

EXPLORATORY GRAPHICS AND GEOTECHNICAL DATA : SOME INTRODUCTORY REMARKS

P.C. Kotzias¹, A.C. Stamatopoulos²
and P.J. Kountouris³

SYNOPSIS

Exploratory Data Analysis (E.D.A.) is a set of statistical techniques for compiling and conveying extensive and manifold data in a concise and lucid form. The data are shown singly or in combination with graphical and tabular displays. Although part of the rapidly advancing discipline of "Statistical/Graphics", the E.D.A. conveys information with no reference to any probabilistic model or statistical test. It makes use of "Robust and Resistant Parameters", which are not affected either by the type of the underlying distribution, its symmetry, skewness or any extreme values. The E.D.A. can be a useful tool for succinctly presenting multiple geotechnical results without sacrificing their salient characteristics, nor marginal and extreme values. Its use is explained with specific examples.

INTRODUCTION

Paraphrasing the celebrated admonition of Professor Terzaghi (1946) to his students "*You should always walk the terrain*" - geotechnical engineers should always "plot, scan and replot their data", whether such data stem from the field, the monitored structure or the laboratory. Geotechnical data are characteristically diversified. They are reported in words, numbers, acronyms, charts and whatnot. Showing such miscellanies in traditional forms, i.e., logs, subsurface sections, tables, maps or simple graphs, has been adequate in routine cases, but could be wasteful, non-discriminating and masked, when the input is intricate, wide-spread or swarming in large quantities. Modern statistical graphics can tame and succinctly phrase multitudinal and diversified results. Applied to geotechnics, as an exploratory and descriptive expedient, it will guide the plotting

¹ Dr. Ing. Geotechnical Engineer, K&S, Isayron 5, Athens GR-11471, Greece

² M.Sc. Geotechnical Engineer, K&S, Athens, Greece

³ Dipl. Ing. Geotechnical Engineer, K&S, Athens, Greece

and will display the message. There is more to that: if properly utilized, statistical graphics can enormously *enhance without trespassing* the geological framework of the material under investigation.

SCOPE : THE E.D.A.

Various phases of statistical graphics and geotechnical data have been previously reported, Stamatopoulos & Kotzias (1975), Kotzias & Stamatopoulos (1978, 1983), Kotzias et al. (1985). This note points at some graphical aspect of one of its recent branches labelled: "*Exploratory Data Analysis*", or simply "E.D.A.", which originated from Tukey (1972, 1977), Mosteller & Tukey (1977) and Hoaglin et al. (1983). The E.D.A. provides an extensive repertoire of methods for detailed exploration of quantitative data. In contrast to classical statistics, which analyses such data under formal and stringent assumptions, the exploration here is informal, flexible, and not necessarily compared to probabilistic model or statistical test. As aptly stated by Tukey (1977), it is "a graphical detective work". Complementary processes of confirming this informal approach, viz. "*Confirmatory Data Analyses*" (Hoaglin et al. 1983) will not be discussed, but certain notions, intrinsic to E.D.A., such as "Outliers", "Robustness", "Resistance", etc., (Mosteller & Tukey (1977), Velleman & Hoaglin (1981)), are only introduced in Appendix (A).

Two principal tools will be stressed. The "Stem and Leaf" and the "Box Plot" (*). There are numerous others, equally ingenious, but can not be covered in this short note.

THE "STEM AND LEAF"

This is a variation of the traditional histogram and displays the distribution of numerical values. Although a semi-graphical tool, the "Stem and Leaf" can be written out in the typewriter, is just as informative as the histogram, and reveals features usually blurred by the histogram, either in its simple or its cumulative form. It is an informal, simple and compact display, combining the

(*) Sometimes labelled as the "Box and Whisker Plot". The E.D.A. literature impresses the reader, at first sight, with its idiosyncratic terminology, such as "batch", "trimean", "hinge", "fence", "count", "depth". It is considered by its originators to be more to the point than the counterpart terminology of classical statistics (Average, Lot, Quantile, Range, etc.).

