Berry, J.S. et al. (eds.) (1984):
Teaching and Applying Mathematical Modelling
Chichester: Ellis Horwood Ltd.

Exploratory Data Analysis—A New Field of Applied Mathematics

R. Biehler
Universität Bielefeld

1. INTRODUCTION

This chapter deals with Exploratory Data Analysis (EDA) and is based on a detailed theoretical analysis of the latter (cf. Biehler, 1982). On the one hand, EDA is controversially discussed among statisticians because it questions quite a lot of principles that underlie traditional statistical work. On the other hand, EDA begins to play quite a role within discussions on the teaching of probability and statistics because it seems to be both rather elementary in a mathematical respect as well as widely applicable and close to practical work with real data (cf. for example Gnanadesikan *et al.* 1983). These features will probably, or so we hope, make EDA interesting enough for a conference devoted to problems of teaching applications of mathematics. The paper is not concerned with questions of teaching directly. As EDA is as yet not well known or, sometimes, is apparently known but misunderstood, the paper concentrates on clarifying general principles of EDA, on discussing its relation to subject matter fields and to traditional statistics. Subsequently, some propositions are formulated which argue that the results of studying EDA might be more generally important for a theory of mathematics education which is—to formulate this as a challenge—not only concerned with teaching mathematical modelling, but also with teaching applied mathematics.

Exploratory Data Analysis is a recent scientific development attained within the efforts to find new tools and principles for the practical analysis of data. EDA and its rise are closely linked to the textbook bearing the same title published in 1977 by J. W. Tukey, from whose introductory lectures at Princeton University it evolved. The textbook contains techniques for the handling and the representation of data, in which probabilistic concepts have

only a subordinate role. The focus of EDA is on the exploration of data, i.e. on the search for peculiarities and structures in data sets, and for simple comprehensive descriptions of the phenomena discovered. Graphical displays are the main tools for this activity. In the next section, an example will be discussed to illustrate EDA's approach. It is primarily a vehicle for general ideas and should not be misunderstood as an example which can directly be transferred into the classroom.

2. SOME PRINCIPLES AND TOOLS OF EXPLORATORY DATA ANALYSIS—DISCUSSED IN THE CONTEXT OF AN EXAMPLE CONCERNING SUICIDE RATES

The example is concerned with a set of data representing suicide rates or, more precisely, death rates due to suicide. It has been taken from Erickson and Nosanchuk (1977, pp. 14) and is also discussed by Biehler (1982, pp. 70). Fig. 1 contains a table of suicide rates related to sex, country, and age.

The aim of EDA is to explore data in order to reveal structure and anomalies, so-called 'indications'. Indications are something to be given to the subject matter expert, who is to think about their interpretation and relevance in connection with his knowledge and with problems of the respective subject matter, i.e. the sociological problem of suicide in this case. It is claimed that the results of EDA can contribute independent new perspectives to the analysis of subject matter issues. This claim is justified because general experience with data analysis in very different fields is incorporated in the general tools and principles of EDA. By applying EDA methods, this experience is implicitly used to analyse new data. It is not claimed, however, that the results have some objective truth or meaning independent of the concrete context. Rather the results are thought to be instruments used to explore the respective context further.

Returning to our concrete example, we can get quite a lot of information from Fig. 1 by visual scanning of the data, e.g. the information that generally male suicide rates are higher than female ones or that suicide rates seem to increase with age etc. The question is how this activity can be made more efficient. The general approach of EDA consists of developing new graphical representations for data to facilitate their exploration. Also, general orientations about what to look for in such displays are given. This approach can be understood if we assume that representations have a double function: they do not only store known information, but they also have an exploratory function, i.e. they are tools to develop knowledge. Different representations usually differ widely with regard to their sensitivity to certain relations and structures in data sets. That is why we find two principles underlying the practice of EDA. Firstly, a *principle of multiplicity of representations*, respectively a *principle of varying representations* is applied, which allows one to explore data from different

