Exploratory Data Analysis – A New Approach to Modelling

Manfred Borovcnik, Klagenfurt, Austria

At first sight, Exploratory Data Analysis (EDA) only seems to be a change in techniques compared to classical statistics. Somehow, EDA has sprung off from a "back to basics" movement, back to simple and easily understandable methods requiring hardly any assumptions.

However, there is more to EDA than new techniques. EDA is an approach to modelling, with almost no assumptions, which is possibly too good to be true. It is characterized by an interactive search for a model, a process driven by substantial knowledge of the subject matter.

There will be a general discussion about relations between theory and real world, as well as an illustration of the innovative kind of modelling that is facilitated by EDA's interactive approach.

Paper presented at the Seventh International Conference on
Teaching Mathematical Modelling and Applications (ICTMA 7). Belfast, 1995.

Exploratory Data Analysis – A New Approach to Modelling

Manfred Borovcnik, Klagenfurt, Austria

At first sight, Exploratory Data Analysis (EDA) only seems to be a change in techniques compared to classical statistics. Somehow, EDA has sprung off from a "back to basics" movement, back to simple and easily understandable methods requiring hardly any assumptions. However, there is more to EDA than new techniques. EDA is an approach to modelling, with almost no assumptions, which is possibly too good to be true. It is characterized by an interactive search for a model, a process driven by substantial knowledge of the subject matter. There will be a general discussion about relations between theory and real world, as well as an illustration of the innovative kind of modelling that is facilitated by EDA's interactive approach.

1 Introduction

This section gives a brief sketch of the history of EDA. The early 70's witnessed a rise of applications of statistics emerging from the growing availability of computers and statistical software. The trend can be summarized ironically as "the fashion of racing the car (the statistical package) as fast as possible – no matter where to (what the results could stand for)". Tukey strived for a simplification of methods since the late 60's; as a result of this endeavour he published "his bible" in 1977 on Exploratory Data Analysis: The new ideas were welcomed immediately in nearly all branches of applications including those in industry; scientists and statisticians at universities were more reluctant and followed with some delay. By the early 90's, EDA has attained a renowned position as a special tool for applications and as a means to teach applied statistics at schools and at universities (see, e.g., Biehler, 1995).

Two directions of EDA are now undisputed: first, it is used as a tool for checking the assumptions of classical statistical models of inference; and, second, it is utilised as an interactive model-building device during the pilot phase of applied projects. While the tools are widely applied now, there has been hardly any discussion about the status of mathematics and the relations between a real problem and the model used to model this situation, except within an educational background. This paper intends to clarify these issues and demonstrate the potential of such an innovative approach towards modelling.

2 Views on relations between models and real world

In this part of the paper an outline and critique of the paradigm of the natural sciences is given. To overcome the inherent subjectivity of any application, multiple modelling is advocated here. The discussion will include the usual approaches to deal with the variety of problems, which emerge from the possibility that inherent assumptions of a model may be violated in a special situation.

2.1 The Paradigm of the Natural Sciences

Modelling is usually performed within an approach with a specific view on the links between a situation in the real world and a mathematical model of it; this so-called Paradigm of the Natural Sciences even dominates the application of mathematics in the "soft sciences". According to this naive but wide-spread conception of mathematics that has emerged from classical physics, models are an *objective* description (or at least have to be an objective description) of some sort of "real world" (whatever this means). From a "problem situation" R embedded in the real world, using some assumptions would lead to a model M on the theory side. Within the level of theory, a solution S is derived by applying some additional optimality criteria; see Figure 1 for a schematic description of the interrelations between R, M and S. This solution S is then transferred back to the real world to solve the initial problem R.

The intersubjective agreement about a solution S is mainly based on the unique and objective way to derive S from the model M. However, it is a diffuse and highly subjective task to find a model M that adequately represents the initial situation R. Within this naive paradigm of science, the idea is to repeat the modelling steps iteratively until the fit of the model M to the initial situation R is sufficient. For details of refined approaches of this iteration cycles of applications, or a wider conception of applications see Jablonka (1996).

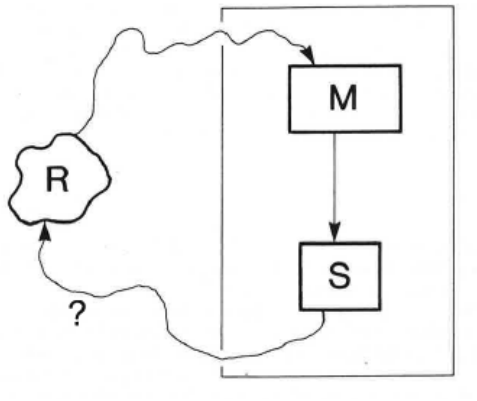

Fig.1: Paradigm of the natural sciences –feedback from solution S to the initial situation R in real world gives rise to a new *cycle of modelling* until the fit of the model M to the problem R is sufficient

2.2 Critique of the Natural Sciences Paradigm

Clearly, the paradigm of applications above is naive. Modern physicists focus much more on *functional* aspects of models instead of its descriptive potential (see Jablonka, 1996, for a discussion of these aspects). However, this paradigm still dominates the great majority of applications and the external opinion of the potential of mathematics. Evaluations are perpetuated as it dominates also the teaching of mathematics and statistics. In what follows, some constraints on objectivity for modelling a real-world problem are discussed; see Borovcnik (1992) for a more detailed critique.

- The initial problem R and the model M are at epistemologically different levels. How can these different entities be compared?
- The modelling process – how to get from the problem situation to the model – is highly dependent on models available to or known by the modeller.

- The background theory used affects the view on real-world problems, e.g., a classical approach to statistics refers to random samples, or uses the frequentist interpretation of probability while a Bayesian approach centres the modelling phase on the idea of weighing the uncertainty.
- There seems to be an attractive but wrong vision of a unique initial problem and a unique model.
- The check for assumptions underlying the finally chosen model is difficult if not impossible.