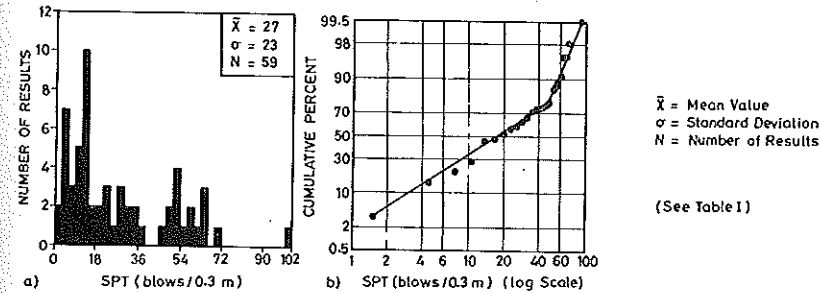


Fig. 1. Results of Standard Penetration Test : In Histograms

positive aspects of a table and a graph. The example data shown in Fig. 1 are used later to compare the "Stem and Leaf" with its counterpart, the Histogram.

THE "BOX - PLOT"

Gives a birds-eye view over a set of data and highlights salient structural features. With some familiarity, one can readily discern the shape of their underlying distribution regarding symmetry, skewness, modality, spread of tails and outlying or "wild" points. This effective presentation is straightforward, conveys extensive information at a glance, and gives powerful descriptions for single, as well as for various sets of numerical data.

APPLICATIONS

Both tools are better seen by actual examples. The raw values for the examples appear in Tables 1, 2 and 3, which record the result of standard penetration tests, quality control in an earth dam (compaction of core), and several grain size analyses on a specific mass of soil. They illustrate respectively: the compilation of a single set, the contrast of two sets, and the juxtaposition of multiple sets. The Stem & Leaf and the Box-Plot are shown here in their simplest forms. More advanced versions are not presented (Gunter (1988a, 1988b), Heyes (1988), McGill et al. (1978)).

Table 1: Results of Standard Penetration Test (Blows / 0.3 m)

2	2	3	3	3
3	4	4	4	5
7	8	10	10	10
11	11	12	12	12
12	12	12	13	13
14	14	16	17	18
20	21	22	22	25
28	28	29	30	32
33	35	36	46	49
50	51	53	53	53
54	57	59	61	63
64	64	71	100	

Table 2: Results of Controls in the Core of an Earthdam

S = PERCENT COMPACTION

REJECTS INCLUDED - RETESTS EXCLUDED (S%)

82.9	90.5	92.9	94.0	94.3
94.4	94.8	94.9	95.0	95.4
95.6	96.3	96.4	96.6	97.0
97.9	98.3	98.5	98.5	98.6
98.9	99.0	99.2	99.3	99.5
99.8	100.0	100.3	100.4	100.4
100.6	100.8	100.8	101.1	101.2
101.2	101.7	101.8	102.0	102.4
102.7	103.1	103.4	103.5	103.9
105.2	105.2	105.3	105.3	105.6

REJECTS EXCLUDED - RETESTS INCLUDED (S%)

93.6	97.4	97.5	98.3	98.5
98.5	98.6	99.0	99.0	99.1
99.2	99.3	99.4	99.8	99.9
100.0	100.3	100.6	100.8	100.8
101.1	101.2	101.2	101.7	101.8
102.0	102.1	102.4	103.1	103.4
103.4	103.5	103.6	103.9	104.5
104.8	105.0	105.2	105.2	105.3
105.3	105.6	106.0	106.1	106.1
106.5	107.8	108.3	109.0	

Table 3: Results of Gradation Analysis

PASSING SIEVE No 1/2" (%)

55	60	65	72	77
80	81	84	85	89
90	91	98	99	100
100	100	100		

PASSING SIEVE No 4 (%)

38	39	46	53	56
60	63	70	73	73
88	88	90	93	98
98	98	100		

Table 3: Results of Gradation Analysis (con.)