		Age				
Country	Sex	25–34	35–44	45–54	55–64	65–74
Canada	M	21.6	27.3	31.1	33.5	23.5
	F	7.8	11.5	14.8	12.3	9.2
Israel	M	9.4	9.8	10.2	14.0	27.3
	F	7.6	4.2	6.7	22.9	19.1
Japan	M	21.5	18.7	21.1	31.1	48.7
	F	14.0	10.3	13.2	21.0	40.1
Austria	M	28.8	40.3	52.3	52.8	68.5
	F	8.4	16.4	22.4	21.5	29.4
France	M	16.4	25.2	36.1	47.3	56.0
	F	6.6	8.9	13.0	16.7	18.5
Germany	M	28.3	34.6	41.3	49.1	51.8
	F	11.3	15.6	24.2	25.6	27.3
Hungary	M	48.2	65.0	84.1	81.3	107.4
	F	12.7	18.4	26.9	34.7	47.9
Italy	M	7.1	8.3	10.8	17.9	26.6
	F	3.5	3.7	5.5	6.7	7.7
Netherlands	M	7.8	10.6	17.9	20.2	28.2
	F	4.7	8.2	10.5	15.8	17.3
Poland	M	26.2	29.1	35.9	32.3	27.5
	F	4.4	4.7	6.6	7.3	7.0
Spain	M	4.1	7.0	9.6	15.7	21.9
	F	1.4	1.6	3.8	5.4	5.7
Sweden	M	27.6	40.5	45.7	51.2	35.1
	F	13.0	17.5	19.6	22.4	17.1
Switzerland	M	21.7	33.6	41.1	50.3	50.8
	F	10.4	15.9	18.2	20.1	20.6
UK	M	9.6	12.7	14.6	17.0	21.7
(England and Wales)	F	5.1	6.5	10.7	13.0	14.1
United States	M	19.6	22.2	27.8	32.8	36.5
	F	8.6	12.1	12.5	11.4	9.3

(Table taken from: Erickson and Nosanchuk, 1977, p. 14; their source of data: *World Health Statistics Annual 1971*, vol. 1; World Health Organization, 1974.)

Fig. 1. Suicide rates 1971; unit: number per 100.000

points of view. Secondly, a *principle of making representations more efficient* underlies the construction of new and the transformation of old schemes of representation. The latter principle takes into account both the human mind's specific but limited capability to recognize patterns and the specific structures for which the representations should be sensitive. The latter reflects general practical experience as to which structures seem to be most relevant in exploring data. Below, several representations of our data set will be constructed to illustrate use and usefulness of these principles.

In the following, we shall concentrate on male suicide rates. First, we will abstract from country-dependence and treat the data in each age category as a one-dimensional data set. For the exploration of one-dimensional data, Tukey has invented a new display, the stem-and-leaf display. Fig. 2 shows such a

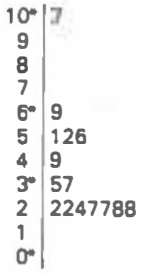

Fig. 2. Male suicide rates, age 65–74; stem-and-leaf display

display for the data of age group 65–74. The data have been sorted. The first digit has been put in the so-called 'stem', the second digit in the so-called 'leaf'. The third line from the bottom, for example, represents the (rounded) data 22, 22, 24, 27, 27, 28, 28. In fig. 2, several indications can be seen. Among others, it indicates where the values are centered, that the distribution is highly skewed and that there is an extreme outlier (value 107). All these indications can hardly be seen in the original table. Obviously, the stem-and-leaf display is related to the histogram. The following advantages of the new display are notable.

The stem-and-leaf display enables us to perceive the distribution of data values within each interval. Also, it is simpler to proceed from a value in the display to the datum that produced it. For example, it is easy to identify the country belonging to the outlier (Hungary) and to the lowest values. The change to the stem-and-leaf display reflects a change in goals with regard to the exploration of data. Usually, a histogram displays relative frequencies and is used to get a first impression of the underlying probability distribution. The individual data values do not matter and are not 'distinguishable' because it is usually assumed that the data have been generated by independent repetitions of an experiment under similar conditions. EDA is concerned with situations where these conditions are not satisfied and where it might be important to get more information about individual data, e.g. about the circumstances that have produced an outlier.

Stem-and-leaf displays can also be used to compare several sets of data. Fig. 3 shows a so-called 'back-to-back' stem-and-leaf for two age groups. The back-to-back arrangement of the stem-and-leaf displays follows the principle of making representations more efficient, because it would be more difficult to compare two distributions if the two displays were arranged, for instance, one beneath the other.