Consistently, any application is to some degree subjective. As a due consequence, it is an essential question on how to deal with such subjective features of applications. Since we are considering EDA, we will compare EDA with classical approaches in statistics.

2.3 The approach of Multiple Modelling

In reply to this critique, this author has elaborated on a multiple modelling approach. Borovcnik (1986, 1992) develops these ideas in a reconstruction of the famous historical problem of the "Division of Stakes" by using both a classical frequentist and a Bayesian analysis. Accordingly, an initial problem situation is analysed from the perspective of *different* models (see Figure 2 for a schematic diagram). This attitude towards modelling should help us to cope with several drawbacks of the modelling process:

- The initial problem situation is too vague.
- The process of modelling is accompanied by subjective decisions.
- The discrepancy between a problem and its model cannot be judged easily.
- The cyclic "improvement" of the distance between the entities of problem situation and model might blur the difference between them as the final model *fits* to the resulting view of the problem situation; but it fits to a situation that may have nothing in common with the *initial* problem.

To view a problem from various perspectives might yield a holographic 3D image of it, which could enable insights. The insights from a visual inspection may be deepened by an analysis of the differing assumptions of the various models and the impact of these assumptions on the results. The idea here is no longer to analyse the fit of the various models to the problem *separately* but to study, in which respect these models embody different views of the problem. Revealing these tacit assumptions and inherent steps of modelling should highlight the subjective character of any application. Once such restrictions are known, their impact on the interpretation of the various results may be judged more easily and compared to the alternative choices in the modelling process and the models derived.

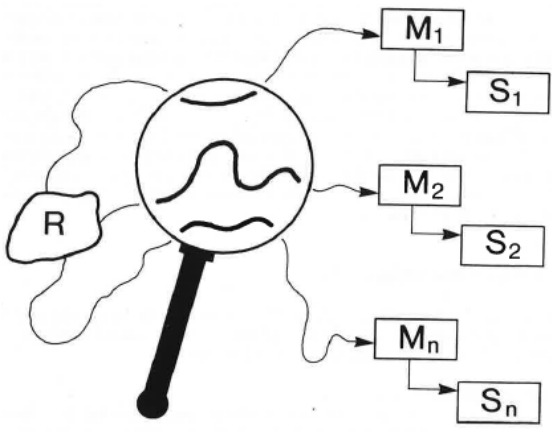

Fig.2: Multiple modelling – instead of analysing the fit of various models separately, it is investigated how and why the models differ – R real world, M models, S related solutions

2.4 Dealing with underlying assumptions

There are various schools of statistics (see e.g., Barnett, 1982) that explicitly address the problem of the assumptions underlying any model. Amongst others, these schools differ in their perception of randomness.

- Classical statistics refers to the assumption of an underlying family of distributions (e.g., normal distributions) and relies on a random sample (of independent data). It is claimed that such assumptions can be tested (however, this conflicts with the paradigm of classical statistics).
- Robust statistics investigates whether the effect of slight deviances from assumptions (e.g., the normal distribution) is small and strives for methods that "guarantee" *robust* results.
- Nonparametric statistics uses simpler assumptions (models) based on random samples from a finite population.
- The Bayesian approach refines the model by modelling the unknown parameters as random variates, which requires qualitative information to fully specify the model.
- Exploratory Data Analysis uses simpler concepts and an interactive approach to "scan" various models for the *exploration* of an initial problem situation.

This author has elaborated on the relative advantages of the various approaches and how they deal with assumptions that are required to use a model for analysing a problem (Borovcnik, 1992). Checking the assumptions by statistical tests within the *classical* framework of statistics lacks in rationality insofar as the key features of this framework cannot be applied.

For example, there is no possibility to control for the power of a test for a specific family of distributions for a variable in the model. First, there is no clear-cut idea, as to which distributions should be considered as the alternative, and second, the alternatives are *distributions* and cannot be ordered by their distance to the null distribution; the situation is similar if the assumption of independence is tested.

Different alternatives have been developed in order to avoid or control various assumptions:

- Robustness and Nonparametrics seek to avoid uncontrollable assumptions, either by investigating systematically the impact of deviations from assumed relations or by reducing the amount of assumptions via simpler models.
- The Bayesian approach introduces a new kind of information to refine models to get a better fit to the problems; this information, however, requires an interpretation of probability with subjective connotations. While it is supported by Bayesian statisticians, it is rejected by the so-called classical statisticians.
- EDA removes the assumption of random sampling – one of the key hypotheses to all the other approaches – by a *new type* of argument *to justify the generalization of results* based on samples to wider populations. The idea is to get a direct insight into "structures" of the data, which tries to corroborate the results by knowledge from the context that works like an aha insight.

All these approaches to deal with assumptions in specific ways use varying views on randomness, which lead to a very divergent scientific status (quality) of potential solutions of the individual problems that are analysed: objective (and generally approved scientific) status in classical, robust and nonparametric statistics; qualitative (personal bound) status in Bayesian statistics; insight from knowledge and subjective experience about the context in EDA.

3 EDA – a new conception of modelling

In this section, the specific character of Exploratory Data Analysis is discussed. The approach avoids separating models and real world situations completely as is done within the classical paradigm; its key features are an interactive perception of the situation by repeated steps of modelling the situation, interpreting the results, and modelling it again and so forth.

3.1 EDA equations

Regression theory describes a dependent variable or a phenomenon by an equation that reveals its dependence on various influential factors. In the spirit of EDA to use simple concepts, this practice could be used as a *metaphor* to "describe" the phenomenon of EDA as a target variable by the following equation (even though the equation sign is used symbolically and no attempt is made to develop units of measurement for the variables):

EDA = VIS + MA + MF + INT

The influential factors that establish the character of EDA via this structural equation are:

- VIS – visual analysis: the tools within EDA are mainly graphical as many people have an easier access to graphical representations than to mathematical formulae and abstract models.
- MA – multiple analyses: a special feature of the EDA approach is to use various representations and different levels of data reduction in order to search for potential inherent structures.
- MF – model-free: the approach is intended to put as few assumptions as possible into the analysis of the problem, i.e., rather than using complicated models for the data it works with direct graphical representations of the plain data.
- INT – interactive: the approach resembles playing with data; the interpretation of intermediate results decides about what to do next. Interactively, both the contextual knowledge of the modeller and the interpretation of patterns in the previous representation based on this knowledge guide the next steps of the analysis.

