad (\mathrm{mod}\ 2^{31}).$$

But then

$$X_{n+2} - 6X_{n+1} + 9X_n \equiv 0 \quad (\text{mod } 2^{31}).$$

Defining $U_n = X_n/2^{31}$ gives

$$U_{n+2} - 6U_{n+1} + 9U_n \equiv 0 \quad (\text{mod } 1)$$

so that $U_{n+2} - 6U_n + 9U_n$ is always an integer between -5 and +9 (inclusive). Thus the triples (U_n, U_{n+1}, U_{n+2}) or (X, Y, Z) all lie on 15 planes in the unit cube.

We took 1000 triples from RANDU and plotted them using CLOUDS (Exhibit 1). A rotation in the X, Y plane was performed to place the

EXHIBIT 1. One thousand triples from the random number generator RANDU as plotted by CLOUDS.

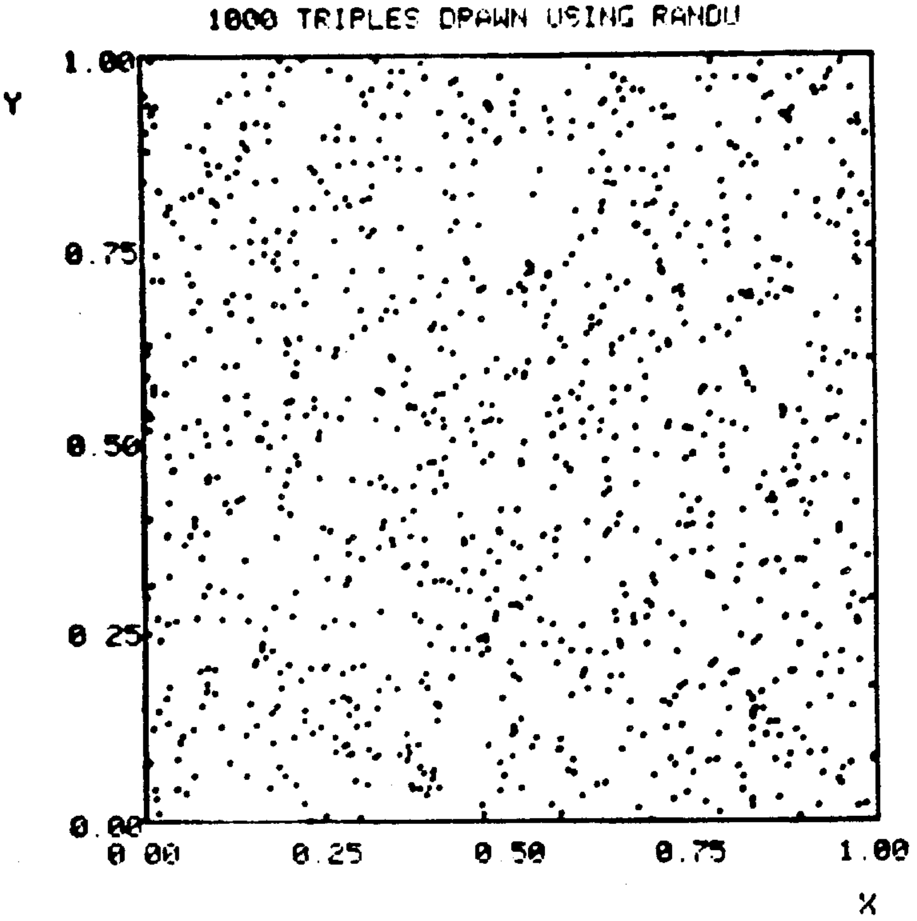
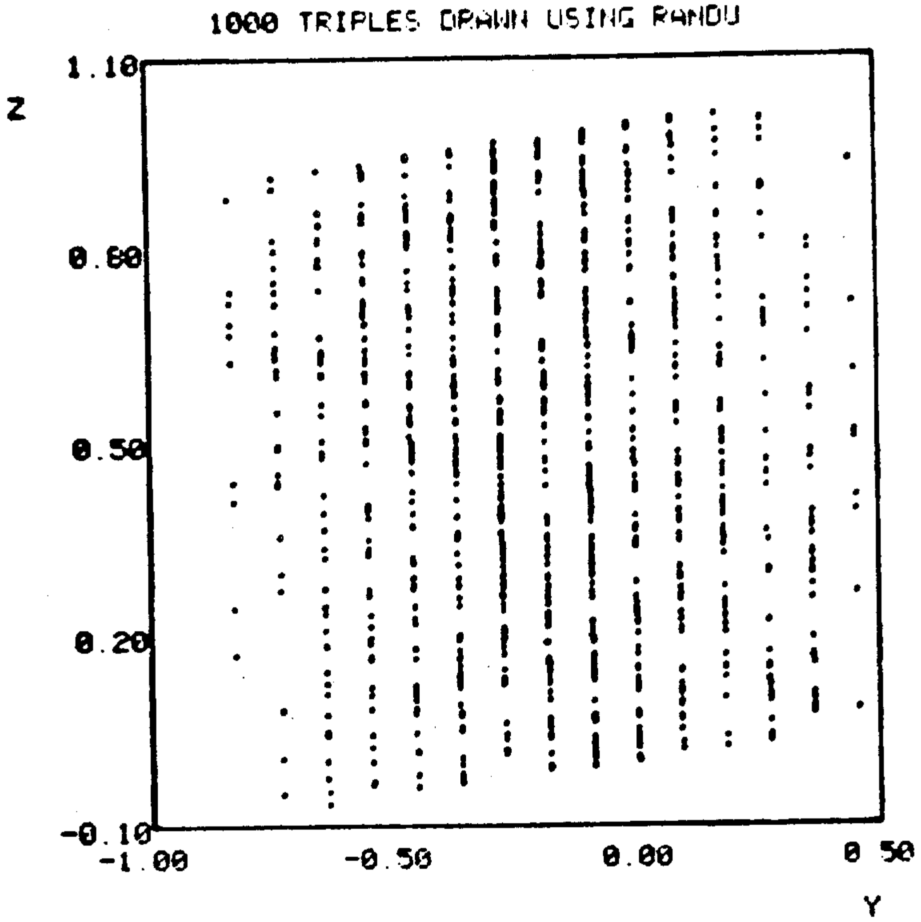