Several new phenomena can be seen in the new representation:

— the average level of suicide rates is higher in the older age group,
— the spread of the data has increased,

age 25-34		age 65-74
	10*	7
	9	
	8	
	7	
	6*	9
	5	126
8	4	9
	3*	57
98862220	2	2247788
60	1	
9874	0*	

Fig. 3. Male suicide rates, age 25–34 and age 65–74; 'back-to-back' stem-and-leaf display

— both data sets contain one upper outlier, which seems to be more extreme in the older age group.

One might have expected the rise of level after having explored the table of Fig. 1. The increase in spread is a new discovery. One may begin to speculate about the interpretations of the indications revealed. One reason for the greater diversity of suicide rates for older men might be that perhaps the experience of old age and the older people's values related to suicide differ more strongly from country to country than those of younger men. This may be due to the fact that they grew up in earlier periods when countries were less 'homogenized' by industrialization and higher levels of international communication (cf. Erickson and Nosanchuk 1977, pp. 18). We shall not follow these lines of research, which are the responsibility of the subject matter expert, the sociologist, but shall continue to explore the data in order to reveal more indications.

Fig. 4 is another version of the stem-and-leaf display, in which the numbers have been substituted by coded country names. Again, several new indications can be noticed:

—the outlier is Hungary in both age-groups
— some countries have hardly changed their rank and still cluster together
— France, for instance, has dramatically changed its rank.

These indications could be used as a starting point for detailed investigations into differences and similarities among the different groups of countries in order to see which variables might affect suicide rates. We shall now try to compare all the five age groups by their stem-and-leaf displays. The general increase of level can be seen in Fig. 5, but it is difficult to get a clear impression with regard to spread and shape. Besides, the huge variety of numbers impairs visual comparison. EDA has developed some new graphical displays to

age 25-34		age 65-74
	10*	HUN
	9	
	8	
	7	
	6*	AUS
	5	SWI, GER, FRA
HUN	4	JAP
	3	SWE, USA
AUS, SWE, GER, POL, CAN, JAP, SWI, USA	2*	SPA, UK, CAN, ISR, ITA, NET, POL
FRA, UK	1	
ISR, NET, ITA, SPA	0	

Fig. 4. Male suicide rates,
age 25–34 and age 65–74;
'back-to-back' stem-and-leaf display
with additional information (country codes)

	25-34	35-44	45-54	55-64	65-74
10					7
9					
8			4	1	
7					
6		5			9
5			2	310	621
4	8	01	161	79	9
3		54	166	4123	57
2	22986802	9275	18	0	4778822
1	60	0913	01805	4867	
0	9784	87			

Fig. 5. Male suicide rates;
5 stem-and-leaf displays

optimize such comparisons. In Fig. 6 we see five so-called box plots for the suicide data. The line in the middle of each box marks the median of the data, the two lines at the end of each box mark the lower quartile q_l and the upper quartile q_u. Thus, the length of the box represents the interquartile range, which is used as a measure of spread. The nonsymmetry of the box gives a rough indication of skewness. The meaning of the lines starting from the box can be explained as follows: A rule of thumb for outlier identification is used. First, so-called 'fences' are defined by

$$\text{upper fence } f_u = q_u + 1.5\,(q_u - q_l)$$
$$\text{lower fence } f_l = q_l - 1.5\,(q_u - q_l)$$

Values outside the fences are called outliers and have been marked by a small circle in Fig. 6. The most extreme values inside the fences are called 'adjacent