Following the core idea of EDA to strive for simpler descriptions, one could also arrive at the following simplified model equation for EDA:

EDA = VIS

This equation highlights that the *key* feature of EDA is the graphical and visual orientation of the analysis. The background hypothesis is that such methods put the least amount of assumptions into an analysis. Clearly, even if the graphical tools of EDA are intended to be simple, they still have to be learned like one has to learn to use any tool.

So far, these specific characteristics are ingredients of the use of EDA in the practice of statistics and in the educational discussion about the relative merits of EDA. The driving force is to simplify the ways to use statistics. For statistical practice, EDA techniques provide an easy check for the fit of classical models, e.g., the various plots for the inspection of residuals in regression problems. For educational purpose, conceptually easy techniques simplify learning and applications considerably as the learned methods do not require much pre-knowledge of a sophisticated background theory.

As an example, the median is preferred to the mean for the following reasons. The mean is an effective tool only for describing one-peaked and nearly symmetric data; or, the mean is used in inferential methods only because complex models are available for judging the significance of the

difference between the means of several sets of data (like t-test or ANOVA). Without similar restrictions, the median facilitates us to describe data sets and differences between several data sets.

Multiple analyses, a model-free approach, and interactively playing with data lead to some degree of arbitrariness and a lack of criteria, as to which technique to use and which results to judge as relevant. This prompts the usual critique against EDA, which may be summarized by the saying "If you look at some tiles carefully and long enough, then you will see *any* pattern". To justify results derived by EDA techniques, the relations between the real problem and models have to be studied at length and the context of the problem gets a vital role. A specific type of argument has to be elaborated to generalize results from an analysis; EDA develops an argument for this purpose, which is necessarily different from significance statements associated with statistical tests based on the random sample argument for the underlying data.

3.2 The EDA view on relations between models and real world

In EDA, there is no separation between the real problem part and the modelling side. Furthermore, there is no implicit hypothesis of a unique *initial* problem and no overall inherent assumption that it is possible to judge the fit of a model to this real problem. Both initial problem and model are revised by the repeated steps of analysis. Which parts of the real world situation are selected and transferred to the model and which techniques are used is decisively influenced by an interplay between intermediate results and the contextual knowledge of the analyst: This leads to an interactive progression of understanding the initial situation R and the intermediate models M'. The concepts used for modelling must, therefore, allow for an immediate and easy interpretation of preliminary results in the context as further steps of analysis would become completely arbitrary without such a feedback.

A substantial part of the insight gained from the analysis emerges from the comparison of different approaches to the initial problem and from checking how the potential results fit into the existing network of knowledge about the context of the problem prior to the analysis.

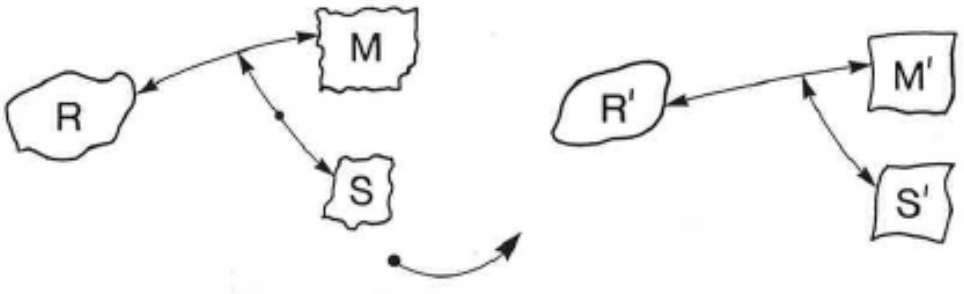

Fig.3: Interactive revision of models and real world in EDA: R, M, S initial problem in the real world, first vague model M and solution S – R', M', S' revised entities possibly less vague

A model M is simply a graphical representation of the data with a special "perspective". On the one hand, the interaction between a situation R in the real world and the model M is influenced by the selection of various views on the data (investigate only one variate at a time or two variates together; compare specific subgroups, check for a third variable, etc.) and by the choice of the perspective (i.e., the specific type of representation); on the other hand, the interaction is "driven" by contextual knowledge. This knowledge guides *what to analyse and how*; the intermediate results are interpreted within the context and the present interpretation will decide about the next steps of the analysis.

The interaction is driven from both sides leading to various steps of R' and M' revising both the model M *and* the real world situation R. The interpretation of the solution S' with respect to solving R' is not based on a judgement of some sort of fit of the model M' to the real world situation R (or, to its actual view R') as would be done within the paradigm of the natural sciences. The final result resembles *a potential view* on the real problem rather than a unique solution. It gains its justification mainly from an insight, which might be shared by the modeller and the potential users of the model.

3.3 Interactive modelling in EDA

The key feature of interactive modelling in EDA is that the real world R and model M are *not* separated as happens within the natural sciences paradigm. Both mathematical knowledge about the tools and the contextual knowledge of the modeller establish the driving force of the interpretation of previous results preparing the decision on further steps of modelling. To enhance the potential of the approach, the tools of analysis should have special characteristics. Hence this is more directly dependent on the prior (possibly subjective) experience and knowledge of the modeller about the context.

In fact, EDA tools *intentionally* support the interaction.

- Simple concepts enhance a direct interpretation of intermediate results.
- Flexible models allow nearly all types of questions to ask from the models.
- Visual representations of the data facilitate the direct (model-free) detection of inherent patterns.

