TEACHING DATA ANALYSIS USING INTERACTIVE APL-BASED GRAPHICS PACKAGES

I.G. O'Muircheartaigh
University College
Galway, Ireland

1. Introduction

The present paper is essentially a preliminary report on the author's experience of teaching a course in Data Analysis to students at the Naval Postgraduate School, Monterey, California. At the time of writing (which is uncomfortably close to the deadline for receipt of papers for inclusion in the Conference Proceedings) that course is in progress. The main emphasis in this paper is on the use of graphics packages as a teaching tool, and, in particular, on how these packages can assist the student (and the teacher) to achieve greater insight into both the data analytic and methodological aspects of our discipline. A brief outline of the paper is as follows: in Section 2, we discuss various graphical methods of analyzing "batches" of univariate data and in Section 3, we present some aspects of the analysis of bivariate and multivariate data. In our concluding Section 4, we discuss the importance of the graphics package as a pedagogical tool in relation to each of the preceding areas of data analysis.

2. Graphical Methods for Invariate Data

2.1 Single Batches

Almost any statistical analysis of data should, in the first instance at least, involve examining each variable in the data set separately, from an exploratory point of view. The use of an interactive graphics package greatly facilitates this exercise. For univariate data all of the following plots can be informative: (1) histograms, (2) stem and leaf plots, (3) one-dimensional scatter plots, (4) quantile plots empirical and (5) cumulative distribution functions and empirical density functions. Use of a versatile graphics package enables us to explore the strengths and weaknesses of each type of plot, and how varying the parameters of a plot can affect its efficiency (e.g. the choice of interval width for histogram or for empirical density estimation). In the case of the Data Analysis course at the Naval Postgraduate School, we are using Chambers, Cleveland, Kleiner and Tukey (1983), as one of our texts and the 33 data sets provided in the Appendix to the text have been made available (on both PC and mainframe) to the students. This enables them to explore interactively the nature of a large number of data sets. It also facilitates such exploration in class. We have also provided some additional data sets not in the text. Figures 2.1 -2.9 indicate respectively,

1. the effect of choice of interval width on the histogram of 136 Stanford ozone measurements (data set 1, Chambers et al (1983)) (Figures 2.1 - 2.3)