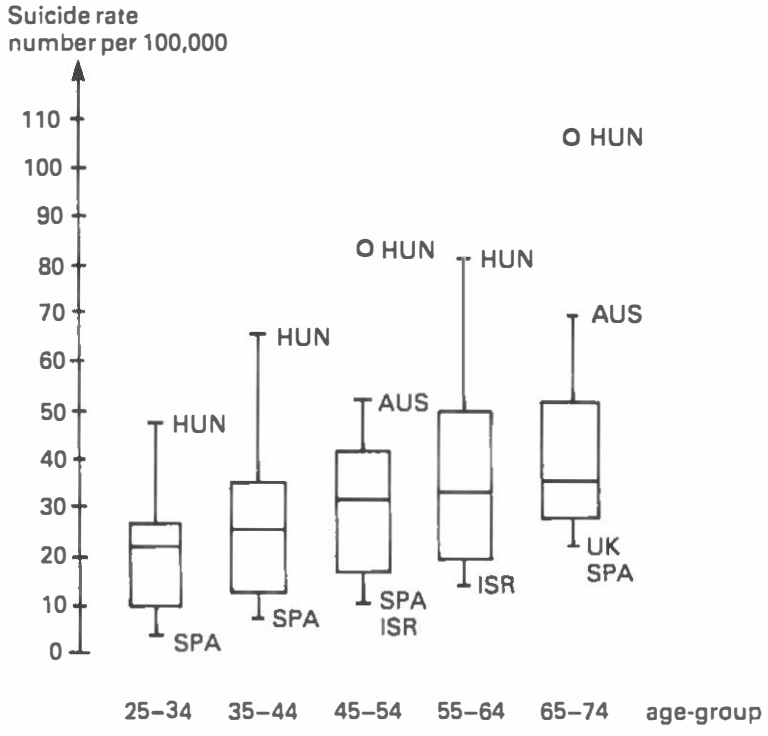

Fig. 6. Male suicide rates; 5 box plots

values'. The lines go from the end of the box to the respective adjacent values. Both outliers and adjacent values have been identified in Fig. 6 by the first three letters of the country they belong to, because general experience suggests that they deserve special attention. At this place, it is not possible to tell more about the genesis and the justification of this rule of thumb; for more details see Tukey (1970, ch. 5, pp. 14) and Biehler (1982, pp. 285).

Box plots are visual displays of numerical summaries of data. They reflect an important change as compared to traditional descriptive statistics. The arithmetical mean as a measure of location and the standard deviation as a measure of spread are no longer applied in EDA to summarize data. Instead, the median and the interquartile range are used because they are simply calculable *resistant* measures of location and spread. Resistance means, roughly speaking, resistance with regard to outliers. The traditional measures are strongly affected if there are outlying observations. To use resistant measures and to define outliers with regard to resistant measures of location and spread is a double strategy which serves to achieve a clear separation of the

data in a summarized 'main group' and in a set of outliers. This is particularly important in exploratory work, where so-called 'dirty data', which usually contain outliers, are to be analysed. The special attention paid to outliers is based on the experience that outliers are frequently interesting indications of relevant variables which have not been thought of in advance. It should be noted that the concept of resistance is related to the statistical concept of robustness, but it is conceptually distinct from that concept (cf. Biehler 1982, pp. 57).

Now let us return to Fig. 6. Among other things, we see the following new indications:

— the spread (height of the box) shows an upward trend; the oldest age group, however, does not fit into this pattern;
— the type of skewness changes systematically from one direction to the other;
— the character of Hungary as an 'overall outlier' shows up very clearly;
— Spain is always at the bottom of the distribution.

All these indications raise a lot of questions. We shall explore only one of them because the data can be exploited even more in this respect. The discovery that the oldest age group does not follow the trend of spread pertaining to the other ones can be interpreted as follows. The group 65–74 is an outlier on a "higher level". As the general strategy of EDA in dealing with outliers is to search for discriminating variables, this line of enquiry seems to be promising. A peculiar feature of the group 65–74 is that 65 is the age of retirement in most countries. Hence, might there be a particular retirement effect, which decreases variation? To explore this hypothesis, which has typically been generated by the combined effort of EDA and subject matter considerations, we shall construct two more representations of our data. Pursuing the principle of varying representations one step further, they will contribute still another perspective on our data.

Figs. 7 and 8 show profile displays. To avoid visual confusion, i.e. following the principle of making representations more efficient, separate pictures for countries with and without strictly increasing suicide rates have been constructed. Even in those countries with strictly increasing suicide rates 'something happens' in connection with retirement. In Germany and Switzerland, for instance, the 'rate of change' decreases. In some other countries, it increases. Most interestingly, some countries in Fig. 8 show even a pronounced absolute decline of suicide in the oldest age group. Besides, Fig. 8 shows another anomaly which could not be seen in the other displays. Japan differs from all other countries because we perceive a decrease from the first to the second age group. This particular outlying character of Japan surely deserves special considerations as to its social structure.