The background idea of "a structure within the data" differs from the classical paradigm. In classical statistics it is hypothesized that one model fits and explains the data by its inherent assumptions; consistently, data not fitting to this model is discarded as outliers. In contrast to it, the aim of EDA is to split the data into one component comprising the fit of the main model and a further component representing the residuals; the modeller has to focus on an interpretation of the fit *and* the residuals simultaneously. In many cases, much more insight may be gained from an investigation of the residuals asking for reasons why the related objects are "outliers". The basic modelling equation in EDA thus covers two components, the main model or fit *and* the residual:

$$\textbf{Data} = \textbf{Fit} + \textbf{Residual}$$

More suggestively written as

$$\textbf{D} = \textbf{F} + \textbf{R}$$

The exploratory approach may be characterized as an interactive search for a fit (or for several fits) and an explanation of the residuals.

4 Innovative modelling features inherent to EDA techniques

This section is devoted to discuss the techniques of EDA in the light of the idea to search for patterns – the model – in the data *and* to interpret the deviations from the model by contextual arguments. In the description of the techniques, elements are highlighted which improve or facilitate the interactive modelling task. Therefore, the exposition is not reduced to a mere introduction to the techniques of EDA which is found elsewhere to a broader extent (Tukey, 1977; Velleman and Hoaglin, 1981; or, Cleveland, 1993).

On the contrary, the background idea of this paper is that EDA signifies *a new style* of data analysis. This refers to the inner mechanism of the characteristics of EDA and cannot be marked by a mere change in techniques. Consistently, the techniques are explained as briefly as possible and as is necessary to communicate features of this style of EDA. Elements of this style may be also read off from basic literature on EDA. Here, the accent is put on the inner mechanism that may be characterized by questions like "Why a specific approach is undertaken?" and "What are the aims and merits of this approach?" We present and discuss the following techniques:

Techniques for exploring distributions:

- Stem-and-leaf diagram – discover shape and subgroups.
- Median – facilitate to see a centre and detect extreme values.
- Box plot – inspecting the distribution from various angles.

Techniques for inspecting interrelations between variables:

- Scatter plot – inspecting for homogeneity assumptions and subgroups by third variables.
- The window technique – wandering through a scatter plot to detect changes and trends.
- Three-groups line – detect the type of trend and summarize it.

4.1 Techniques for exploring distributions

Stem-and-leaf diagram, median, and box plot are standard techniques of EDA to inspect the distribution of single variables. These techniques are illustrated by examples with respect to their potential to reveal the general pattern and the individual deviations in a data set. These deviations will be the focus of further investigations and will become the key basis for the task of generating insights into the results of the analysis.

a. Stem-and-leaf diagram – discover shape and subgroups

The potential of this representation of data is shown within an example of investigating the bed & breakfast places in Klagenfurt; data are displayed on the left-hand side of Figure 4. The units are H, T, U for hundreds, tens, and units. Hence most accommodation is for under 80 people.

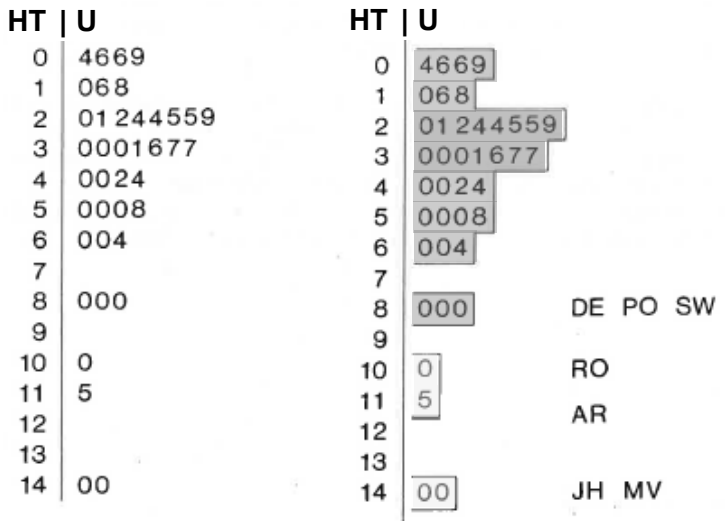

Fig.4: Stem-and-leaf diagram and shadow of it with extremes encoded – data on bed & breakfast places in Klagenfurt 1986

The “dark shadow” of the stem-and-leaf to the right-hand side of Figure 4 reveals the fit, which is slightly skewed *and* – apart from it – the residuals in the “light shadow” which seem to establish a second group. A typical EDA technique for such cases is to use codes to identify this group of deviating data in order to search for contextual knowledge for an adequate interpretation of the phenomenon. The code JH stands for the local youth hostel, which explains its size. Looking at all the other hotels from the second group of the more extreme data, shows that they are traditional hotels in Klagenfurt, which have been founded in the early 1900’s when hotels used to be much bigger than nowadays.

This insight is derived firstly on a formal split of the data into a group of ordinary data (the fit, the general model) and the extreme group and secondly on a follow-up interpretation of the extreme group by the context.

b. Median – facilitate to see a centre and detect extreme values

The median halves the distribution and its value is *not* influenced by extreme data, which is further outside. Figure 5 illustrates how the median marks the position of a centre of a distribution.

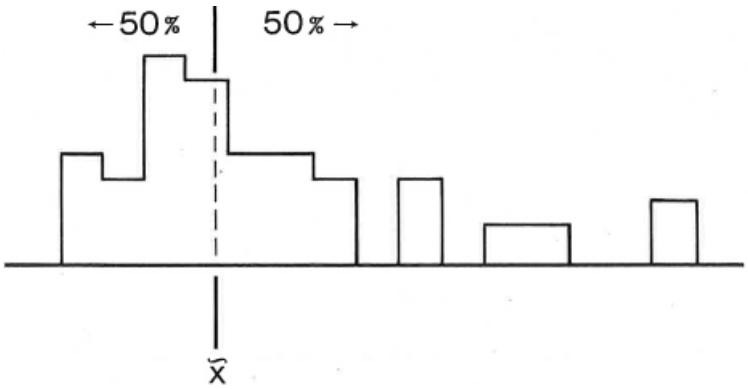

Fig.5: Median – halving the pile of points into a lower and an upper pile
– bed & breakfast data

It is important to recognize that the median is very simple; anyone can see and therefore understand immediately that the pile of data is split into two piles of equal size, the “lower” and the “upper” pile. Furthermore, the median requires only the assumption that the variable under scrutiny has to be measured on an ordinal scale. In contrast to this, it is not always obvious what an arithmetic mean could signify for a set of data, especially if the data is highly skewed such as for income. Ideally, applying an arithmetic mean is restricted to the assumption of a markedly peaked (one peak only) and symmetric distribution of the given data; in such a case, the mean marks the centre of symmetry around which the single data are scattered.