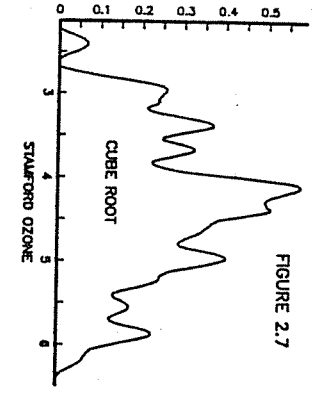


FIGURE 2.7

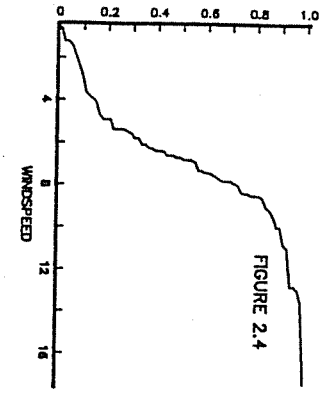


FIGURE 2.4

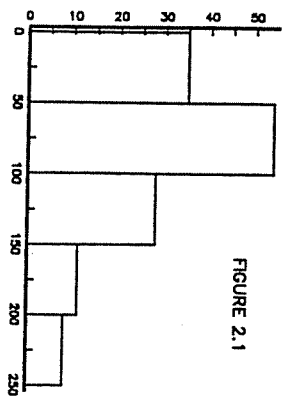


FIGURE 2.1

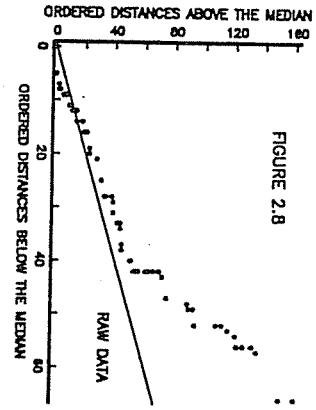


FIGURE 2.8

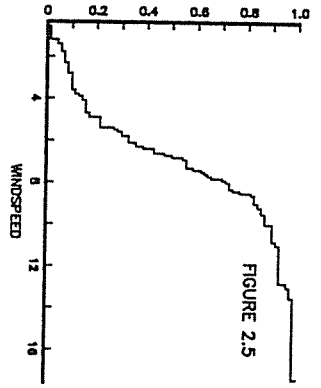


FIGURE 2.5

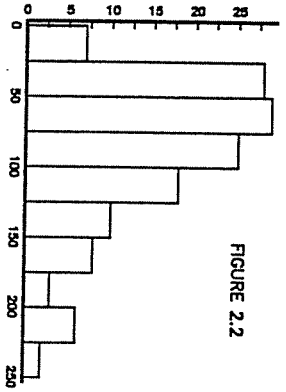


FIGURE 2.2

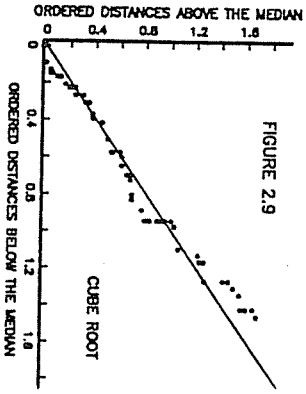


FIGURE 2.9

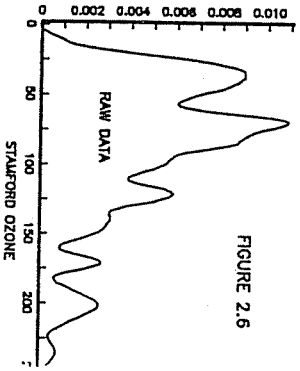


FIGURE 2.6

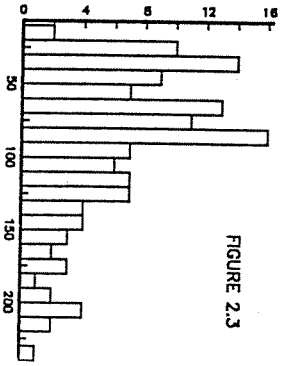


FIGURE 2.3

2. the very close relationship between a quantile plot and an empirical cdf for some wind speed data (Monahan and O'Muircheartaigh, 1986), (Figures 2.4 - 2.5)

3. the effect of different power transformations on
   a) the shape of the empirical density (Figures 2.6 - 2.7) and
   b) a custom built symmetry plot, (Figures 2.8 - 2.9) for the ozone data.

Students are also encouraged to generate their own data from different models and examine the data using the same techniques, but keeping in mind their knowledge of the (real) underlying distribution.

The major advantage of these techniques (in particular the graphical aspects) is the facility they provide to interactively learn (and teach) what information the various techniques contain; the facility to do this for both real and simulated data is a further advantage. Parallel use of different techniques can provide complementary/corroborative insights.

2.2 Comparing Several Batches of Univariate Data

Once familiarity has been established with the various exploratory techniques for dealing with univariate data, these can then be very easily used to compare two or more batches (samples) of such data. Among the techniques used in this section of course, empirical Q - Q plots, multiple (and notched) box plots and multiple density traces were all found by the students to be complementary in building up their picture of multiple batches of data. Comparing Q - Q plots with 2-D scatter plots (when both types of plots are relevant for a particular data set) can be instructive to the student (in terms of learning what the plots are telling us). See Figures 2.10 - 2.12.

2.3 Assessing Distributional Assumptions

Graphical techniques are particularly appropriate and illuminating when one is interested in evaluating how well a particular underlying model describes a specific data set. This type of plot can be very insightful, and can enable the student to better understand

a) what is meant by saying that a set of data comes from a specific distribution.

b) how to visually assess the fit of the distribution to the data

c) how transformations change the shape of a distribution, and a feel for which transformation might change the shape of a distribution towards some specified distribution (such as the normal).

