European Journal of Operational Research 23 (1986) 5-13
North-Holland

Invited Review

Exploratory data analysis

Chris CHATFIELD
School of Mathematics, Bath University, Bath, Avon, BA2 7AY, United Kingdom

Abstract: It is usually wise to begin any statistical analysis with an informal, exploratory examination of the data, and this is often called exploratory data analysis (abbreviated EDA). The ingredients of EDA are discussed, and two main objectives are delineated, namely data description and model-formulation. It is suggested that it is important to see EDA as an integral part of statistical inference, and several examples are presented to show how EDA is used in model-building in regard to such topics as queueing, reliability, stock control, regression and forecasting. Some drawbacks to Tukey's (1977) approach to EDA are discussed, and an alternative title is suggested, namely the initial examination of data (IDA).

Keywords: Descriptive statistics, stem-and-leaf plot, box plot, tables, model formulation, regression, forecasting, queueing, stock control

1. Introduction

Many experienced analysts like to get a 'feel' for a given set of data by starting with an informal, exploratory examination. The idea is to clarify the general structure of the data, obtain simple descriptive summaries and perhaps get ideas for a more sophisticated analysis. This type of approach is often called Exploratory Data Analysis (abbreviated EDA), after the title of the book by Tukey (1977), but is in the same sort of spirit as Preliminary Data Analysis (Cox and Snell, 1981; Chatfield, 1982), the initial examination of data (Chatfield, 1985) and the Cross-Examination of Data (Rao, 1983).

This paper reviews the ingredients and objectives of EDA, and strongly encourages its use for data-description and model-formulation. However certain aspects of Tukey's (1977) approach are criticized. In particular it is argued that it is essential to integrate EDA into statistical analysis and model-building and not to regard it as a separate subject.

The virtues of EDA in Operational Research are illustrated with examples from several areas, including queueing theory, reliability and time-series forecasting.

Finally the reader should note that the exact scope and role of EDA is the subject of some controversy amongst statisticians. The personal views expressed in this paper are, I believe, shared by many other statisticians, but a broader view of the current 'state of the art' can be obtained by reading the long discussion following Chatfield (1985).

2. The objectives of EDA

It is impossible to define EDA exactly, but in broad outline it includes checks on data quality, the calculation of summary statistics, the plotting

I would like to thank Mr. M.E. Brigden for constructive comments on an earlier draft.

Received June 1985

of appropriate graphs, and perhaps the use of more complicated data-analytic techniques such as principal component analysis. These procedures are outlined in more detail in Section 3.

The two main objectives of EDA are (a) data description and (b) model formulation. The first is 'well known' but the second is not always explicitly recognized despite its importance in both Statistics and OR.

As regards data description, it is obviously essential to begin by summarizing the data and picking out the more important features. What is not always realised, is that there are situations where this is all that is desirable, as for example if the data quality is too poor to justify inferential methods. Inference is also unnecessary with entire population, as opposed to sample, data. Furthermore in Section 4 we note that EDA may obviate the need for significance tests, and Examples 1 to 4 in Chatfield (1985) illustrate the important but neglected possibility that EDA may be all that is required.

However in OR it is probably true to say that the second objective of EDA, namely model formulation, is the more important. There are many situations where EDA is vital in generating hypotheses, building a suitable model and suggesting an appropriate statistical procedure to analyse a given set of data.

It may be helpful to regard model building as having three stages, namely model formulation, estimation, and model validation. Traditionally statisticians have devoted much (excessive?) energy to the second stage, namely parameter estimation for an *assumed* model. However the model builder's main problem is often not how to fit a model but rather how to formulate it in the first place. In a similar vein, the OR literature rather gives the impression that more effort is devoted to manipulating an assumed model rather than to formulating it (though hopefully this is not so in practice?). In recent years, statisticians have shown increased interest in model validation with emphasis on diagnostic checks, residual analysis and the like, but model formulation still seems rather neglected.

The general principles of model formulation are probably more familar to OR analysts than they are to statisticians and will not be described here (e.g. see Cox and Snell, 1981, Chapter 4; Gilchrist, 1984). In brief they include the need for collaboration, for the incorporation of background theory, and for looking at the data and incorporating their main features. In practice most statistical models are based on EDA to some extent, and, after reviewing the techniques of EDA in Section 3, some brief examples of model formulation will be discussed in Section 4.