Unlike the mean, the median is not influenced by data far off the “normal” pile of data; it summarizes only the bulk of ordinary data in the centre. Consistently, the residuals, the deviations of single data from the mean tend to be larger irrespective of the circumstance whether the data are in the centre or in the outer region of the distribution. As the median is not influenced by extreme data, it better separates the two clusters of centre and extreme data; the residuals of the data from the median allows a magnifying view on the candidates for extreme data as is illustrated by the plot in Figure 6.

This is due to the fact that residuals from the median are small for the ordinary data (for which only it is a representative) but are extremely large for peculiar data (which do not influence the median at all). This is what was named as "magnifying the difference between the centre and the extreme values". This magnifying potential of the median facilitates the two different tasks of an exploratory data analysis:

- To find representative statistical regularities, which can be analysed by the methods of classical statistics, i.e., by methods for mass data.
- To identify and to explain peculiarities, which have to be analysed by individual explanation by the knowledge of the data generation and the context.

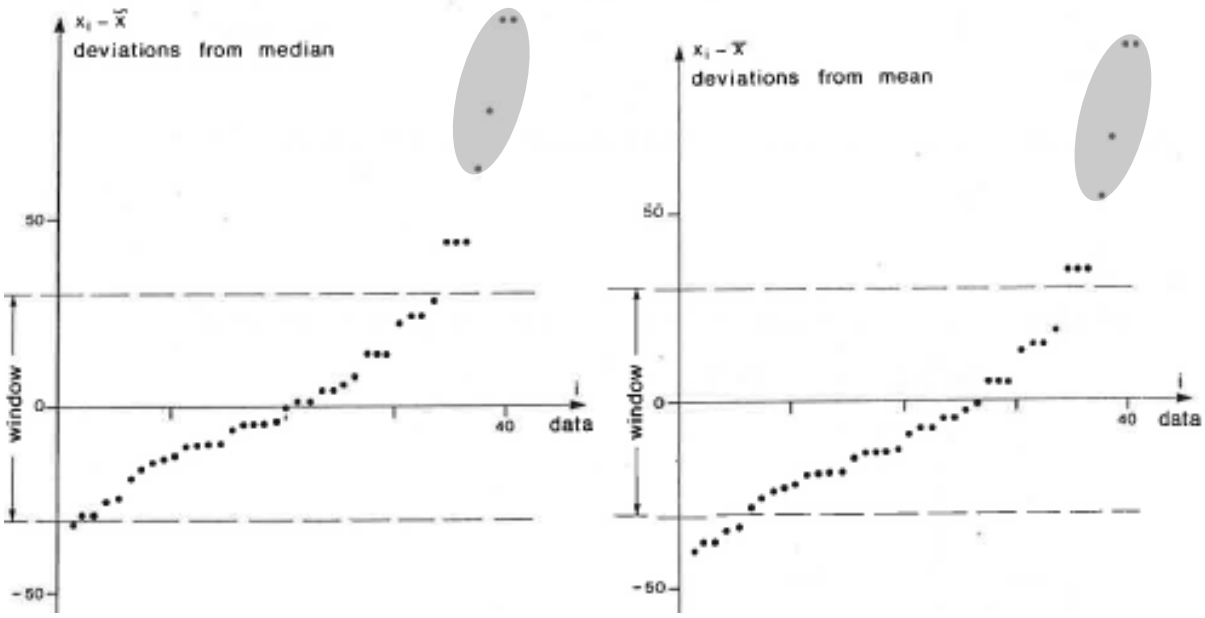

Fig.6: Residual plots – residuals of the data from their median resp. from their mean – the median magnifies the residuals for the extreme values – bed & breakfast data

c. Box plot – inspecting the distribution from various angles

The box plot can be seen as the graphical result of an iterative procedure of splitting the pile of data into two halves, see Figure 7. The lower und upper 25 % of data are usually not split into halves again but are split into an extreme region of 5 (or 10, or 1) % in which single data are singly displayed with a code allowing for identification and the so-called whiskers which allow judging the shape and spread of the outer "normal" region. The core of the inner 50 % is visualized by a box whence the name box plot, sometimes box-and-whiskers plot. Usually, the inner core and the whiskers summarize the fit of the bulk of ordinary data whereas the encoded single points mark the (candidates of) outliers. Again, a separate analysis from the context is essential. A sophisticated evaluation of the shape of the core of ordinary data is possible and may be backed by knowledge or insights from the context.

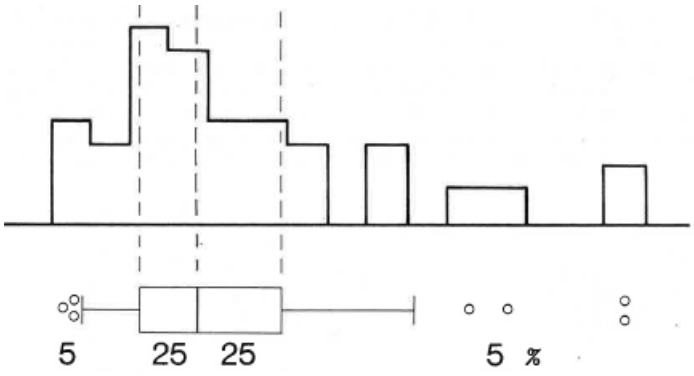

Fig.7: Box plots – the box contains the core of the inner 50 % of data,
the end of whiskers mark the outer extreme 5 (or 1) % of points – bed & breakfast data

The comparison of box plots of different groups or variables is easy and often quite conclusive. The method is illustrated in Figure 8, which shows the box plot of the residuals of the data from their mean respectively from their median in the bed & breakfast example. From the representation it is clearly visible that the median magnifies the deviation of those data that refer to more extreme data and thus facilitates the detection of outliers.