PASSING SIEVE No 16 (%)				
22	24	30	34	42
47	49	61	64	66
69	86	86	89	94
97	97	100		

PASSING SIEVE No 50 (%)				
11	14	18	25	33
38	38	53	54	55
56	78	81	82	86
94	94	95		

PASSING SIEVE No 200 (%)				
7	8	12	18	25
27	27	39	41	42
45	59	64	66	69
72	78	84		

STANDARD PENETRATION TEST

Results in Classical Display

All values of Table 1 are compiled in the frequency diagrams of Fig. 1, to be later compared with their E.D.A. counterparts. Here the histogram is markedly left-skewed (Fig 1a), and can be approximated by two log-normal distributions indicated in Fig. 1b. The classical measures of location and spread; viz. the mean value \bar{x} and the standard deviation σ , are also shown.

Results in Stem & Leaf

When numerical values are reported as in Tables 1, 2 or 3, the "coarse" values (STEM) and the "fine" values (LEAVES) can be split and respectively placed on the left and on the right of a vertical scale. E.g. the numbers 30, 30, 35, 36 are stated as: 3/0056 and the numbers 0.0, 0.0, 0.1, 0.6, 0.7, as: 0/00167. Fig. 2b records the results of Table 1 in "Stem and Leaf" which is compared to its equivalent histogram of Fig. 1a.

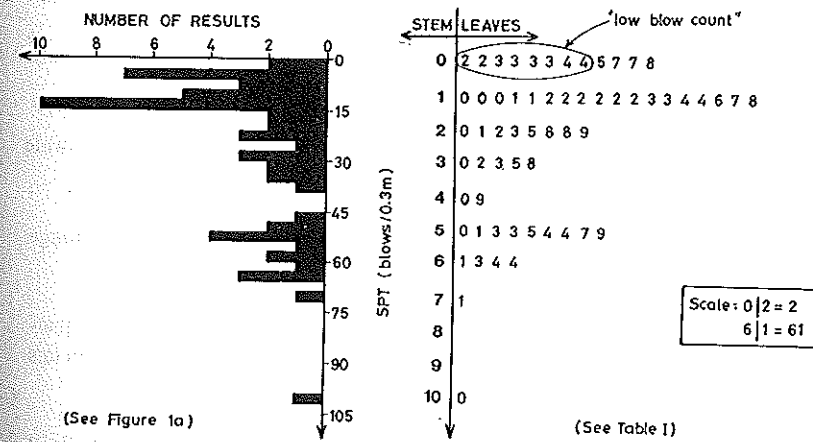


Fig. 2 Histogram Versus Stem & Leaf

Following desirable characteristics are prominent in this display:

- a) The data can be rapidly ordered and presented.
- b) The actual digits of the data are retained (Fig. 2b).
- c) Grouping problems associated with the histogram are avoided. In our case the crucial values of low blow counts viz. 2, 4 blows/0.3 m are explicitly shown in Fig. 2b, but are unavoidably blurred in the histograms regardless of the cell width (Fig. 1 and 2a).
- d) There is no distortion in the shape of the distribution. The skewness, due to a logarithmic-normal shape (manifest in the histograms of Fig. 1b) is just as well retained in the semitabular stem & leaf of Fig. 2b, if the stem value is carefully chosen for the particular set.
- e) This compact display fits conveniently in any text.

Results in Box Plot

The SPT scores of Table 1 can be compiled also in the pithy display of Fig. 3. A detailed explanation is given in Appendix A. Here the *median and the quartiles are always used, instead of the mean value and the standard deviation.*

Following features of the frequency distribution (Fig. 1 and 2) are prominent in the Box-Plot: viz: The skewness and the wide spread, of the data.

The skewness is signalled by the median in two respects:

1. It is off-center with respect to the surrounding box (viz. the Interquartile range)* as well as the "adjacent values".
2. It diverges from the "trimean"* which is given as a dot within the range of Fig. 3b.

The wide spread of the data is reflected by the width of the box (interquartile range) and the length of the upper and lower "adjacent values" which in this case - stretch from 2 to 100 blows/0.3 m.