EXHIBIT 2. Rotating the triples so that the normal vector lies on the Y axis. You can see that the triples lie on 15 planes perpendicular to the normal vector.



normal vector (9, -6, 1) in the Y, Z plane. Then another rotation in the Y, Z plane makes the planes orthogonal to the Y axis (Exhibit 2). A mask is then created to isolate one plane (Exhibit 3) and the earlier rotations are reversed giving Exhibit 4. Related studies of other random number generators are planned when the projection pursuit algorithms are installed.

In a recent study of robust estimators of location, Andrews et al. [1972] listed estimated variances for 65 estimators on Cauchy samples of

EXHIBIT 3. Isolating one of the planes formed by the triples.

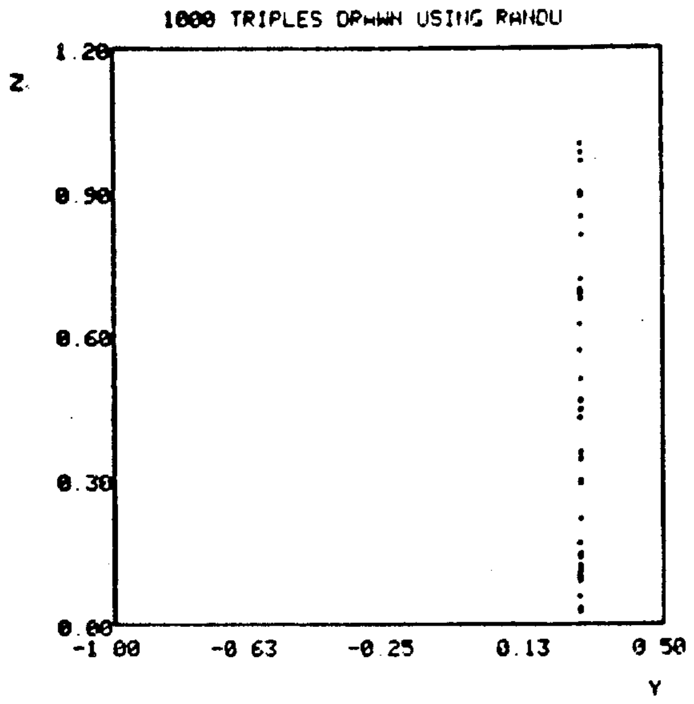
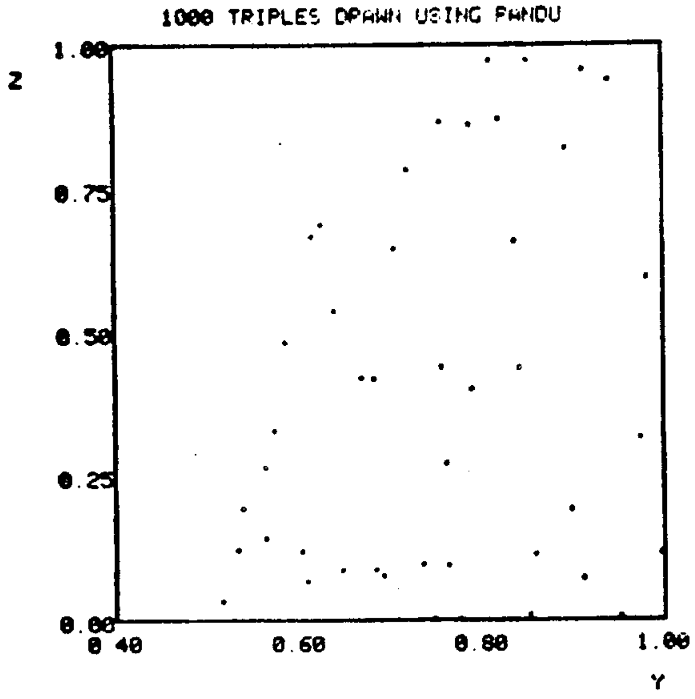


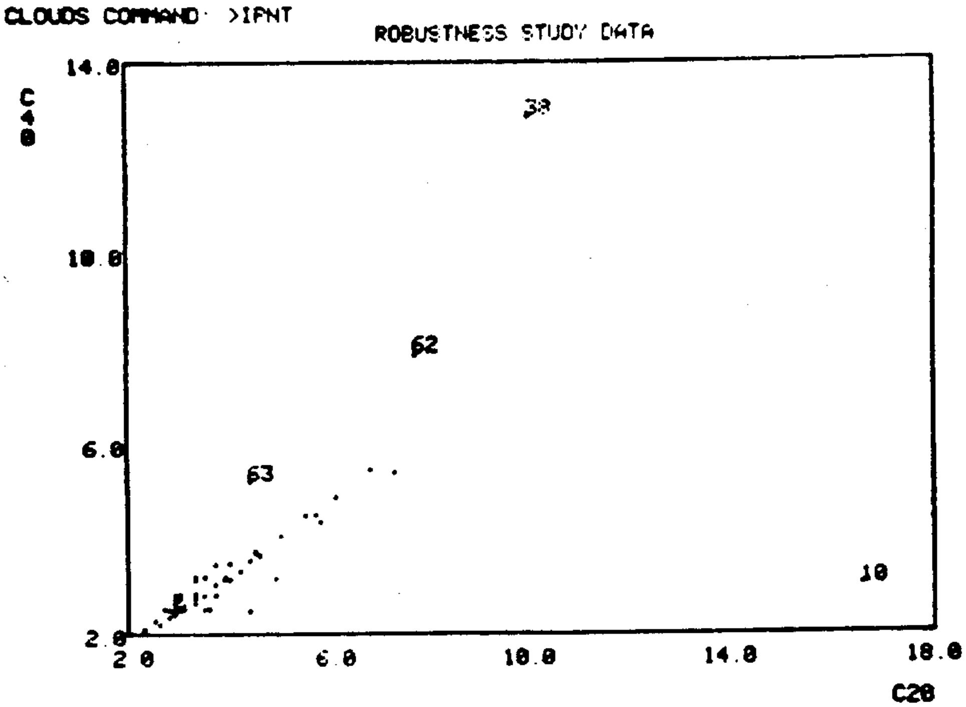
EXHIBIT 4. Triples from one plane as seen in their original state.