To identify potential outliers clearly is the first step; the second step is to explain them by knowledge of the context. Such an explanation is the essential criterion to generalize the results of a specific problem with EDA and replaces the random sampling argument, which is usually applied to generalize findings from data.

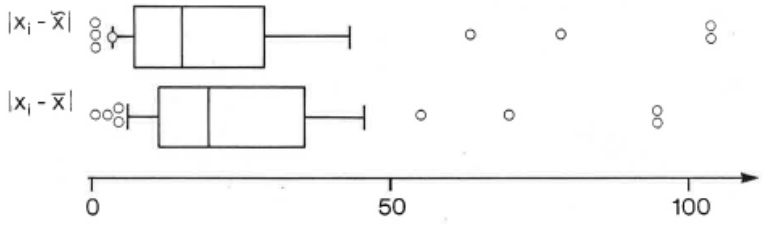

Fig.8: Box plots of residuals of the data from the median and from the mean –
the median magnifies the residuals of the extreme points – bed & breakfast data

Tukey specifies a different rule for drawing the whiskers according to which the whiskers are drawn from the quartiles with a preliminary length of 1.5 times the interquartile range (the width of the box). The actual length of the whiskers is reduced if there are no more data (they extend only to the “last” data within the 1.5 interquartile distances). Such a rule hinders a direct interpretation of the whiskers but has advantages in comparing the box plot to the normal (Gaussian) distribution.

4.2 Techniques for exploring interrelations between variables

Scatter plot, the window technique, and the three-groups line are standard techniques of EDA for the inspection of the joint distribution of two variables. Again, the techniques help to reveal the general pattern *as well as* to highlight the individual (extreme) deviations of such a pattern. The strategy to magnify and code the extreme points helps to study the interrelations between the two variables. This method supports the recall of contextual knowledge in order to explain the reasons for such deviations.

a. Scatter plot – inspecting for homogeneity assumptions and subgroups by third variables

No specific example will be given here. Clearly, an inspection of the scatter plot (like one in Figure 9) may indicate whether the assumptions of an intended classical regression analysis are fulfilled or not.

However, the EDA technique is much more powerful insofar as extreme points or groups of points are encoded in order to facilitate the retrieval of contextual knowledge. Or, subgroups according to other variables may be represented by a different symbol for the points. There is no limit to the variations of the technique in order to identify possible relations between the variables shown and to make hidden variables visible and to detect deviations from such patterns. Encoding is always a helpful strategy for explaining the results in the context.

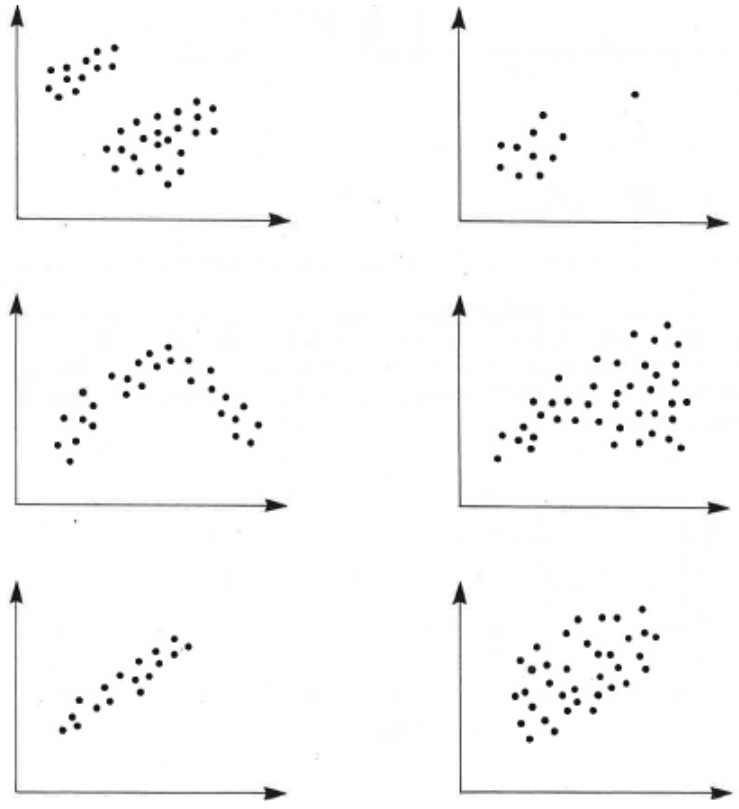

Fig.9: Inspection of scatter plots reveals different cases of relations between the variates; some violate the assumptions for the usual least squares regression

b. The window technique – wandering through a scatter plot to detect changes and trends

The window technique is a special technique to focus the attention in inspecting a scatter plot. The points are separated into horizontal or vertical stripes to increase the visual power to see the differences in the single stripes; a monotonic development of the cloud of points would be magnified thereby. For the following deliberations, the context of meteorology is used. Data are from Borovcnik and Ossimitz (1987) and represent mean daily temperatures of various meteorological stations according to their sea level and geographical altitude. The scatter plot in Figure 10 is divided into separate windows to highlight the monotonic development of the scatter plot and clearly shows the relation between northern latitude and temperature residuals.

