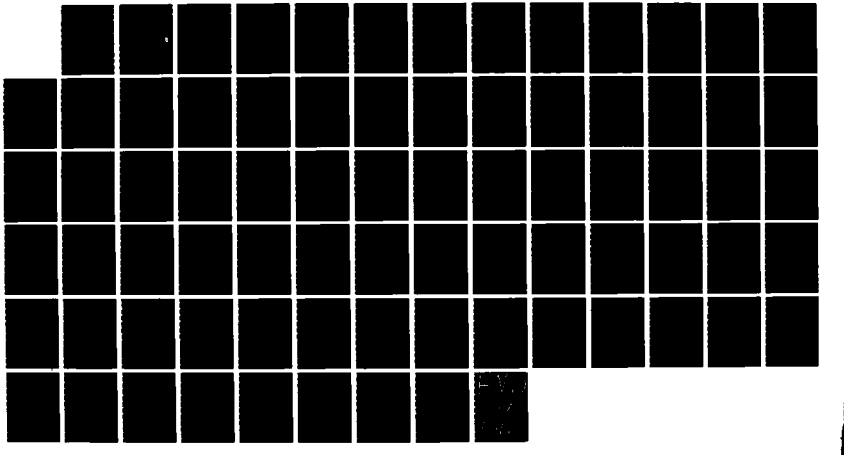
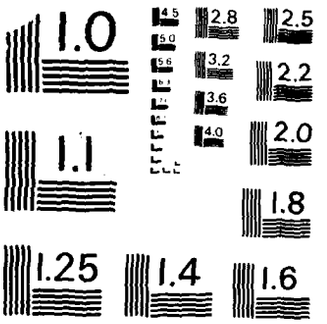
AD-A175 146 BOXPLOTTED TABLES AND OTHER GRAPHICAL TECHNIQUES FOR EXPLORATORY DATA ANALYSIS(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA J M ISUSI SEP 86 1/1

UNCLASSIFIED F/G 9/2 NL





NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

BOXPLOTTED TABLES AND OTHER
GRAPHICAL TECHNIQUES
FOR EXPLORATORY DATA ANALYSIS

by

Juan M. Isusi

September 1986

Thesis Advisor: P. A. W. Lewis

Approved for public release; distribution is unlimited.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a NAME OF PERFORMING ORGANIZATION Naval Postgraduate School	6b OFFICE SYMBOL (*If applicable*) Code 55	7a NAME OF MONITORING ORGANIZATION Naval Postgraduate School
6c ADDRESS (*City, State, and ZIP Code*) /Monterey, California 93943-5000		7b. ADDRESS (*City, State, and ZIP Code*) Monterey, California 93943-5000
8a NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (*If applicable*)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (*City, State, and ZIP Code*)		10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO

11 TITLE (*Include Security Classification*)
BOXPLOTTED TABLES AND OTHER GRAPHICAL TECHNIQUES FOR EXPLORATORY DATA ANALYSIS

12 PERSONAL AUTHOR(S)
Isusi, Juan M.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (*Year, Month, Day*)	15 PAGE COUNT
Master's Thesis	FROM _____ TO _____	1986 September	76

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	Boxplots, Scatter Plots, Draftsman Plots, Star Plots, Profile Plots, APL, GRAFSTAT

19 ABSTRACT (*Continue on reverse if necessary and identify by block number*)

This thesis presents several interactive computer programs for the analysis of multivariate data. A special case is that of panel data; multiple time series of short length. The first program, BOXPLOTAB, handles this type of multivariate data; it is an enhancement on an existing graphical technique for exploratory data analysis know as BOXPLOTS. The program works by appending boxplots as column dividers in a table of the raw data which originates the box plots. This combination of the raw data and the graphical representation of that data improves the understanding of the characteristics of the data in exploratory and descriptive applications; differencing and tracing of data through the table is alos implemented. This thesis also presents and explores the use of other graphical techniques for exploratory data analysis of multivariate data such as STAR plots, PROFILE plots, CODED SCATTER plots and CODED DRAFTSMAN plots. These techniques are examined and implemented in a series of computer programs which produces these graphical displays. A technical description of each computer program is presented

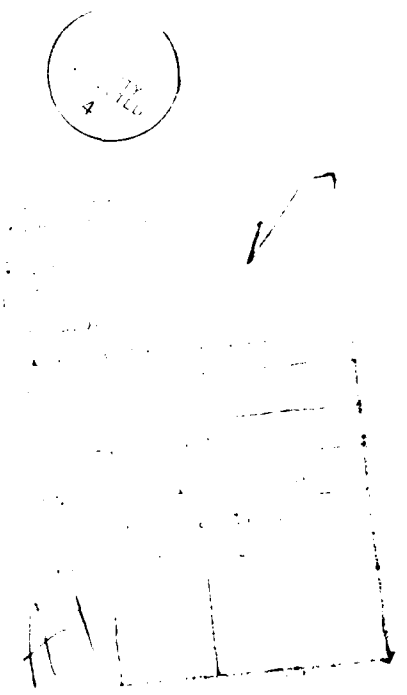
20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL P. A. W. Lewis	22b TELEPHONE (*Include Area Code*) (408) 646-2283	22c OFFICE SYMBOL Code 55Lw

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

19. ABSTRACT

and user implementation procedures are discussed. The programs are implemented in APL and run in conjunction with the experimental IBM APL Graphics program GRAFSTAT. To demonstrate the use of these techniques, an analysis is conducted on several sets of multivariate data.



Approved for public release; distribution is unlimited.

Boxplotted Tables and Other
Graphical Techniques
for Exploratory Data Analysis

by

Juan M. Isusi
Major, Peruvian Air Force
B.S., Air Force Academy (Peru), 1973

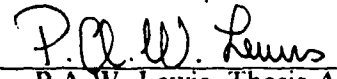
Submitted in partial fulfillment of the
requirements for the degree of

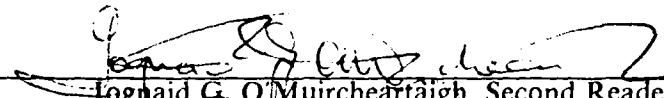
MASTER OF SCIENCE IN OPERATIONS RESEARCH

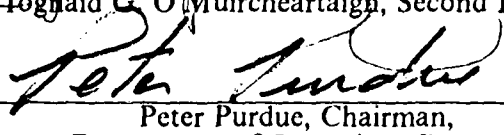
from the

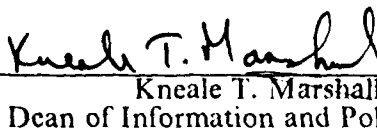
NAVAL POSTGRADUATE SCHOOL
September 1986

Author: 
Juan M. Isusi

Approved by: 
P.A.W. Lewis, Thesis Advisor


Iognaid G. O'Muircheartaigh, Second Reader


Peter Purdue, Chairman,
Department of Operations Research


Kneale T. Marshall,
Dean of Information and Policy Sciences

ABSTRACT

This thesis presents several interactive computer programs for the analysis of multivariate data. A special case is that of panel data; multiple time series of short length. The first program, BOXPLOTAB, handles this type of multivariate data; it is an enhancement on an existing graphical technique for exploratory data analysis known as BOXPLOTS. The program works by appending boxplots as column dividers in a table of the raw data which originates the box plots. This combination of the raw data and the graphical representation of that data improves the understanding of the characteristics of the data in exploratory and descriptive applications; differencing and tracing of data through the table is also implemented. This thesis also presents and explores the use of other graphical techniques for exploratory data analysis of multivariate data such as STAR plots, PROFILE plots, CODED SCATTER plots and CODED DRAFTSMAN plots. These techniques are examined and implemented in a series of computer programs which produces these graphical displays. A technical description of each computer program is presented and user implementation procedures are discussed. The programs are implemented in APL and run in conjunction with the experimental IBM APL Graphics program GRAFSTAT. To demonstrate the use of these techniques, an analysis is conducted on several sets of multivariate data.

THESIS DISCLAIMER

The reader is cautioned that computer programs developed in this research may not have been exercised for all cases of interest. While every effort has been made, within the time available, to ensure that the programs are free of computational and logic errors, they cannot be considered validated. Any application of these programs without additional verification is at the risk of the user.

TABLE OF CONTENTS

LIST OF FIGURES

I. INTRODUCTION

A. PREFACE

One of the main problems in experimental statistics and experimental design is the exploratory analysis of raw data. This problem is greatly enlarged when the data presented to the statistician comes from unknown multiple populations or so called multivariate data. A special case is that of panel data; multiple time series of short length. The initial purpose of the data analysis is to try to capture the most important distributional characteristics of the data such as the range, location and spread of the data points. For the experimental statistician the main tool available for the analysis of the data is the graphical display of the marginal distributions of the data in order to visualize and gain better understanding of these characteristics and to compare them against those of the different populations. Following this, interactions or dependencies can be examined, and this is the domain of multivariate data analysis.

B. PURPOSE

The purpose of this thesis is twofold: first, to add to the tabular display of the original multivariate data an existing graphical technique known as the BOXPLOT (see [Ref. 1]). This addition can be done in several alternating ways and is done in order to better understand the populations and the relations between the different populations. The second purpose of the thesis is to make available different computer programs to exploit several other enhanced statistical graphical techniques for multivariate data analysis. These techniques are: STAR plots, PROFILE plots, CODED SCATTER plots and CODED DRAFTSMAN plots.

C. BACKGROUND

Presently, the BOXPLOT technique is one of the most common graphical techniques used by data analysts, both outside and at the Naval Postgraduated School (NPS). There are different software packages that provide these plots, some of which are in the experimental IBM APL GRAFSTAT program and some in the IBM Mainframe NONIMSL library.

One of the most important limitations of this graphical display technique is the absence of the raw data in the display; this absence is critical in the special case of

multiple box plots for the comparison of multiple sets of data. This would provide an inmediate identification of peculiar characteristics of the data, such as pin-pointing the outliers and the variability and/or relation of a sample datum with respect to other samples. Once the BOXPLOTS are displayed on the screen (or printed on a graph), the analyst has to go back to the original raw data in order to identify these data points. This limitation is overcome by the new technique presented herein, which is called the BOXPLOTTED TABLES. This technique has already been implemented, and can be used in the NPS IBM 3033 computer using an APL (A Programming Language) program, which make use of the graphical capabilities of the IBM experimental GRAFSTAT software package. In GRAFSTAT, an interactive technique for identifying odd or outlying points is given. This implementation highlights the importance of a technique for data point identification. However, one does not always have access to such a program and the ability to do this identification on a printed page is important to a data analyst. The BOXPLOTTED TABLES do precisely this. Note too that a primary concern in multivariate data analysis is to get as much information on a two dimensional page as possible. Thus having tabular and distributional data together on one page, as in the BOXPLOTTED TABLES, is a step in this direction.

There are other graphical techniques commonly used by data analysts such as : SCATTER plots, STAR plots, PROFILE plots, CODED SCATTER plots and CODED DRAFTSMAN plots (see [Ref. 1]). These techniques are mainly used to enhance the interpretation and understanding of displayed multivariated data. Out of these, the DRAFTSMAN and SCATTER plot techniques (without coded symbols) are the only ones available at NPS up to this point. These other techniques are used to display the data points in many different forms, giving a new perspective to the interpretation of the original data.

This thesis present a group of APL functions that will make possible the use of these graphical techniques to the experimental statistician at the NPS. Various examples that show how to use this software to analyze and graphically display sample data will be shown in the following chapters of this thesis.

D. ORGANIZATION

This thesis consists of three main blocks. The first one, Chapter Two, is dedicated to explain the technical aspects of these graphical techniques; the mathematical and

statistical attributes of each technique are treated in this chapter. The second block is intended to introduce the user to the APL software code used to implement these techniques. In Chapter Three, both the user and system requirements are explained; and for the more technical oriented reader, a listing of the APL code is listed in Appendix A. In addition, several examples of program execution are listed in Appendix B. The last block, composed of Chapter Four and Appendix C, is dedicated to the exploratory analysis of several sets of sample data to demonstrate some of the potential applications of these graphical approaches to statistical analysis.

II. GRAPHICAL TECHNIQUES

A. BOXPLOTTED TABLES

1. Overview

The BOXPLOT graphical technique was first conceived by Tukey as a method to display an almost one-dimensional summary of the distribution characteristics of a set of data, Chambers [Ref. 1] provides an excellent analysis of this technique. This display shows some of the most prominent characteristics of the sample distribution such as the median, mean, the inter-quartile range and the outliers, if there are any. In the case where the sample comes from multivariate data, the BOXPLOT is used not only to show the individual characteristics of each subsample, but, in addition, to compare the behavior of these characteristics with respect to other samples (see [Ref. 1: p. 89]). Figure 2.1 shows a BOXPLOT display. The BOXPLOT's almost one-dimensional character, as opposed to the two-dimensional character of the familiar histogram, facilitates comparison of the marginal properties of multivariate data sets.

One of the limitations of the BOXPLOT is that of the identification of specific values of interest such as outliers; if the identification of the outliers is the prominent feature, the statistician must make reference to the original data in order to identify which data point the outlier correspond to.

A solution to this limitation, suggested by Professor P.A.W. Lewis in an unpublished work, is to show the original data and the BOXPLOT in the same graphical tabular display. In this case, the BOXPLOTS are shown as dividers of the original tabulated data, so that aberrations are readily apparent and checkable (see Figure 2.2). This technique clearly requires the availability of high resolution graphics and a sophisticated plotting and data manipulation package. This requirement is met by the experimental APL GRAFSTAT program from IBM Research which is being used at the NPS on a test bed basis.

2. Technical details of BOXPLOTTED tables

In the BOXPLOT the top and bottom of the rectangle represent the upper and lower quartile of the data respectively. Therefore, the length of the rectangle represent the inter-quartile range ($Q(.75) - Q(.25) = IQR$), where $Q(\alpha)$, for $0 < \alpha < 1$ is the α-quantile of the sample. The mean of the data sample is shown by a small

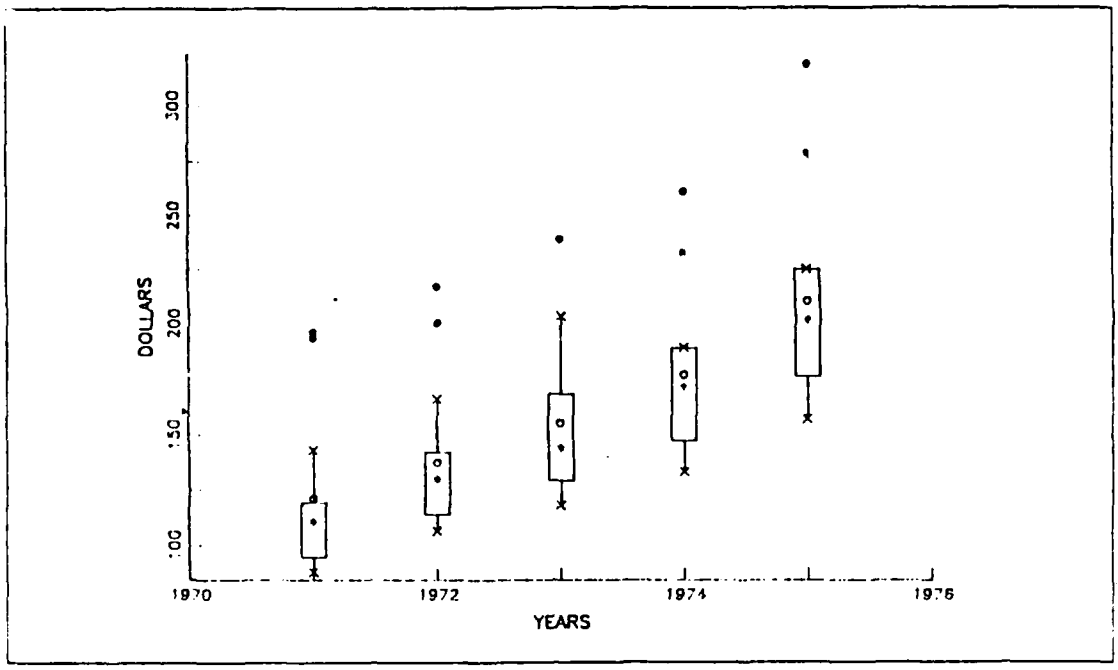


Figure 2.1 BOXPLOT of California Hospital Data (Per Capita, Hospital Expenses, Years 1971-1975, in 14 Health Service Areas).

OBSERVATION		1971	1972	1973	1974	1975
HOSPITAL NO. 4	1	100.02	217.08	239.02	261.08	310.77
HOSPITAL NO. 3	2	103.48	200.37	203.48	232.08	279.44
HOSPITAL NO. 11	3	142.08	188.08	200.11	231.01	277.38
HOSPITAL NO. 10	4	110.06	130.38	183.21	177.04	207.06
HOSPITAL NO. 9	5	110.02	133.15	160.04	160.20	183.28
HOSPITAL NO. 12	6	110.08	142.06	160.20	160.06	235.08
HOSPITAL NO. 8	7	110.04	150.00	148.04	170.08	200.12
HOSPITAL NO. 5	8	110.02	117.03	130.10	140.77	184.38
HOSPITAL NO. 7	9	100.33	130.07	144.20	170.07	200.01
HOSPITAL NO. 6	10	103.03	114.46	127.47	147.30	170.40
HOSPITAL NO. 14	11	94.83	100.27	120.41	134.04	167.02
HOSPITAL NO. 1	12	90.01	107.00	117.00	133.00	167.00
HOSPITAL NO. 13	13	92.13	100.01	141.10	171.70	202.10
HOSPITAL NO. 2	14	97.04	113.07	131.00	140.02	162.43
RANK CORR.		.9267	.7890	.8650	.8725	
MEAN		121.0014	137.4457	156.3500	177.2179	210.0857
VARIANCE		1187.8042	1184.0004	1271.8934	1850.8047	2448.3453
MEDIAN		113.2850	130.2300	146.4800	172.9380	204.8900

Figure 2.2 BOXPLOTTED Table of California Hospital Data (Per Capita Hospital Expenses, Years 1971-1975, in 14 Health Service Areas).

circle, and the median by an asterisk inside the rectangle. The solid lines going out from the top and bottom of the rectangle represent the adjacent values. These values are defined as those data points greater or equal than Q(.75) and less or equal than Q(.75) + IQR for the upper line, and those values less or equal Q(.25) and greater or equal than Q(.25) - IQR for the bottom line. Those data points that fall in the range of [(Q(.25) - IQR) , (Q(.25) - IQR*1.5)] or [(Q(.75) + IQR) , (Q(.75) + IQR*1.5)] are called outliers and are represented by small light circles. The data points that fall beyond the ranges of these outliers are called extreme outliers and are represented by small black circles. As an example, for normally distributed data, approximately 5 percent of the points should be outliers and marked with light circles and only about 0.5 percent should be extreme outliers (see [Ref. 2]).

To obtain a BOXPLOTTED table, a tabular display of the data is added to the BOXPLOT display. At the bottom of each column the estimates for the mean, median, variance and the rank correlation between that column and the next column to the right are listed. The estimators for these parameters are defined as follows :

Let X_{ij} be the entry in the i^{th} row and j^{th} column, and let n be the number of values in each column. Then

$$\text{Mean}_j = \bar{X}_j = \sum_i X_{ij} / n. \tag{2.1}$$

The Median is defined as follows. Let MID_j be n/2 if n is even, and the largest integer smaller than n/2 if n is odd. Then

$$\text{Median}_j = X^*_j(MID_j), \text{ if n is odd, or} \tag{2.2}$$

$$\text{Median}_j = X^*_j(MID_j) + X^*_j(MID_j + 1) \ / \ 2, \tag{2.3}$$

if n is even, where X^*_j represents the j^{th} column sorted in descending order. The estimator for the variance is

$$\text{Variance}_j = \sum_i (X_{ij} - \bar{X}_j)^2 / (n - 1). \tag{2.4}$$

Finally, the Spearman's ρ (RHO) Rank Correlation coefficient is defined as follows: let X_i and Y_i be two sets of data, and let $[R(X_i)]$ and $[R(Y_i)]$ be the ranks of X_i and Y_i as compared to the others X and Y values respectively, for $i = 1,2,...,n$. $R(X_i) = 1$ if X_i is the smallest of $X_1, X_2, ..., X_n$, and $R(X_i) = 2$ if X_i is the second smallest, and so on, with rank n being assigned to the largest of the X_i. The same applies for $R(Y_i)$. When assigning the ranks, if a tie is found (when two or more sample values are exactly equal to each other, they are tied), assign to each tied value the average of the ranks that would have been assigned if there had been no ties (see [Ref. 3: p. 252]). Then the estimator is

$$\rho_j = \sum_i [R(X_i) - (n+1)/2][R(Y_i) - (n+1)/2] \,/\, [(n(n^2 - 1))/12], \tag{2.5}$$

if there are no ties in the data. If there are ties in the data, then the estimator is

$$\rho_j = \frac{\sum_i R(X_i)R(Y_i) - n((n+1)/2)^2}{[\sum_i R(X_i)^2 - n((n+1)/2)^2]^{1/2}\ [\sum_i R(Y_i)^2 - n((n+1)/2)^2]^{1/2}}. \tag{2.6}$$

Once the BOXPLOTTED tables for the original data are obtained, it is then possible to join with lines, values with the same rank (order in magnitude) within their corresponding columns. The statistician may select to use this technique when it is desirable to study any possible relation with respect to time among variables (as in the case of multiple short time series), or with respect to magnitudes (as in the case of data with mixed qualitative and quantitative information).

If the data are ordered (in descending order on the first column), then the outlier in the first boxplot corresponds to the first value in the table. However, this ordering may be lost in the second column, so that it is not clear that an outlier in the second boxplot corresponds to the first value in the second column and so forth. Thus if a line is drawn linking the largest value in each column, two extreme results are informative. If the line is straight (or almost straight), it means that the outliers in successive columns come from the same source. If the line wanders, then there is no structural relationship along columns (or time, if one considers panel data). As an example, the study of health care expenses is a good prototype of multiple short time series analysis.

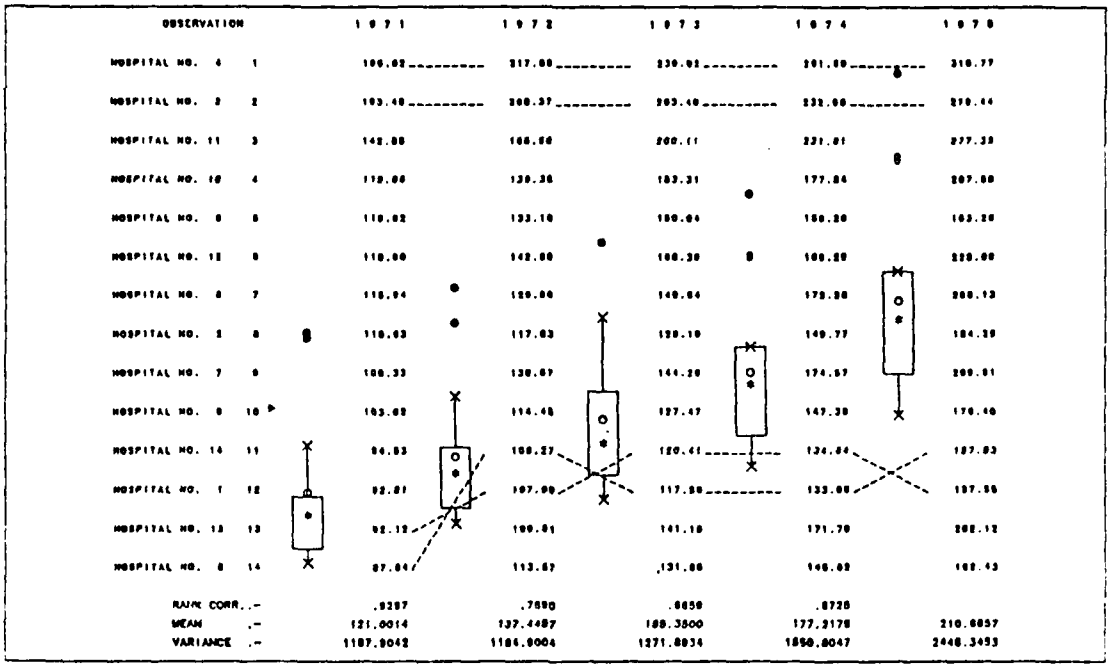


Figure 2.3 BOXPLOTTED Table with Joining Lines
(Per Capita, California Hospital Expenses, Years 1971-1975).