size 20 (C20), Cauchy samples of size 40 (C40), Student's t with three degrees of freedom on 20 samples (T3.20), and the Gaussian over independent uniform on 20 samples (SLASH20). These numbers were used to create a four dimensional point cloud with 65 points (estimators with large variances like the mean were given the value NA).

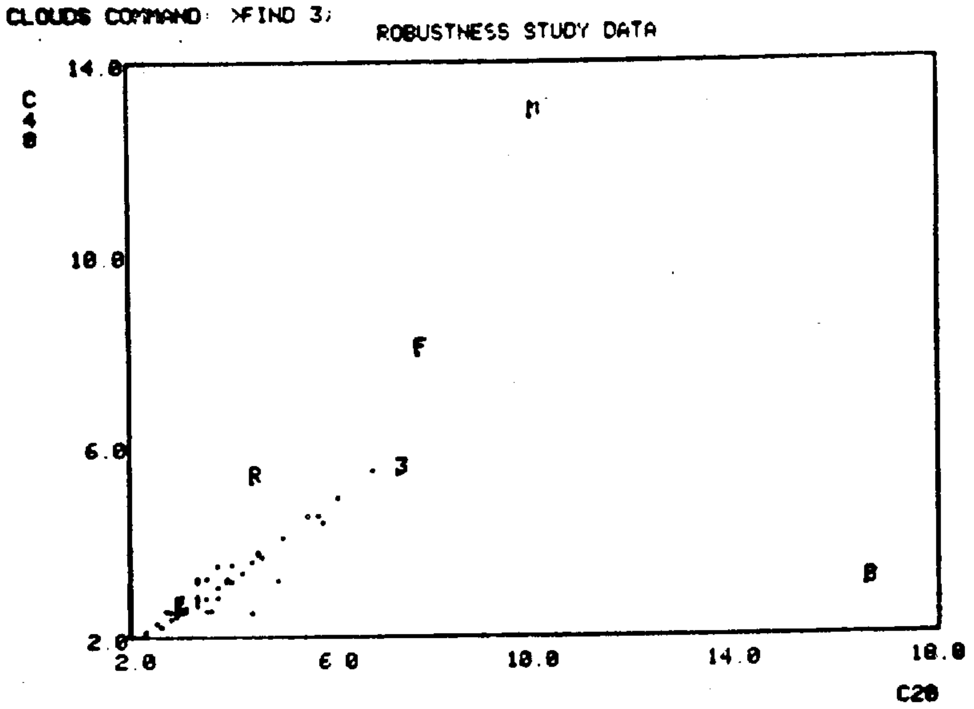
Exhibit 5 is a plot of C20 with C40. Outliers are noted and they are identified using IPNT. The outliers are then labeled so that they can be identified in future plots. Because the 10% trimmed mean (#3) was of interest we used the FIND command to locate it in Exhibit 6, but it was not labeled for future reference. The ZOOM command was used in Exhibit 7 to get a full screen picture of the lower left portion of Exhibit 6. Finally we put four projections on the screen at once in Exhibit 8.

EXHIBIT 5. Variances for 65 robust estimators as plotted by CLOUDS. Outliers are identified.



```
CLOUDS COMMAND >IPNT
```

EXHIBIT 6. Labelling outliers of variances for robust estimators.



```
CLOUDS COMMAND: >FIND 3;
```

EXHIBIT 7. Closeup of robust estimators.