After having explored Figs. 7 and 8 we shall be surer that something particular happens in connection with retirement and that the decrease of spread noticed in Fig. 6 is probably no mere chance fluctuation. Above that, the

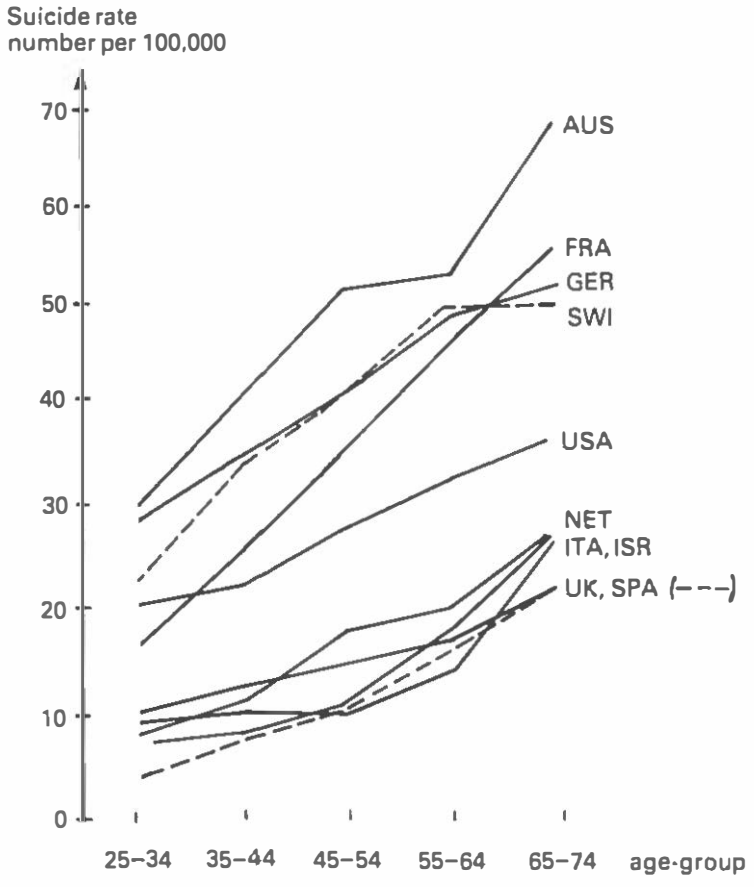

Fig. 7. Male suicide rates;
countries with strictly increasing rate;
profile displays

last two figures have revealed some further structure in the data. They allow us to classify the countries in a new way, namely with regard to the type of profile and not only with regard to the average level of suicide rates. This might open some new lines of further investigation. Such investigations may be carried through on the subject matter side, but one might also think of exploring the female rates or data from other years than 1971 along the lines suggested by the results of the above analysis.

Let me summarize the general aspects of EDA which have been illustrated by our example.