Such explicit features provide a rapid overview of the distribution, without

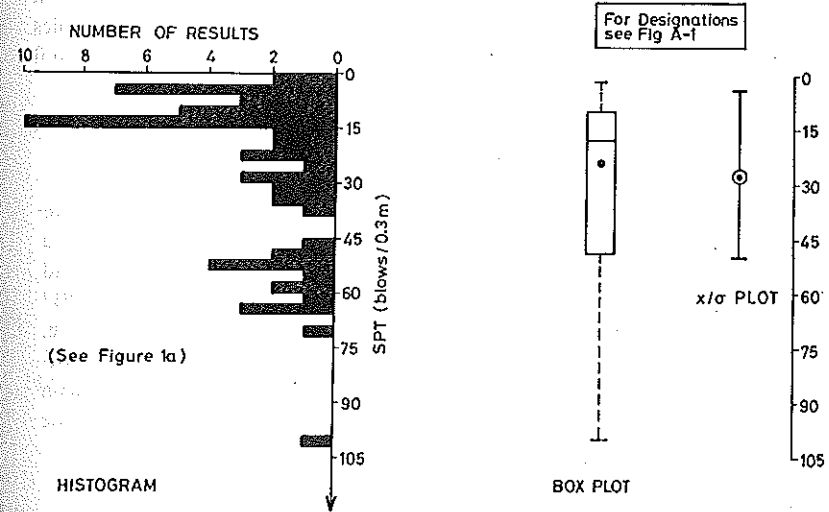


Fig. 3. Histogram - Box And \bar{x}/σ -Plot

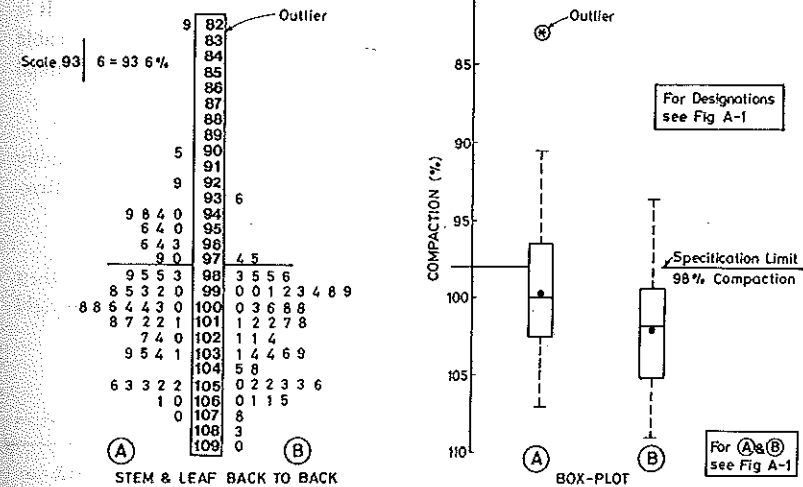


Fig. 4 Contrast of Two Distributions For Percent Compaction

* See Fig. A-1 in Appendix A

resorting to original data, whether in tables or in frequency diagrams. As shown in the following paragraphs, placing two or more Box-Plots in parallel, enables a comparison and a contrast of various distributions with respect to central values, spread as well as skewness and outliers.

TWO SETS OF DATA

Results of field compaction control, in the case of an Earth Dam, are contrasted as Stem and Leaves and as Box-Plots in Fig. 4. (The original data are given in Table 2). Group "A" reports the controls without field correction (and subsequent retesting) viz. "Rejects included-Retests excluded", and Group "B" gives the finally accepted values viz. "Rejects excluded - Retests included i.e. the "as built" results*. Both graphics indicate explicitly the shape of the distributions, particularly their skewness as well as effects of truncation imposed on group "B". Moreover the Box-plot shows distinctly extreme values as well the outliers.**

SUCCESSIVE SETS OF DATA

Numerous distributions can be compiled and readily compared by placing their Box-Plots side by side. Such a juxtaposition of EDA plots conveys at a glance variations from set to set with regard to their structural features. It is illustrated by a simple drawing in Fig. 5, where the eighteen test results of Table 3 are summarized. The same results are compiled in Fig. 6 but in the traditional \bar{x}/σ plot, explained in Appendix A. A comparison follows:

The skewness and the spread of the underlying distributions is readily discerned in Fig. 5 (e.g. No 1/4", No 4, No 16). The extreme values of the Box-Plots coincide with the outer ranges of the data.