3. The ingredients of EDA

This short section can only provide a brief introduction to EDA, and so should be read in conjunction with Chatfield (1985, Section 3) and the references therein, as well as with the books on EDA referred to in Section 5.

After clarifying the objectives of the investigation and getting sufficient background information, the analyst should start by assessing the *structure* of the data, as the analysis will depend crucially, not only on the number of observations, but also on the number of variables and whether they are continuous, discrete, qualitative, binary or whatever. If the analyst was not responsible for collecting the data, then it is important to find out how this was done. The data quality should be checked particularly in regard to errors, outliers and missing observations. Outliers (e.g. Barnett and Lewis, 1985) are observations which do not seem to be consistent with the rest of the data and can create severe problems. Various tests are available for 'rejecting' outliers, but the inexperienced analyst is advised that such tests are less important than advice from people 'in the field' and that it is sometimes worth repeating an analysis with and without a suspect observation.

After screening the data, the analysis will usually continue with what is often called 'Descriptive Statistics'. Summary statistics should be calculated for the data as a whole and for important subgroups. These usually include the mean and standard deviation for each variable although the range is sometimes preferred to the standard deviation as a descriptive measure for comparing the variation in samples of roughly equal size.

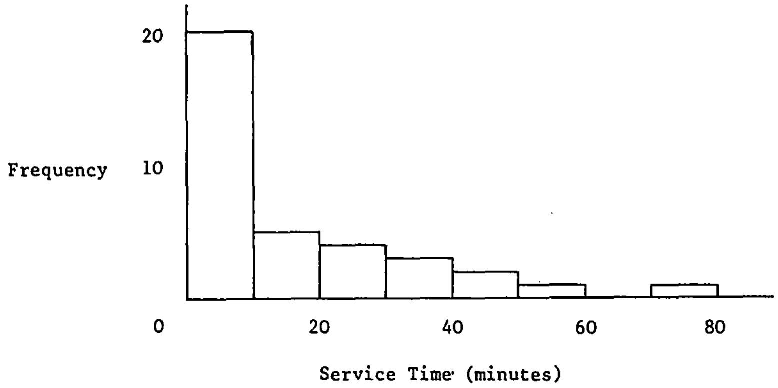
It is also a good idea to plot the data in whatever way seems appropriate (e.g. Chambers et al., 1983; Tufte, 1983). The histogram is the usual way of presenting a frequency distribution of observations on a single variable, though some statisticians now prefer to use a graph called a *stem-*

and-leaf plot. The latter is like a histogram on its side except that the 'leaves' (or columns of the histogram) show the next significant digit and hence give extra information. Figure 1 shows the histogram and stem-and-leaf plot for a random sample of 36 observations from an exponential distribution mean 15 used in simulating a queueing system, and the reader can judge for himself which he prefers. Further information may be obtained in many recent statistics books (e.g. Chatfield, 1983, Appendix D4; Chambers et al., 1983), and many computer packages (e.g. MINITAB) will now produce this and other types of graph very easily.

Another modern type of graph is the box-plot or box-and-whisker plot) which can also be used for displaying a distribution, but is particularly useful for comparing several groups of observations. Figure 2(a) shows the same data as in Figure 1. The rectangle covers the interquartile range and the median is at the centre vertical line. The 'whiskers' project to the two extreme points indicating the overall range. Figure 2(b) also shows the box plot of a sample of 36 observations from a normal distribution (mean 15, standard deviation 5) for comparison.

The reader should note how easy it is to see differences between the two samples particularly in regard to the *shape* of the distribution. While the normal is symmetric, the exponential is skewed to the right. Indeed the highest observation in Figure 2(a) is something of an outlier and could be connected to the second largest observation by a dotted line. (There are several refinements to stem-and-leaf and box plots not discussed here). When two groups of observations have similar symmetric shape, it is also easy to compare the mean values and it is for example useful to draw appropriate box plots before carrying out a one-way ANOVA to compare several groups of measurements.