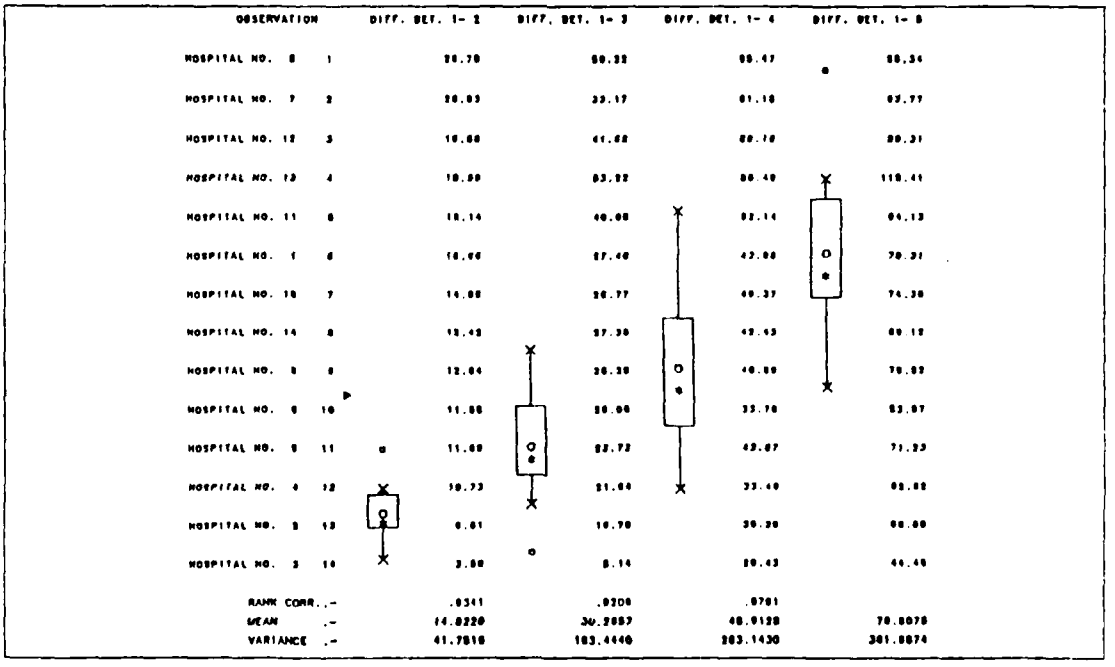


Figure 2.4 BOXPLOTTED Table of Relative Differences, First col.
(Per Capita, California Hospital Expenses, Years 1971-1975).

In Figure 2.2 one could draw a line joining the highest or lowest expenses through time and see to which hospitals they belong. The reader could now make reference to Figure 2.3. In this figure one could see the transition of health care cost for the most and least expensive health service areas in California through the period of 1971 to 1975. A plot with the connections is shown in Figure 4.2.

In addition to this option, the user can display the differences between the values of the columns. These differences could be relative to the first column (base column) or with respect to the previous one. When the statistician is dealing with panel data, it is desirable to study the trend of relative (or absolute) rate of change in the data points. Again, in the analysis of health care expenses, one may want to study the relative (or absolute) rate of change in this cost through a given period. In Figure 2.4, it is possible to infer that the relative change of health care expenses is not linear within the period of study. This inference would be enhanced by a plot of differences, as is done in Chapter Four. It is also possible to readily identify those health services areas that had the extremes rates of change. An analysis of this data is presented in Chapter Four.

B. STAR PLOTS

1. Overview

In working with multivariate data, one of the key problems is how to represent more than two variables (dimensions) in a single display. There are several graphical approaches to deal with this problem, as mentioned in Chambers [Ref. 1]. Four of these techniques are treated in this thesis : STAR plots, PROFILE plots, CODED SCATTER plots and CODED DRAFTSMAN plots.

In the STAR plot each subpopulation is displayed by a *star* in which each arc (or ray) represents a variable of interest. The value of each variable is coded by the length of the corresponding arc; to avoid overlapping between arcs, these are portrayed symmetrically about the origin. This can be seen in Figure 2.5 and Figure 2.6, in which some characteristics of automobile data are displayed (a complete description of the data is presented in Chapter Four). In Figure 2.5, twelve variables of interest are assigned to the rays of the star (i.e., price, length, etc.). In Figure 2.6, the same representation is used to portray the same information but for several automobile subpopulations (models). It is now easy to graphically compare these characteristics (by the corresponding length of each ray) among the different subpopulations (models).

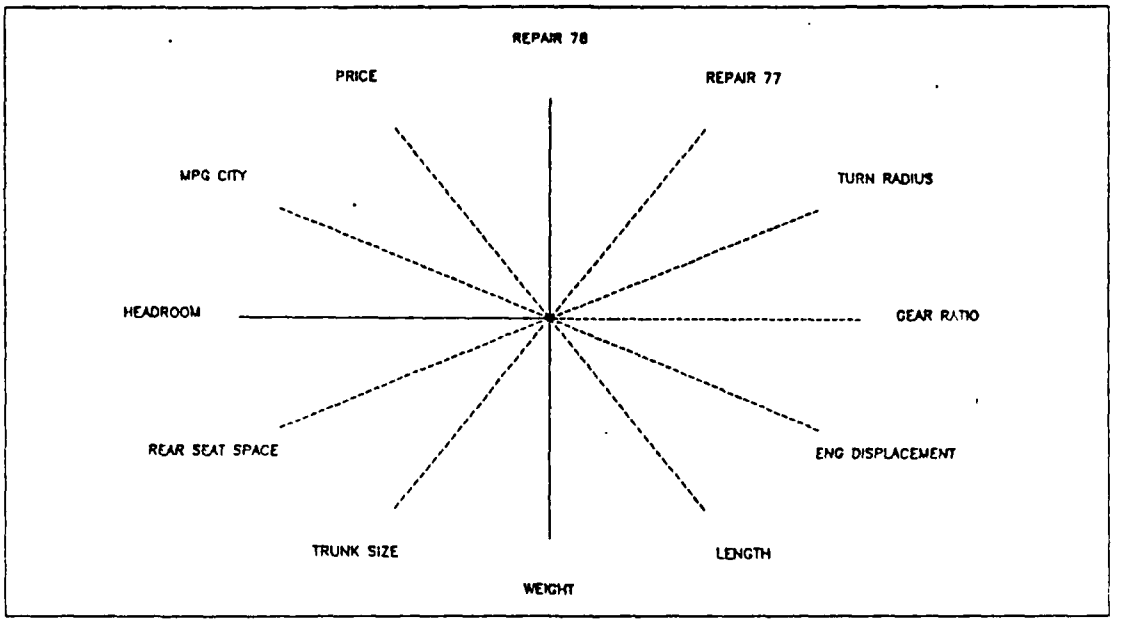


Figure 2.5 Assignment of Variables to the Rays of the STAR (Automobile Data).

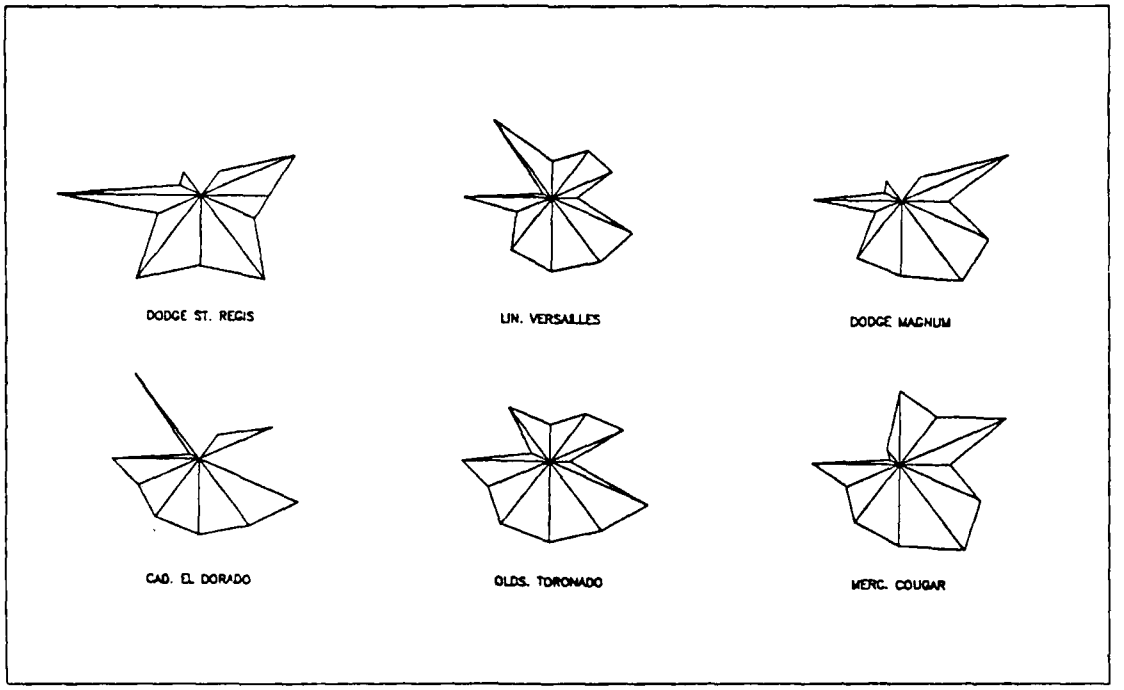


Figure 2.6 STAR Plot of the Automobile Data.

2. Technical details of STAR plots

There are two essential features in the construction of the STAR plot; the lengths of the rays and the angle between the rays. As stated in the last section, the value of the variables are represented by the length of the rays; therefore, these values should be non-negative and be represented using a similar scale. This is accomplished by initially rescaling the value of the variables using the following formula :

$$X'_{ij} = [(1\text{-}c)(X_{ij} - \min_j) / (\mathrm{Max}_j - \min_j)] + c , \qquad (2.7)$$

where c is a constant and is usually given a value of zero; and X_{ij} represents the i^{th} observation of the j^{th} variable. The coefficients $\min_j$ and Max_j represent the minimum and maximum value of the j^{th} variable respectively. Once the rescaling of the variables is performed, the angle between the rays must be determined. The first variable (variables are enumerated in increasing order) is plotted on the horizontal axis at an angle of zero degrees. Then the j^{th} angle between the remaining variables is calculated using the following formula :

$$\omega_j = 2\pi (j - 1) / n , \qquad (2.8)$$

where n represents the number of variables (parameters), and j is the j^{th} variable. The rays are then enumerated from 2 to n and displayed counterclockwise. Finally, the *star* is constructed by joining the end points of the n rays. The end point of each ray is calculated by the following formula :

$$P_{ij} = (X'_{ij} R \cos \omega , X'_{ij} R \sin \omega), \qquad (2.9)$$

where j = 1,2,..., n^{th} variable, and R is the maximum allowable radius of the star.

C. PROFILE PLOTS

1. Overview

The PROFILE technique is similar in nature to the STAR plot, the only difference is that in the PROFILE plot the rays are displayed by equidistant vertical lines arising from a common horizontal axis. In fact, as stated in Chambers [Ref. 1: p. 159], the STAR plots are actually PROFILE plots conceived in polar coordinates. In

the PROFILE plots, the values of the variables are used to control the length of the ends of the connected line segments (see Figure 2.7 and Figure 2.8).

One of the possible advantages of the PROFILE plots over the STAR plots is that in the former it is possible to represent variables with negative values. Since in the PROFILE plots all value-vectors are displayed with respect to a horizontal axis, it is then easy to show variables with negative values. The base line, the horizontal axis, is used to represent zero and negative values are displayed by lines dipping below this line. In the STAR plots this is not possible.

2. **Technical details of PROFILE plots**

In the PROFILE plot the rescaling is performed using the same formula as for the STAR plot. Negative values of the variables are allowed by using the following rescaling formula:

$$X^{*}_{ij} = (X_{ij} / Max_{j}). \qquad (2.10)$$

D. CODED SCATTER PLOTS

1. **Overview**

The CODED SCATTER plot is an enhancement of the most commonly used technique namely a SCATTER plot for two variables. Using this coding technique it is possible to represent more than two variable (dimensions) in the same display. Different symbols, and sizes and colors of these symbols, are used to represent three or higher dimensional data. The size and color of the symbols could be used to control different ranges of data values.

2. **Technical details of CODED SCATTER plot**

The CODED SCATTER plot uses essentially the same plotting technique as the usual SCATTER plot; only coded symbols, sizes and colors are added. This is in line with the need to represent as many dimensions as possible from a multivariate data set on a two dimensional graph. Thus, in a CODED SCATTER plot the position of the points in the graphical plane are represented by the (X,Y) values of the two variables. Then, if X is the miles per gallon variable in a data set and Y is the price of the car, plotting Y vs X shows how gas consumption increases or decreases with increasing cost of a car, or that there is a much more complex relationship between the two variables, or that there is even no relationship at all. However, there are other

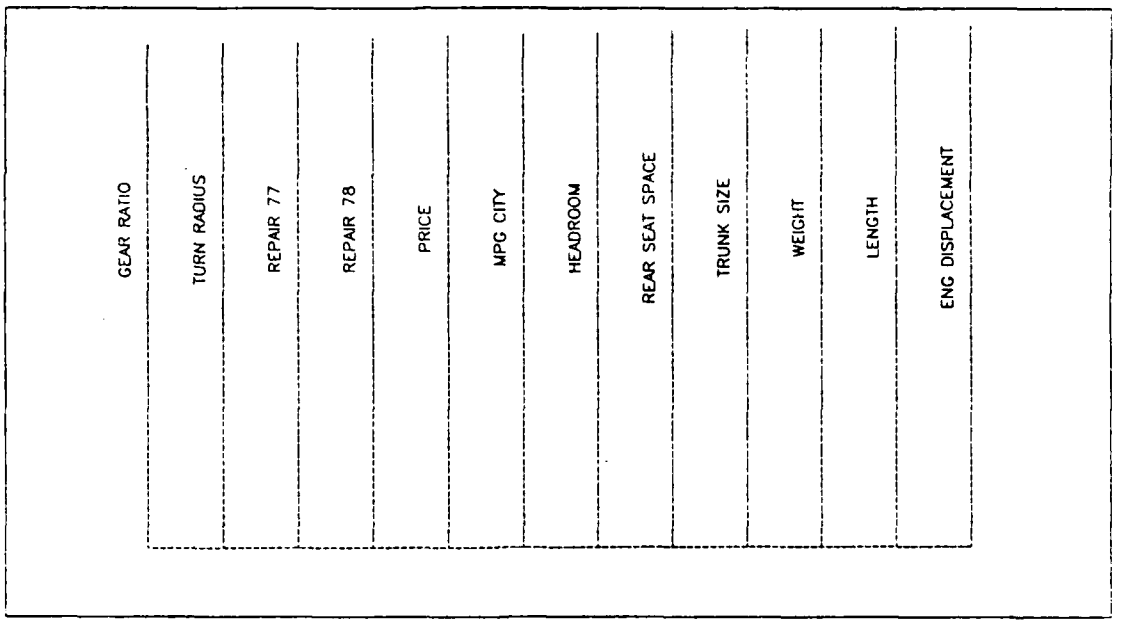


Figure 2.7 Assignment of Variables to the Lines of the PROFILE Plot (Automobile Data).

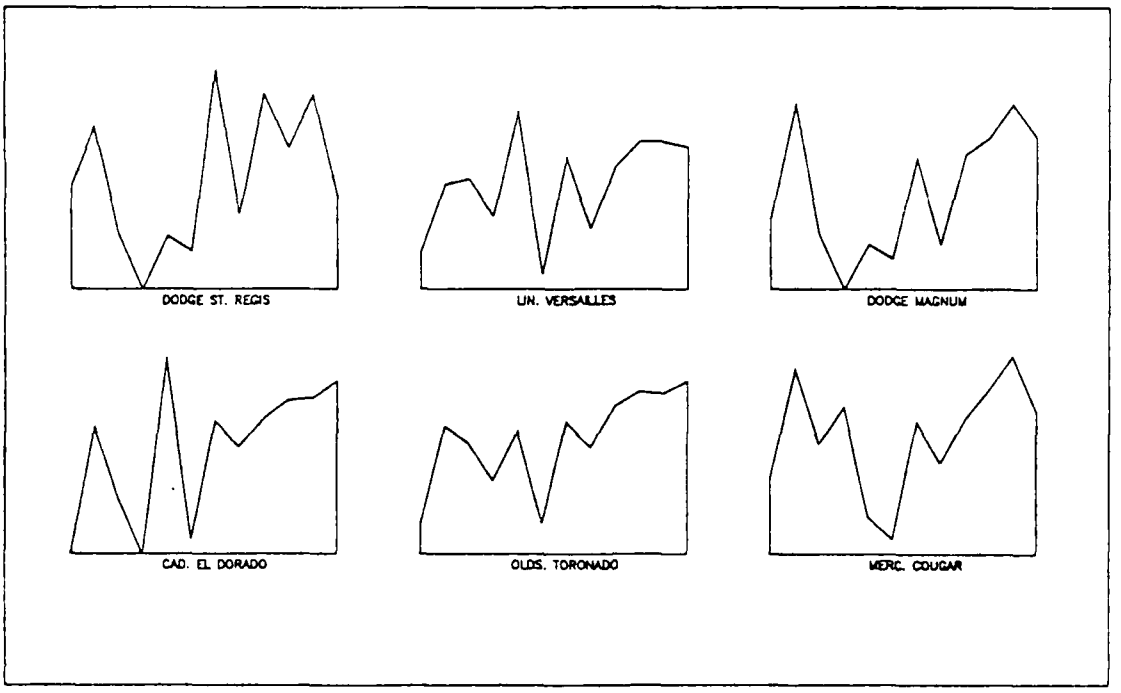


Figure 2.8 PROFILE Plot of the Automobile Data.

variables or factors involved in the relationship. These may be either continuous, discrete or categorical factors. An example of the first is the weight of the car, an example of the second is the number of cylinders in the car. A categorical factor is origin, i.e. whether the car is domestically produced or not.

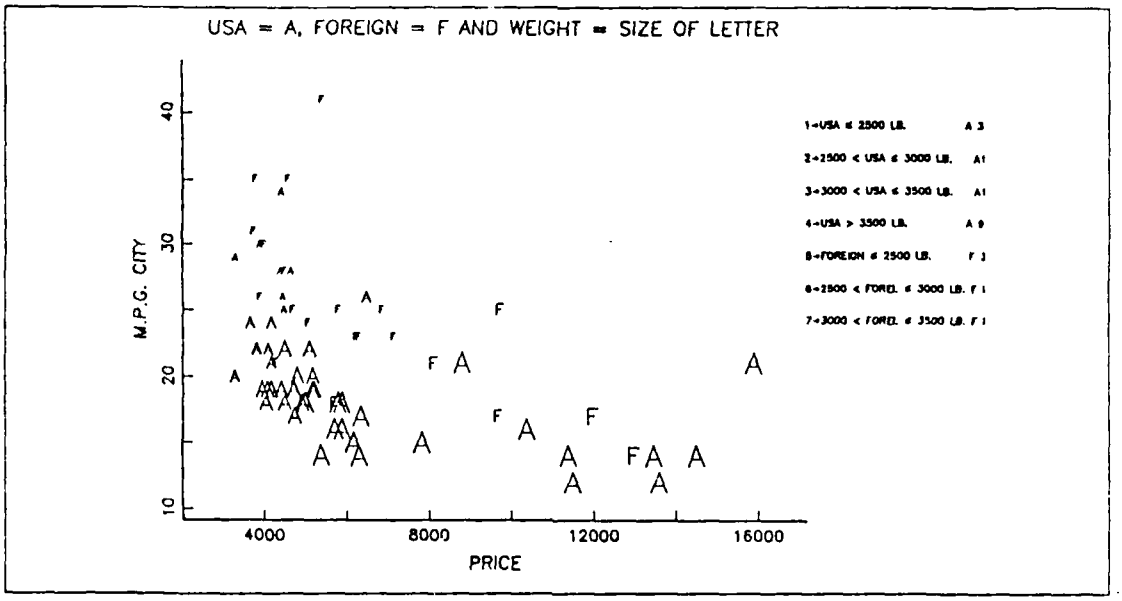


Figure 2.9 CODED SCATTER Plot of the Automobile Data. (Price vs M.P.G. City).

Figure 2.9 shows a CODED SCATTER PLOT of the car price variable, X, versus the miles per gallon variable, Y. The best way to code the origin of the car is by using colors; however, due to reproduction problem, this has been encoded as symbol type. The weight of the car is coded as the size of the symbol. In Figure 2.9, one can see that increasing price gives lower m.p.g., although the relationship is far from linear. The other factor is weight; weight clearly increases with price, also m.p.g decreases with weight. Again, referring to the categorical variable, origin, American cars cost more than foreign cars, weigh more and get less mileage, although there is interaction and overlap between all of these variables. There are also a few outliers. The very high mileage, low cost, and light weight car is the VW Rabbit (Diesel) and the very heavy, low mileage, and high cost car is the Cadillac Seville.

E. CODED DRAFTSMAN PLOTS

1. Overview

The DRAFTSMAN plot is an arrangement of SCATTER plots in which any adjacent pair of plots have a common axis (see [Ref. 1: p. 145]). In this way, the practitioner can observe the relationships of the variables within a specific plot and, in addition, can follow any particular observation (or group of observation) through the sequence of plots. Therefore, the analysis of multiple interactions among the variables is possible. This DRAFTSMAN plot can further be enhanced by portraying one or several additional variables by the assignment of symbols, sizes of the symbols, and colors to the already displayed variables. This is the main idea behind the CODED DRAFTSMAN plot, in which the techniques used in both DRAFTSMAN and CODED SCATTER plots are combined to display a single plot.

2. Technical details of CODED DRAFTSMAN plot

The CODED DRAFTSMAN plot, as mentioned earlier, can be seen as a displayed array of several CODED SCATTER plots. In certain situations the SCATTER plots may be deceiving due to the overlapping of data points. This situation may require to *jitter* one or more variables in order to alleviate this problem. Also, the practitioner may want to *transform* the data in order to achieve a simple and more understandable picture and in this way facilitate the analysis of the relationships among the variables (see [Ref. 4]). Another technique used by statisticians to reduce the spread in the data, and to enhance the visual interpretation of the plots is to *smooth* the data , relying on the *Moving Averages* technique or the *Locally Weighted Regression* (*LOWESS*) for this purpose. For further explanation about these two techniques see Moran [Ref. 5]. These techniques (*jitter*, *transformation*, and *smoothing*), are included in the CODED DRAFTSMAN program presented in this thesis, and they can be invoked interactively.

III. COMPUTER PROGRAMS : USER INSTRUCTIONS AND TECHNICAL DESCRIPTION

A. GENERAL

This chapter provides detailed instruction on how to use the computer programs presented in this thesis. These programs were written in APL and are designed to be used in conjunction with the experimental IBM graphical software GRAFSTAT. All these programs are interactive and all user defined parameters and options selections are entered in response to program queries. Although no APL skills are required to operate these programs, it is recommended that the user becomes familiar with APL system commands and procedures to load and copy workspaces, groups and variables, and to understand the meaning of *workspace*, *variables*, *groups* and *vectors* in the APL terminology. The user should read VS APL AT NPS , [Ref. 6] before attempting to use these programs. For the experienced APL user it will be easy to make changes to these programs in order to accommodate any additional requirement.

These programs were designed to be used on the IBM 3033 computer and to be executed using an IBM 3277/TEK 618, 3278/3279 or 3179G2 graphic display terminals using a memory capacity of at least 2 Megabytes.

All of these programs are contained within an APL workspace called APLGRAFS, and are organized in different *groups* of functions (each group contain those functions related to a specific program application). The reader can find a list of Groups and the content of each group of functions in Appendix A2. In order to make use of these programs, the user must have access to this workspace and to the APL workspace called GRAFSTAT.

There are two ways of executing these programs. The first one is described in the following steps :

1) LOGON to the system.
2) Once in CMS, enter *APLGST*.
3) At the prompt CLEAR WS, type *)LOAD GRAFSTAT*.
4) Enter *) PCOPY APLGRAFS groupname*, where groupname is one of the groups listed in Appendix A2.
5) Enter the name of the desired program to be executed (i.e., *STARPLOT*) and then answer the queries.

The second mode is more user-friendly. The steps that must be followed are :

1) LOGON to the system.

2) Once in CMS, type *APLGRAFS*, (this will cause the execution of the macro APLGRAFS EXEC), then you will see a menu describing all the available programs (see Figure 3.1).

3) After you enter the number corresponding to the selected program, you only have to follow the instructions given on the screen (this macro will execute steps 2), 3), and 4) of the previous list for you).

```
FILE: MENU     MENU     A1


YOU HAVE THE FOLLOWING PROGRAMS TO USE

(1) STAR AND PROFILE PLOTS
(2) BOX PLOTTED TABLES
(3) SYMBOLIC SCATTER PLOTS
(4) DRAFTSMAN DISPLAY
(5) LOWESS
(6) EXPLANATION ON THESE FUNCTIONS
(7) QUIT

TYPE THE NUMBER CORRESPONDING TO THE PROGRAM YOU WANT
```

Figure 3.1 Menu Presented by APLGRAFS EXEC.