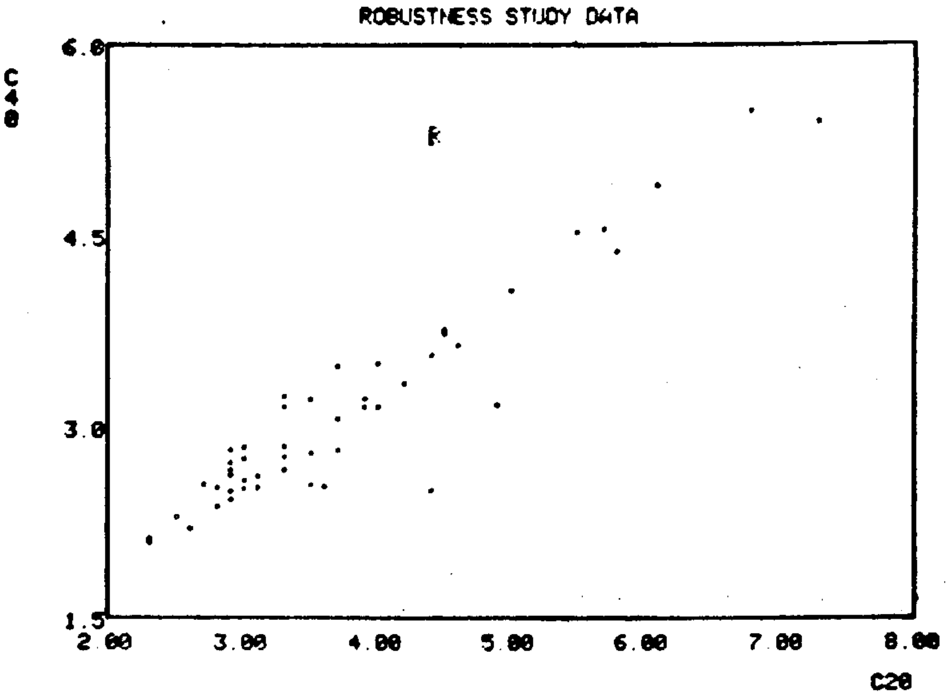
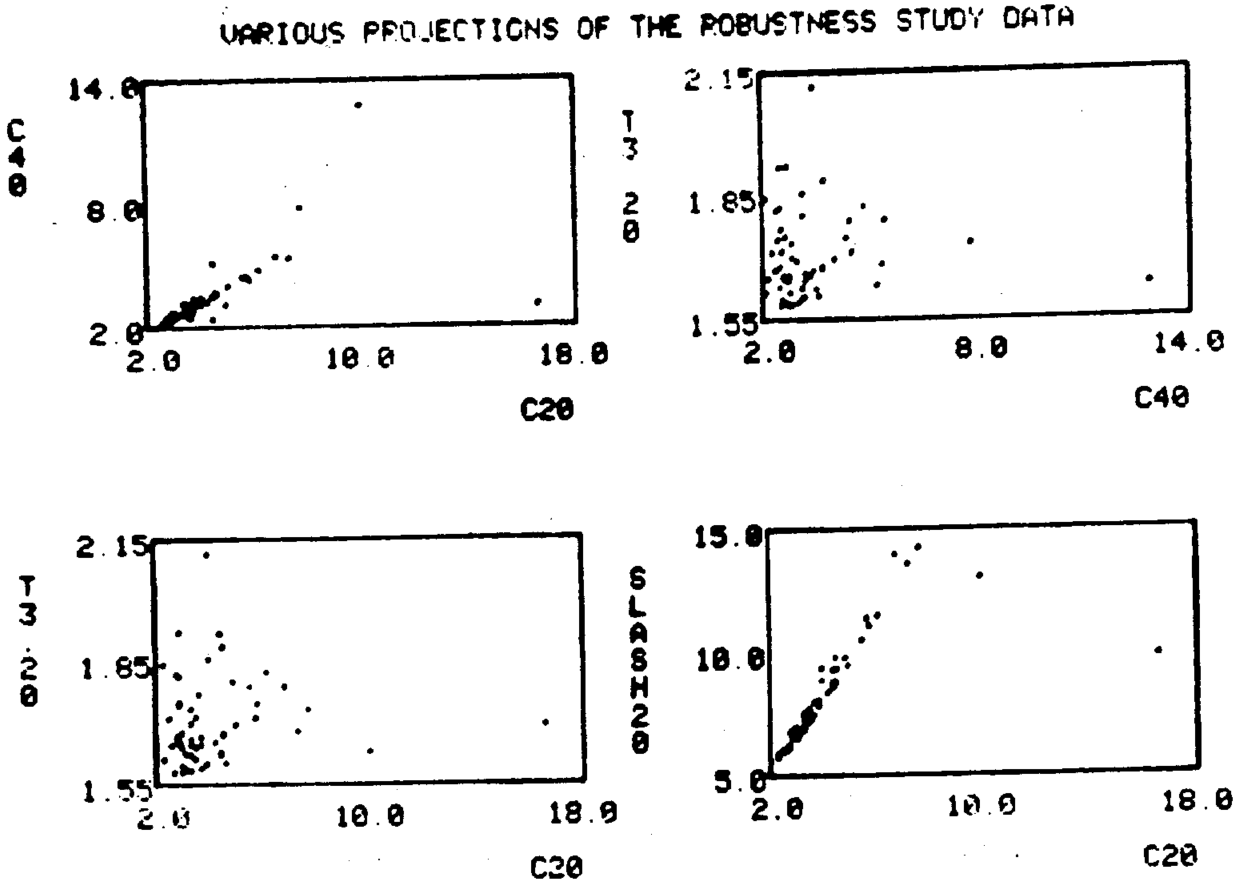


EXHIBIT 8. Four projections of variances for robust estimators on screen at once.



The next examples show how CLOUDS has proved to be a useful form of graphic language.

The robustness study data mentioned above was clustered using a hierarchical clustering program provided by Don Olivier. A macro PLTREE uses CLOUDS commands to obtain a plot of the tree representing the clusters (Exhibit 9). Tree endpoints can be identified (IPNT) and other endpoints located using the FIND command. The ZOOM command provides a way to examine a portion of the tree in more detail (Exhibit 10).

EXHIBIT 9. Using CLOUDS to plot a tree representing hierarchical clusters.