All underlying distributions in Fig. 6 are assumed symmetrical about their mean. Here the extreme values (usually set at $> 3 \sigma$ about the mean) would yield unrealistic results - far beyond any boundary or limit.

* Elaboration on this breakdown of quality control results is given by the authors (Kotzias & Stamatopoulos, 1978). Here we are interested only in the form of the distributions.

** Outliers are discussed in Appendix A.

EXPLORATORY GRAPHICS

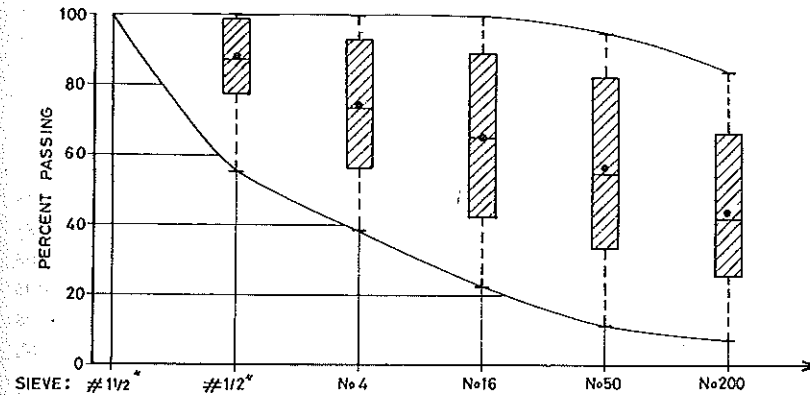


Fig. 5. Compilation of Grainsize Distribution Curves In Box - Plots

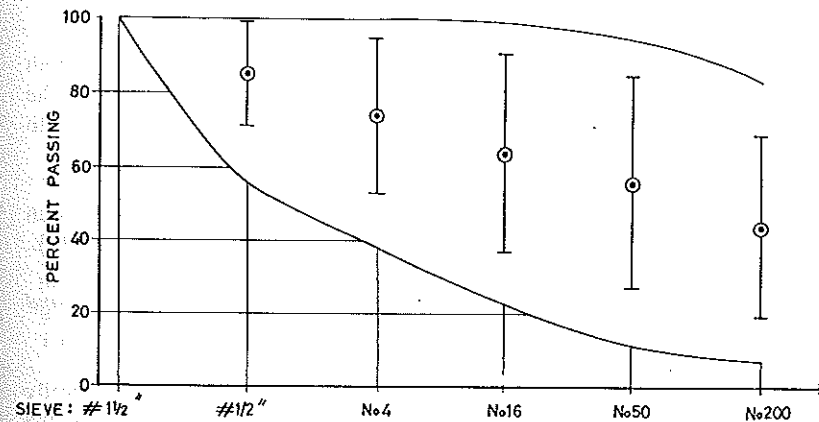


Fig. 6. Compilation of Grainsize Distribution Curves In \bar{x}/σ Plots

This comparison shows that, for the compilation of, for example, numerous gradation curves, the Box Plot is more versatile and informative than the traditional \bar{x}/σ plot.

CONCLUDING REMARKS

A brief introduction was made on the use of Exploratory Data Analysis (E.D.A.) by describing two of its current tools: The Stem and Leaf and the Box-Plot. Both can be instrumental in compiling and succinctly reporting numerous geotechnical data, without tacitly imposing a probabilistic model or a rigorous statistical test. There are additional simple and powerful E.D.A. techniques, as well as informal statistics displays, which can reveal and impart the message, otherwise latent in a maze of results.

The present paper - *dedicated to Professor Victor De Mello* - will hopefully instigate a closer look into this unexplored interplay between geotechnical data and modern Statistical Graphics.