First, some new displays have been introduced, such as the stem-and-leaf and

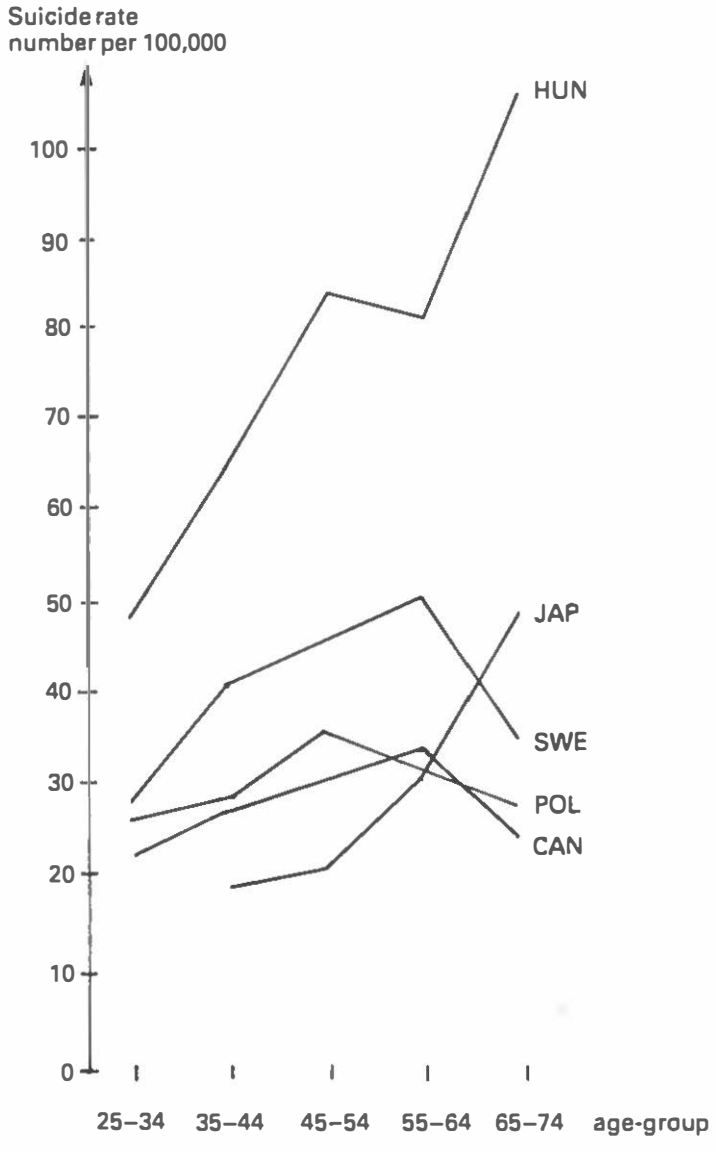

Fig. 8. Male suicide rates; other countries; profile displays

the box plot. It has been argued in which respect they are results of the attempt to make representations more efficient.

Second, the usefulness of the principle of multiplicity of representations has been illustrated.

Third, it has been shown what it means to discover indications in data sets, which function they have for providing new perspectives for the investigations of the subject matter expert. It must be added that the textbooks on EDA develop a systematic knowledge about the question, which indications in the different displays should be looked for. This is part of the general competency of the EDA-expert.

Fourth, the fundamental role of graphical tools for exploring data has become clear. This is mainly due to the fact that graphical displays are particularly suited for using the pattern recognition capability for discovering new phenomena (see MacDonald-Ross, 1979). It is only a slight exaggeration if we state that EDA is an important element in a recent historical revolution concerning the importance of graphical displays for the analysis of data (cf. Beniger and Robyn, 1978; Fienberg, 1979). In practical work with large data sets, the computer plays an important part in calculating and in constructing graphical displays on the basis of the latter. This development can be interpreted as organizing a particular man-machine interaction. It is not aimed at a complete automatisation of data analysis, rather, the exploration of graphical displays and the decisions what to do next are genuine human activities, which are not formalized.

3. EXPLORATORY DATA ANALYSIS AND TRADITIONAL STATISTICS

I shall now turn to the question of the relation of EDA and traditional probabilistic statistics. The example of suicide rates will provide an illustrative background for the following general considerations. There are two related, but different approaches to this problem:

— the type of data which can be analysed;
— the type of problems with regard to data which can be solved and the role of data in the analysis.

From the first perspective, it has to be stated that, roughly speaking, classical objectivist statistics has to presuppose that data are a random sample from a wider population. As our example illustrates, EDA has a wider range of applicability. Its objects are data sets which may be whole populations or samples for which the broader population and the way of sampling is not known. One motive for developing EDA was to construct general methods for the analysis of such so-called 'dirty data'. EDA represents a constructive answer to practical demands which can be distinguished from two other complementary ways of responding. There are statisticians who accept the limited applicability of probabilistic statistics and deny the possibility of general methods for the analysis of dirty data. Similarly, there are subject matter

experts who favour subject specific intuition where statistical methods cannot be applied.

Second, the general questions answered by statistical methods are problems of parameter estimation, confidence intervals and testing of hypotheses. Even if the data are a random sample, certain conditions as to the state of knowledge have to be fulfilled before statistical methods can be applied. For example, a set of hypotheses which should be tested has to be known or an estimator must have been chosen whose reliability should be determined. Such 'pre-data-decisions' usually rely heavily on subject matter knowledge and on considerations about which aspects of the data might be relevant for the subject matter problem. Now, EDA has developed methods which are useful in situations where such definite assumptions and questions do not exist but rather should be generated by means of exploring the data. Hence, even if the suicide rates had been a random sample, it would not have been reasonable, for instance, to apply analysis of variance techniques to estimate and test for effects. For it could not be assumed that the data sets only differ with respect to means, or that such a difference would be the most relevant one for the problem. Thus, the second difference between EDA and classical statistics is not an ontological one, i.e. not related to different types of data, but an epistemological one, i.e. related to the level of pre-data knowledge. It is central for EDA that the pre-data knowledge or hypotheses do not determine the data analysis completely as in traditional statistics; rather, the data themselves influence their analysis with the effect that the pre-data knowledge may be transformed and enriched by unforeseen elements. This strong claim becomes clearer when we recall the principle of classical statistics which forbids that hypotheses be formulated after data inspection because that would render the significance levels invalid. So, EDA breaks with the statistical tradition that error probabilities should be controlled, re-establishing the positive principle that revealed indications should be interpreted by the subject matter expert and evaluated on the basis of other data and in other experiments. EDA has been compared to the activity of a detective who collects indications that later are to be carefully considered by the court of statistical inference (cf. Tukey 1977, p. 1). This metaphor is a bit misleading as it seems more promising in many applications of EDA, just as in our example, to relate the indications to the context of the data or to use them as guidelines for analysing similar data. The direct test of observed differences with regard to statistical significance is only one option among several and cannot have the conclusive power present in the case of pre-data hypotheses.