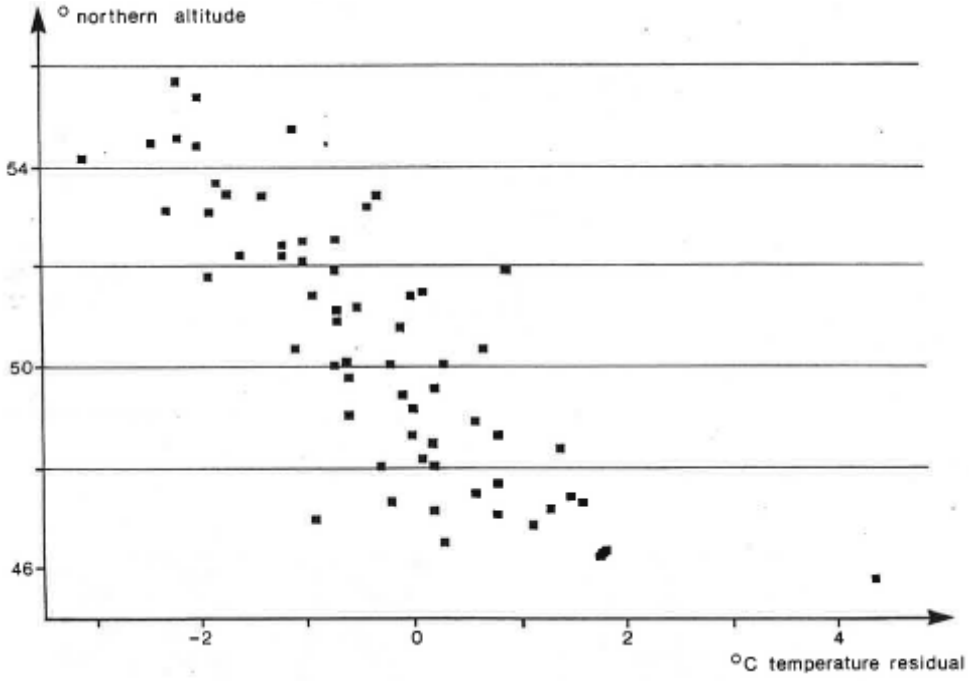

Fig.10: Window technique – residuals of temperature adjusted for sea level and geographical altitude

These *temperature residuals* have already been *adjusted for* the height above *sea level*, i.e., they are the residuals of a regression line (or some comparable technique from EDA as the three-groups line) that describes how daily temperatures of a location ideally vary with the sea level. If such an adjusted residual still is very low (i.e., large but negative) this may be explained by the circumstance that the location is far north. This relationship may be studied from the scatter plot in Figure 10.

The window technique can be refined to summarize the distribution of the temperature residuals in the various stripes separately by methods for the study of single variables. The fit may be described by a median curve, which connects the median points in the various stripes. The extreme data could be separated from the bulk of ordinary data by the quartile curves, which connect the 25 % (or 75 %) points of the data in the single stripes. Again, encoding of the extreme data might be helpful to link the points back to their context. The technique of wandering box plots is a slightly different way to summarize and compare the distribution of the temperature residuals in the single stripes and may be inspected from left to right (or upwards here) like the movement in an animated film. The diagram in Figure 11 clearly shows how northern latitude explains for low temperature residuals.

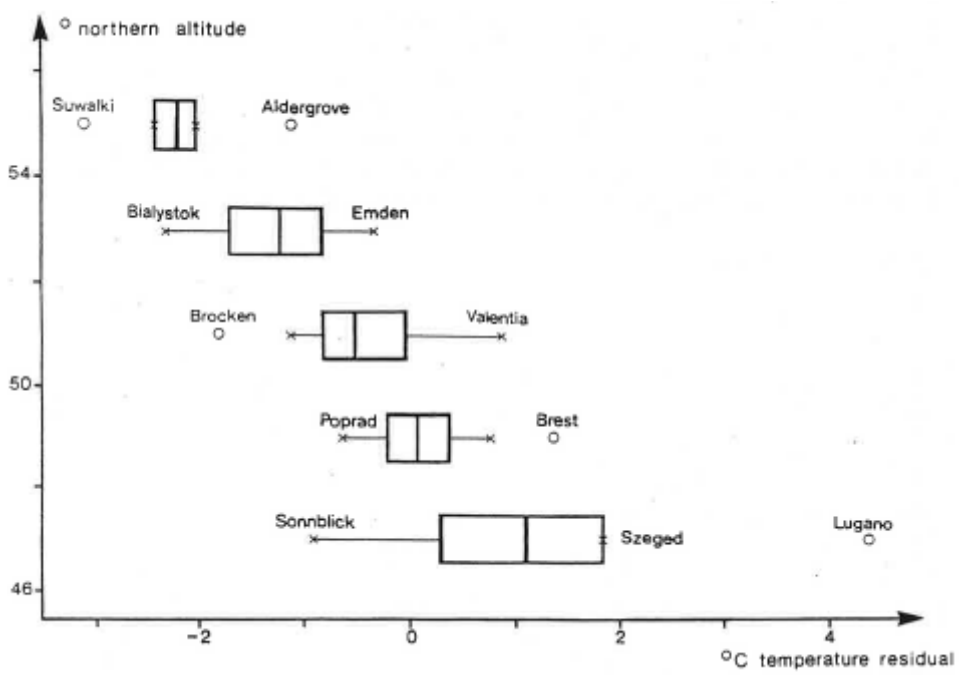

Fig.11: Wandering box plots – temperature data

To encode the still extreme points in the various stripes enables explaining them as it reveals their geographical clustering. Marking the locations on a map of Europe shows the climate zones in Europe (see Figure 12). Maritime climate reinforced by the Gulf Stream is responsible for the large positive residuals; rough continental climate in Middle and Eastern Europe is the reason for large negative residuals; two other points indicating Mediterranean climate and the climate of the Southeast; the remaining point is the Sonnblick that represents high mountains in the core of the eastern Alps.