APPENDIX A: REMARKS ON THE STRUCTURE OF THE BOX AND THE \bar{x}/σ PLOT

The Box and the \bar{x}/σ Plot

Fig. A-1 explains the various symbols in the Box as well as the \bar{x}/σ plot. Both summarize an underlying frequency distribution. The location and spread in the Box-Plot is measured by the Quartiles. A Quartile is explained in Fig. A-2 with reference to its cumulative frequency.

On the other hand the \bar{x}/σ (Fig. A-1) is always symmetrical about the mean value and measures the spread using with the standard deviations as a yardstick.

Why Use Medians and Quartiles in the Box Plot Instead of Mean Values and Standard Deviations

The aforementioned \bar{x}/σ plot is a popular way of summarizing graphically a frequency distribution. It is currently appearing in quality control (e.g. the Shewhard Chart) and in many reports, where batches of data are compiled in sequence. Such a presentation carries a tacit assumption that the summarized distribution is symmetrical about its mean, and can be satisfactorily approxi-

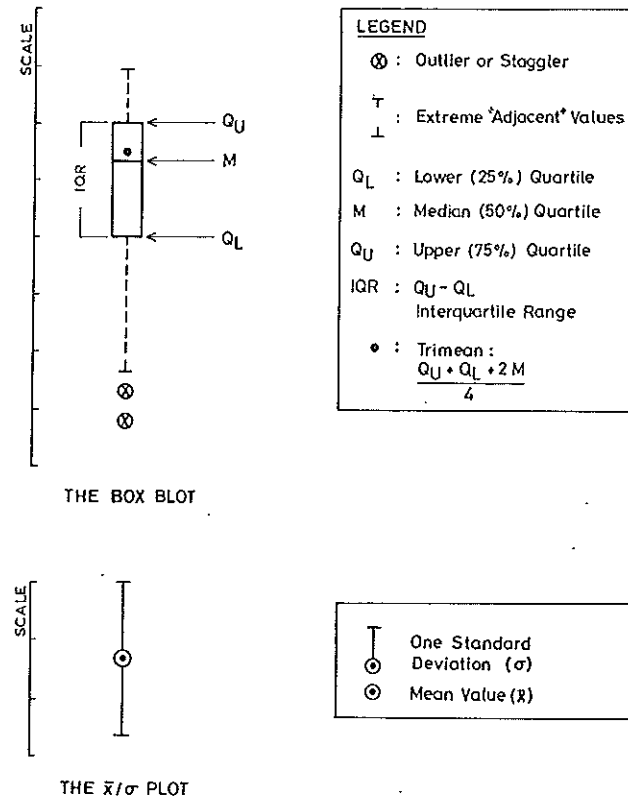


Fig. A-1 Figures And Designations In Summary Plots

mated by a normal or gaussian model. The Box-Plot, being unattached to a gaussian or any other model, is flexible and adjustable to the underlying data. This is attained by employing the quartiles and designating the outliers.

Assymetry and Spread

To Stress this point further, the two plots are compared in Fig. A-3. Both summarize the non-symmetrical underlying distribution. *Its assymetry* is reflected in that the median in the box-plot is off center with regard to the quartiles,