Two other important types of graph are the *scatter diagram* for plotting one variable against another, and the *time plot* for plotting a time series against time. The former is important in *regres-*



```
        0 | 1 1 2 2 3 3 3 3 3 4 5 5 5 6 6 7 8 8 9 9
        1 | 0 2 3 5 8
        2 | 2 3 4 6
        3 | 1 3 5
        4 | 2 5
        5 | 0
        6 |
        7 | 9
  The STEM            The LEAVES
Units = 10 minutes   Units = 1 minute
```

Figure 1. A histogram and stem-and-leaf plot of 36 observations from an exponential distribution, mean 15 minutes

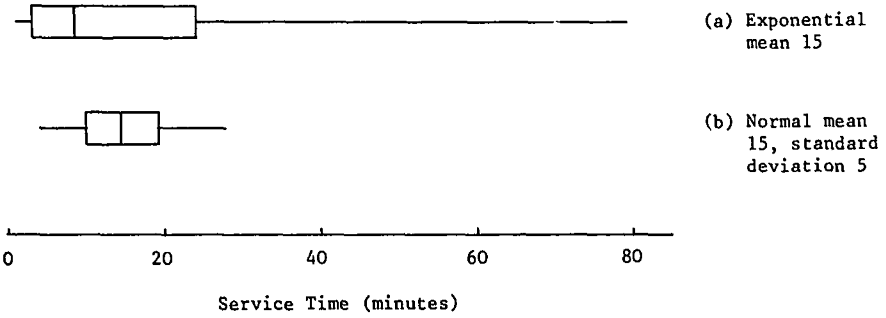


Figure 2. Box plots of an exponential distribution (above, mean 15) and a normal distribution (below, mean 15, standard deviation 5)

sion, and the latter (which is really a special type of scatter diagram) is important in time-series-analysis and *forecasting*—see Section 4.

It is obviously important to present graphs and tables in a clear self-explanatory way (e.g. see Ehrenberg, 1982, Chapters 15–17), and it is a sad commentary on much published work that one needs to mention 'simple' guidelines such as (a) summary statistics should be suitably rounded (probably using Ehrenberg's (1982, Chapter 15) two-effective-digits rule), (b) graphs should have a clear title with carefully labelled axes, (c) tables should have a clear self-explanatory title and units of measurement should be stated.

The treatment of multivariate data is more complex. Summary statistics should include correlations between each pair of variables, and it is usually a good idea to plot at least some scatter diagrams of pairs of variables. However while scatter diagrams are often helpful, there are occasions when it can be misleading to collapse multivariate data onto two dimensions (see the remarks on multiple regression in Section 4). Thus traditional descriptive statistics may only tell part of the story in three or more dimensions.

Chatfield (1985, Section 3.3) suggests that various multivariate data-analytic techniques can also arguably be seen as part of EDA. By 'data-analytic', I follow modern usage in applying the adjective to techniques which do not depend on a formal probability model. These techniques include principal component analysis, many forms of cluster analysis and multi-dimensional scaling. The main objective is often to plot higher-dimensional data in two dimensions, and in particular a graph of the first two principal component scores may be a good way of revealing clusters and outliers. Various examples of the use of data-analytic-techniques are given by Everitt (1978). These techniques are much more sophisticated than earlier descriptive techniques and should not be undertaken lightly.

4. Examples of model formulation

This section describes some illustrations of the use of EDA in model formulation.

One of the simplest types of problem is concerned with assessing a suitable distributional form for a single variable of interest. This is usually done by looking at a histogram of relevant data, although a stem-and-leaf plot or box plot may also be helpful. For example the study of queueing systems requires assumptions to be made about the distribution of service times and of inter-arrival times, and appropriate histograms should indicate what assumptions are reasonable. In reliability studies, the distribution of failure times of a given item may need to be assessed. Figure 3 shows the histogram of the failure times (in hours) of 23 refrigerator motors and the analyst may ask if it is normal, exponential, Weibull or what? Now the distribution is not symmetric (and hence not normal), but neither is it anywhere near as skewed as the exponential distribution in Figure 1 (which is more commonly described as reverse J-shaped). Instead the Weibull of gamma distribution should be considered for data like that in Figure 3. Both give a reasonable fit here, as may be confirmed in other ways, for example by probability plotting methods using the appropriate graph paper.