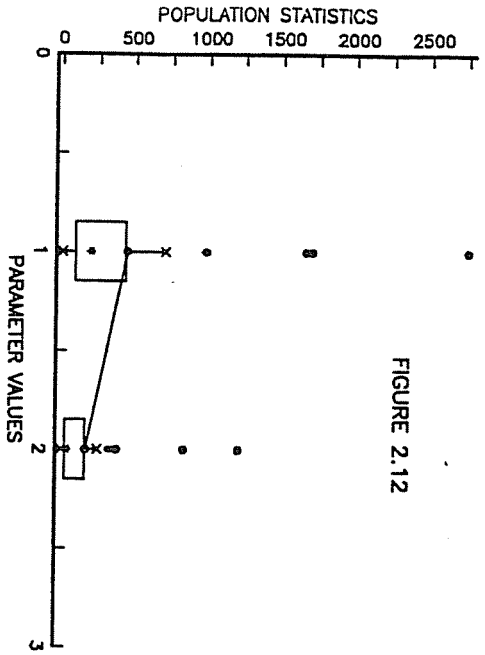


FIGURE 2.12

POPULATIONS= RS;RC

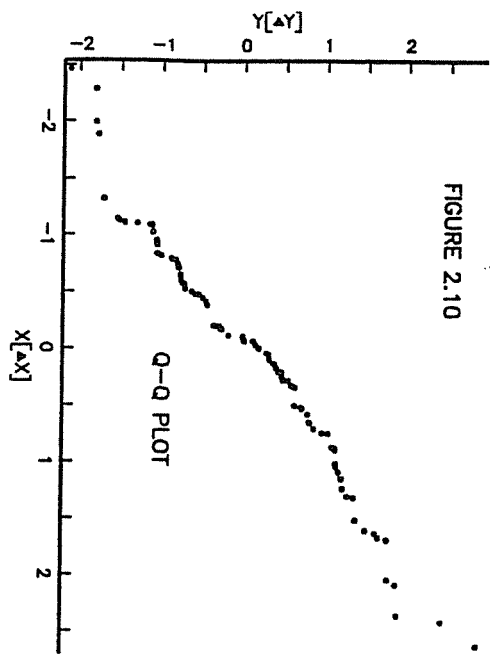


FIGURE 2.10

X,Y ~ INDEPENDENT SAMPLES FROM N(0,1)

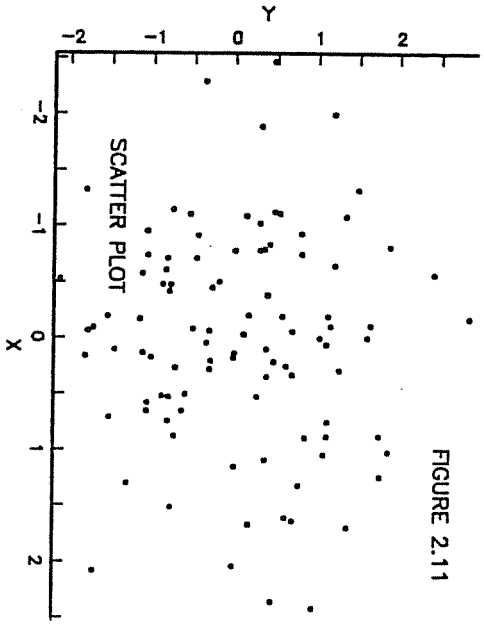


FIGURE 2.11

3. Bivariate/Multiple Data

The analysis of bivariate data is greatly facilitated by the use of graphical methods. 2-D scatter plots are, of course, the standard tool, but with regression models we can add residual plots, probability plots (for residuals) etc. which are very helpful from a teaching viewpoint. The ability to examine such plots, to interactively experiment to determine appropriate transformations and compare the results is most helpful. Figures 3.1 and 3.2 show some scatterplots for the wind/whitecap data before (3.1) and after (3.2) a log transformation is applied. The apparent outlier of Figure 3.1 has disappeared in Figure 3.2.

For multivariate data, the number of graphical techniques available is not as large, although the residual plotting methods used referred to above for simple regression also extend, of course, with equal effect to the case of multiple regression. Simultaneous plots of residuals (against fitted values, against y, and probability plots) can provide substantial insights into the error distribution and the validity of the underlying model.

4. Conclusion

The use of graphics aids in teaching a course on Data Analysis can be of considerable advantage in assisting the student to understand the nature of the data and the native of underlying models. Two graphics packages (both APL-based, and quite similar to each other) were used in teaching this course. One, STATGRAPHICS (1985) was used on IBM PC's and the other GRAFSTAT (see acknowledgements) was operated on the mainframe. The fact that both packages were APL-based greatly facilitated interactive data transformation and manipulation. Taken in conjunction with standard statistical methodology, graphics can be used to emphasize the inherent uncertainly of all statistical techniques, and the fact that, at best, all models are only an approximation to reality.

5. Acknowledgements

The author is grateful to Professor Peter Lewis of the U.S. Naval Postgraduate School, Monterey, CA for his considerable assistance in developing the course to which this paper refers, and also for introducing the author to GRAFSTAT, an experimental APL package from IBM Research, with which the figures in the paper were created, and which is being used in teaching the course discussed in the paper.