B. PROGRAM DESCRIPTION

In order to use any of these programs, the user will need a matrix containing the sample data. This matrix could be in a CMS file, or could be a character array in an APL workspace. The programs will accept the data in either way; just follow the instructions given by the program as to the location of the data set. In addition, the user will need an APL two-dimensional character array containing the *names* of the variables which will appear in the display. These *names* are the labels which will be shown on the axis of the plots or in the rows and columns of the tables as in Figure 2.2 and 2.4. If the user has not previously created this array, the programs will allow the user to enter the labels directly in response to a sequential series of queries. At this point, the user is ready to execute any of the programs.

When answering the queries, if the user enter an erroneous response, the program will prompt the user to enter the correct response only in the case of a YES or NO question, a range question (i.e., 3,4, or 5 plots), if the name of any APL matrix does not exist in the workspace, etc.; in all other cases, the program does not have any means to know the validity of the response so the program will accept any response as

a correct one. When using these programs, if the user wants to cancel the execution at any time during the execution, the user must hit the PA2 key.

1. BOXPLOTTED TABLES

This program is executed by entering the command *BOXPLOTAB*. Once this command is entered, the program will start running by prompting the user with a sequence of queries indicating the user to enter the input arrays and to select the different available options (see Appendix B1, for an example of program execution).

a. Input requirement.

(1) The array containing the sample data, the array containing the *names* (labels) of the columns , and the array containing the *names* (labels) of the rows.

(2) The title of the display.

b. Options.

(1) The data could be displayed as originally entered or could be displayed ordered (sorted) by the first column.

(2) Once the BOXPLOTTED Tables are shown on the screen , the user will be prompted as to whether or not he or she wants to join the data points of the same position with lines. The position is given by the order of the data points of the first column (see Figure 2.3).

(3) After finishing with the previous display, the user will be prompted whether or not he or she wants to see BOXPLOTTED TABLES of the differences between columns; and if it is so, whether absolute or relative differences are desired. The difference could be calculated as follows : difference between all other columns with respect to the first one, or difference between adjacent columns (see Figure 2.4).

2. STAR PLOTS and PROFILE PLOTS

These two programs are executed by entering the command *STARPLOT*. The program will start running and the user will be asked to enter the desired function : (S) for STAR PLOT or (P) for PROFILE PLOT (see Appendix B2, and B3 for an example of the execution of this program).

a. Input Requirements.

(1) Same as for the BOXPLOTTED TABLES.

b. Options.

(1) Whether the whole original data is to be used or just a subsample of the data. The subsample could be constructed by selecting specific columns and or rows.

(2) The user will be asked how many plots per screen are desired. This could be 3,4 or 5 plots per screen.

3. CODED SCATTER PLOT

The execution and the input requirements of this program are similar to that of the STAR PLOT. To execute the program enter the command *SCATPLOT* (see Appendix B4 for an example of the execution of the program).

a. Options.

(1) The user will be asked to enter the title for the screen and the title for each plot on the screen.

(2) For each plot, the user must enter the column to be used on the X and Y axis, and whether or not the entire data or a subsample of it is desired.

(3) Another option is to select whether the data is to be jittered or if a transformation of the data (specified by the user) is desired.

(4) Following this option, the user must select the position of each plot on the screen (1, 21, 22, ..., etc.).

(5) Finally, the user must specify the symbols, colors and sizes of these symbols that will be used to represent an specific subset of the data. If the user selected one plot per screen, the program will ask for a small description for each one of these subsets or categories. These subsets are defined using APL statements. The user can specify more than one subset in the same plot and more than one plot per screen (see Figure 2.9).

4. CODED DRAFTSMAN PLOTS

This program is executed by entering the command *DRAFTSMAN*. Once this command is entered, the program will start by prompting the user with a sequence of queries indicating the user to enter the input arrays and to select the different available options (see Appendix B5, for an example of program execution).

a. Input requirement.

(1) The array containing the sample data, and the array containing the *names* (labels) of the columns.

b. Options.

(1) The data could be used as originally entered, or could be *jittered* or *transformed*. The user must select the desired option.

(2) Select whether or not a smoothed curve will be fitted to the data in all plots on the screen. If the smoothed curve is selected, the user must indicate whether the *Moving Average* or *LOWESS* technique will be used.

(3) Select between using the CODED DRAFTSMAN or the regular DRAFTSMAN plot. If the former is selected, the user must enter an APL expression, a symbol and its size, and the color for each category to be represented.

(4) Select the number of plots desired per screen (the available options are 3,4 or 5 rows and columns of plots per screen).

(5) Once the display is shown on the screen,and if the answer to option (2) was no, the user now has the alternative of fitting a smooth curve to the data of an specific plot.

IV. DATA ANALYSIS

A. GENERAL

The primary purpose of this chapter is to demonstrate the applications of the six graphical techniques presented in this thesis, namely BOXPLOTTED tables, STAR plots, PROFILE plots, CODED SCATTER plots and CODED DRAFTSMAN plots in the analysis of multivariate data. An attempt is made to highlight different peculiarities on the sample data that could be found when the practitioner uses these techniques; therefore, a full analysis of the various samples is not envisioned. However, it will be seen, that in using this techniques one can draw solid conclusions about certain behavior of the characteristics of the population from which the sample is drawn.

B. AN ANALYSIS OF HEALTH CARE EXPENSES

The following type of data represent a good example for which the statistician can make use of the BOXPLOTTED tables and the PROFILE plots. This is a sample of panel data and represents the health care cost (per capita hospital expenses) of 14 health service areas through the State of California from the years of 1971 to 1975. Figure 4.1 is a BOXPLOTTED table which displays the average health care expenses of the areas.

The data was formatted as a two dimensional array of 14 rows and 5 columns. Each row of the array represents the average health care expenses of a given area, and each column corresponds to the average expense for a given year. The data have been ordered in decreasing order by the first column (year), i.e., the service area with higher health care expenses on the first year correspond to the first row and so on. In general, the boxplots in Figure 4.1 give an initial impression of the distribution of each subsample data. Notice that during the first three years the tendency of the distribution is definitely skewed to the right, caused by some possible outliers, indicating that some service areas far exceed the average health care expenses. However, in the last two years the tendency is the opposite, with again the exception of some possible outliers.

The initial impression given by the boxplots could further be exploited by an analysis of the flow of the data in order to study the trend of the mean health care costs, the variance of health care cost, and to identify the occurrence, or recurrence, of

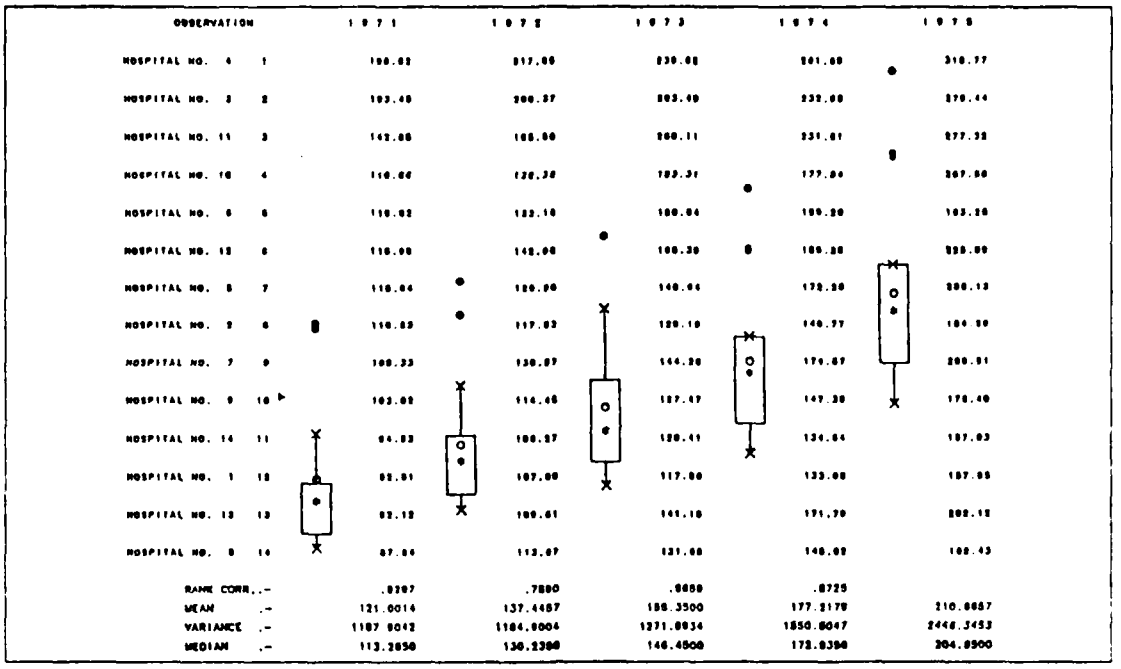


Figure 4.1 Per Capita Health Care Expenses in California Health Service Areas.

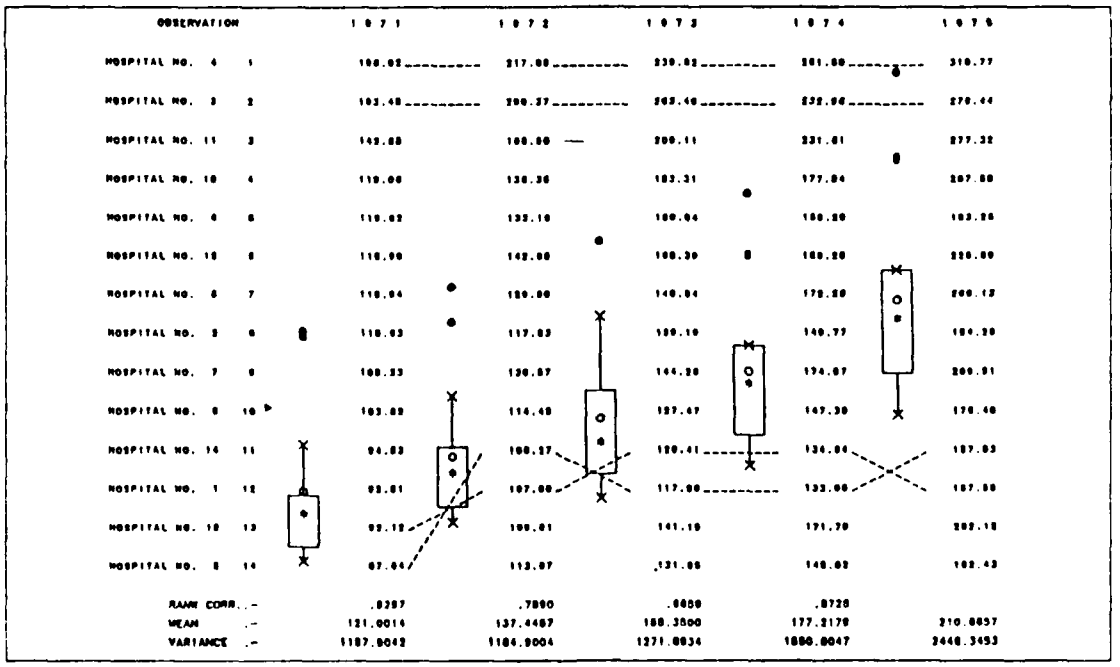


Figure 4.2 Per capita Health Care Expenses in California Health Service Areas (Most and Less Expensive Areas).

possible outliers. Another possible trend that could be study is that of the relative difference in health care expenses. One further area of interest is to study the trend of the most and less expensives service areas through this period.

The tendency in the average health care cost during this period was expected to be an increasing one (during this time, among other things, the inflation rate was starting to increase very rapidly). This tendency can be seen in Figure 4.1. Notice that this change is apparently quite linear up to the year of 1974; in 1975 there is a big jump in the average cost which probably indicates that, overall, the trend in the average health expense through this period was not linear. This same tendency is present in the variance of health care cost, which seems to confirm the nonlinearity in the average health care expense during this period. In this figure and in Figure 4.2, where lines are used to trace the flow of the 1971 high and low cost areas through subsequent years, it is also possible to readily pinpoint those service areas of extreme average health care cost (possible outliers, as defined in Chapter II). Notice that, the health service area number 4 is shown as a possible outlier through all years; it is always at least 2.2σ from the mean cost. The service area number 3 has the same tendency. These two areas are then the possible cause in the high variation observed in the health care cost through this period. They actually represent the Los Angeles and San Francisco metropolitan areas. In Figure 4.2, one could also follow those service areas with lower average cost (these areas are joined by line segments at the bottom of the display), it seems that these areas (number 1 and 14) had the lowest cost through this period, with the exception of service area no. 8 which has the lowest cost in 1971.

One could further follow the trend in the change of health care expense for each respective service area by using the PROFILE plots. In Figure 4.4 each one of the profile plots portrays the values of each row (health care service area), and the values are ordered by the magnitude of the first column, as in the *BOXPLOTTED tables*. The values of each column are represented in each profile according to the assignment given in Figure 4.3. In Figure 4.4 it is quite easy to identify the health service areas that had the highest and lower health care expenses during this period; as it was seen in the BOXPLOTTED tables, these areas are number 4 and 3, and number 1 and 14 respectively. One could also readly pinpoint the area with more variability in health care expenses, in this case notice that area number 13 has greater change in health expenditure than areas number 1, 14 and 4 (this last being the most expensive). Notice that the highest variation in expenditure in area number 13 takes place during 1972 and 1973. This fact could also be capture in Figure 4.5 in columns 2 and 3.

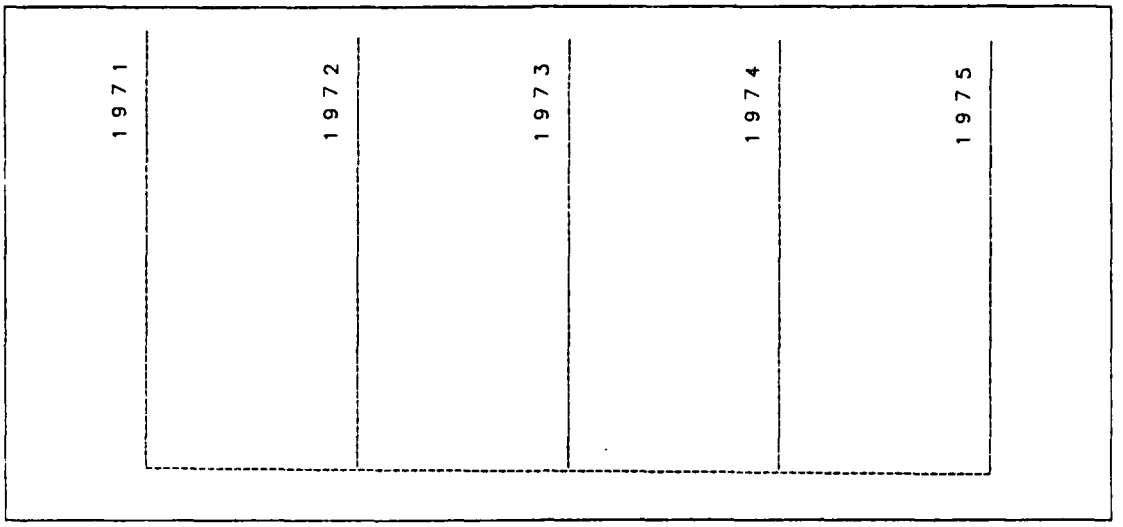


Figure 4.3 Assignment of Variables to the Profile of the Per Capita Health Care Expenses.

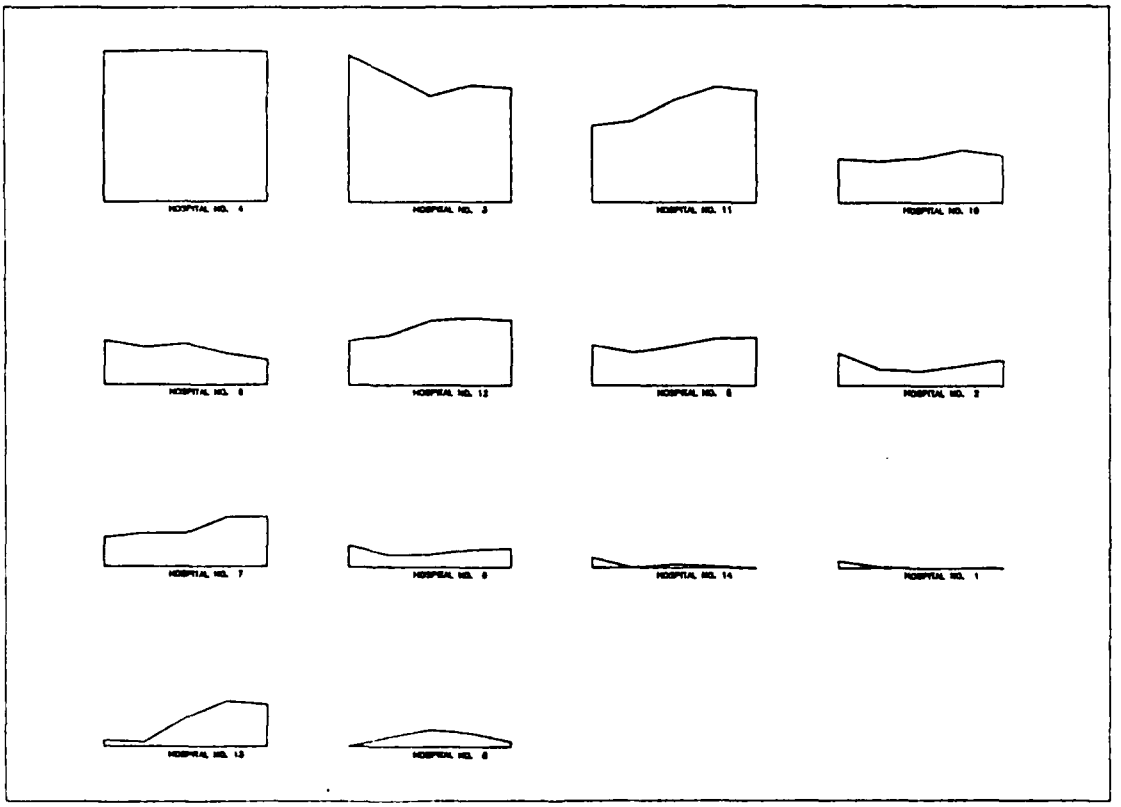


Figure 4.4 Profile Plot of the Per Capita Health Care Expenses California Health Service Areas.

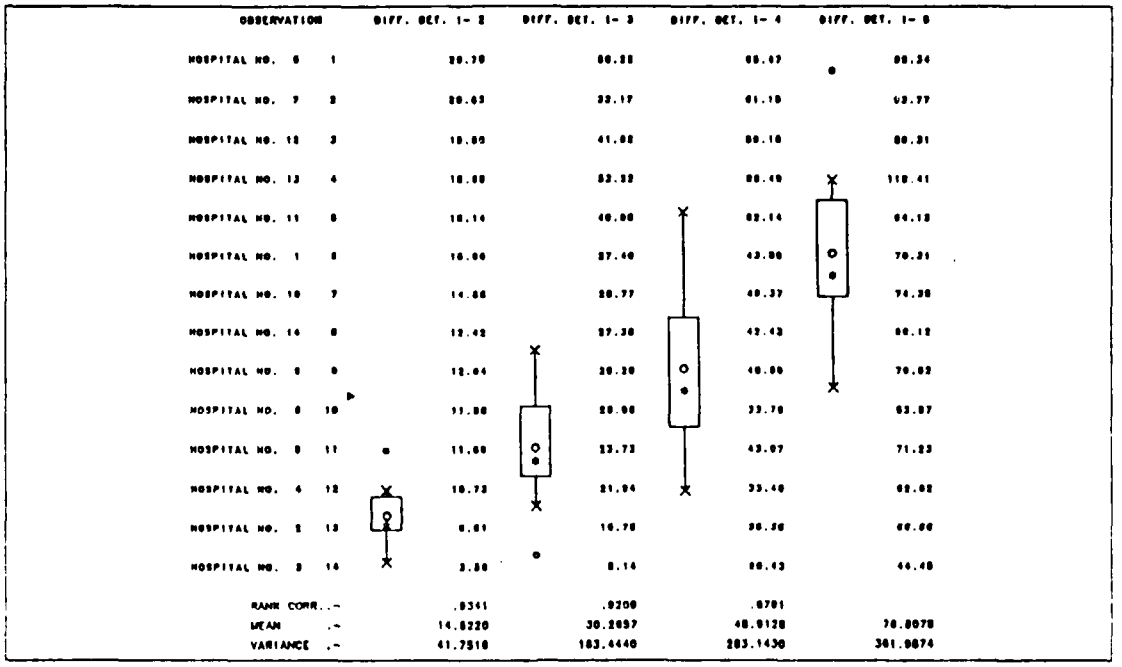
Now one could also reinforce the statement about the nonlinearity in the change of the health expenditure by looking at the relative difference in this variable through this period. Figure 4.5 portrays the trend in the relative difference in health expenditure with respect to the first year of study (1971). In Figure 4.6 one could see the some trend but now with respect to the previous year. Figure 4.6 definitely shows that the change in health care expenses has a nonlinear behavior. It is changing linearly during the first three years, and then at an accelerated, possible quadratic rate, from then on. The same trend seems to be shown in Figure 4.5, this trend is highlighted in the last column, where the mean of the relative differences jump from 48.91 to 76.50.

In Figures 4.5 and 4.6 it is also possible to identify those services areas that have the maximum and minimum relative change. As an example from 1971 to 1972 area number 8 had the maximum positive increase and from 1972 to 1973 the area with the maximum positive change was area number 4.

C. AN ANALYSIS OF THE NEW YORK STOCK EXCHANGE

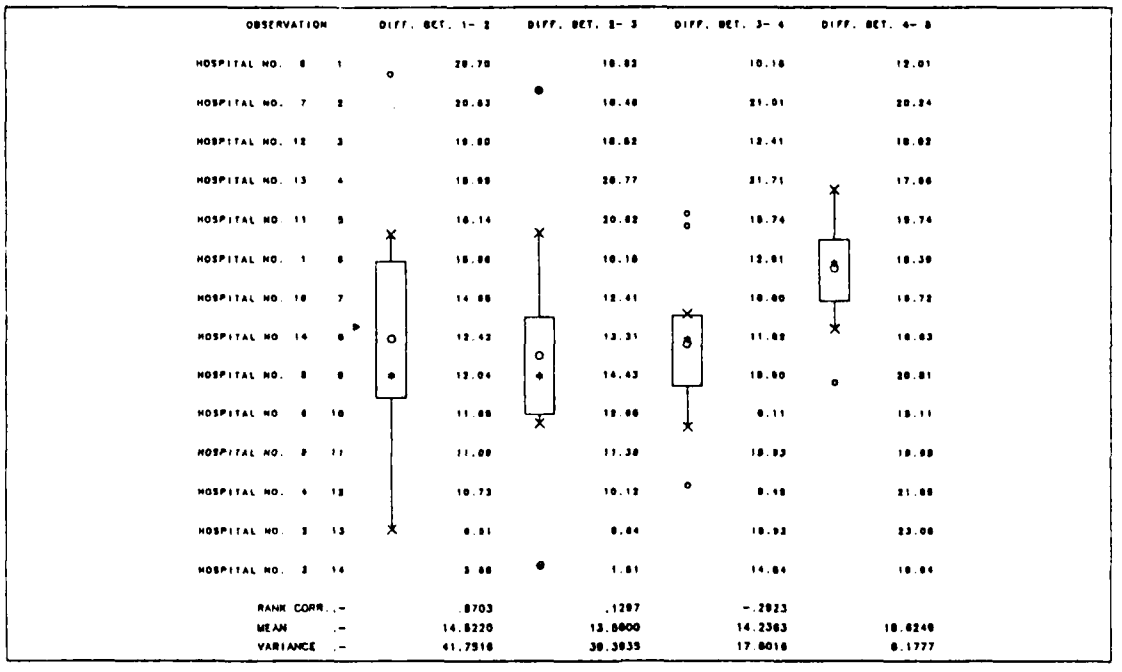
In the previous analysis, the data considered consisted of the same type of commensurable values; i.e., dollars through a period of time (one could consider this as being multiple short time series data). In contrast with this type of data, the practitioner can encounter multivariate data that represent different qualitative and quantitative magnitudes. One example of this type is the data obtained from the stock markets in the United States. Here again, the practitioner can make use of the BOXPLOTTED tables as a tool for data analysis. The data to be analyzed was extracted from the New York Times, representing the most active stocks (measured by the number of shares traded) in the New York Stock Exchange for the week ended on August 8, 1986. The data is initially formatted as a two dimensional array consisting of 40 rows (representing each of the different trading companies) and 6 columns. Each column correspond to the following variables.

(1) Volume of shares traded during the week (in 100,000 units).
(2) Closing price at the end of the week (in dollars).
(3) Price change during the week (in percentage).
(4) Price change during the last 12 months (in percentage).
(5) Earnings per share during the last 12 months (in dollars).
(6) Earnings per share during the last 12 months (in percentage).