The reader should realise that, although tne latter type of approach is graphical, it may be rather different in spirit to the informal examination of a histogram. If the analyst already has a

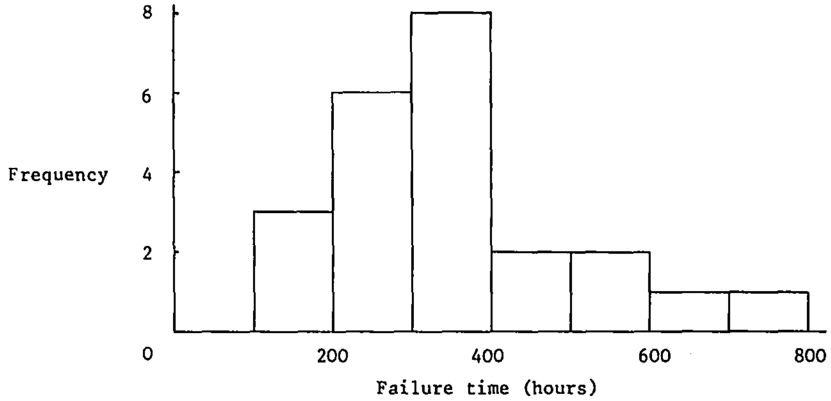


Figure 3. Histogram of failure times of 23 refrigerator motors

specific type of distribution in mind (e.g. the Weibull) and wants to see if it gives a good fit, then a model is being checked rather than formulated and this is not really EDA. However probability plotting may highlight possible outliers as well as departures from assumptions and sometimes various probability plots may be tried in an exploratory way. This exemplifies the blurred nature of the limits of EDA.

In regard to the exponential distribution, it is perhaps worth noting that, although it describes failure times for complex equipment and although it gives rise to many simple, elegant, theoretical results, it only arises empirically on rather infrequent occasions and should not be assumed without evidence. The Weibull and gamma distributions seem more common.

A final example of a situation requiring the assessment of distributional form arises in stock control. Here there are two variables of particular interest, namely the demand per unit time and the time required to make a delivery, and the distribution of each variable should be investigated.

Another important class of problems concerns regression, where the analyst should start by plotting the scatter diagram of the data. This important aspect of EDA should give guidance on a suitable shape for the curve (e.g. linear, quadratic, or whatever) as well as giving guidance on what Cox and Snell (1981, p. 18) call the secondary aspects of the model (are the 'errors' normal, independent, with constant variance, etc.?). Thus EDA is vital in selecting a sensible model and is arguably the most important and difficult stage of the analysis now that it is so easy to actually fit a regression model using a computer package. Despite the emphasis in statistics books on fitting polynomial functions, the analyst should bear in mind the possibility of fitting other types of curve such as the Gompertz curve.

Multiple regression seeks to establish a relationship between a response variable, y, and several regressor variables, x_1, x_2,.... This technique is widely overused and misused, and the tendency to 'throw in' as many x-variables as one can think of, should be resisted. An initial look at the data is still strongly advisable. Plotting the scatter diagrams of y against each x is usually worth doing to give some indication of the appropriate model, and in particular to spot non-linearities and possible outliers. However when there are interactions, or correlations between the x-variables, these graphs will only tell part of the story and may on occasion be misleading. Thus scatter diagrams of pairs of x-variables should also be examined if there are not too many x-variables and it is certainly worth calculating the correlations between pairs of x-variables. It is generally inadvisable to fit an equation where some of these correlations are 'large'.

In time-series analysis and forecasting it is also essential to start by plotting the observations against time. This will show up features such as trend, seasonality, discontinuities and outliers and give guidance on a suitable time-series model (see Raftery, 1985, Section 4.1). For example the time plot will show if seasonal variation is present and, if so, whether it is additive (of constant size from year to year) or multiplicative. There is a well-known adage that if you think you can get good forecasts 'by eye', then any sensible forecasting procedure will also give good forecasts. But if you can't, then it won't! Figure 4 shows a time series discussed by Chatfield (1978) in which there is a sudden jump in sales near the end of the series as a result of a sales drive. Here the time plot tells us that we cannot expect any univariate model to adequately describe the whole series, and, although rather negative, this information is invaluable. In a more positive role, the time plot should help in formulating a trend-and-seasonal model for other time series and Chatfield (1978, Figure 2) shows a series which is seasonal for the first 4 or 5 years but non-seasonal thereafter. Better forecasts are obtained by fitting a non-seasonal model solely to the latter part of the data, and the possibility of ignoring the earlier observations (which horrifies some people!) arises directly from the time plot.