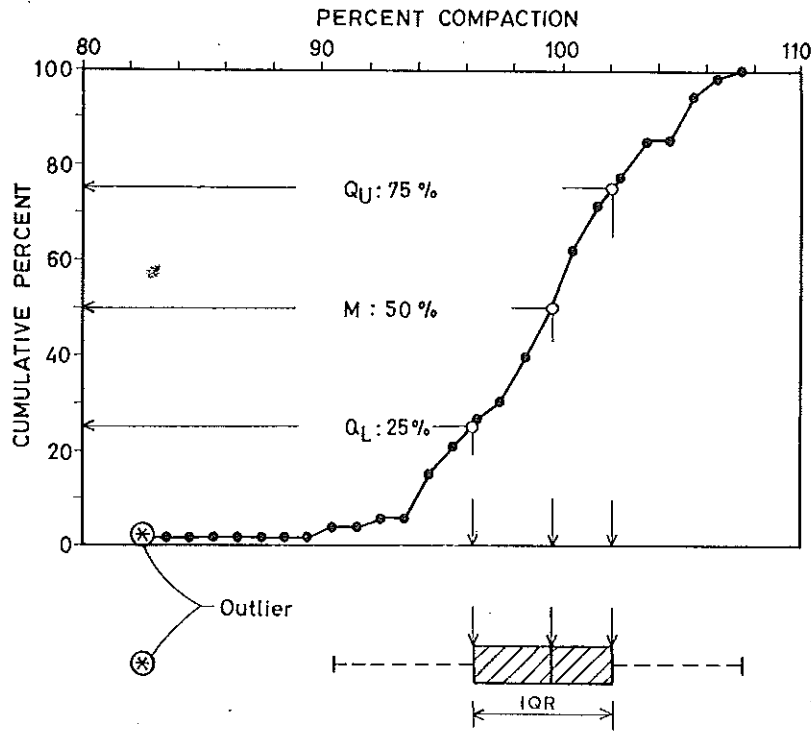


Fig. A-2 The Quartiles And The Outlier

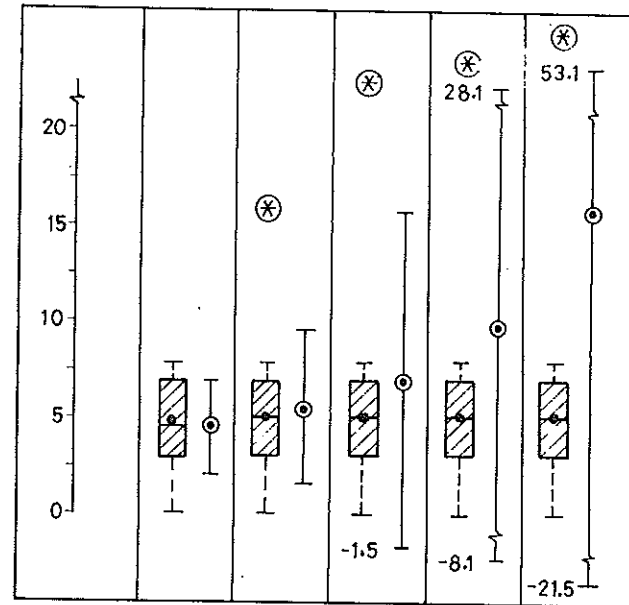
diverge from "trimean" indicated as a dot within the IQR box. Its spread is designated by the stretch of the Box by of the length between the "adjacent" values. On the other hand, the \bar{x}/σ presentation remains rigidly symmetrical with regard to the pronounced left-skewness, and with regard to the spread of the data.

Outliers

An important issue in the E.D.A. is the isolation of outliers (or stragglers) from the main body of data. Fig. 4 shows an outlier in the Box Plot. A useful criterion is given by Velleman & Hoaglin (1981). We first compute the

Batch : 0 3 3 3 4 5 6 7 7 8

Outlier (*)	—	(16)	(32)	(64)	(128)
\bar{x} :	4.6	5.6	7.1	10.0	15.8
σ :	2.5	4.1	8.6	18.1	37.3
2σ :	5.0	8.2	17.2	36.2	74.6
Median :	4.5	5.0	5.0	5.0	5.0
Q_U :	7.0	7.0	7.0	7.0	7.0
Q_L :	3.0	3.0	3.0	3.0	3.0
IQR :	4.0	4.0	4.0	4.0	4.0



For Designations see Fig A-1

Fig. A-3 Effect of An Outlier On Quartiles And \bar{x}/σ

interquartile range (the IQR). The "upper adjacent value" (the extreme upper) is defined to be the largest observation that is less or equal to the Upper Quartile plus (+) 1.5 x IQR. Similarly the "lower adjacent value" (the extreme lower) is defined to be the smallest observation that is greater than or equal to the lower quartile minus (-) 1.5 x IQR. Any values beyond these boundaries are outliers. This criterion, (which holds only for designating outliers and not for extreme or "adjacent" values,) is employed in Figs. 3, 4 and 5.