OBSERVATION		DIFF. BET. 1- 2	DIFF. BET. 1- 3	DIFF. BET. 1- 4	DIFF. BET. 1- 5
HOSPITAL NO. 6	1	20.70	60.82	60.07	60.34
HOSPITAL NO. 7	2	20.63	32.17	61.10	62.77
HOSPITAL NO. 12	3	10.80	41.82	60.10	60.31
HOSPITAL NO. 13	4	10.80	52.52	60.40	110.41
HOSPITAL NO. 11	5	10.14	40.00	62.14	64.10
HOSPITAL NO. 1	6	10.00	37.40	42.00	70.21
HOSPITAL NO. 10	7	14.00	30.77	40.37	74.30
HOSPITAL NO. 14	8	12.42	27.30	42.43	60.12
HOSPITAL NO. 8	9	12.04	20.20	40.00	70.02
HOSPITAL NO. 9	10	11.00	20.00	33.70	63.07
HOSPITAL NO. 2	11	11.00	33.73	43.07	71.23
HOSPITAL NO. 4	12	10.73	21.04	33.40	60.02
HOSPITAL NO. 5	13	0.01	10.70	30.00	60.00
HOSPITAL NO. 3	14	3.00	0.14	00.43	44.40
RANK CORR.		.8341	.0200	.0701	
MEAN		14.0220	30.2057	40.0120	70.0070
VARIANCE		41.7510	103.4440	203.1430	301.0074

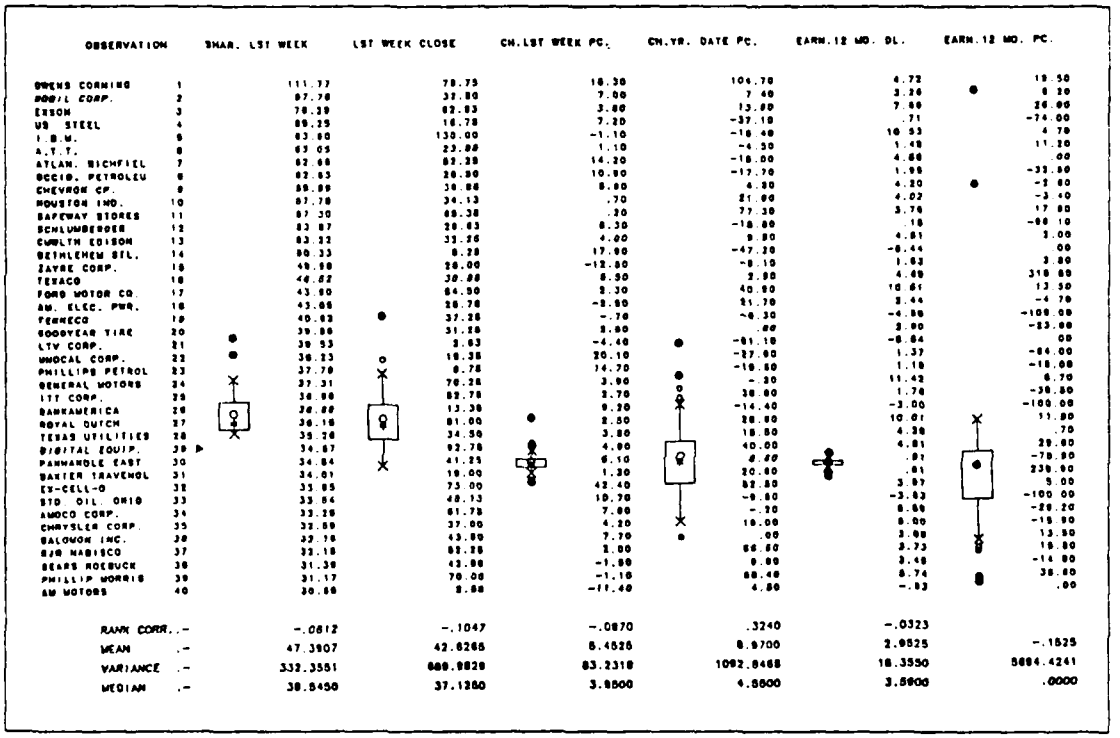
Figure 4.5 Relative Differences in the Health Care Cost California Health Service Areas (First Year).



OBSERVATION		DIFF. BET. 1- 2	DIFF. BET. 2- 3	DIFF. BET. 3- 4	DIFF. BET. 4- 5
HOSPITAL NO. 6	1	20.70	10.02	10.10	12.01
HOSPITAL NO. 7	2	20.03	10.40	21.01	20.24
HOSPITAL NO. 12	3	10.00	10.02	12.41	10.02
HOSPITAL NO. 13	4	10.00	20.77	21.71	17.00
HOSPITAL NO. 11	5	10.14	20.02	10.74	10.74
HOSPITAL NO. 1	6	10.00	10.10	12.01	10.30
HOSPITAL NO. 10	7	14.00	12.41	10.00	10.72
HOSPITAL NO. 14	8	12.42	13.31	11.02	10.03
HOSPITAL NO. 8	9	12.04	14.43	10.00	20.01
HOSPITAL NO. 9	10	11.00	10.00	0.11	10.11
HOSPITAL NO. 2	11	11.00	11.30	10.03	10.00
HOSPITAL NO. 4	12	10.73	10.12	0.10	21.00
HOSPITAL NO. 5	13	0.01	0.04	10.02	23.00
HOSPITAL NO. 3	14	3.00	1.01	14.04	10.04
RANK CORR.		.0703	.1207	-.2023	
MEAN		14.0220	13.0000	14.2363	10.0240
VARIANCE		41.7510	30.3035	17.0010	0.1777

Figure 4.6 Relative Differences in the Health Care Cost California Health Service Areas (Prev. Year).

Figure 4.7 shows the initial distributional characteristics (in the form of boxplots) of each subsample data. One of the first visual messages from these plots are the outliers in each column. Here is where the power of this new graphical technique lies: one can easily identify those possible outliers by looking at the tabular data adjacent to the boxplots; although this is easier in the first column since the data is ordered in that column. As an example, looking at the first boxplot and the first column, it is easy to identify Owen Corning and the Mobil Corp. as those companies traded by these two companies is greater that 2.7σ of the average column traded.



OBSERVATION		SHAR. LST WEEK	LST WEEK CLOSE	CH.LST WEEK PC.	CH.YR. DATE PC.	EARN.12 MO. DL.	EARN.12 MO. PC.
OWENS CORNING	1	111.72	78.75	18.38	104.70	4.72	19.50
MOBIL CORP.	2	87.78	32.88	7.00	7.40	3.26	8.20
EXXON	3	78.39	62.63	3.88	13.88	7.88	26.88
US STEEL	4	88.25	18.78	7.28	-37.18	.71	-74.88
I.B.M.	5	83.88	138.88	-1.18	-18.48	10.53	4.78
A.T.T.	6	83.85	23.88	1.18	-4.58	1.48	11.28
ATLAN. RICHFIEL	7	82.68	82.38	14.28	-18.88	4.88	.88
OCCID. PETROLEU	8	82.63	28.88	18.88	-17.78	1.98	-33.88
CHEVRON CP.	9	88.88	38.88	8.88	4.88	4.28	-8.88
HOUSTON IND.	18	87.78	34.13	.78	81.88	4.82	-3.48
SAFEWAY STORES	11	87.38	48.38	.28	77.38	3.78	17.88
SCHLUMBERGER	12	83.87	28.83	8.38	-18.88	.18	-88.18
CMBLTH EDISON	13	83.32	31.58	4.88	9.88	4.81	3.88
BETHLEHEM STL.	14	88.33	8.38	17.88	-47.28	-8.44	.88
ZAYRE CORP.	15	48.88	28.88	-12.58	-8.18	1.83	3.88
TEXACO	16	48.82	38.88	8.58	2.88	4.88	318.88
FORD MOTOR CO.	17	43.88	84.58	3.38	48.58	18.81	13.58
AM. ELEC. PWR.	18	43.88	28.78	-8.88	81.78	3.44	-4.78
FEMNECO	19	48.83	37.28	-.78	-8.38	-4.88	-188.88
GOODYEAR TIRE	28	39.88	31.28	3.88	.88	2.88	-83.88
LTV CORP.	21	38.53	3.83	-4.48	-81.18	-8.84	.88
UNOCAL CORP.	22	38.33	18.38	28.18	-37.88	1.37	-84.88
PHILLIPS PETROL	23	37.78	8.78	14.78	-18.88	1.18	-18.88
GENERAL MOTORS	24	37.31	78.28	3.88	-.28	11.42	8.78
ITT CORP.	25	36.88	82.78	3.78	38.88	1.78	-38.58
BANKAMERICA	26	36.88	13.38	8.28	-14.48	-3.88	-188.88
ROYAL DUTCH	27	36.18	81.88	8.58	28.88	18.81	11.88
TEXAS UTILITIES	28	35.78	34.58	3.88	18.58	4.38	.78
DIGITAL EQUIP.	29	34.87	82.78	4.88	48.88	4.81	28.88
PANHANDLE EAST	38	34.84	41.25	8.18	.88	.81	-78.58
BAXTER TRAVENOL	31	34.81	18.88	1.38	28.88	.81	238.88
EX-CELL-O	32	33.85	73.88	42.48	82.58	3.87	5.88
STD OIL OHIO	33	33.84	48.13	18.78	-8.88	-3.83	-188.88
AMOCO CORP.	34	33.28	61.75	7.88	-.28	8.88	-28.28
CHRYSLER CORP.	35	32.88	37.88	4.28	18.88	8.88	-18.88
SALOMON INC.	38	32.78	43.88	7.78	.88	3.88	13.58
RJR NABISCO	37	32.18	88.75	3.88	88.88	3.73	18.88
SEARS ROEBUCK	38	31.38	43.88	-1.88	8.88	3.48	-14.88
PHILLIP MORRIS	39	31.17	78.88	-1.18	88.48	8.74	38.88
GM MOTORS	48	38.88	8.88	-11.48	4.88	-.83	.88
RANK CORR.	.-	-.8812	-.1847	-.8878	.3248	-.8323	
MEAN	.-	47.3187	42.6265	8.4528	8.9788	2.8525	-.1525
VARIANCE	.-	332.3581	888.8828	83.2318	1892.8468	18.3558	5884.4241
MEDIAN	.-	38.8458	37.1288	3.8588	4.8588	3.8488	.8888

Figure 4.7 40 Most Active Stocks for the Week Ended August 8, 1986 (New York Exchange).

Another observation that can be made from this figure is the absence of statistical correlation among the variables when these are compared in the order shown in Figure 4.7. In this case, the sample serial rank correlation are obtained by comparing the adjacent columns. Looking at the sample serial rank correlation, one could conclude that there is no statistical relationship between, as an example, the volume of shares traded and the price at which the share closed at the end of the week.

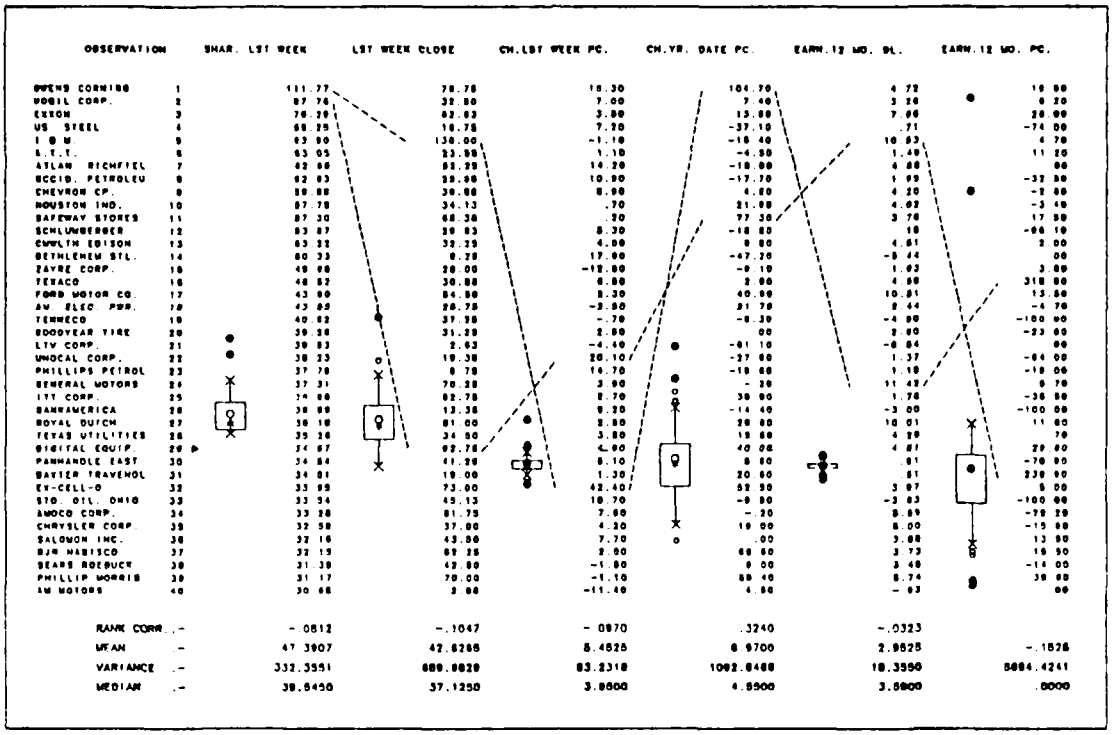


Figure 4.8 40 Most Active Stocks for the Week Ended August 8,1986 (New York Exchange) with Lines Connecting the Two Higher Stocks.

The only positive indication is a relationship between percentage change during the last week (column 4) and percentage change in the last year (column 5). One can visually confirm this lack of correlation by identifying the maximum and minimum values of adjacent columns. For example, the Am Motor Co. shows to have the lowest volume of shares traded during that week but the LTV Corp. had the lowest close price.

Notice that one is not only interested in the stock which is the most active during the week. One is also interested in which stock has the greatest (absolute or relative) change in price, and whether this is related to other factors like earnings (absolute or relative). With this is mind, it is possible to follow those stocks that have the largest value in each of the variables considered in the analysis. Figure 4.8, shows the two stocks which have this characteristic. These stocks are joined by line segments. It is easy to see that the Owens Corning Co. is the stock with the highest volume of shares traded during the week and also the largest price change (in percentage) during the last twelve months; also, the IBM Co. has the highest closed price at the end of the week and has the second largest earnings per share (in dollars) during the last year. Likewise,

the practitioner could follow the trend of those stocks with the lowest value in each of the variables considered, or even the mid-point values.

Differences have no meaning here but it is interesting to trace the movement of the most active stocks (in volume) to other indicators (columns).

D. AN ANALYSIS OF AUTOMOBILE DATA

The purpose of the following analysis is to try to explore some important descriptive characteristics of different types of automobiles and an attempt is made to find any relation between these characteristics. As it is shown in this analysis, the STARPLOTS, the CODED SCATTER plots and the CODED DRAFTSMAN plots techniques are paramount experimental statistical tools in this type of analysis. It is appropriate to mention at this time that one other author has previously made use of the data treated here and has written an outstanding analysis (See [Ref. 4]). The purpose here is to demonstrate how one can convey to the same general conclusion using these new techniques. The new technique is the enhancement of SCATTER and DRAFTSMAN plots by coding in other variables.

The data represent three general categories of quantitative and qualitative characteristics of American and Foreign automobiles of 1979 (the data was obtained from the Consumer Report Review). These categories are: performance, dimension and price. The variables under these categories are as follows.

In category one : mileage in miles per gallon, repair records for 1977 and 1978 (rated on a 5 points scale; 5 = best and 1 = worst), turning diameter (clearance required to make a U-turn) in feet, gear ratio for high gear.

In the second category : headroom in inches, weight in pounds, length in inches, displacement in cubic inches.

And under the last category: price in dollars. This data was initially formatted into a two dimensional array consisting of 74 rows (name of automobiles) and 13 columns. Each column corresponds to each one of the variables mentioned above, and the last column correspond to an ordinal variable to denote American or Foreign car. This variable has been added to the original data to demonstrate one of the many possible application of the CODED SCATTER plot and of the CODED DRAFTSMAN plot introduced in this thesis; as an example, one can readily identify if a certain deviation from a possible pattern is due to American or Foreign cars.

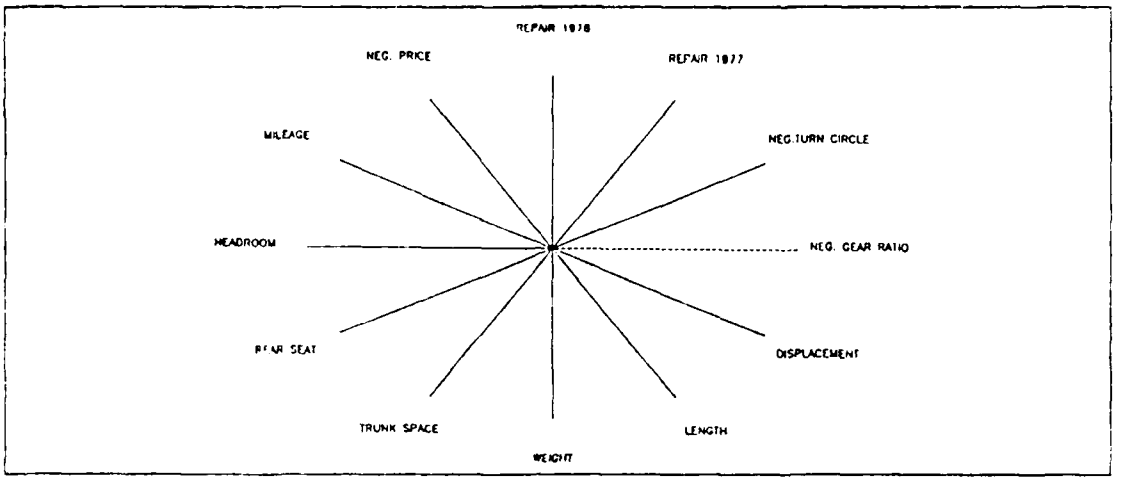


Figure 4.9 Assignment of Variables to the Rays of the Star Automobile Data.

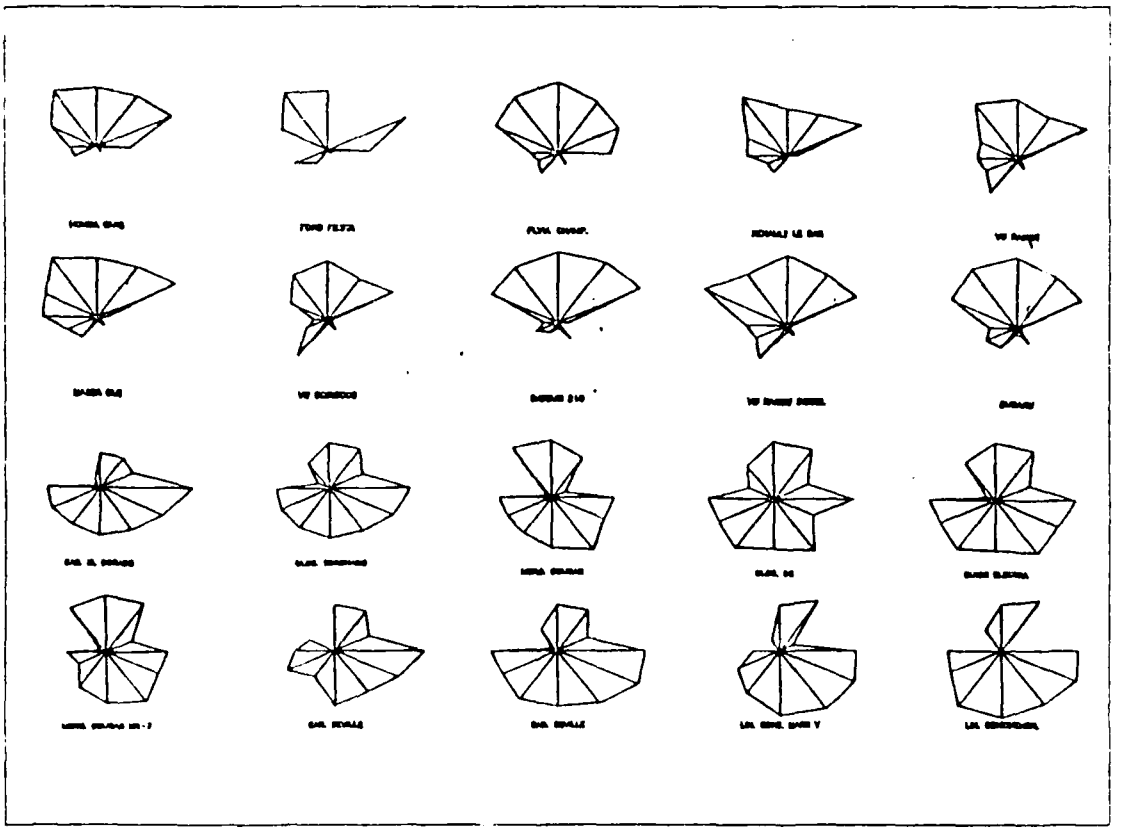


Figure 4.10 STAR Plot of Automobile Data, 10 Lighter and the 10 Heavier Automobiles.

As an initial starting point for this analysis, one could study the characteristics of each one of the individual automobiles. The STAR plot technique was choosed for this purpose. Figure 4.9 shows the assignment of the twelve characteristics to the rays of the star. In the study of this type of data, it is interesting to highlight the *favorable* characteristics of each automobile. So as in Chambers [Ref. 1] the larger the ray of the star the more positive that attribute is to the respective automobile. To make price, turning diameter and gear ratio favorable, these variables were multiplied by -1 (i.e., the larger the ray corresponding to price, the less expensive the car is). The star is arranged in such a way that the statistics corresponding to cost and performance categories are rising upward and horizontally, and those rays pointing downwards correspond to variables closely related to the dimension of the automobiles. Appendix C shows the complete STAR Plots for the 74 automobiles.

The array of stars are ordered by the weight, the first and last stars corresponding to the lightest and heaviest automobiles respectively. Figure 4.10 displays a summary of Appendix C, showing the 10 heaviest and 10 lightest automobiles. The idea behind this arrangement is, as commonly accepted, that weight is positively correlated to safety. Note the switch between the first (Honda Civic) and last (Lincoln Continental). For the first of these all positive values are above the line; for the latter this is switched. Note too that the variable of greatest interest to Consumers Reports, Repair 78, is the vertical ray.

In Figure 4.10, it is easy to see that nine out of the ten lighter cars, in the top panel, are of foreign make, the exception being the Ford Fiesta. Also, that the 10 heavier cars are Americans. From the STAR plot of Figure 4.10, one could also compare other characteristics among these automobiles. As an example, in terms of price variable alone, notice where in this case, that among the 10 heavier cars there are 4 American cars that are inexpensive compared with most of the lighter foreign cars (these American cars are the Mercury Cougar and Cougar XR-7, Buick Electra and, in lesser way the Oldsmobile 98). Also in terms of repair records (of 1977 and 1978), those American cars among the heavier ones compare with those foreign cars among the other group. The information is abundant in these plots. However, when there are many variables involved in the analysis it is questionable whether the practitioner can actually capture the behavior of one variable alone or the joint behavior of two or more variables. As in this case one would like to see if there is any relation (linear or other type) between price and weight or, say, displacement and price (it is difficult to

identity any deviation, if it exists, from a possible relation in the STAR plot). In this type of situation, once the practitioner has an initial impression of the data, it is now the time to make use of other exploratory data analysis technique, such as CODED SCATTER plots and CODED DRAFTSMAN plots.

To continue the analysis it was desired to study any possible relation among the price, mileage per gallon, weight and displacement of the automobiles and to compare how American cars stand against the Foreign cars. The relations between these variables are examined in Figure 4.11 by using a CODED DRAFTSMAN plot. It was expected to see positive correlation between displacement and weight. This can easily be seen in the plot position 2,2 of Figure 4.11. The two possible outliers in plot 2,2 of Figure 4.11 show that there are two American cars that stand favorably among all others. They are lighter cars with high displacement. From the figures in Appendix C these two automobiles were identified as the Chevrolet Chevette and the Buick Opel. In terms of price, it is also possible to conclude from plot position 3,3 of Figure 4.11 that there is a negative relation, as expected, between price and weight. Notice, in the plot position 2,1 that there seems to be two types of subsamples within the data, one of foreign cars and the other of American cars (the foreign cars standing favorably against the American ones); however, both subpopulations have the same trend, namely, that weight increases with price. There are a couple of interpretations of this plot, beside the obvious dichotomy between American and Foreign autos. One is that if you want a heavy car, you will have to pay more if you also want it foreign made.

An expansion of Figure 4.11 is given by the CODED SCATTER plot, which have additional variables coded in as symbol type and size. Looking at price versus m.p.g., in the CODED SCATTER plot of Figure 4.12, one can confirm the idea that the higher the price of the automobile the less miles per gallon is expected. Notice, that with this figure, it is possible to analyze four variables at the same time : price and miles per gallon being the axis and weight and nationality the coded variables. It is interesting to notice that one American and one Foreign car tend to deviate from the norm. The American cars is the Cadillac Seville, with very high price, quite heavy, but a good relative mileage; the Foreign, being the V.W. Rabbit (Diesel) is at the opposite site of the spectrum. In the middle of the plot is a medium price, foreign car with very good mileage. This is a BMW 320i.