As our last example of model-formulation, it is worth stressing that EDA can be a useful preliminary to many types of significance test, both in generating hypotheses and in assessing suitable secondary assumptions. On the latter point, the EDA should indicate whether assumptions such as normality and constant variance are at least not unreasonable (and the double negative here is intentional). As to the generation of hypotheses, an important unresolved question is whether and when it is reasonable to generate and test hypotheses on the same set of data. Finally we note that EDA may indicate that a test is unnecessary or undesirable because (a) the results are 'clearly significant', (b) the differences are not large enough to be interesting, or (c) the data are unsuitable for formal testing because of data inadequacies.

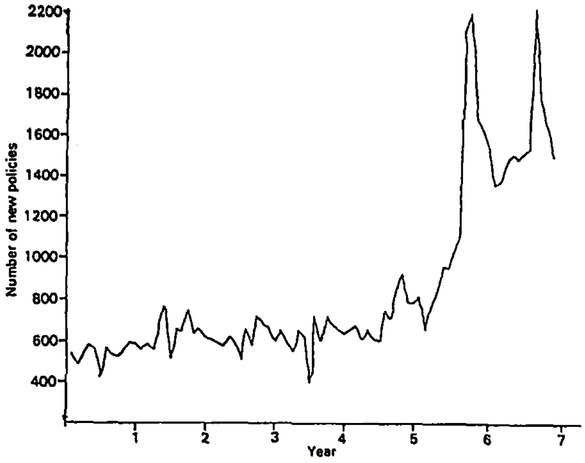


Figure 4. Numbers of new insurance policies issues by a particular life office

5. Comments on Tukey's approach

The phrase 'Exploratory Data Analysis' was promoted by Tukey (1977), who describes a variety of arithmetical and graphical techniques for carrying out what he calls numerical detective work. They include methods for summarizing a single group of observations, for assessing the relationship between variables, for analysing two-way and three-way tables, and for summarizing count data in distributional form.

There is no doubt that Tukey's (1977) book has provided a major stimulus to the sort of approach outlined earlier. There is also no doubt that Tukey has provided several useful additions to the analyst's toolkit, notably the stem-and-leaf plot and the box plot. It may therefore seem churlish to criticize the book, but, as anyone who has tried to read it will know, the book is very hard to 'read'. This is partly because Tukey invents new jargon, not only for new methods, but also to replace well-established statistical terms, so that, for example, a quartile becomes (almost) a hinge. Indeed the reader may get desperate for the sight of a familiar word or phrase. More serious is the choice of techniques and the overall general approach, and it is in these two important matters that my own approach differs markedly from that of Tukey (1977), as discussed here and in Chatfield (1985, Section 5).

On detailed techniques, Tukey chooses to omit some standard tools, such as the (arithmetic) mean and the standard deviation. These surprising omissions are not justified. Instead Tukey prefers the median as a measure of location throughout, and he devotes considerable space to describing a number of techniques based on medians, such as a method for analysing two-way tables called *median polishing*. These techniques are examples of a class of procedures described as *resistant* (e.g. Erickson and Nosanchuck, 1977) which means that the results will not be unduly affected by outliers. While this sort of method has its place, there is no attempt to compare them with standard methods and they arguably receive more attention than they deserve.

On more general matters, the most serious criticism of Tukey (1977) is that he makes little attempt to integrate EDA into mainstream statistical inference. For example Tukey (1977) says little about data collection, data quality and model formulation and there is little or no attempt to demonstrate how his exploratory techniques should complement and augment the more standard statistical procedures. The rest of Statistics might just as well not exist for all the attention it receives, and the exploratory techniques appear to have become an end in themselves. As one illustration, I note that Tukey (1977, Chapter 5) discusses relationships between two variables but incredibly makes no mention of regression and correlation. As a second illustration, Tukey (1977, Chapter 7) presents several resistant procedures for smoothing time series based on running medians. The phrase 'time series' is mentioned once in passing but there is no attempt to compare the approach with the standard time series smoothing procedure called linear filtering. No doubt Tukey himself is aware of the more general perspective which should be taken of his work (e.g. Tukey, 1980) but it is a pity that it does not appear in his book.