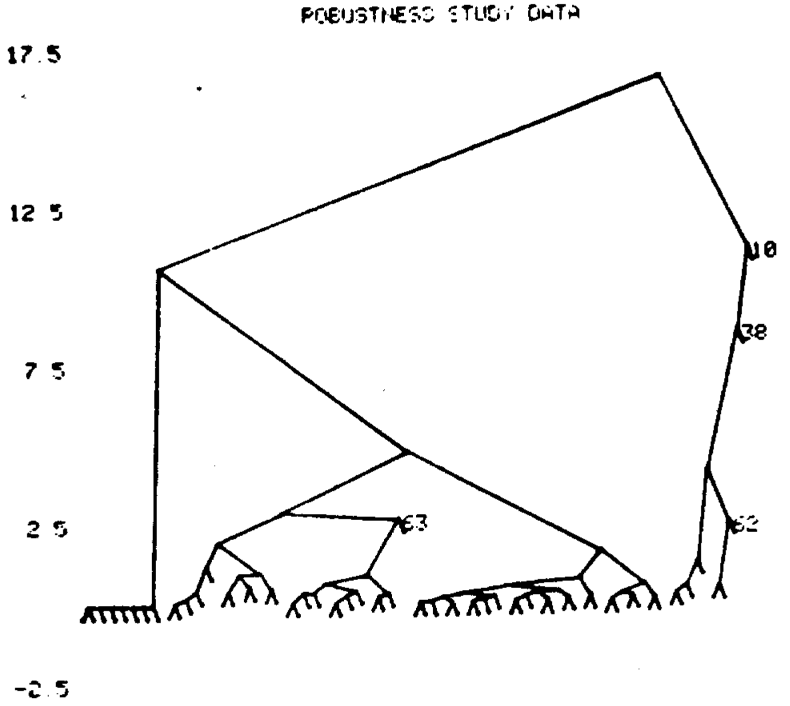
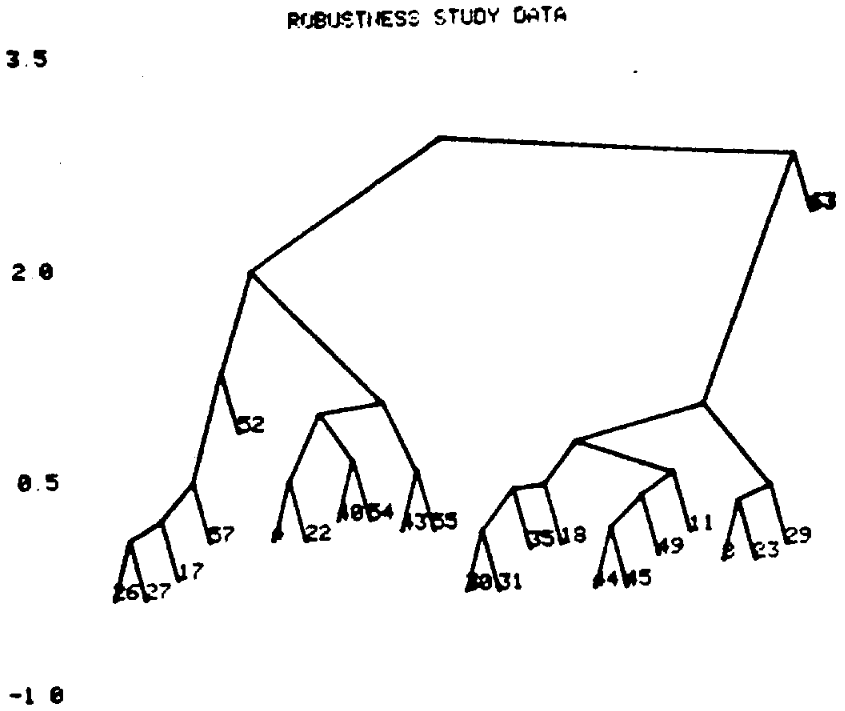


EXHIBIT 10. Closeup of a portion of a tree representing hierarchical clusters.

```
PLTREE COMMAND: >C FIND ALL.
```



Recently, there has been a great deal of attention focused on the ridge regression techniques of Hoerl and Kennard [1970]. The ridge estimates for a centered and scaled linear model are given by

$$\hat{\beta} = (X^TX + kI)^{-1}X^TY$$

where $\hat{\beta}$ is $p\times 1$, X is $n\times p$, and Y is $n\times 1$. The vector $\hat{\beta}$ can be plotted as a function of k. An example is given in Exhibit 11 for a set of data due to Longley [1967]. Such a plot may aid in choosing k, but we can at least look at the sensitivity of $\hat{\beta}$ to changes in k and get some idea of the difficulties that may be caused by a nearly singular X^TX matrix.

Huber [1973] has proposed several ways to perform robust regression. One of these is to replace the least squares loss function (or minus the log of the Gaussian likelihood) with

$$\rho(t) = \begin{cases} t^2 & |t| \leq k \\ 2k|t| - k^2 & |t| > k \end{cases} .$$

By varying k a range of estimates for β can be obtained including least squares ($k=\infty$) and least absolute deviations ($k\doteq 0$). Exhibit 12 plots $\hat{\beta}$ against $r = 1/1+k$.

Both the ridge and Huber traces are produced via a macro of CLOUDS commands which provide the special symbols, axis scales, and ability to overlay the plot of each coefficient. A special option in OUTOPT allows points to be connected by lines.

EXHIBIT 11. Ridge trace of six coefficients. Produced by a macro formed from CLOUDS commands. Six plots are overlayed.