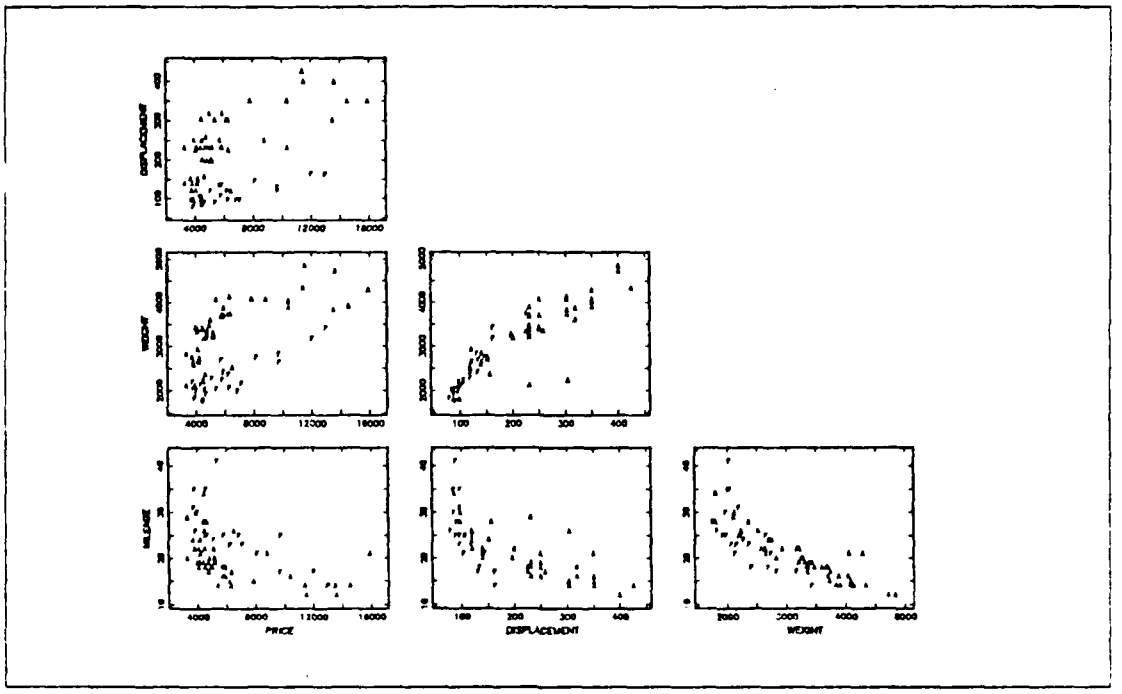


Figure 4.11 CODED DRAFTSMAN Plot of Automobile data
A = American, F = Foreign.

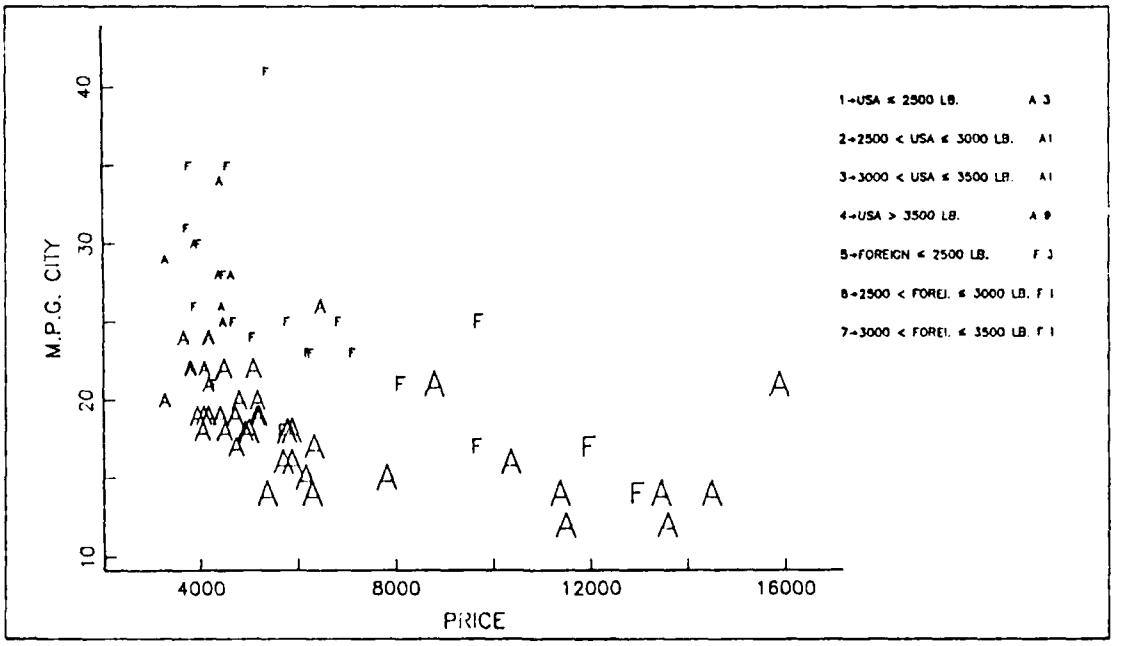


Figure 4.12 CODED SCATTER Plot of Automobile Data
Price vs MPG (A = American, F = Foreign and Size = Weight).

E. AN ANALYSIS OF CONTRACT DATA

The purpose of this analysis is to demonstrate other possible applications of the CODED SCATTER plot graphical technique as a tool in the exploratory data analysis. It is also appropriate to emphasize that the data considered in this section has been amply analyzed by other authors ([Ref. 4]) and again the purpose is only to highlight the use of this mentioned graphical technique.

The data consisted of 177 contracts (rows), which were authorized by the Department of Defense during the period of 1949 through 1963. The columns consist of 11 variables of possible interest to the Department of Defense on how they have interfaced on a contractual level with the private sector of manufacturers. The data represent contracts let with 23 major contractors during this period, and includes information concerning 7 types of manufacturer products, ranging in complexity from drone aircrafts to missiles and helicopters. The 11 variables are listed below:

(1) Deviation from target cost (percent).
(2) Months to comply a contract.
(3) Target profit of manufacturer (percent).
(4) Sharing ratio (percent).
(5) Ceiling price (percent of target price).
(6) Target cost.
(7) Number of items produced in the contract.
(8) Number of contracts let that year.
(9) Year the contract was signed.
(10) Contractor awarded the contract.
(11) Type of system.

Due to the diversity of the data and the purpose of this section, it was decided to narrow the objective of the analysis to a single issue, which is probably the most important to the Department of Defense: an attempt will be made to see if there is any increase (or decrease) in the deviation from the manufacturer target cost through time. The one deviation that is considered to be the most significant will be the positive one, since this phenomenum would represent additional expenditure to the government. Thus, the task is to try to find a possible cause to this increase.

Among the other 10 factors, it was hypothesized that the year in which the contract was signed and the time (in months) to complete the contract had significant influence on the deviation from the manufacturer original target cost. The variable year

signed was considered since the period of study includes an event that had significant impact on the US economy: the Korean War; therefore, it was expected that smaller contractors, not really prepared to react to the contingency of war production, would be less capable of making accurate predictions. Figure 4.13 shows the display of the year the contract was signed versus the deviation from target cost. It was also expected that the contractor, increasing from normal productions, would also be affected in their prediction capabilities. The hypothesis about the time to complete the contract is based on a simple idea : the wider the interval of time for which the prediction is made, the less is the probability of asserting the prediction. It was also desired to see if the major trend in this deviation of the major contracts, since these were probably of greatest interest to the government. In Figures 4.13 and 4.14 three major contractors were selected as been of relative importance: Lockeed, Douglas and Grumman. These three are coded by the initial letter. It is easy to see that the actual year that the contract was signed does not really influence the deviation from target cost; the deviation are evenly distributed across the period of interest. However, notice that during 1951 and 1952 (Korean War period) the deviation are mainly on the negative side (probably the significance of patriotism) and thereafter are evenly distributed.

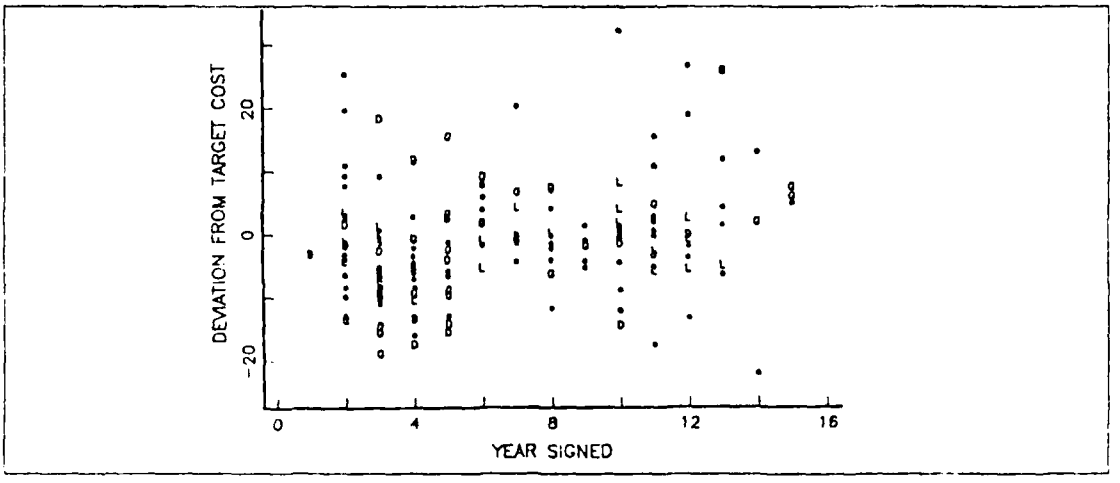


Figure 4.13 Year Signed vs Dev. From Target Cost, Contract Data (L = Lockeed, G = Grumman, D = Douglas, o = Others).

The other variable of interest was then considered, namely the time to complete the contract. The range of this variable is from around 15 months to 130 months.

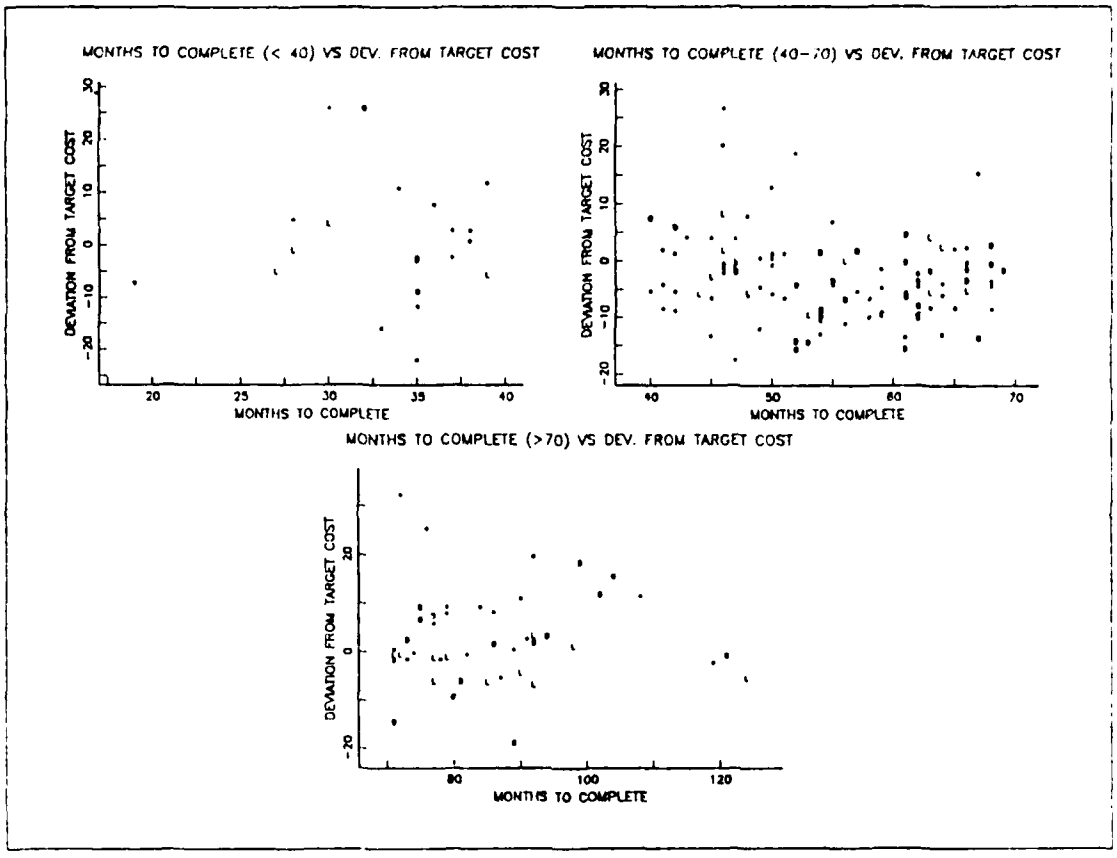


Figure 4.14 Months to Complete vs Dev. Target Cost, Contract Data (L = Lockeed, G = Grumman, D = Douglas, o = Others).

Figure 4.14 shows the plot of those contracts that took less than 40 months, 40 months and less than 70, and 70 or more months versus cost deviation respectively. It is clear that there is some form of positive relation between cost deviation and those contracts that took more than 70 months to be completed; confirming the initial hypothesis. There are some exception to this conclusion, and these are mainly contracts that were given to two of the largest contractors, Lockeed and Grumman, and possibly two smaller ones.

These plots, in a clear way, demonstrate the care that must be taken in the analysis of single scatter plots, one scatter plot portrays only isolated relationship of two variables and may not indicate a casual relationship. One should make use of different exploratory data analysis techniques in an attempt to discover possible trends in the data being analyzed.

APPENDIX A
COMPUTER PROGRAMS

1. APLGRAFS EXEC.

This exec program present a menu with all the programs available in the APL workspace APLGRAFS, after the selection is made the exec will load the necesary workspaces and will prompt the user to enter the name of the selected program.

```
&TRACE
SET BLIP *
-ONE
CLRSCRN
&TRACE
&TYPE
&TYPE YOU HAVE THE FOLLOWING PROGRAMS TO USE
&TYPE
&TYPE       (1)   STAR AND PROFILE PLOTS
&TYPE       (2)   BOX PLOTTED TABLES
&TYPE       (3)   SYMBOLIC SCATTER PLOTS
&TYPE       (4)   DRAFTSMAN DISPLAY
&TYPE       (5)   LOWESS
&TYPE       (6)   EXPLANATION ON THESE FUNCTIONS
&TYPE       (7)   QUIT
&TYPE
&TYPE  TYPE THE NUMBER CORRESPONDING TO THE PROGRAM YOU WANT
&READ VAR &OPT
&IF &OPT =  7 &GOTO -FINAL
&IF &OPT < 1 &GOTO -ERROR1
&IF &OPT > 6 &GOTO -ERROR1
* CP DEFINE STORAGE 2048K
* &STACK I CMS
CP TERMINAL APL ON
&STACK )LOAD GRAFSTAT
&IF &OPT = 2 &GOTO -TWO
&IF &OPT = 3 &GOTO -THREE
&IF &OPT = 4 &GOTO -FOUR
&IF &OPT = 5 &GOTO -FIVE
&IF &OPT = 6 &GOTO -SIX
&STACK 'NOW LOADING , DONT TOUCH YOUR KEYBOARD'
&STACK )PCOPY APLGRAFS GSTARPLOT GDEMO
&STACK )PCOPY 990 CMSIO
&STACK ' '
&STACK 'FOR A DESCRIPTION OF THESE FUNCTIONS TYPE  : '
&STACK ' '
&STACK '   INSTRUCTIONIONS '
&STACK ' '
&STACK ' '
&STACK '  TO EXECUTE THE FUNCTION STARPLOT  TYPE  : '
&STACK ' '
&STACK '   STARPLOT    '
&STACK ' '
&GOTO -SEVEN
-TWO  &IF &OPT > 2 &GOTO -THREE
&STACK 'NOW LOADING , DONT TOUCH YOUR KEYBOARD'
&STACK )COPY APLGRAFS GBOXPLOTAB GDEMO
&STACK )PCOPY 990 CMSIO
&STACK ' '
&STACK 'FOR A DESCRIPTION OF THESE FUNCTIONS TYPE  : '
&STACK ' '
&STACK '   INSTRUCTIONIONS '
```

```
&STACK ' '
&STACK ' '
&STACK '  TO EXECUTE THE FUNCTION BOXPLOTAB TYPE  : '
&STACK ' '
&STACK '   BOXPLOTAB  '
&STACK ' '
&GOTO -SEVEN
-THREE &IF &OPT > 3 &GOTO -FOUR
&STACK 'NOW LOADING , DONT TOUCH YOUR KEYBOARD'
&STACK )PCOPY APLGRAFS GSCATPLOT GDEMO
&STACK )PCOPY 990 CMSIO
&STACK ' '
&STACK 'FOR A DESCRIPTION OF THESE FUNCTIONS TYPE  : '
&STACK ' '
&STACK '   INSTRUCTIONIONS '
&STACK ' '
&STACK ' '
&STACK '  TO EXECUTE THE FUNCTION SCATPLOT  TYPE  : '
&STACK ' '
&STACK '   SCATPLOT   '
&STACK ' '
&GOTO -SEVEN
-FOUR &IF &OPT > 4 &GOTO -FIVE
&STACK 'NOW LOADING , DONT TOUCH YOUR KEYBOARD'
&STACK )PCOPY APLGRAFS GDRAFTSMAN GDEMO
&STACK )PCOPY 990 CMSIO
&STACK ' '
&STACK 'FOR A DESCRIPTION OF THESE FUNCTIONS TYPE  : '
&STACK ' '
&STACK '   INSTRUCTIONIONS '
&STACK ' '
&STACK ' '
&STACK '  TO EXECUTE THE FUNCTION DRAFTSMAN TYPE  : '
&STACK ' '
&STACK '   DRAFTSMAN  '
&STACK ' '
&GOTO -SEVEN
-FIVE &IF &OPT > 5 &GOTO -SIX
&STACK 'NOW LOADING , DONT TOUCH YOUR KEYBOARD'
&STACK )PCOPY APLGRAFS GLOWESS GDEMO
&STACK )PCOPY 990 CMSIO
&STACK ' '
&STACK 'FOR A DESCRIPTION OF THESE FUNCTIONS TYPE  : '
&STACK ' '
&STACK '   INSTRUCTIONIONS '
&STACK ' '
&STACK ' '
&STACK '  TO EXECUTE THE FUNCTION LOWESS TYPE  : '
&STACK ' '
&STACK '   LOWESS  '
&STACK ' '
&GOTO -SEVEN
-SIX CLRSCRN
&CONTROL OFF
&BEGTYPE -PAPA

    THIS WORKSPACE CONTAINS PROGRAMS THAT MAY BE USED AS EXPLORATORY
DATA ANALYSIS TOOLS. PROGRAMS THAT ARE USED TOGETHER ARE
CONTAINED IN GROUPS.

    THE GROUPS CURRENTLY AVAILABLE ARE GDEMO,GSCATPLOT,GBOXPLOTAB,
GSTARPLOT, GDRAFTSMAN AND GLOWESS, WHERE THE G STANDS FOR GROUP.

    IF YOU HAVE COPIED THE WHOLE WORKSPACE APLGRAFS YOU CAN SEE A
LIST OF THESE GROUPS AT ANY TIME BY DROPPING INTO APL AND TYPING :


          )GRPS .
```

```
GROUPS :
======

        GDEMO..... THIS GROUP CONTAIN SOME DATA SETS TO BE USED
                           FOR ILLUSTRATION BY THE PROGRAMS IN THIS WS.

        GSCATPLOT...... THIS GROUP CONTAIN ALL OF THE PROGRAMS REQUIRED
                           TO PRODUCE SYMBOLIC SCATTER PLOT OF TWO OR MORE
                           DIMENSIONAL DATA. A BASIC DISCUSSION OF THESE
                           DISPLAYS IS CONTAINED IN 'GRAPHICAL METHODS FOR
                           DATA ANALYSIS' BY CHAMBERS (PAGE 157) .
                           TO EXECUTE THIS PROGRAM TYPE :

                                          SCATPLOT

                           AND THEN ANSWER THE QUESTIONS.
                           YOU WOULD NEED THE FOLLOWING TWO DIMENSIONAL
                           ARRAY :
                                    - ARRAY OF DATA   ( IN APL INSIDE THE
                                      WS OR OUTSIDE AS A FORTRAN FILE)

                           FOR A DEMO USE THE FOLLOWING ARRAY :

                                 DATA --------> CALHOS

                           CALHOS CONSISTS OF COST PER PATIENT IN 14 GEO-
                           GRAPHICAL DISTRICTS (ROWS) OF CALIFORNIA OVER 5
                           YEARS (COLUMNS).

        GBOXPLOTAB..... THIS GROUP CONTAINS ALL OF THE PROGRAMS REQUIRED
                           TO PRODUCE BOX PLOTTED TABLES (A COMBINATION OF
                           BOX PLOTS AND A TABLE WITH THE ORIGINAL DATA ON
                           THE SAME DISPLAY).TO EXECUTE THIS PROGRAM TYPE :

                                          BOXPLOTAB

                           AND THEN ANSWER THE QUESTIONS.
                           YOU WOULD NEED THE FOLLOWING TWO DIMENSIONAL
                           ARRAYS :
                                    - ARRAY OF DATA   ( IN APL INSIDE THE
                                      WS OR OUTSIDE AS A FORTRAN FILE ).
                                    - ARRAY OF NAMES OF COLUMNS (AN ARRAY OF
                                      DIMENSION [NCOL,20])
                                    - ARRAY OF NAMES OF ROWS (AN ARRAY OF
                                      DIMENSION [NROW,20])

                           IF YOU DONT HAVE THE ARRAYS OF NAMES THE PROGRAM
                           WILL ASK YOU TO ENTER THE NAMES ONE BY ONE.

                           FOR A DEMO USE THE FOLLOWING ARRAYS :

                                 DATA --------> CALHOS
                                 ROW NAMES ---> CALHOSR
                                 COL NAMES ---> CALHOSC

        GSTARPLOT...... THIS GROUP CONTAINS ALL OF THE PROGRAMS REQUIRED
                           TO PRODUCE STAR AND PROFILE PLOTS OF TWO OR MORE
                           DIMENSIONAL DATA. A BASIC DISCUSSION OF THESE
                           DISPLAYS IS CONTAINED IN 'GRAPHICAL METHODS FOR
                           DATA ANALYSIS' BY CHAMBERS (PAGES 158-163)
                           TO EXECUTE THIS PROGRAM TYPE :

                                          STARPLOT

                           AND THEN ANSWER THE QUESTIONS.
                           YOU WOULD NEED THE FOLLOWING TWO DIMENSIONAL
                           ARRAYS :
                                    - ARRAY OF DATA   ( IN APL INSIDE THE
```

WS OR OUTSIDE AS A FORTRAN FILE)
- ARRAY OF NAMES OF COLUMNS (AN ARRAY OF DIMENSION [NCOL,20])
- ARRAY OF NAMES OF ROWS (AN ARRAY OF DIMENSION [NROW,20])

IF YOU DONT HAVE THE ARRAYS OF NAMES THE PROGRAM WILL ASK YOU TO ENTER THE NAMES ONE BY ONE.

FOR A DEMO USE THE FOLLOWING ARRAYS :

DATA --------> CARS
ROW NAMES ---> CARSR
COL NAMES ---> CARSC

CARS IS THE CAR REPAIR DATA GIVEN BY CHAMBERS, ET D.

GDRAFTSMAN ..... THIS GROUPS CONTAINS ALL OF THE PROGRAMS REQUIRED TO PRODUCE DRAFTSMAN DISPLAYS OF TWO OR THREE DIMENSIONAL DATA. A BASIC DISCUSSION OF THESE DISPLAYS IS CONTAINED IN 'GRAPHICAL METHODS FOR DATA ANALYSIS' BY CHAMBERS (PAGES 136-140) DETAILED EXPLANATIONS OF THESE PROGRAMS ARE CONTAINED IN 'DRAFTSMAN DISPLAY ; A GRAPHICAL EXPLORATORY DATA ANALYSIS TECHNIQUE' AN NPS THESIS BY CAPT. MALCOLM JOHNSON, USA.
THESE PROGRAMS ARE COMPLETELY INTERACTIVE AND CAN BE INITIATED BY TYPING :

DRAFTSMAN

AND THEN ANSWER THE QUESTIONS.
YOU WOULD NEED THE FOLLOWING TWO DIMENSIONAL ARRAYS :
- ARRAY OF DATA (IN APL INSIDE THE WS OR OUTSIDE AS A FORTRAN FILE)
- ARRAY OF NAMES OF COLUMNS (AN ARRAY OF DIMENSION [NCOL,20])
- ARRAY OF NAMES OF ROWS (AN ARRAY OF DIMENSION [NROW,20])

IF YOU DONT HAVE THE ARRAYS OF NAMES THE PROGRAM WILL ASK YOU TO ENTER THE NAMES ONE BY ONE.