FIGURE 3.1

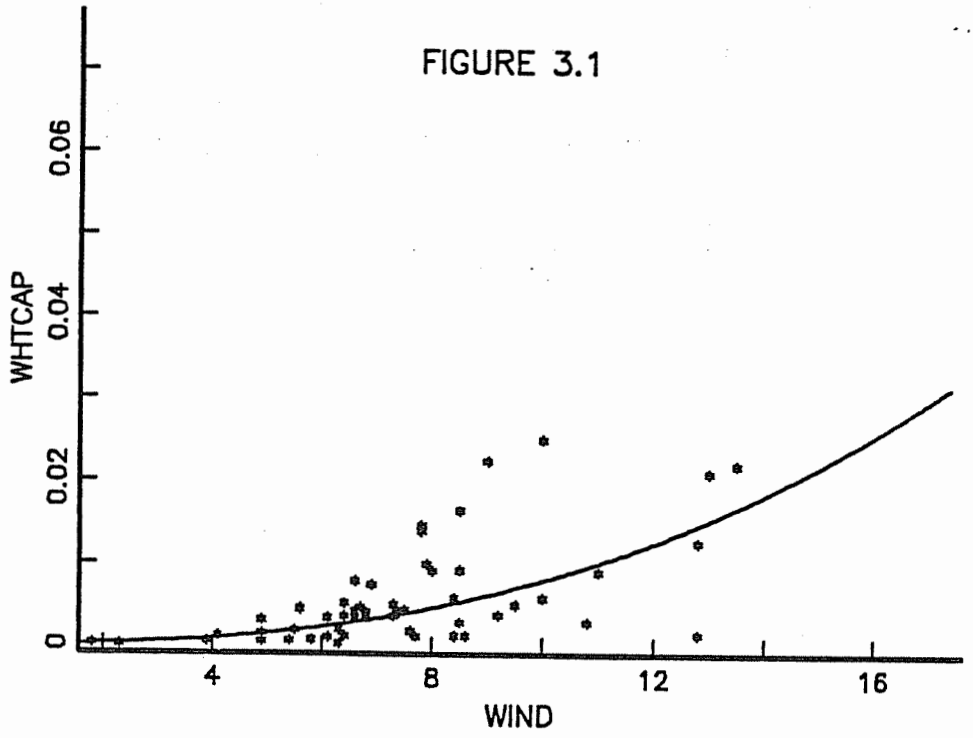
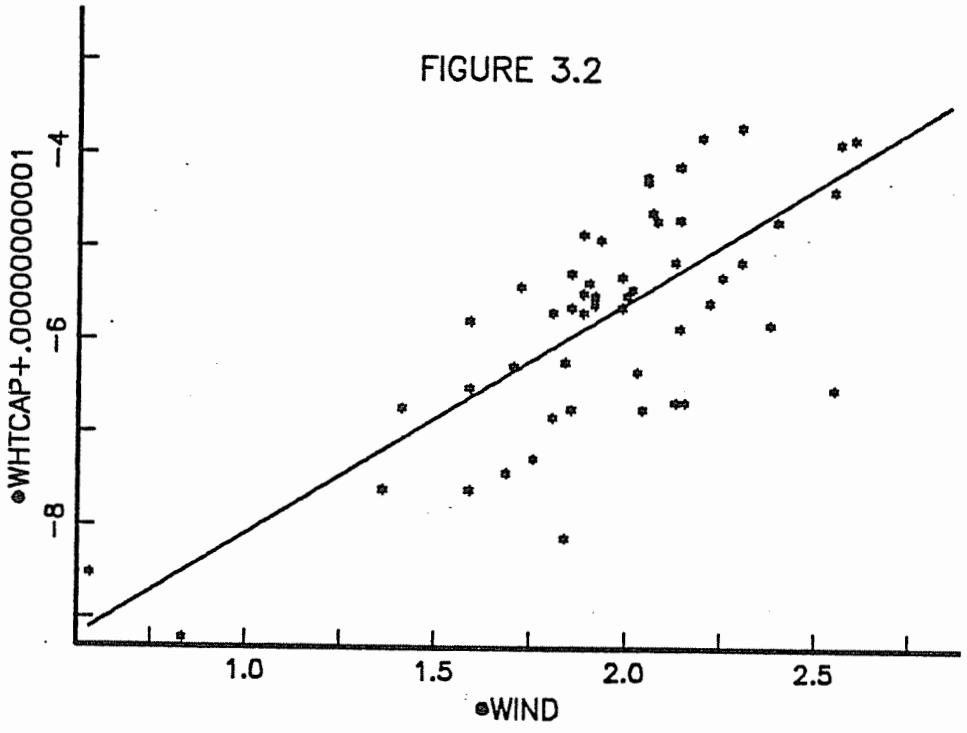


FIGURE 3.2



6. References

Chambers, J.M., Cleveland, W.S., Kleiner, B., & Tukey, P.A. (1983). Graphical methods for data analysis. Boston: Durby Press.

Monahan, E.C., & O'Muircheartaigh, I.G. (1986). Whitecaps and the passive remote sensing of the ocean surface. Int. J. Remote Sensing, 627-642.

STATGRAPHICS (1985). Statistical Graphics System. STSC, Inc., 2115 East Jefferson Street, Rocksville, Maryland 20852.