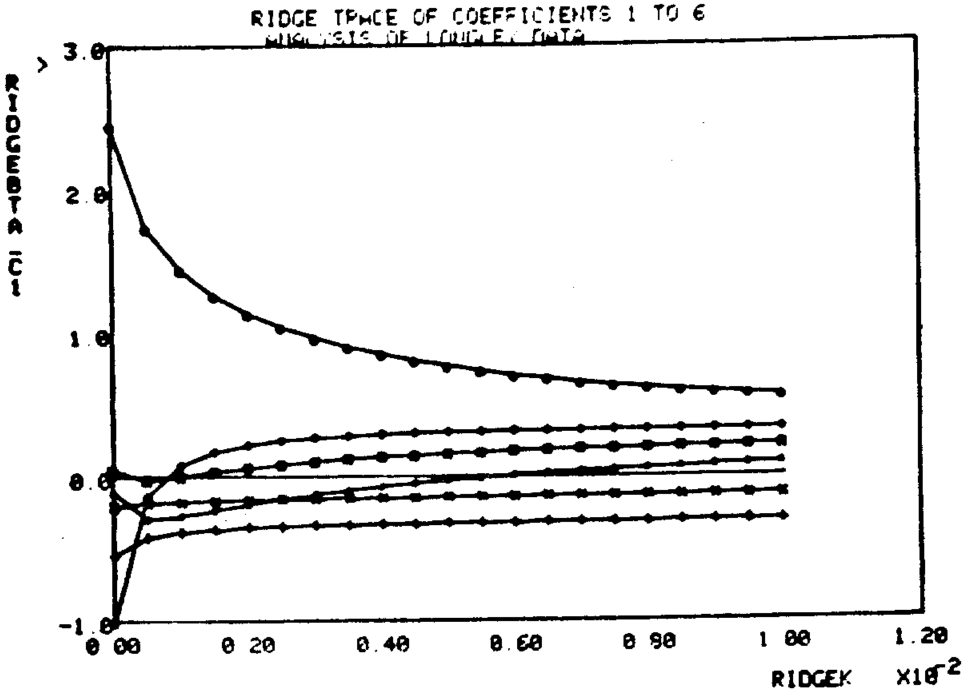
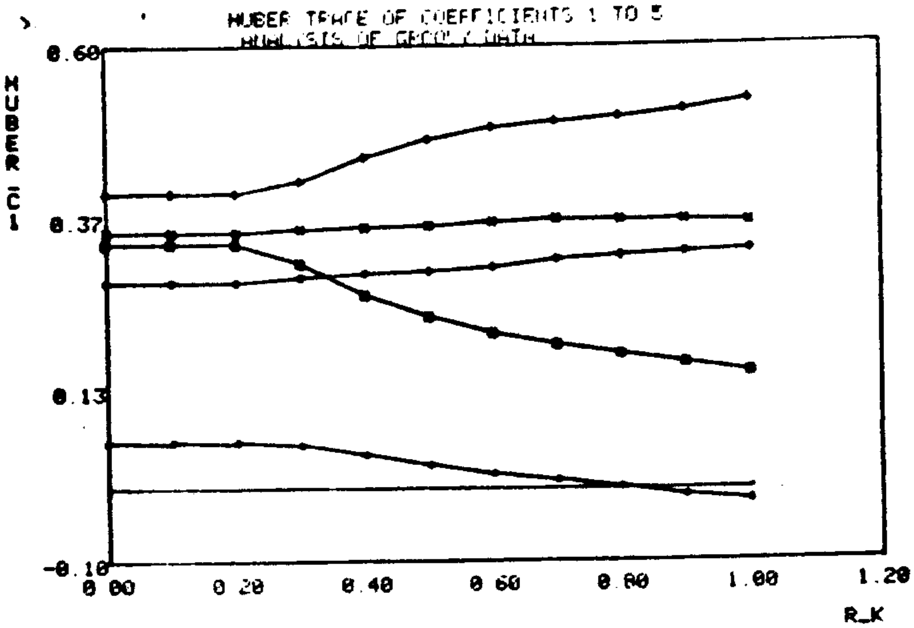


EXHIBIT 12. Huber trace of five coefficients. Produced by a macro formed from CLOUDS commands. Five plots are overlayed.



The TROLL plotting package can be used with macros to make many of the plots in Tukey [1971, 1972]. For example, Exhibit 13 shows a schematic plot of the variances of 65 estimators for some of the situations used in Andrews et al. [1972]. Exhibit 14 is a plot of fit plus residuals for a two-way table. Both of these plots make extensive use of the idea that we can ask that points be connected by lines except when an NA is encountered. The NA's (missing values) are used to pick up the "pen" where appropriate.

FACES and STARS

In order to enable TROLL users to experiment with some recent proposals for plotting multivariate data, two other systems were created--FACES and STARS.

The FACES system allows the user to display multivariate data using cognitive correlates (or icons)--in this case cartoon faces as suggested by Chernoff [1973]. An example showing the variances of estimators 37 to 45 from the robustness data is contained in Exhibit 15.

Up to eighteen separate variables (data files) can be used to control the features of the face (width, shape, mouth position, eyebrow slant, etc.). Separate scaling is allowed and any particular feature can be set to a fixed value or defaulted to a standard value. Features can be completely removed in order to speed up plotting if fewer than eighteen variables are needed. A maximum of nine faces can be plotted at one time on the CRT. For large amounts of data, the feature selection can be done on the scope with subsets of the data, and then offline commands can be used to plot all of the data.

EXHIBIT 13. Schmatic plot of variances of 65 estimators. Produced by a macro using CLOUDS commands.