FOR A DEMO USE THE FOLLOWING ARRAYS :

DATA --------> CARS
ROW NAMES ---> CARSR
COL NAMES ---> CARSC

GLOWESS..... THIS GROUP CONTAIN ALL OF THE PROGRAMS REQUIRED TO USE THE ROBUST LOCALLY WEIGHTED REGRESSION SCATTER PLOT SMOOTHING TECHNIQUE DESCRIBED IN 'GRAPHICAL METHODS FOR DATA ANALYSIS' BY CHAMBERS (PAGE 121). DETAILED EXPLANATION OF THESE PROGRAMS IS PRESENTED IN 'LOCALLY WEIGHTED REGRESSION AND SCATTER PLOT SMOOTHING; A GRAPHICAL EXPLORATORY DATA ANALYSIS TECHNIQUE' AN NPS THESIS BY CDR GARY W MORAN, USN.
THESE PROGRAMS ARE COMPLETELY INTERACTIVE AND CAN BE IMPLEMENTED BY TYPING :

LOWESS

AND THEN ANSWER THE QUESTIONS.
YOU WOULD NEED THE FOLLOWING TWO DIMENSIONAL ARRAYS :
- ARRAY OF DATA (IN APL INSIDE THE WS OR OUTSIDE AS A FORTRAN FILE)

```
                                    - ARRAY OF NAMES OF COLUMNS (AN ARRAY OF
                                      DIMENSION [NCOL,20])
                                    - ARRAY OF NAMES OF ROWS (AN ARRAY OF
                                      DIMENSION [NROW,20])

                                IF YOU DONT HAVE THE ARRAYS OF NAMES THE PROGRAM
                                WILL ASK YOU TO ENTER THE NAMES ONE BY ONE.

                                FOR A DEMO USE THE FOLLOWING ARRAYS :

                                      DATA --------> CARS
                                      ROW NAMES ---> CARSR
                                      COL NAMES ---> CARSC
-PAPA
&GOTO -ONE
-SEVEN EXEC APLGST
&EXIT 100
-ERROR1 &TYPE YOUR VALUE HAS TO BE BETWEEN 1 AND 6 TRY AGAIN
&GOTO -ONE
-FINAL &EXIT 100
```

2. APLGRAFS VSAPLWS

The following is a description of the content of the APL workspace APLGRAFS VSAPLWS, which contains all the functions needed to use the programs described in this thesis. This workspace contains several *groups*, each groups is related to an especific program, and contains the functions required to execute that program.

Following is a list of groups and functions inside those groups.

```
Group GBOXPLOTAB
Functions BOXPLOTAB        ADMI          BOXLINES


Group     GDRAFTSMAN
Functions DRAFTSMAN        DRASYM        DRAFT
          REPEATCK         MINMAX        TRANSFORM
          JJITTER          SUB           ADMINS
          LABELS           REGRES        REGRES2
          LOWS             YMAVS         MOVS
          MMOVAV           GARY          GARY2


Group     GLOWESS
Functions REPEATCK         LOWESS        REGRES2
          LOWS             REGRES        PLOTQUERY
          DATAINPUT


Group     GSCATPLOT
Functions MINMAX           TRANSFORM     JJITTER
          SCATPLOT         ADMI


Group     GSTARPLOT
Functions TRANSFORM        STARPLOT      ADMI
```

3. APL PROGRAMS.

This section contains the program listings of the APL programs written for this thesis, and the modified version of some existing APL programs taken from Johnson [Ref. 4] and from Moran [Ref. 5].

a. BOXPLOTTED tables (Program BOXPLOTAB)

```
[0] BOXPLOTAB;DATAO;DATA;IPL;NROW;NNCOL;NNNCOL;NCOL;PLO;
      UIND;YL;S20;DIF;LHEAD;YN;UIND1;UIND2;ORD;ORDEN;SORT;
      ORD1;XL;BAS1;COL20;XX;MEA;VAR;MED;DATA2;SCRE;TX;LX;
      SIZ;YY;COA;DATAO1;ORD1;BAS0;CONT;DATA1
[1] ADMI
[2] DATAO←DATA
[3] IPL←0
[4] (PRCD≠'Y')/0
[5] NROW←1↑DATA
[6] (NROW≤50)/L01
[7] 'THE MAXIMUM NUMBER OF ROWS ALLOWED IS 50 , TRY AGAIN'
[8] 0
[9] L01:NNNCOL←NNCOL←¯1↑DATAO
[10] L02:PLO←UIND←XL←YL←S20←20 DIF←' '
[11] 'ENTER THE SCREEN LABEL'
[12] LHEAD←⍞
[13] 'DO YOU HAVE A (NCOL 20 CHARS) MATRIX WITH THE NAMES
      OF COLUMNS Y/N?'
[14] YN←1↑⍞
[15] →(YN≠'Y')/L03
[16] 'ENTER THE NAME OF THE MATRIX'
[17] UIND1←⍞
[18] UIND1←,((NNCOL,5)' '),(UIND1[;15])
[19] →L015
[20] L03:I←0
[21] UIND1←UIND2←''
[22] L04:I←I+1
[23] 'ENTER THE LABEL FOR COLUMN NUMBER ',⍕(I)
[24] UIND1←UIND1,(20↑(S20,⍞))
[25] →(I<NNCOL)/L04
[26] L015:I←0
[27] 'DO YOU HAVE A (NROW 15 CHARS) MATRIX WITH THE NAMES
  OF ROWS Y/N?'
[28] YN←1↑⍞
[29] →(YN≠'Y')/L014
[30] 'ENTER THE NAME OF THE MATRIX'
[31] UIND2←⍞
[32] UIND2←UIND2[;15]
[33] →L055
[34] L014:I←I+1
[35] 'ENTER THE LABEL FOR ROW NUMBER ',⍕(I)
[36] UIND2←UIND2,(15↑(⍞,S20))
[37] →(I<NROW)/L014
[38] UIND2←(NROW,15)UIND2
[39] L055:ORD←NROW
[40] 'DOU YOU WANT THE DATA ORDERED BY THE FIRST COLUMN? Y/N'
[41] →('N'=YO←1↑⍞)/L056
[42] →('Y'≠YO)/L055
[43] ORD←⍒DATAO[;1]
[44] DATAO←DATAO[ORD;]
[45] UIND2←UIND2[ORD;]
[46] L056:I←1
[47] JOR←DATAO
[48] JOR[;1]←ORD
[49] L057:I←I+1
[50] JOR[;I]←⍒DATAO[;I]
[51] →(I<NNCOL)/L057
[52] L05:→(NNCOL≤6)/L06
```

```
[53] IPL←IPL+1
[54] NCOL←6
[55] NNCOL←NNCOL-6
[56] →L07
[57] L06:NCOL←NNCOL
[58] IPL←IPL+1
[59] NNCOL←0
[60] L07:DATA←DATAO[;(((IPL-1)6)+NCOL)]
[61] JER←JOR[;(((IPL-1)6)+NCOL)]
[62] CONT←,⍉(NROW,(NCOL))((NCOL)+1)
[63] DATA1←,⍉DATA
[64] BAS0←'⍝⍝50∇CONT∇DATA1∇0∇Δ∇BOX;0 1 .2∇N∇PLO∇LHEAD∇XL
     -YL-'
[65] BAS0←BAS0,'.16 .20 .92 .85∇∇LIN 1.8 7.2∇LIN∇1 1 0∇
     0 1 0 0∇'
[66] RUN BAS0
[67] COL20←(NCOL+1)20
[68] XX←COL20
[69] I←1
[70] RHO←(NCOL-1)0
[71] LRHO1:I←I+1
[72] TIE←+⌿(JOR[;I-1]=JOR[;I])
[73] →(TIE>(NROW 2))/LRHO2
[74] RHO[I-1]←1-(6(+⌿((JOR[;I-1]-JOR[;I])*2))(NROW
     ((NROW*2)-1)))
[75] →LRHO3
[76] LRHO2:N1←NROW(((NROW+1)2)*2)
[77] RHO[I-1]←(((+⌿(JOR[;I-1]*2))-N1)*0.5)(((+⌿(JOR[;I]
     *2))-N1)*0.5)
[78] RHO[I-1]←((+⌿(JOR[;I-1]JOR[;I]))-N1)RHO[I-1]
[79] LRHO3:→(I<(NCOL))/LRHO1
[80] CORR←(COL20,1)((⍕20↑S20,'RANK CORR..-'),(20 4⍕RHO),
     (20''))
[81] MEA←(COL20,1)((⍕20↑S20,'MEAN .-'),(20 4⍕MEAN DATA))
[82] VAR←(COL20,1)((⍕20↑S20,'VARIANCE .-'),(20 4⍕VARIANCE
     DATA))
[83] MED←(COL20,1)((⍕20↑S20,'MEDIAN .-'),(20 4⍕(COL20
     MEDIAN DATA)))
[84] DATA2←⍉(CORR,MEA,VAR,MED)
[85] UIND←'  OBSERVATION ',UIND1[((IPL-1)120)+(NCOL 20)]
[86] UIND←(COL20,1)UIND
[87] SCRE←0,0.85,0.98,0.9
[88] TX←1
[89] LX←0
[90] SIZ←6
-91- BAS1-'--10-XX-YY---UIND;;SIZ---S20-----SCRE--LIN 0'
-92- BAS1←BAS1,'140-LIN LX TX∇1 0 0∇0 1 0 0∇'
[93] YY←COL20 1
[94] RUN BAS1
[95] SCRE←0,0.05,0.98,0.15
[96] TX←4
[97] I←0
[98] SIZ←5
[99] M00:I←I+1
[100] (I>4)/M001
[101] UIND←(COL20,1)DATA2[(5-I);]
[102] YY←COL20 I
[103] RUN BAS1
[104] M00
[105] M001:I←0
[106] TX←NROW
[107] SIZ←2
[108] LX←1
[109] SCRE←0,0.2,0.98,0.85
[110] M01:I←I+1
[111] UIND←(COL20,1)((UIND2[I;]),(5 0⍕ORD[I]),
      (20 2⍕DATA[I;]))
[112] YY←COL20((NROW+1)-I)
[113] RUN BAS1
```

```
[114] (I<NROW)/M01
[115] (YO≠'Y')/M02
[116] PAUSE
[117] 'DO YOU WANT TO JOIN WITH LINES DATA POINTS OF THE
      SAME POSITION'
[118] ('Y'≠1↑⍞)/M02
[119] BOX:'ENTER THE POSITION OF THE DATA POINT (ENTER 0
      TO FINISH)'
[120] (0=DP←⎕)/M02
[121] ZZ←JER[DP;] BOXLINES DATA
[122] BOX
[123] M02:PAUSE
[124] (NNCOL>0)/L05
[125] ('Y'=1↑DIF)/0
[126] TUMA:'DO YOU WANT TO SEE THE DIFFERENCES BETWEEN
      COLUMNS Y/N?'
[127] ('Y'≠DIF←1↑⍞)/0
[128] IPL←0
[129] NNCOL←NNNCOL-1
[130] COA←(NNCOL,1) NNCOL
[131] 'DO YOU WANT ABSOLUTE DIFFERENCES (A) OR RELATIVE
      DIFF. (R)'
[132] ('A'≠DIF1←1↑⍞)/TUMA1
[133] LHEAD←'ABSOLUTE DIFFERENCES BETWEEN COLUMNS'
[134] 'THE DIFFER RELATIVE TO THE FIRST COLUMN (F) OR THE
      PREVIOUS (P) ?'
[135] ('P'≠1↑DF←⍞)/TUMA01
[136] DATAO←|(DATAO[;NNCOL]-DATAO[;(1+NNCOL)])
[137] TUMA2
[138] TUMA01:DATAO1←⍉(NNCOL,NROW) DATAO[;1]
[139] DATAO←|DATAO1-DATA[;(1+NNCOL)]
[140] COA←(NNCOL,1) 1
[141] TUMA2
[142] TUMA1:LHEAD←'RELATIVE DIFFERENCES BETWEEN COLUMNS'
[143] 'THE DIFFER. RELATIVE TO THE FIRST COLUMN (F) OR THE
      PREVIOUS (P) ?'
[144] ('P'≠1↑⍞)/TUMA11
[145] DATAO←|((DATAO[;NNCOL]-DATAO[;(1+NNCOL)])
      DATAO[;NNCOL]) 100
[146] TUMA2
[147] TUMA11:DATAO1←⍉(NNCOL,NROW) DATAO[;1]
[148] DATAO←|((DATAO1-DATAO[;(1+NNCOL)]) DATAO1) 100
[149] COA←(NNCOL,1) 1
[150] TUMA2:AA←(NNCOL,15)' DIFF. BET.'
[151] UIND1←,AA,(2 0 ⍕(COA)),((NNCOL,1) '-'),(2 0
      ⍕((NNCOL,1) (1+NNCOL)))
[152] UIND2←UIND2[ORD;]
[153] L055
```

b. STAR plots and PROFILE plots (Program STARPLOT)

```
[0] STARPLOT;PRCD;ANS;SCOL;SROW;NCOL;NUP;INC;MAX;SINT
    ;COST;M;BAS;I;MIN;SPA;TC;PI;ONE;R;C;POSN;XAXIS;P;
    BAS0;BAS1;N;X;XX;TYP
[1] J0:'TYPE (S) FOR STAR PLOT OR (P) FOR PROFILE PLOT '
[2] TYP←1↑⍞
[3] ((TYP≠'S')∧(TYP≠'P'))/J0
[4] ADMI
[5] (PRCD≠'Y')/0
[6] ONE←I←0
[7] NCOL←¯1↑(DATA)
[8] NROW←1↑(DATA)
[9] 'DO YOU HAVE A (NROW 20 CHARS) MATRIX WITH THE NAMES OF
    ROWS Y/N?'
[10] →('Y'≠(1↑⍞))/J00
[11] 'ENTER THE NAME OF THE MATRIX OF NAMES'
```

```
[12] N←⍬
[13] →J01
[14] J00:I←I+1
[15] 'ENTER THE NAME FOR ROW NUMBER ',⍕(I)
[16] N←N,(20↑(⍞,(20 ' ')))
[17] →(I<NROW)/J00
[18] J01:I←0
[19] 'DO YOU HAVE A (NCOL 20 CHARS) MATRIX WITH THE NAMES OF
     COLUMNS Y/N?'
[20] →('Y'≠(1↑⍞))/J02
[21] 'ENTER THE MATRIX WITH THE NAMES'
[22] NC←⍬
[23] →J03
[24] J02:I←I+1
[25] 'ENTER THE NAME FOR COLUMN NUMBER ',⍕(I)
[26] N←N,(20↑(⍞,(20 ' ')))
[27] →(I<NCOL)/J02
[28] J03:'DO YOU WANT ALL COLUMNS OF YOUR MATRIX OR
      SELECTED COL. ALL/SEL?'
[29] ANS←1↑⍞
[30] →(ANS≠'S')/K01
[31] 'ENTER AS A VECTOR THE SELECTED COLUMNS '
[32] DATA←DATA[;SCOL←⍬]
[33] K01:'DO YOU WANT ALL THE ROWS OF YOUR MATRIX OR
      SELECTED ROWS (ALL/SEL)'
[34] ANS←1↑⍞
[35] →(ANS≠'S')/K02
[36] 'ENTER AS A VECTOR THE SELECTED ROWS '
[37] DATA←DATA[SROW←⍬;]
[38] K02:TRANSFORM
[39] NCOL← 1↑(DATA)
[40] NROW←1↑(DATA)
[41] CON1:'ENTER NUMBER OF PLOTS PER SCREEN  (3 4 OR 5) '
[42] NUP←⍬
[43] →CON ((NUP>2)∧(NUP<6))
[44] 'NUMBER OF PLOTS MUST BE 3 4 OR 5 , TRY AGAIN'
[45] →CON1
[46] CON:INC←0.95 NUP
[47] ONE←(NCOL,1)⍴1
[48] MAX←MIN←NCOL⍴0
[49] →(TYP='S')/L0
[50] XX←(NCOL,1)⍴X←((1 NCOL)⍴(NCOL)-1))
[51] M←((NCOL,1)⍴0),XX,ONE,ONE,XX,((NCOL,1)⍴0)
[52] M←(((2 NCOL),3)⍴M),[1]((2,3)⍴1,X[NCOL],0,1,0,0))
[53] →LL0
[54] L0:SINT←1○(○2÷(NCOL)-1))×NCOL
[55] COST←2○(○2÷(NCOL)-1))×NCOL
[56] M←ONE,((NCOL,1)⍴(((0.8×COST)+1)÷2)),((NCOL,1)
      ⍴(((0.8×SINT)+1)÷2))
[57] M←(((2 NCOL),3)⍴(M,((NCOL,3)⍴1,0.5,0.5))))
[58] LL0:BAS←'⎕⎕12∇M∇3∇.0 .0 RP∇RP∇OFF∇1∇LIN∇LIN∇OFF∇'
[59] RUN BAS
[60] ANG←I←0
[61] →(TYP='P')/LL01
[62] M←((NCOL,1)⍴((COST+1)÷2)),((NCOL,1)⍴((SINT+1)÷2))
[63] →L00
[64] LL01:M←((NCOL,1)⍴(((1 NCOL)⍴(NCOL)-1)))-0.02),
      ((NCOL,1)⍴0.5)
[65] ANG←90
[66] L00:I←I+1
[67] NAM←NC[I;]
[68] POSN←(⍕(M[I;1],M[I;2])),' RP'
[69] BAS←'⎕⎕2∇NAM∇0∇C∇ANG∇6∇NO∇NO∇',POSN,'∇FF∇'
[70] RUN BAS
[71] MIN[I]←DATA[1↑⍋DATA[;I];I]
[72] MAX[I]←DATA[(⍴1)↑⍒DATA[;I];I]
[73] →(I<NCOL)/L00
[74] PAUSE
[75] SPA←''
```

```
[76] I←0
[77] TC←-NUP
[78] LOOP3:TC←TC+NUP
[79] PI←0.05,(1-(INC-(INC 6))),(0.05+(INC-(INC 6))),1
[80] R←0
[81] LOOP2:R←R+1
[82] C←0
[83] LOOP1:C←C+1
[84] I←I+1
[85] POSN←PI+((INC,(-INC),INC,(-INC))((C-1),(R-1),(C-1),
      (R-1)))
[86] XAXIS←N[I;]
[87] P←(DATA[I;]-MIN)(MAX-MIN)
[88] →(TYP='S')/LOOO
[89] M←(1,0,0),[1](ONE,XX,P),[1]((3,3)(1,X[NCOL],0,1,0,0))
[90] →MOOO
[91] LOOO:M←ONE,((NCOL,1)(((P COST)+1)2)),((NCOL,1)
      (((P SINT)+1)2))
[92] M←M,[1](((2 NCOL),3)(M,((NCOL,3)(1,0.5,0.5))))
[93] MOOO:BASO←'⍝⍝12∇M∇1∇.0 .0 RP∇RP∇OFF∇POSN∇LIN∇LIN∇OFF∇'
[94] BAS1←'⍝⍝2∇XAXIS∇0∇C∇0∇6∇YES∇NO∇.5 0.07 RP∇ON∇'
[95] RUN BASO
[96] RUN BAS1
[97] →(I≥NROW)/ENDO
[98] →((TC+C)≥NROW)/END
[99] →((C<NUP)∧((TC+C)<NROW))/LOOP1
[100] (R<NUP)/LOOP2
[101] END:PAUSE
[102] ((TC+C)<NROW)/LOOP3
[103] ENDO:PAUSE
```

c. CODED SCATTER plots (Program SCATPLOT)

```
[0] SCATPLOT;QUE1;CX;CY;I;DATA1;LHEAD;LPLOT;NCOL;LABX
    ;LABY;EXX;EXY;POSN;POSI;EXPRE;SYM;COL;SYZ;DESCRI;
    POSLEG;A1;X;Y;SPA;X;Y;PLOT1;PLOTO;PLOTLEG
[1] POSI←0
[2] ADMI
[3] DATA1←DATA
[4] →END (PRCD≠'Y')
[5] →ONE (DIM=2)
[6] 'YOUR DATA IS NOT A TWO DIMENSIONAL ARRAY, SCATPLOT BEING
 TERMINATED'
[7] 'PLEASE REFORMAT YOUR DATA AND START AGAIN'
[8] →END
[9] ONE:DATA←DATA1
[10] SPA←' '
[11] NCOL←¯1↑(DATA)
[12] (POSI≠0)/ONEOO
[13] 'ENTER THE SCREEN HEADER '
[14] LHEAD←⍞
[15] ONEOO:'ENTER THE PLOT HEADER '
[16] LPLOT←⍞
[17] 'ENTER THE VARIABLE (COLUMN) FOR THE X AXIS'
[18] X←DATA[;CX←⍞]
[19] 'ENTER THE LABEL FOR THE X AXIS '
[20] LABX←⍞
[21] 'DO YOU WANT ALL THE VALUES OF X OR JUST A SUBSAMPLE
      OF IT (ALL/SUB)'
[22] QUE1←1↑⍞
[23] TWO (QUE1='A')
[24] 'ENTER AN APL EXPRESSION WITH THE RANGE OF VALUES FOR X'
[25] 'E.G. (DATA[;',(⍕CX),']≥500)∧(DATA[;',(⍕CX),']≤1000)'
[26] EXX←⍞
[27] DATA←EXX/DATA
[28] TWO:'ENTER THE VARIABLE (COLUMN) FOR THE Y AXIS'
```

```
[29] Y←DATA[;CY←⎕]
[30] 'ENTER THE LABEL FOR THE Y AXIS '
[31] LABY←⍞
[32] 'DO YOU WANT ALL THE VALUES OF Y OR JUST A SUBSAMPLE
     OF IT (ALL/SUB)'
[33] QUE1←1↑⍞
[34] TWO1 (QUE1='A')
[35] 'ENTER AN APL EXPRESSION WITH THE RANGE OF VALUES FOR Y'
[36] 'E.G. (DATA[;',(⍕CY),']≥500)∧(DATA[;',(⍕CY),']≤1000)'
[37] EXY←⎕
[38] DATA←EXY⌿DATA
[39] TWO1:Y←DATA[;CY]
[40] X←DATA[;CX]
[41] JJITTER
[42] TRANSFORM
[43] MINMAX
[44] I←0
[45] 'ENTER THE POSITION FOR THE PLOT E.G. 1 21 22 ... '
[46] POSI←POSN←⎕
[47] LOOP1 (POSI>1)
[48] POSN← 0.1 0.1 0.8 0.8
[49] LOOP1:I←I+1
[50] ⍞←30 (⍕I)
[51] ' '
[52] 'ENTER IN AN APL EXPRESSION FOR THIS CATEGORY'
[53] 'I.E. (DATA[;4]≤.5)∧(DATA[;8]=5) '
[54] 'USE DATA AS THE NAME OF YOUR VECTOR '
[55] EXPRE←⍞
[56] 'ENTER THE SYMBOL '
[57] SYM←⍞
[58] 'ENTER THE COLOR , I.E. BLUE'
[59] COL←⍞
[60] 'ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND
     12 (BIG)'
[61] SYZ←⍞
[62] SYMBOLS←SYM,';',COL,';',SYZ
[63] FOUR (I=1)
[64] PLOT1←'⍝⍝10⍛X⍛Y⍛⍎EXPRE⍛',SYMBOLS,'⍛SPA⍛SPA⍛SPA⍛SPA'
[65] PLOT1←PLOT1,'⍛POSN⍛⍛P⍛P⍛1 0 0⍛0 1 0 0⍛'
[66] RUN PLOT1
[67] FIVE
[68] FOUR:PLOT0←'⍝⍝10⍛X⍛Y⍛⍎EXPRE⍛',SYMBOLS,'⍛LPLOT⍛LHEAD⍛
     LABX'
[69] PLOT0←PLOT0,'⍛LABY⍛POSN⍛⍛LIN LX TX⍛LIN LY TY⍛1 1 1
     ⍛0 1 0 0⍛'
[70] RUN PLOT0
[71] FIVE:→SIX (POSI>1)
[72] 'ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25
     CHARS.)'
[73] DESCRI←25↑⍞,' '
[74] DESCRI←(⍕I),'→',DESCRI,SYM,' ',(⍕SYZ)
[75] POSLEG←0.8,(0.75-(I 5)100)
[76] PLOTLEG←'⍝⍝2⍛DESCRI;',COL,'⍛0⍛L⍛0⍛3⍛YES⍛NO⍛POSLEG RS
     ⍛ON⍛'
[77] RUN PLOTLEG
[78] SIX:'DO YOU WANT ANOTHER CATEGORY (YES/NO)'
[79] QUE1←1↑⍞
[80] LOOP1 (QUE1='Y')
[81] (POSI>1)/QUE01
[82] PAUSE
[83] END
[84] QUE01:'DO YOU WANT ANOTHER PLOT (YES/NO)'
[85] QUE1←1↑⍞
[86] (QUE1='Y')/ONE
[87] PAUSE
[88] END:
```