Effect of Outliers - Resistance

Even a single outlier can distort the mean value and the standard deviation, especially when the number of results is relatively small. In contrast, the median and the quartiles are resistant to outliers. This is illustrated in Fig. A-3 which shows a batch of ten random numbers. The effect of an outlier on the shape of these summary graphs is compared in Fig. A-3. Following values are used for the outliers: (16), (32), (64) and (125). The classical measures of location and spread viz \bar{x}/σ are strongly effected; whereas the equivalent E.D.A. measures are "robust" to outliers (Tukey (1977), Mosteller & Tukey (1977), Hoaglin et al. (1983)).

Flexibility of the Box Plot

A realistic comparison of numerous sets of data (e.g. Figs. 5 and A-1) is readily effectuated by the box plot because:

- a) The outliers are specified, isolated and displayed
- b) The spread of the distribution is explicitly shown by the upper and lower adjacent values
- c) The measures of location and spread viz. the Median and the Quartiles are resistant to outliers.

REFERENCES

- GUNTER, B. (1988a): "Subversive Data Analysis, Part I: The Stem and Leaf Display, *Quality Progress*, September 1988, pp. 88-89.
- GUNTER, B. (1988b): "Subversive Data Analysis Part II: More Graphics, Including my Favorite Example". *Quality Progress*, November 1988, p. 77-78.

- HEYES, G. (1988): "The Box Plot" *Quality Progress*, December 1985, pp. 12-17.
- HOAGLIN, D., MOSTELLER, F., & TUKEY, J.W. (1983): "*Understanding Robust and Exploratory Data Analysis*, J. Wiley, pp. 1-31, 58-75, 301-302.
- KOTZIAS, P., & STAMATOPOULOS, A. (1978): "Statistical Quality Control at Kastraki Earth Dam" *Journal of the Geotechnical Engineering Division, A.S.C.E.*, September 1978.
- KOTZIAS, P., & STAMATOPOULOS, A. (1983): "Graphic Statistics for Multiple Geotechnical Data". *Proc. 4th Intern. Conference on Applications of Statistics and Probability in Soil and Structural Engineering*, Florence, Vol. II, pp. 1003-1015.
- KOTZIAS, P., STAMATOPOULOS, A., & KOUNTOURIS, P. (1985): "Statistical Graphics for Estimating computation Parameters". *Proceedings, 2nd Hellenic Colloquium on Geotechnics*, Athens, pp. 331-343.
- McGILL, R., TUKEY, J., & LARSEN, W. (1978): "Variations of Box Plots". *The American Statistician*, February 1978, Vol. 32, No. 1.
- MOSTELLER, F., & TUKEY, J. (1977): "*Data Analysis and Regression*" Addison Wesley, Reading Mass., pp. 12-17 and Chapter 13.
- STAMATOPOULOS, A., & KOTZIAS, P. (1975): "The Relative Value of Increasing the Number of Observations" *Proceedings IInd International Conference on Application of Statistics and Probability in Soil and Structural Engineering, Aachen*, Vol. II pp. 495-510.
- TUKEY, J. (1972): "Some Graphics and Semigraphic Displays" in T.A. Bancroft Ed. *Statistical Papers in Honor of George W. Snedecor, Ames IA*, Iowa State University Press, pp. 293-316.
- TUKEY, J. (1977): "*Exploratory Data Analysis*", Addison Wesley Publishing Company, pp. 7-55.
- TERZAGHI, K. (1946): "*Notes on Engineering Geology*" : Harvard Graduate School of Engineering, Spring Term 1946.
- VELLEMAN, P., & HOAGLIN, D. (1981): "*Applications, Basics and Computing of Exploratory Data Analysis*" Duxbury Press, North Situate, Mass., (Chapters 2, 3, 6).
- VELLEMAN, P., & HOAGLIN, D. (1981): "*ABC's of EDA*". Duxbury Press.