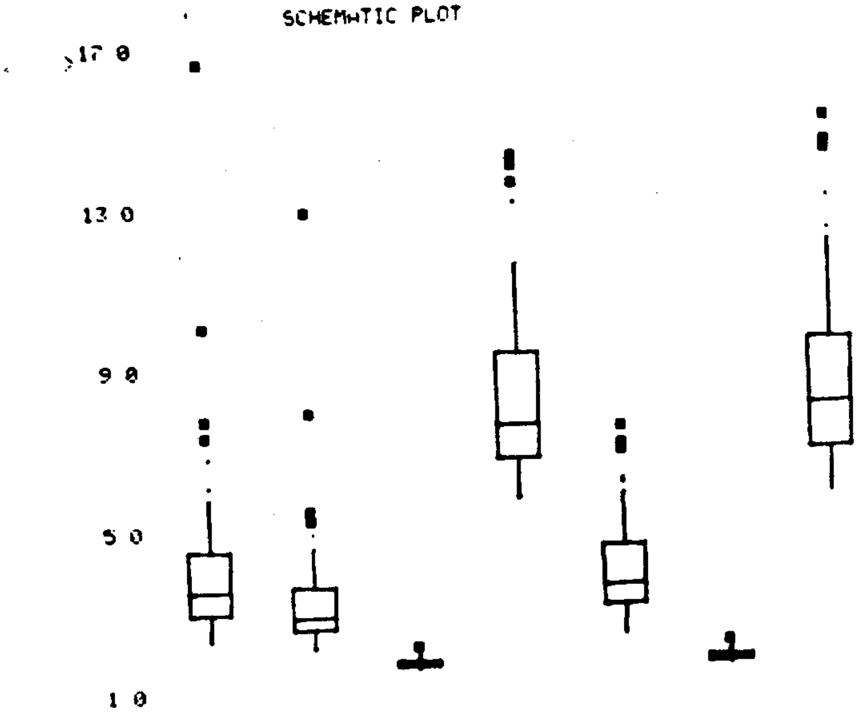


EXHIBIT 14. Plot of fit plus residuals for a two-way table. Produced by a macro using the CLOUDS commands.

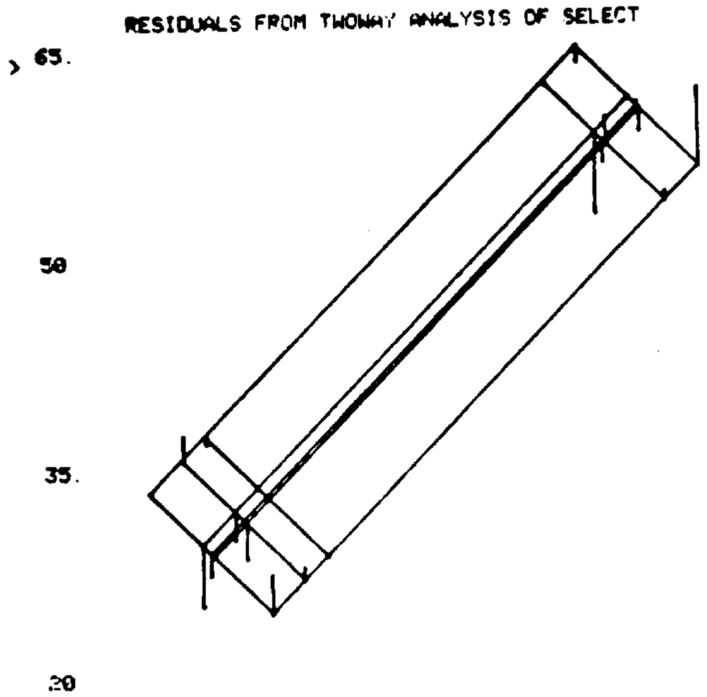
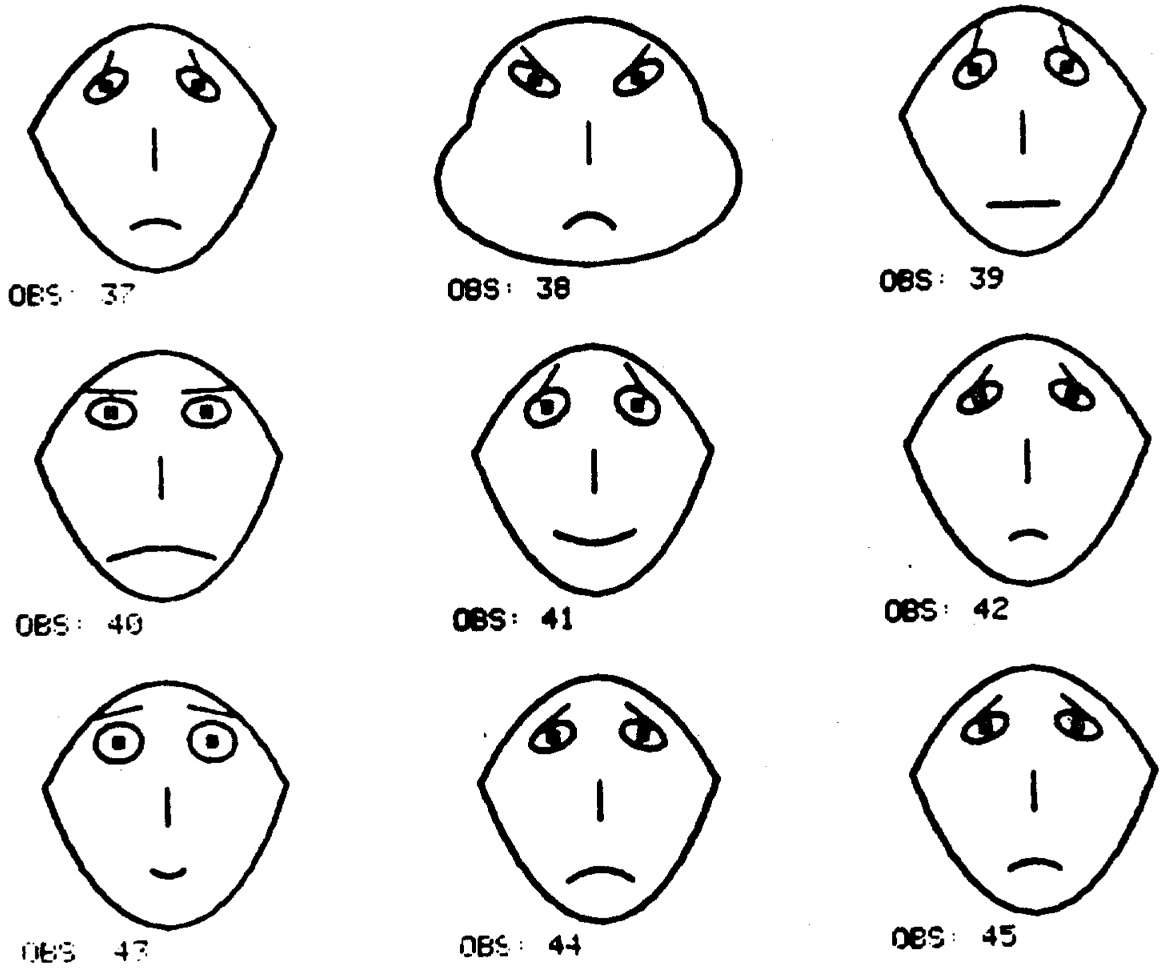


EXHIBIT 15. Variances for robust estimators as plotted by FACES.