d. **CODED DRAFTSMAN plots (Program DRAFTSMAN)**

```
[0] DRAFTSMAN;NCOL;PI;R;C;Y;TN;T2N;XAXIS;YAXIS;X;LX;TX;LY;
    TY;ANS;F;ROB;Y1;X1;YS;M;NUM;PRCD;DIM;YM;XM;UM;SKP
[1] ADMINS
[2] SPA←SYMBOLS←LPLOT←LHEAD←XAXIS←YAXIS←' '
[3] SYMBOLS←'o'
[4] EXP←'Δ'
[5] →LP1 (PRCD='Y')
[6] →0
[7] LP1:→LP2 (DIM>3)
[8] →(LP2,LP3,LP4)[DIM]
[9] LP2:'YOUR DATA SET IS NOT A TWO OR THREE DIMENSIONAL
    ARRAY.'
[10] 'DRAFTSMAN IS BEING TERMINATED. PLEASES REFORMAT YOUR
     DATA AND'
[11] 'REINITIATE DRAFTSMAN'
[12] 0
[13] LP4:N DRAFT DATA
[14] 0
[15] LP3:
[16] NCOL←¯1↑(DATA)
[17] JJITTER
[18] TRANSFORM
[19] GARY
[20] 'DO YOU WANT A SYMBOLIC DRAFTSMAN (YES/NO)'
[21] QUE1←1↑⍞
[22] CON1 (QUE1≠'Y')
[23] XX←DATA
[24] NCOL←DRASYM DATA
[25] LHEAD←' '
[26] LPLOT←' '
[27] 'YOU HAVE NOW ',(⍕NCOL),' BASIC VARIABLES TO PLOT'
[28] CON1:'ENTER NUMBER OF PLOTS PER SCREEN (3 4 OR 5)'
[29] NUP←⍞
[30] CON ((NUP>2)∧(NUP<6))
[31] 'NUMBER OF PLOTS MUST BE 3 4 OR 5 , TRY AGAIN'
[32] CON1
[33] CON:TR←-NUP
[34] INC←0.95÷NUP
[35] LOOP4:TR←TR+NUP
[36] TC←-NUP
[37] LOOP3:TC←TC+NUP
[38] WI←0.05,(1-(INC-(INC 6))),(0.05+(INC-(INC 6))),1
[39] R←0
[40] LOOP2:R←R+1
[41] C←0
[42] Y←DATA[;(TR+R)]
[43] LOOP1:C←C+1
[44] X←DATA[;(TC+C)]
[45] ((TR+R)=(TC+C))/SKIP
[46] POSN←WI+((INC,(-INC),INC,(-INC))((C-1),(R-1),(C-1),
     (R-1)))
[47] XAXIS←N[(TC+C);]
[48] YAXIS←N[(TR+R);]
[49] ((C=1)∧((R=NUP)∨((TR+R)=NCOL)))/GRAPH
[50] XAXIS←''
[51] (C=1)/GRAPH
[52] XAXIS←N[(TC+C);]
[53] YAXIS←''
[54] ((R=NUP)∨((TR+R)=NCOL))/GRAPH
[55] XAXIS←YAXIS←''
[56] GRAPH:MINMAX
[57] (ANS≠'Y')/FIN
[58] (SMT='M')/MOV
[59] X LOWS Y
[60] SMOOTH←'⍝4⍛X⍛Y;YS⍛0 1⍛1⍛.⍛SPA⍛SPA⍛XAXIS⍛YAXIS⍛POSN'
[61] SMOOTH←SMOOTH,'⍛LIN LX TX⍛LIN LY TY⍛1 1 1⍛10 11 0 0'
```

```
[62] RUN SMOOTH
[63] SKIP
[64] MOV:M MOVS Y[⍋X]
[65] YM←UM
[66] M MOVS X[⍋X]
[67] XM←UM
[68] SMOOTH1←'⍞⍞▽X;XM▽Y;YM▽0 1▽1▽.▽SPA▽SPA▽XAXIS▽YAXIS▽POSN▽'
[69] SMOOTH1←SMOOTH1,'LIN LX TX▽LIN LY TY▽1 1 1▽10 11 0 0'
[70] RUN SMOOTH1
[71] SKIP
[72] FIN:BAS←'⍞⍞10▽X▽Y▽',EXP,'▽',SYMBOLS,'▽LPLOT▽LHEAD▽
      XAXIS▽'
[73] BAS←BAS,'YAXIS▽POSN▽▽LIN LX TX▽LIN LY TY▽1 1 1
      ▽10 11 0 0▽'
[74] RUN BAS
[75] SKIP:→(((TR+R)≥(NCOL))∧((TC+C)≥NCOL))/END
[76] ((C<NUP)∧((TC+C)<NCOL))/LOOP1
[77] ((R<NUP)∧((TR+R)<NCOL))/LOOP2
[78] END:→(ANS='Y')/SKIP1
[79] WI GARY2 INC
[80] SKIP1:PAUSE
[81] ERASE
[82] ((TC+C)<NCOL)/LOOP3
[83] ((TR+R)<NCOL)/LOOP4
```

e. Suporting Sub-programs

```
FUNTION ADMI

[0] ADMI;QR1;QR2
[1] ⍝
[2] ⍝   FUNCTION ADMI  CALLED BY FUNCTION SCATPLOT, USES
[3] ⍝   FUNCTION CMSREAD, THIS FUNCTION IS A MODIFIED
[4] ⍝   VERSION OF THE FUNCTION ADMINS FROM DTNLFNS VSAPLWS.
[5] ⍝
[6] PRCD←'Y'
[7] 'IS YOUR DATA SET LOCATED IN THIS WORKSPACE? (YES/NO)'
[8] QR1←1↑⍞
[9] LP1 (QR1≠'Y')
[10] GO
[11] LP1:'IS YOUR DATA SET LOCATED:'
[12] '(1) IN AN APL WORKSPACE LOCATED ON THIS DISK OR ON A DISK'
[13] '  THAT YOU ARE LINDED TO'
[14] '(2) IN A CMS FILE ON THIS DISK OR ON A DISK THAT YOU ARE'
[15] '  LINKED TO'
[16] '(3) NIETHER (1) OR (2) ABOVE'
[17] 'ENTER (1,2 OR 3)'
[18] QR2←⍞
[19] (LP2,LP3,LP4)[QR2]
[20] LP2:'TO TRANSFER YOUR DATA TO THIS WORKSPACE:'
[21] '(1) TYPE . . . )PCOPY (WS NAME) (DATA SET NAME)'
[22] '  EXAMPLE: )PCOPY DTNLDATA CARS'
[23] ' '
[24] '  DATE AND TIME SAVED INFORMATION IS DISPLAYED'
[25] '  WHEN THE TRANSFER IS COMPLETE. THEN ENTER GO'
[26] '  TO PROCED WITH SCATPLOT'
[27] SΔADMI←GO
[28] GO:'ENTER THE NAME OF YOUR DATA SET '
[29] DATA←⍞
[30] DIM← DATA
[31] END
[32] LP3:'TO TRANSFER YOUR CMS DATA FILE TO THIS WORKSPACE'
[33] 'ANSWER THE FOLLOWING QUESTIONS ABOUT YOUR DATA SET'
[34] DATA←CMSREAD
[35] DIM← DATA
[36] END
```

```
[37] LP4:'YOUR DATA SET MUST BE STORED IN AN APL WORKSPACE OR '
[38] 'IN A CMS FILE LOCATED ON THIS DISK OR ON A DISK TO WHICH '
[39] 'YOU ARE LINKED. SCATPLOT IS BEING TERMINATED. PLEASE '
[40] 'COMPLY WITH CONDITIONS (1) OR (2) AND REINITIATE SCATPLOT'
[41] PRCD←'N'
[42] END:
```

FUNTION BOXLINES

```
[0] DAT←JQR BOXLINES DATA
[1] NCOL← 1↑ DATA
[2] NROW←1↑ DATA
[3] MX←⍳(1+ (NCOL-1))
[4] MX←1 ,MX←, ⍳(2, MX) (MX,MX)
[5] JM←(( 1↑MX),1) (0 1)
[6] JOR←,(1-((1 (NROW-1))JOR←JOR[MX]-1))
[7] MX←((NROW ( 1↑MX)),1) ((MX (2 14))+(1↑MX) ((21 140),
    (11 140)))
[8] JM←(( MX) JM),MX,(( MX) JOR)
[9] BAS2←'⎕⎕12▽JM▽3▽.0 .0 RP▽RP▽ON▽ 0 .20 .98 .85 ▽LIN▽LIN▽OFF▽'
[10] RUN BAS2
[11] END:DAT←0
```

FUNTION DRASYM

```
[0] NCOL←DRASYM MATRIX;CI;CV;I;SYM;COL;SYZ;ANS
[1] 'ENTER AS A VECTOR THE VARIABLES (COLUMNS) THAT YOU
     WHISH TO HAVE'
[2] 'IN THE X AND Y AXIS (THE FIRST AND SECOND DIMENSION
     FOR THE PLOT)'
[3] CI←⍞
[4] N←N[(CI);]
[5] DATA←MATRIX[;CI]
[6] NCOL← CI
[7] I←0
[8] EXP←SYM←COL←SYZ←' '
[9] 'NEXT, YOU HAVE TO ENTER APL EXPRESSION FOR EACH
     CATEGORY (CODE)'
[10] 'USE XX AS THE NAME OF YOUR ARRAY'
[11] ' '
[12] ' I.E. (XX[;I]>100)∧(XX[;J]=400) '
[13] ' '
[14] 'WHERE I AND J REPRESENT COLUMN NUMBERS BETWEEN 1
     AND ',(⍕ CI)
[15] 'BE CAREFULLY NOT TO OVERLAP VALUES '
[16] ' '
[17] 'WHEN THE PROGRAM ASK FOR SYMBOLS TYPE ANY (ONE)
     CHARACTER'
[18] 'FOR COLORS TYPE THE NAME OF THE COLOR I.E. BLUE OR RED'
[19] 'WITH SIZES 1 REPRESENT SMALL AND 12 BIG'
[20] ' '
[21] LOOP1:I←I+1
[22] 'ENTER THE APL EXPRESSION FOR THE CATEGORY (CODE)
     NUMBER ',(⍕I)
[23] EXP←EXP,':',⍞
[24] 'ENTER THE SYMBOL'
[25] SYM←SYM,⍞
[26] 'ENTER THE COLOR'
[27] COL←COL,' ',⍞
[28] 'ENTER THE SIZE'
[29] SYZ←SYZ,' ',⍞
[30] 'DO YOU WHISH ANOTHER CATEGORY  (YES/NO)'
[31] ANS←1↑⍞
[32] LOOP1 (ANS='Y')
[33] EXP←2↓EXP
```

```
[34] SYMBOLS←(1↓SYM),';',(2↓COL),';',(2↓SYZ)


FUNTION DRAFT

[0] N DRAFT M;DATA;NCOL;TR;TC;PI;R;C;Y;TN;T2N;XAXIS;YAXIS;
    LX;TX;LY;TY;ANS;F;ROB;Y1;X1;YS;M;NUM;NPAG;VAR;MORE;XU;
    X;POSN;YU
[1] ⍝⍝⍝ DO NOT MOVE OR ERASE; GRAFSTAT FUNCTION HEADER
[2] ⍝⍝⍝ GRAFSTAT WILL NOT ADD A LINE TO THIS FUNCTION
     WITHOUT THIS HEADER
[3] 'THE THREE DIMENSIONAL DRAFTSMAN DISPLAY IS BUILT
     ONE VARIABLE AT A'
[4] 'TIME. THE PROGRAM WILL ASK YOU WHICH VARIABLE YOU
     WANT TO LOOK AT'
[5] 'EACH TIME IT IS READY FOR A NEW ONE. THE DISPLAY
     PRESENTED FOR EACH'
[6] 'VARIABLE REPRESENTS THAT VARIABLE PLOTTED AGAINST
     ALL OTHER'
[7] 'VARIABLES PAGE BY PAGE. THAT IS, THE FIRST ROW
     REPRESENTS THE FIRST'
[8] 'PAGE OF DATA , THE SECOND ROW REPRESENTS THE SECOND
     PAGE AND SO ON'
[9] DATA←M
[10] SPA←''
[11] NCOL←2↓⍴DATA
[12] NPAG←1↑⍴DATA
[13] LOOP5:'WHAT VARIABLE DO YOU WANT TO LOOK AT?'
[14] ((⍕(NCOL,1) NCOL),[2](⍕(NCOL,1)' ')),[2] N
[15] VAR←⍞
[16] XU←XU+0.1 XU←⌈/⌈/DATA[;;(VAR)]
[17] (⍕N[(VAR);]),' WILL BE PLOTED AS THE INDEPENDENT (X
     VARIABLE)'
[18] 'AND ALL OTHERS WILL BE PLOTTED AS DEPENDENT (Y
     VARIABLES).'
[19] JITTER
[20] TRANSFORM
[21] GARY
[22] CON:'ENTER # OF PLOTS PER SCREEN (3,4 OR 5)'
[23] NUP←1↑⍞
[24] ((NUP<3)∨(NUP>5))\CON
[25] INC←0.95 NUP
[26] PI←0.05,(1-(INC+(INC 6))),(0.05+(INC-(INC 6))),1
[27] TR← NUP
[28] LOOP4:TR←TR+NUP
[29] TC← NUP
[30] LOOP3:TC←TC+NUP
[31] R←0
[32] LOOP2:R←R+1
[33] C←0
[34] X←DATA[(TR+R);;VAR]
[35] LOOP1:C←C+1
[36] Y←DATA[(TR+R);;(TC+C)]
[37] YU←YU+0.1 YU←⌈/⌈/DATA[;;(TC+C)]
[38] ((VAR)=(TC+C))/SKIP
[39] POSN←PI+((INC,(¯INC),INC,(¯INC))((C-1),(R-1),(C-1),
     (R-1)))
[40] XAXIS←N[(TC+C);]
[41] YAXIS←N[(VAR);]
[42] ((C=1)∧((R=NUP)∨((TR+R)=NPAG)))/GRAPH
[43] XAXIS←''
[44] (C=1)/GRAPH
[45] XAXIS←N[(TC+C);]
[46] YAXIS←''
[47] ((R=NUP)∨((TR+R)=NPAG))/GRAPH
[48] XAXIS←YAXIS←''
[49] GRAPH:MINMAX
[50] (ANS≠'Y')/FIN
```

```
[51] (SMT='M')/MOV
[52] X LOWS Y
[53] SMOOTH3←'⍞4⍛X⍛Y;YS⍛0 1⍛1⍛.⍛SPA⍛SPA⍛XAXIS⍛YAXIS⍛POSN⍛
     LIN LX XU⍛'
[54] SMOOTH3←SMOOTH3,'LIN LY YU⍛1 1 1⍛10 11 0 0'
[55] RUN SMOOTH3
[56] SKIP
[57] MOV:M MMOVAV Y[⍋X]
[58] YM←YMAV
[59] M MMOVAV X[⍋X]
[60] XM←YMAV
[61] SMOOTH13←'⍞4⍛X;XM⍛Y;YM⍛0 1⍛1⍛.⍛SPA⍛SPA⍛XAXIS⍛YAXIS⍛POSN
     ⍛LIN LX XU⍛'
[62] SMOOTH13←SMOOTH13,'LIN LY YU⍛1 1 1⍛10 11 0 0'
[63] RUN SMOOTH13
[64] SKIP
[65] FIN:BASIC3←'⍞4⍛X⍛Y⍛0⍛1⍛.⍛SPA⍛SPA⍛XAXIS⍛YAXIS⍛POSN⍛
     LIN LX XU⍛'
[66] BASIC3←BASIC3,'LIN LY YU⍛1 1 1⍛10 11 0 0'
[67] RUN BASIC3
[68] SKIP:→(((TR+R)≥(NPAG))∧((TC+C)≥NCOL))/END
[69] ((C<NUP)∧((TC+C)<NCOL))/LOOP1
[70] ((R<NUP)∧((TR+R)<NPAG))/LOOP2
[71] END:→(ANS='Y')/SKIP1
[72] GARY2
[73] SKIP1:PAUSE
[74] ERASE
[75] ((TC+C)<NCOL)/LOOP3
[76] ((TR+R)<NPAG)/LOOP4
[77] 'DO YOU WANT TO LOOK AT ANOTHER VARIABLE?'
[78] MORE←1↑⍞
[79] LOOP5 (MORE='Y')
```

FUNTION GRAPHER

```
[0] GRAPHER;GR1;GR2;GR3;ANS3;YS;Y1;X1;X;Y;ANS3;PRCD;
     DIM;N;REG
[1] ⍞⍞⍞ DO NOT MOVE OR ERASE; GRAFSTAT FUNCTION HEADER
[2] ⍞⍞⍞ GRAFSTAT WILL NOT ADD A LINE TO THIS FUNCTION
     WITHOUT THIS HEADER
[3] ADMINS
[4] NNN←N
[5] →LP1 (PRCD='Y')
[6] →0
[7] LP1:→LP2 (DIM>3)
[8] →(LP2,LP3,LP4)[DIM]
[9] LP2:'YOUR DATA SET IS NOT A TWO OR THREE DIMENSIONAL
     ARRAY.'
[10] 'GRAPHER IS BEING TERMINATED. PLEASE REFORMAT YOUR DATA
     AND'
[11] 'REINITIATE GRAPHER'
[12] 0
[13] LP4:NNN GRAPHER3 M
[14] 0
[15] LP3:
[16] NCOL←¯1↑(DATA)
[17] JITTER
[18] TRANSFORM
[19] RR:'DO YOU WANT TO CONTINUE AND PLOT? (ENTER Y OR N )'
[20] GR1←⍞
[21] (GR1≠'Y')/0
[22] 'WHAT MATRIX POSITION ARE YOU REPRODUCING?'
[23] GR2←⍞
[24] 'WHAT POSITION ON THE SCREEEN?'
[25] GR3←⍞
[26] GARY3
[27] SPA←''
```

```
[28] (ANS3≠'Y')/L1
[29] GRFRS←'⍞4⍛X⍛Y;YS⍛0 1⍛1⍛.⍛SPA⍛SPA⍛NNN[(GR2[2]);]⍛
 NNN[(GR2[1]);]'
[30] GRFRS←GRFRS,'⍛GR3⍛LIN⍛1 1 1⍛10 11 0 0'
[31] RUN GRFRS
[32] RR
[33] GRFR←'⍞4⍛DATA[;(GR2[2])]⍛DATA[;(GR2[1])]⍛.⍛1⍛∘⍛
      SPA-SPA-'
[34] GRFR←GRFR,'⍛NNN[(GR2[2]);]⍛NNN[(GR2[1]);]⍛GR3⍛
      LIN-LIN
      ⍛1 1 1⍛10 11 0 0'
[35] L1:RUN GRFR
[36] RR
```

FUNTION GRAPHER3

```
[0] NNN GRAPHER3 M
[1] ⍞⍞⍞ DO NOT MOVE OR ERASE; GRAFSTAT FUNCTION HEADER
[2] ⍞⍞⍞ GRAFSTAT WILL NOT ADD A LINE TO THIS FUNCTION
     WITHOUT THIS HEADER
[3] DATA←M
[4] NCOL← 1↑(DATA)
[5] JITTER
[6] TRANSFORM
[7] RR:'DO YOU WANT TO CONTINUE AND PLOT? (ENTER Y OR N )'
[8] GR1←⍞
[9] →(GR1≠'Y')/0
[10] 'WHAT MATRIX POSITION ARE YOU REPRODUCING?'
[11] GR2←⍞
[12] LIMITS
[13] 'WHAT POSITION ON THE SCREEEN?'
[14] GR3←⍞
[15] GARY3
[16] SPA←''
[17] (ANS3≠'Y')/L1
[18] GRFRS←'⍞4⍛X⍛Y;YS⍛0 1⍛1⍛.⍛SPA⍛SPA⍛NNN[(GR2[2]);]
      ⍛NNN[(GR2[1]);]'
[19] GRFRS←GRFRS,'⍛GR3⍛LIN⍛LIN⍛1 1 1⍛10 11 0 0'
[20] RUN GRFRS
[21] RR
[22] L1:GRFR←'⍞4⍛DATA[;(GR2[2])]⍛DATA[;(GR2[1])]⍛.⍛1
      ⍛∘⍛SPA⍛SPA⍛'
[23] GRFR←GRFR,'NNN[(GR2[2]);]⍛NNN[(GR2[1]);]⍛GR3⍛LIN
      ⍛LIN⍛1 1 1⍛10 11 0 0'
[24] RUN GRFR
[25] RR
```

FUNTION PLOTQUERY

```
[0] PLOTQUERY
[1] ''
[2] SPA←''
[3] 'DO YOU WANT A PLOT OF YOUR LOWESS SMOOTHED CURVE?'
[4] '(YES OR NO) ..... ENTER NO IF NOT USING GRAFSTAT'
[5] PT←1↑⍞
[6] →END (PT≠'Y')
[7] 'INPUT X AXIS LABEL'
[8] XAXIS←⍞
[9] 'INPUT Y AXIS LABEL'
[10] YAXIS←⍞
[11] PL1 (ROB≠'Y')
[12] PHDR←'ROBUST LOWESS SMOOTHING; F = ',⍕F
[13] RPLT←'⍞4⍛X1⍛Y1;YS⍛0 1⍛1⍛.⋆+ ∇⍙∘⍟⍋⍛⍛SPA⍛PHDR⍛XAXIS⍛
      YAXIS⍛21⍛'
[14] RPLT←RPLT,'LIN⍛LIN⍛1 1 1⍛0 1 0 0'
```

```
[15] RUN RPLT
[16] VIEW
[17] PL2
[18] PL1:PHDR←'NON-ROBUST LOWESS SMOOTHING; F = ',⍕F
[19] NRPLT←'⎕4▽X1▽Y1;YS▽0 1▽1▽.×+ ∇Δ○⊗⍋⍒▽▽SPA▽PHDR▽XAXIS▽
     YAXIS▽21▽'
[20] NRPLT←NRPLT,'LIN▽LIN▽1 1 1▽0 1 0 0'
[21] RUN NRPLT
[22] VIEW
[23] PL2:'DO YOU WANT A PLOT OF |RESIDUALS| VS X?'
[24] '(YES OR NO)'
[25] QS5←1↑⍞
[26] END (QS5≠'Y')
[27] 'DO YOU WANT THIS PLOT SMOOTHED?'
[28] '(YES OR NO)'
[29] QS6←1↑⍞
[30] XRESID←'|RESIDUALS|'
[31] PL3 (QS6≠'Y')
[32] X LOWS(|RESY)
[33] SRESPLT←'⎕⎕1▽X▽(|RESY);YS▽0
 1▽1▽.×+ ∇Δ○⊗⍋⍒▽▽SPA▽SPA▽XAXIS▽XRESID▽'
[34] SRESPLT←SRESPLT,'22▽LIN▽LIN▽1 1 1▽0 1 0 0▽'
[35] RUN SRESPLT
[36] PAUSE
[37] END
[38] PL3:RESPLT←'⎕⎕1▽X▽(|RESY)▽0▽1▽.×+ ∇Δ○⊗⍋⍒▽▽SPA▽SPA▽XAXIS▽
     XRESID▽'
[39] RESPLT←RESPLT,'22▽LIN▽LIN▽1 1 1▽0 1 0 0▽'
[40] RUN RESPLT
[41] PAUSE
[42] END:
```

APPENDIX B

SAMPLE PROGRAM EXECUTION

1. BOXPLOTTED TABLES

This program, as mentioned in Chapter III is executed by typing BOXPLOTAB, and answering the queries as follows :

```
        BOXPLOTAB

IS YOUR DATA SET LOCATED IN THIS WORKSPACE? (YES/NO)
YES
ENTER THE NAME OF YOUR DATA SET
□:
        STOCK

ENTER THE SCREEN LABEL
ACTIVE STOCKS FOR THE WEEK ENDED AUG. 8, 1986

DO YOU HAVE A (NCOL×20 CHARS) MATRIX WITH THE NAMES OF
COLUMNS Y/N?
YES

ENTER THE NAME OF THE MATRIX
□:
        STOCKC

DO YOU HAVE A (NROW×15 CHARS) MATRIX WITH THE NAMES OF
ROWS Y/N?
YES

ENTER THE NAME OF THE MATRIX
□:
        STOCKR

DOU YOU WANT THE DATA ORDERED BY THE FIRST COLUMN ? Y/N
YES


 (AT THIS POINT THE BOXPLOTTED TABLES ARE DISPLAYED ON THE
SCREEN)


 ENTER Q TO QUIT
 ENTER E TO ERASE AND CONTINUE
 ENTER C TO COPY AND CONTINUE
 ENTER CE TO COPY, ERASE AND CONTINUE
 PRESS ENTER ONLY TO CONTINUE
CE

 DO YOU WANT TO JOIN WITH LINES DATA POINTS OF THE SAME
 POSITION

 YES

 ENTER THE POSITION OF THE DATA POINT (ENTER 0 TO FINISH)
 □:
        1

 ENTER THE POSITION OF THE DATA POINT (ENTER 0 TO FINISH)
 □:
        0
```

```
ENTER Q TO QUIT
ENTER E TO ERASE AND CONTINUE
ENTER C TO COPY AND CONTINUE
ENTER CE TO COPY, ERASE AND CONTINUE
PRESS ENTER ONLY TO CONTINUE
Q
```