There are several alternative books which are concerned with Tukey's EDA to a greater or lesser extent, and they include McNeil (1977), Tukey and Mosteller (1977), Velleman and Hoaglin (1981), and Hoaglin et al. (1983). These books, while easier to follow than Tukey (1977), still tend to follow the latter's approach so that the authors come over as Tukey's disciples rather than as critical admirers. What seems to be needed are text books which integrate EDA into Statistics, by including the better aspects of EDA as well as more traditional Statistics and showing how the different approaches complement each other. One reasonably successful book in this regard is Erickson and Nosanchuck (1977) which discusses the interdependence of the exploratory and confirmatory approaches as applied to most of the standard topics including regression and the analysis of variance. Another attempt to integrate the beneficial aspects of EDA into mainstream Statistics is made in a paper by Chatfield (1985). The latter avoids the term EDA in order to distance himself from some aspects of Tukey's approach. Instead Chatfield uses the term 'The Initial Examination of Data' (abbreviated to Initial Data Analysis or IDA) which is an exact, neutral description of the important area of Statistics which is the subject of this review. It is possible that the term IDA will replace EDA in the long term, but it is also possible that the title EDA will continue to be used even though it may become increasingly less identified with all aspects of Tukey's approach.

The failure to put EDA in a suitable perspective is also noted by Stuart (1984) in a review of Hoaglin et al. (1983). Stuart (1984) also points out that, while Tukey has given the term EDA a new and special meaning, others would say that something like it has been practised by proficient applied statisticians for years, though this may not be clear in published work which generally cloaks research in mathematical statistics to make it academically respectable.

Tukey (1977) has also been criticized by Ehrenberg (1979) in a completely different way. Ehrenberg objects to the use of the word 'exploratory' which may be appropriate when analysing a completely new set of data, but is less appropriate when analysing a sequence of similar sets of data. Ehrenberg (1982) favours a *comparative* approach to model building which emphasises the desirability of establishing regular patterns across several or many sets of data, so that Statistics is seen as a search for what John Nelder calls 'significant sameness'. As an example I refer to the general model of consumer purchasing behaviour which has been developed by the author with Professors A.S.C. Ehrenberg and G.J. Goodhardt (see Goodhardt et al., 1984). In contrast to many research studies which are 'one-off' or fairly short-lived, this model has been developed over some 25 years giving results which apply, not just to a single set of data, but generalize to different brands and product fields, in different countries, over different time periods in different years. Faced with a new set of purchasing data, the analyst need only carry out a brief data scrutiny before comparing the new results with previous findings. This is hardly 'exploratory' analysis, and the description 'Initial Data Analysis' seems more appropriate.

In practice, the analyst must be prepared to face both comparative and true exploratory situations and it is fortunate that the techniques described in Section 3 have the virtue of being useful in all types of situation whenever we have to analyse a new set of data.

6. Discussion

In principle most statisticians accept the need for some sort of EDA. However, in practice, the literature suggests that EDA is often undervalued or neglected and sometimes even actively regarded with disfavour. One has only to look at published work to see for example how often tables and graphs are poorly presented. More seriously Example 4 in Chatfield (1985) demonstrates that EDA is sometimes overlooked completely. When this omission occurs, a sophisticated inferential technique may be implemented which is really inappropriate.

Fortunately there are signs that the use of EDA is increasing and Chatfiled (1985, Section 6) refutes a number of arguments against the discussion and use of EDA. This section provides a brief summary.

The first suggestion that might be made is that EDA is straightforward and does not warrant serious discussion. This clearly does not apply to advanced aspects of model formulation, but I would maintain that it does not apply to simple descriptive statistics either. As noted above, the latter are often poorly presented suggesting that 'common-sense' is not common. Indeed I believe that EDA needs to be taught far more thoroughly than it is at present. The idea that most people 'do' EDA but don't bother to talk about it may be true to a limited extent but is not a situation which is either desirable or generally true.