The FACES system can be used separately or as a part of the CLOUDS system so that point clouds and faces may be used to view the same data before and after transformation by CLOUDS commands.

The STARS system is a set of commands to implement the circular graphs suggested by Friedman et al. [1972] and Kolence and Kiviat [1973]. Since starplots use mainly straight lines they can be produced quickly on the scope, an advantage over the faces approach to multivariate data.

Summary

What we have described is an attempt to make graphics a routine part of data analysis and model building. The system is aimed at the broad spectrum of researchers in economics, management science, and statistics. It is not based on expensive hardware and is widely available. We hope that user comments as well as new research on graphics will make it possible to use even more effectively the capabilities of inexpensive graphic terminals. Perhaps we will someday exhaust the possibilities of storage tube graphics, but we have a long way to go.

Appendix: List of CLOUDS Commands

Entering and Leaving the Task

CLOUDS	Initialize the task.
DEFSPACE	Specifies the input DATA files, range of observations, and the scales.
QUIT	Terminates the task.

Point Order

ORDER	Changes the number of points plotted and the order in which points are put on a plot.
SORDER	Saves the current order.
LORDER	Loads an order vector.

Plotting on Projection Planes

PPLANE	Establishes a projection plane.
SCAT	Draw a scatter plot on the established projection plane.

Rotation

RPLANE	Establishes a rotation plane.
ANGLE	Establishes a base rotation angle.
RCENTER	Establishes a rotation center.
ROTATE	Rotates the point cloud.
RTSCAT	Combines ROTATE and SCAT.

Superimposition

ADDSCAT	Adds a plot of the current position to the plot already on the scope without printing new scales or labels.
ARTSCAT	Combines ROTATE and ADDSCAT.
PLTROT	Rotates the point cloud a specified number of times and adds a plot at each rotation.
SCAT NOERASE	Performs a SCAT without first erasing the screen.

Working With Individual Points

APNT	Adds a point to the point cloud.
DPNT	Deletes a point from the point cloud.
MPNT	Moves a point in the point cloud.
IPNT	Identifies a point using the cursor.
FIND	Identifies a point using keyboard input.
LABEL	Labels a point.
DLABEL	Deletes a label.

Masking

MASK	Excludes portions of the current or initial point cloud from further plots.
DMASK	Deletes a mask.
LISTMASK	Lists all masks currently in effect.
PMASK	Prints a mask.
SMASK	Saves a mask.
LMASK	Loads a mask file.
ZOOM	Gives a closeup of a section of a plot.

Saving Position Values

SAVE	Replaces the initial point cloud with the current point cloud.
INIT	Replaces the current point cloud with the initial point cloud.
SVECTOR	Saves the current point cloud in external files.
SRMAT	Saves the current rotation matrix.
LRMAT	Loads a rotation matrix.
PRMAT	Prints the current rotation matrix.
SCENTER	Saves the current rotation center.
LCENTER	Loads a rotation center.
RCENTER	Prints the current rotation center.

Utilities

PLTVECTOR	Draws a plot of CLOUDS dimensions with the observation number along one axis.
MARK	Changes the symbol used for a point in plotting.
NORM	Normalize a file.
SCALE	Specifies scales.
TITLE	Specifies titles.
OUTOPT	Controls plot size, location, and labeling.
LKOUTOPT	Checks output options.
PINFO	Prints information about the current status of CLOUDS.
HARDCOPY	Produces a hard copy of the screen.
ERASE	Erases the screen.

REFERENCES

Andrews, D. F. et al. (1972). *Robust Estimates of Location*. Princeton University Press, Princeton, New Jersey.

Chernoff, H. (1973). Using Faces to Represent Points in k-Dimensional Space Graphically. *Journal of the American Statistical Association*, **68** 361-368.

Friedman, A. P., E. J. Farrell, R. M. Goldwyn, M. Miller, and J. H. Siegel (1972). A Graphic Way of Describing Changing Multivariate Patterns. *Proceedings of the Computer Science and Statistics 6th Annual Symposium on the Interface*, Berkeley, California.

Friedman, J. H. and J. W. Tukey (1973). A Projection Pursuit Algorithm for Exploratory Data Analysis. Stanford Linear Accelerator Center Publication 1312.

Hoaglin, D. C. (1973). Personal Communication.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12** 55-68.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures, and Monte Carlo. *Annals of Statistics*, **1** 799-821.

Kolence, K. W. and Kiviat, P. J. (1973). Software Unit Profiles and Kiviat Figures. *ACM Performance Evaluation Review*, September 1973.

Longley, J. W. (1967). An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. *Journal of the American Statistical Association*, **62** 819-41.

Marsaglia, G. (1968). Random Numbers Fall Mainly in the Planes. *Proc. Nat. Acad. of Sciences*, **60** 25-28.

Tukey, J. W. (1971). *Exploratory Data Analysis*, Limited preliminary edition, Vol. I, II, III. Addison-Wesley, New York.

Tukey, J. W. (1972). Some Graphic and Semi-Graphic Displays. T. A. Bancroft, Ed. *Statistical Papers in Honor of George W. Snedecor* Ames, Iowa: Iowa State University Press, pp. 293-316.