2. STARP PLOTS

This program, as mentioned in Chapter III is executed by typing STARPLOT, and answering the queries as follows :

```
      STARPLOT

TYPE (S) FOR STAR PLOT OR (P) FOR PROFILE PLOT

S

IS YOUR DATA SET LOCATED IN THIS WORKSPACE? (YES/NO)
YES

ENTER THE NAME OF YOUR DATA SET
□:
      AUTOS

DO YOU HAVE A (NROW×20 CHARS) ARRAY WITH NAMES OF ROWS Y/N?
YES

ENTER THE NAME OF THE MATRIX OF NAMES
□:
      AUTOSR

DO YOU HAVE A (NCOL×20 CHARS) MATRIX WITH THE NAMES OF
COLUMNS Y/N?
Y

ENTER THE MATRIX WITH THE NAMES
□:
      AUTOSC

DO YOU WANT ALL COLUMNS OF YOUR MATRIX OR SELECTED COL.
ALL/SEL?
SEL

ENTER AS A VECTOR THE SELECTED COLUMNS
□:
      ι12

DO YOU WANT ALL THE ROWS OF YOUR MATRIX OR SELECTED ROWS
(ALL/SEL)
ALL

HOW MANY VARIABLES DO YOU WANT TO HAVE TRANSFORMED ?
TYPE 0 IF YOU WANT NONE
□:
      0

ENTER NUMBER OF PLOTS PER SCREEN  (3 4 OR 5)
□:
      5

(AT THIS POINT THE STAR PLOT IS SHOWN ON THE SCREEN)
```

```
ENTER Q TO QUIT
ENTER E TO ERASE AND CONTINUE
ENTER C TO COPY AND CONTINUE
ENTER CE TO COPY, ERASE AND CONTINUE
PRESS ENTER ONLY TO CONTINUE

CE
```

3. PROFILE PLOTS

This program, as mentioned in Chapter III is executed by typing STARPLOT, and answering the queries as follows :

```
      STARPLOT

TYPE (S) FOR STAR PLOT OR (P) FOR PROFILE PLOT

P

IS YOUR DATA SET LOCATED IN THIS WORKSPACE? (YES/NO)
YES

ENTER THE NAME OF YOUR DATA SET
□:
      AUTOS

DO YOU HAVE A (NROW×20 CHARS) ARRAY WITH NAMES OF ROWS Y/N?
YES

ENTER THE NAME OF THE MATRIX OF NAMES
□:
      AUTOSR

DO YOU HAVE A (NCOL×20 CHARS) MATRIX WITH THE NAMES OF
COLUMNS Y/N?
Y

ENTER THE MATRIX WITH THE NAMES
□:
      AUTOSC

DO YOU WANT ALL COLUMNS OF YOUR MATRIX OR SELECTED COL.
ALL/SEL?
SEL

ENTER AS A VECTOR THE SELECTED COLUMNS
□:
      ι12

DO YOU WANT ALL THE ROWS OF YOUR MATRIX OR SELECTED ROWS
(ALL/SEL)
ALL

HOW MANY VARIABLES DO YOU WANT TO HAVE TRANSFORMED ?
TYPE 0 IF YOU WANT NONE
□:
      0

ENTER NUMBER OF PLOTS PER SCREEN  (3 4 OR 5)
□:
      5

(AT THIS POINT THE STAR PLOT IS SHOWN ON THE SCREEN)

ENTER Q TO QUIT
```

```
ENTER E TO ERASE AND CONTINUE
ENTER C TO COPY AND CONTINUE
ENTER CE TO COPY, ERASE AND CONTINUE
PRESS ENTER ONLY TO CONTINUE

CE
```

4. CODED SCATTER PLOTS

This program, as mentioned in Chapter III is executed by typing SCATPLOT, and answering the queries as follows :

```
      SCATPLOT
IS YOUR DATA SET LOCATED IN THIS WORKSPACE? (YES/NO)
YES

ENTER THE NAME OF YOUR DATA SET
□:
      AUTOS
FROM NOW ON YOUR DATA SET WILL BE CALLED DATA (IN THIS PROGRAM)

ENTER THE SCREEN HEADER
AUTOMOBILE DATA; PRICE VS M.P.G. CITY

ENTER THE PLOT HEADER
USA = A, FOREIGN = F AND WEIGHT = SIZE OF LETTER

ENTER THE COLUMN NUMBER FOR THE VARIABLE ON THE X-AXIS
□:
      1

ENTER THE LABEL FOR THE X AXIS
PRICE

DO YOU WANT ALL THE VALUES OF X OR JUST A SUBSAMPLE OF IT (ALL/SUB)
ALL

ENTER THE COLUMN NUMBER FOR THE VARIABLE ON THE Y-AXIS
□:
      2

ENTER THE LABEL FOR THE Y AXIS
M.P.G. CITY

DO YOU WANT ALL THE VALUES OF Y OR JUST A SUBSAMPLE OF IT (ALL/SUB)
ALL

HOW MANY VARIABLES DO YOU DESIRE JITTERED?
TYPE 0 IF YOU WANT NONE
□:
      0

HOW MANY VARIABLES DO YOU WANT TO HAVE TRANSFORMED ?
TYPE 0 IF YOU WANT NONE
□:
      0

ENTER THE POSITION FOR THE PLOT E.G. 1 21 22 ...
□:
      1

1111111111111111111111111111111
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
```

```
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]=1)∧(DATA[;8]≤2500)
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
A

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
BLUE

ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
3

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
USA ≤ 2500 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
YES

2222222222222222222222222222
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]=1)∧((DATA[;8]>2500)∧(DATA[;8]≤3000))
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
A

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
BLUE

ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
5

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
2500 < USA ≤ 3000 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
YES

3333333333333333333333333333333
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]=1)∧((DATA[;8]>3000)∧(DATA[;8]≤3500))
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
A

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
RED

ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
7

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
3000 < USA ≤ 3500 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
YES

444444444444444444444444444444444
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]=1)∧(DATA[;8]>3500)
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
A

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
RED
```

```
ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
9

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
USA > 3500 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
YES

5555555555555555555555555555555
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]≠1)∧(DATA[;8]≤2500)
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
F

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
RED

ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
3

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
FOREIGN ≤ 2500 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
YES

6666666666666666666666666666666
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]≠1)∧((DATA[;8]>2500)∧(DATA[;8]≤3000))
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
F

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
RED

ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
5

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
2500 < FOREI. ≤ 3000 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
YES

7777777777777777777777777777777
ENTER IN AN APL EXPRESSION FOR THIS CATEGORY
I.E. (DATA[;4]≤.5)∧(DATA[;8]=5)
USE DATA AS THE NAME OF YOUR VECTOR
(DATA[;13]≠1)∧((DATA[;8]>3000)∧(DATA[;8]≤3500))
ENTER THE SYMBOL (ANY LETTER,NUMBER OR SPECIAL CHARACTER)
F

ENTER THE COLOR (WHITE,GREEN,BLUE,TURQUOISE,RED,YELLOW OR PINK
RED

ENTER THE SIZE , AS A NUMBER BETWEEN 1 (SMALL) AND 12 (BIG)
7

ENTER A LABEL (DESCRIPTION) FOR THIS CATEGORY (MAX 25 CHARS.)
3000 < FOREI. ≤ 3500 LB.

DO YOU WANT ANOTHER CATEGORY (YES/NO)
NO
```

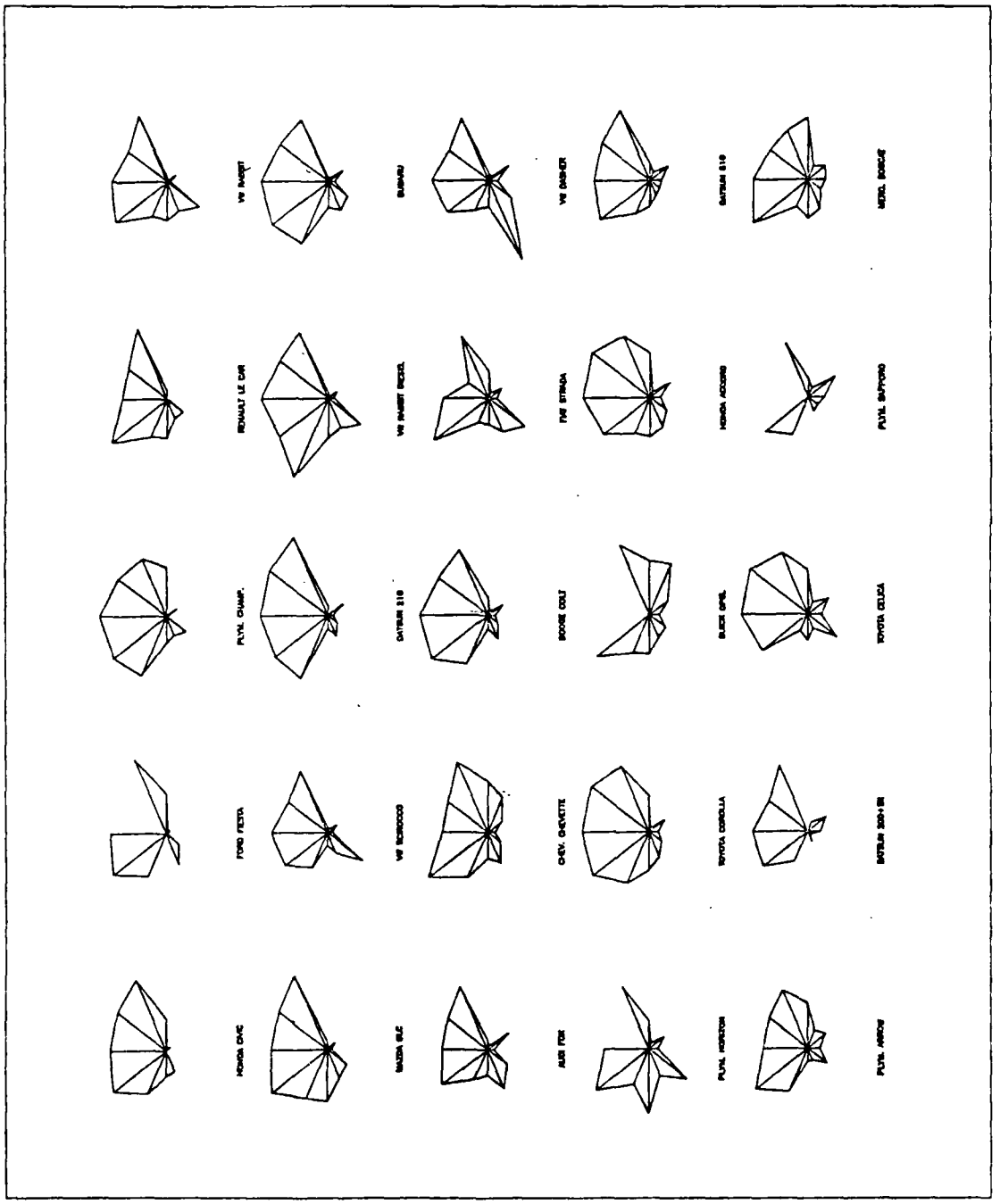
5. CODED DRAFTSMAN PLOTS

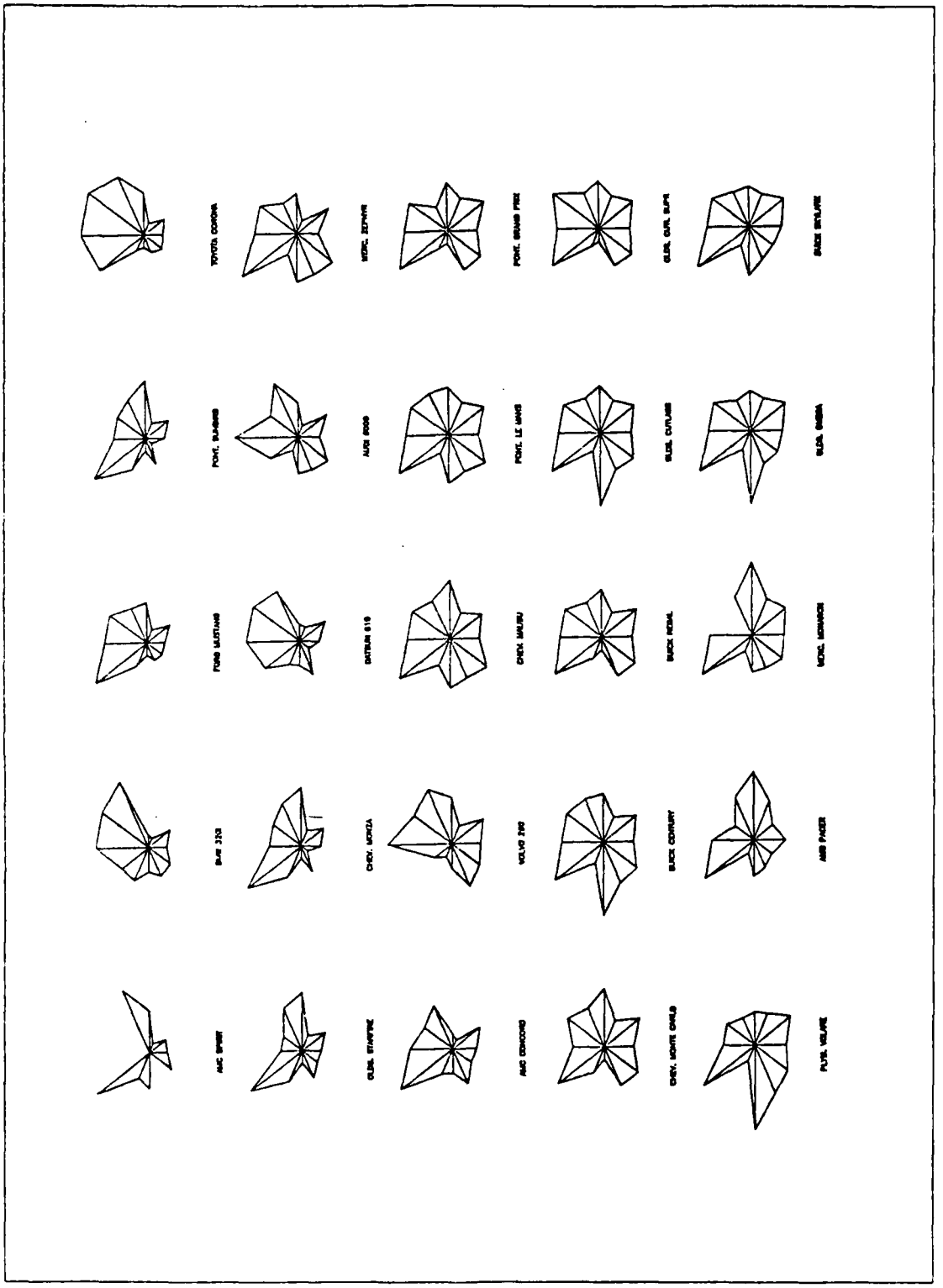
This program, as mentioned in Chapter III is executed by typing DRAFTSMAN, and answering the queries as follows :

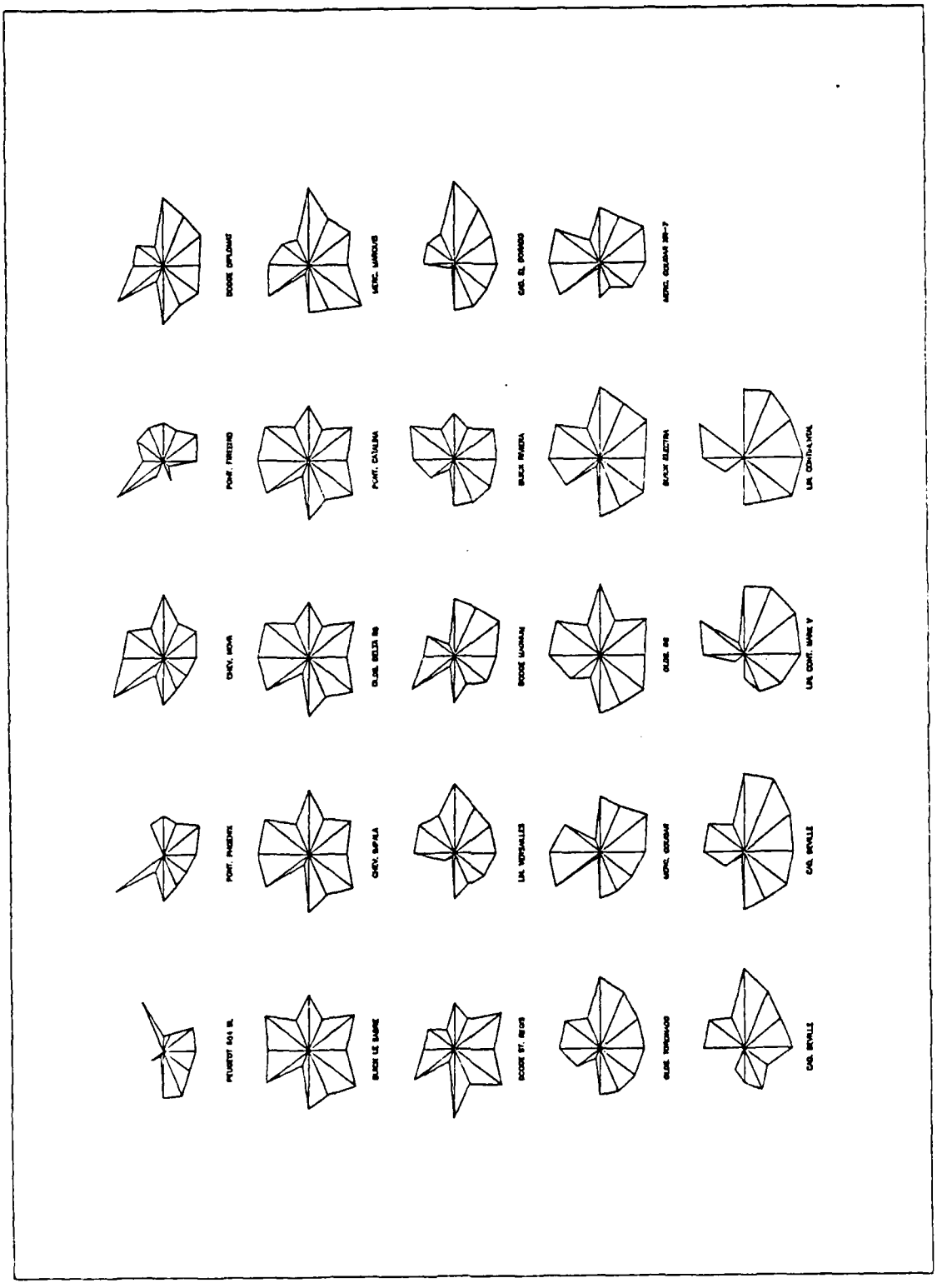
```
      DRAFTSMAN
IS YOUR DATA SET LOCATED IN THIS WORKSPACE?
(YES OR NO)
YES
ENTER THE NAME OF YOUR DATA SET
□:
      AUTOS
DO YOU WANT ALL OF THIS DATA OR JUST A SUBSAMPLE OF IT TO
BE PRESENTED IN THE DRAFTSMAN DISPLAY? ENTER (ALL OR SUB)
ALL
DO YOU HAVE A TWO DIMENSIONAL ARRAY OF NAMES FOR THE DATA
WHICH IS TO BE DISPLAYED? NOTE: THESE NAMES ARE THE NAMES
OF THE VARIABLES REPRSENTED BY THE COLUMNS OF YOUR DATA SET.
(YES OR NO)
YES
WHAT IS THE NAME OF YOUR ARRAY OF VARIABLE NAMES?
□:
      AUTOSC
HOW MANY VARIABLES DO YOU DESIRE JITTERED?
TYPE 0 IF YOU WANT NONE
□:
      0
HOW MANY VARIABLES DO YOU WANT TO HAVE TRANSFORMED ?
TYPE 0 IF YOU WANT NONE
□:
      0
DO YOU WANT TO DO WANT TO FIT A SMOOTHED CURVE
ON ALL DRAFTAMAN PLOTS? ...(YES OR NO)
NO
DO YOU WANT A SYMBOLIC DRAFTSMAN (YES/NO)
YES
ENTER AS A VECTOR THE VARIABLES (COLUMNS) THAT YOU WHISH TO HAVE
IN THE X AND Y AXIS (THE FIRST AND SECOND DIMENSION FOR THE PLOT)
□:
      1 11 8 2
NEXT, YOU HAVE TO ENTER APL EXPRESSION FOR EACH CATEGORY (CODE)
USE XX AS THE NAME OF YOUR ARRAY
  I.E. (XX[;I]>100)∧(XX[;J]=400)
WHERE I AND J REPRESENT COLUMN NUMBERS BETWEEN 1 AND 4
BE CAREFULLY NOT TO OVERLAP VALUES
WHEN THE PROGRAM ASK FOR SYMBOLS TYPE ANY (ONE) CHARACTER
FOR COLORS TYPE THE NAME OF THE COLOR I.E. BLUE OR RED
WITH SIZES 1 REPRESENT SMALL AND 12 BIG
ENTER THE APL EXPRESSION FOR THE CATEGORY (CODE) NUMBER 1
XX[;13]=1
```

```
ENTER THE SYMBOL
A

ENTER THE COLOR
RED

ENTER THE SIZE
4

DO YOU WHISH ANOTHER CATEGORY  (YES/NO)
YES

ENTER THE APL EXPRESSION FOR THE CATEGORY (CODE) NUMBER 2
XX[;13]≠1

ENTER THE SYMBOL
F

ENTER THE COLOR
RED

ENTER THE SIZE
4

DO YOU WHISH ANOTHER CATEGORY  (YES/NO)
NO

YOU HAVE NOW 4 BASIC VARIABLES TO PLOT
ENTER NUMBER OF PLOTS PER SCREEN  (3 4 OR 5)
□:
      4

DO YOU WANT TO FIT A SMOOTHED CURVE
ON SELECTED PLOTS? ...(YES OR NO)
NO

(AT THIS POINT THE CODED DRAFTSMAN PLOT IS SHOWN ON THE SCREEN)

ENTER Q TO QUIT
ENTER E TO ERASE AND CONTINUE
ENTER C TO COPY AND CONTINUE
ENTER CE TO COPY, ERASE AND CONTINUE
PRESS ENTER ONLY TO CONTINUE
CE
```

APPENDIX C
STAR PLOTS OF AUTOMOBILE DATA




TOYOTA CORONA
NISSAN ZEPHYR
PONT. GRAND PRIX
GENL MOTL ELPH
BUICK SKYLARK
PONT. BONNEV
AUDI 5000
PONT. LE MANS
GENL MOTL CUTLASS
GENL MOTL OMEGA
FORD MUSTANG
DATSUN 810
CHEV. MALIBU
BUICK REGAL
MERC. MARQUIS
BMW 320i
CHEV. MONZA
VOLVO 240
BUICK CENTURY
AMC PACER
AMC SPIRIT
GENL MOTL STARFIRE
AMC CONCORD
CHEV. MONTE CARLO
PLYM. VOLARE



LIST OF REFERENCES

1. Chambers, J. M., and others, *Graphical Methods for Data Analysis*, Wadswoth, 1983.

2. Heidelberger, P. and Lewis P.A.W., *Regression-Adjusted Estimates for Regenerative Simulation, with Graphics*, Communications of the ACM, Volume 24, Number 4, April 1981.

3. Connover, W. J., *Practical Nonparametric Statistics, Second Edition*, John Wiley and Sons, 1980.

4. Johnson, Malcolm, D., Jr., *Draftsman Displays, A Graphical Technique for Exploratory Data Analysis*, Master's Thesis, Naval Postgraduate School, Monterey, California, June 1984.

5. Moran, Gary, W., *Locally Weighted Regression Scatter Plot Smoothing (LOWESS): A Graphical Exploratory Data Analysis Technique*. Master's Thesis, Naval Postgraduate School, Monterey, California, September 1984.

6. W.R. Church Computer Center, *VS APL at NPS*, Naval Postgraduate School, July 1982.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 23304-6145	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93943-5002	2
3.	Prof. Peter A. Lewis Naval Postgraduate School (Code 55Lw) Operation Research Department Monterey, California 93943-5000	5
4.	Prof. Iognaid G. O'Muircheartaigh Naval Postgraduate School (Code 55OM) Operation Research Department Monterey, California 93943-5000	1
4.	Mayor FAP Juan M. Isusi Centro de Informatica Fuerza Aerea del Peru Lima, Peru	2

END

2-87

DTIC