4. EXPLORATORY DATA ANALYSIS AND TEACHING APPLIED MATHEMATICS

In this concluding section I shall take a rather abstract view of the problem how to make school mathematics more 'applied'. I interpret the new emphasis given

to mathematical modelling partly as a response to the exaggerated alignment of curricula to pure mathematics. But two problems seem to me to remain unsolved by the new approach. First, this approach does not touch the problem what type of mathematics should be taught in the rest of the curriculum which is not devoted to solving modelling problems. Second, the emphasis on solving realistic problems, i.e. on model formulation, interpretation and evaluation, often faces the problem, at least in contexts of general education, of limited knowledge of subject matter which seems to be highly relevant in these processes. Although I have no firm conclusions to offer, the analysis of EDA leads to some 'indications' which suggest that it might be promising to rethink some issues concerning the desirable type of applied mathematics in the classroom.

EDA has two peculiar features, one concerns its internal structure, its particular type of mathematical activity, the other concerns its relations to subject matter fields. As to the first, some mathematicians have said that EDA does not belong to mathematics. Indeed, EDA is an experimental activity with systems of numbers which is organized by consciously optimized tools and general heuristic principles. Vague mathematical concepts play an important part in exploring displays. No theorems are proved, the ultimate justification of the tools and procedures lies in their success in practice, and it is even more important to have rich experience with applying EDA techniques than to know mathematical properties of the techniques used. Last but not least, graphical methods play a substantial role and they do not fit into the picture of mathematics as a formal science. I agree with this diagnosis but I maintain that EDA is part of mathematics. It is just these particular features of informal and experimental mathematics which, in my opinion, deserve more attention within the efforts of changing curricula towards a more practical or applied interpretation of mathematics.

As to the relation of EDA to subject matter problems, it might be objected that in the textbooks on EDA, just as in our example, we find no genuine modelling problems discussed. Again, I agree with the diagnosis but not with the presumably implied therapy. EDA accepts that thereexists a certain level of division of labour in practice. It starts not 'from scratch' but from a certain state of a problem, then, by means of exploring data, it makes a contribution to an overall problem solution, i.e. the genuine subject matter problem is not *solved*, rather it is *transformed* according to the general experience incorporated in the mathematical techniques of EDA. This attitude, namely not to restrict oneself to a mathematical microworld but to point out new perspectives for the enquiry of subject matter problems—while seeing that an overall problem solution is not attainable in the context of developing a general data-analytical or mathematical competency—would seem to me worth considering when the question how to integrate applications of mathematics in the mathematics classroom is being discussed.

REFERENCES

Beniger J. R., Robyn D. L., 1978, *Amer. Statist.*, **32**, 1.

Biehler R., 1982. Explorative Datenanalyse—Eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie. Universität Bielefeld, IDM-Materialien und Studien Bd. 24.

Erickson B. W., Nosanchuk T. A., 1977, *Understanding Data*, McGraw-Hill.

Fienberg S. E., 1979, *Amer. Statist.*, **33**, 165

Gnanadesikan R., Kettenring J. R., Siegel A. F., Tukey P. A., Symposium on Exploratory Data Analysis, In Proc. Fourth Int. Congress on Mathematical Education (M. Zureng et al., edr), 1983, Birkhäuser, 344.

MacDonald-Ross M., 1979, *Instructional Science*, **8**, 223.

Tukey J. W., 1970, *Exploratory Data Analysis* (limited preliminary edition vols. 1, 2, 3), Addison-Wesley.

Tukey J. W., 1977, *Exploratory Data Analysis*, Addison-Wesley.