Some analysts dislike EDA because of its lack of theory. However this does not make the subject trivial contrary to some expectations. Indeed EDA can be *more* demanding than more advanced inferential methods because of the need for careful subjective judgement and because the analyst cannot rely completely on a computer package.

A further argument against the use of EDA is that it is not model-based and that such an analysis runs the risk of giving invalid conclusions. The implication here is that the analyst may be tempted to think that an EDA is sufficient. While this does carry certain risks, there are occasions when the results are so clear-cut that no further analysis is necessary as pointed out earlier. However I would emphasize that EDA is primarily intended as a preliminary exercise and will normally be followed by a formal model-based analysis. While analyses based on *no* model can be dangerous, it is even more likely that analyses based on the *wrong* model will give invalid conclusions, and, only by looking carefully at the data, can the analyst be sure of using a model which is at least approximately valid. The OR analyst is probably even more inclined to work within the framework of a class of probability models than the statistician, and needs to be aware that the use of EDA in formulating a sensible model can be vital.

7. Summary

Model-building in OR usually relies at least partly on an exploratory examination of a relevant set of data. This is often described as Exploratory Data Analysis (EDA). The main ingredients of EDA have been briefly discussed and it is suggested that the two main objectives of EDA should be data-description and model-formulation. With the aid of some examples, it is emphasized that EDA should be seen as an integrated stage of general statistical inference and not as a separate subject. After discussing some possible drawbacks to Tukey's (1977) important book, the paper advocates an alternative description for EDA, namely the *Initial Examination of Data.*

References

Barnett, V., and Lewis, T. (1985), *Outliers in Statistical Data*, 2nd edn., Wiley, Chichester.

Chambers, J.M. et al. (1983), *Graphical Methods for Data Analysis*, Wadsworth, Belmont, CA.

Chatfield, C. (1978), "The Holt–Winters forecasting procedure", *Applied Statistics* 27, 264–279.

Chatfield, C. (1982), "Teaching a course in applied statistics", *Applied Statistics* 31, 272–289.

Chatfield, C. (1983), *Statistics for Technology*, 3rd edn., Chapman and Hall, London.

Chatfield, C. (1985), "The initial examination of data (with discussion), *Journal of the Royal Statistical Society A* 148 (in press).

Cox, D.R., and Snell, E.J. (1981), *Applied Statistics*, Chapman and Hall, London.

Ehrenberg, A.S.C. (1979), "Book review of Tukey (1977)", *Applied Statistics* 28, 79–83.

Ehrenberg, A.S.C. (1982), *A Primer in Data Reduction*, Wiley, Chichester.

Erickson, B.H., and Nosanchuck, R.A. (1977), *Understanding Data*, McGraw-Hill Ryerson, Toronto.

Everitt, B.S. (1978), *Graphical Techniques in Multivariate Analysis*, Heinemann Educational Books, London.

Gilchrist, W. (1984), *Statistical Modelling*, Wiley, Chichester.

Goodhardt, G.J., Ehrenberg, A.S.C., and Chatfield, C. (1984), "The Dirichlet: A comprehensive model of buying behaviour (with discussion)," *Journal of the Royal Statistical Society A* 147, 621–655.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (eds.) (1983), *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

McNeil, D.R. (1977), *Interactive Data Analysis*, Wiley, New York.

Raftery, A.E. (1985), Time series analysis, *European Journal of Operational Research* 20(2), 127–137.

Rao, C.R. (1983), "Optimum balance between statistical theory and applications in teaching", in: D.R. Gey et al. (eds.), *Proceedings of the First International Conference on Teaching Statistics,* Univ. of Sheffield, Sheffield, 34–49.

Stuart, M. (1984), "Book review of Hoaglin et al. (1983), *Journal of the Institute of Statisticians* 33, 320–1.

Tufte, E.R. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Tukey, J.W. (1980), "We need both exploratory and confirmatory", *The American Statistician* 34, 23–25.

Tukey, J.W., and Mosteller, F. (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

Velleman, P.F., and Hoaglin, D.C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury Press, Boston, MA.