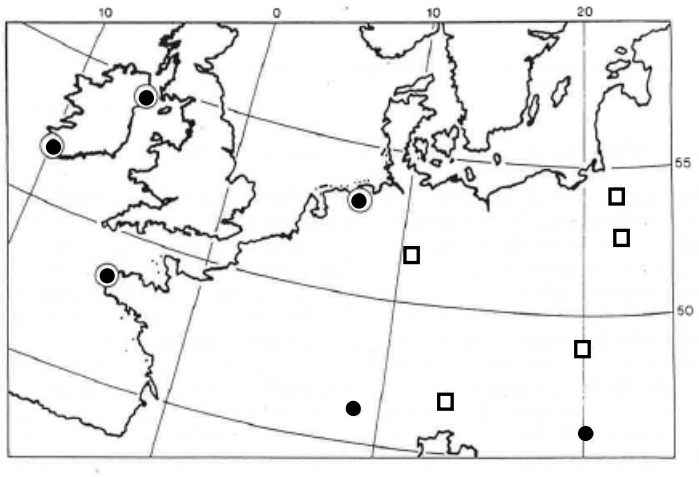

Fig.12: Geographical distribution of extreme temperatures residuals
● extremely high □ extremely low

Here, the residuals of temperature adjusted for sea level are investigated. Further variates were analyzed whether they could explain the patterns of the residuals.

The window technique was used to study the effects of a so-called covariate on the variable of interest; a covariate is either the identity of the units or its measurements on a third variable that may influence a relation between two other variables or may influence the distribution of one variate. If there is no data available for a covariate, which is suspected to exert such an influence during the analysis, it is also called *confounder* as it may change the patterns detected and turn over the results if it were possible to introduce it into the analysis.

To reveal important covariates is a prime aim in many problems of applied statistics and may contribute to a deeper understanding of the context. Ehrenberg (1981) is not in the tradition of EDA but seeks for patterns in the data by investigating further variates to recognize under which conditions a pattern does actually hold. This illustrates how EDA techniques answer directly to needs in applied statistics.

The technique of marking the points in scatter plots by a different shape according to the values of the units on a further variable is helpful in detecting covariates and relevant relations between variables, which may illuminate the context of a problem. It enfolds its full potential with dichotomic variates; with more different types of points it gets harder to inspect for hidden patterns.

c. Three-groups line – detect the type of trend and summarize it

This EDA technique constructs a robust line that summarizes the relation between two variables in a scatter plot. Here, for brevity, the technique will be explained only and no example of its potential will be discussed. It will be clear from the description that this line summarizes only the fit for the bulk of ordinary data and neglects any exceptional (conspicuous) points which might be outliers.

Again, this technique enlarges the deviations of the exceptional points while in general it decreases the residuals of the core of ordinary data. Thus it magnifies the differences between the two groups with respect to the residuals of the fitted line (which represents the model for the relation). The technique facilitates the separation between the group of the ordinary data and the group of the extreme points. Clearly, the exceptional points should be encoded and interpreted by knowledge of the context to study the question as to why they are so far off so that deeper insight into the problem studied may be generated.

Figure 13 shows the steps necessary to construct the three-groups line. First, separate the points of the scatter plot into three vertical stripes of (nearly) equal number of points and replace the single points by a median point (its components are the median of the single points in either axis). Second, connect the median points of the lower and upper stripe by a line. Third, draw the orthogonal distance of the median point of the middle stripe to the line. Fourth, move the line towards the median point of the middle stripe halfway the orthogonal distance. The result is the three-groups line summarizing the linear relation between the two variables under consideration. A “curvature” of the line graph connecting the three median points (as in Figure 13) indicates a nonlinear relation between the variates.

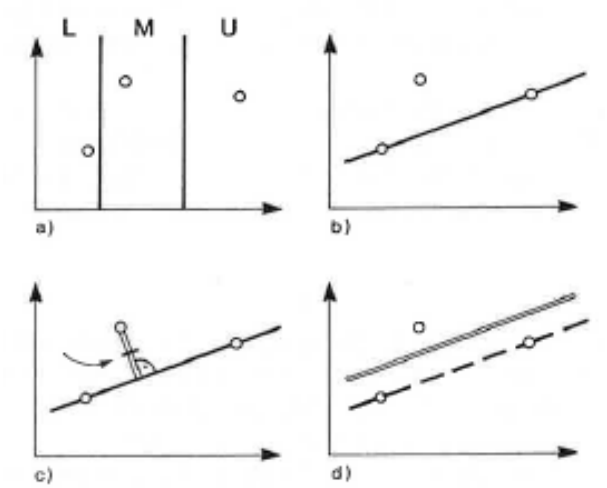

Fig.13: Construction of the three-groups line:
a) draw points of medians in each group
b) connect lower and upper median points by a line
c) draw the distance of the middle median point to the line
d) move the line halfway towards the middle median point

5 Conclusions

EDA differs in style from classical, robust, non-parametric, and Bayesian statistics. Its innovative potential could even reinforce efforts to establish more refined approaches to modelling for mathematics as a whole. A short comparison between EDA and classical statistics emerges directly from the deliberations above.

5.1 Style of EDA

Visual inspection of data is becoming more important in other branches of applications of mathematics. The approach of multiple analyses of a problem with different methods seems to be a promising strategy in modern applied mathematics; these features are inherent to EDA. No application of mathematics to a single problem situation can be free of assumptions. However, the least amount of assumptions necessary to prepare a situation for an analysis is presupposed in EDA. The interactive building of a suitable model, or better, of several suitable models is one of the key features of the exploratory approach to data analysis. In this interaction, the deliberate connection between intermediate results and the knowledge from the context for the decision about further steps is essential for data analysis to be *truly exploratory*. In the classical paradigm, the fit of a model to the problem would be a vital task whereas in EDA the separation of data into the fit and the residual (i.e., the deviations from this model) is the decisive element with the focus on handling both the fit and the residual.

The interactive approach to modelling signifies the style of EDA; the focus of modelling lies on the search for patterns *and* deviations from the patterns, or, on the search for the fit *and* residuals, or, on the search for the model *and* singularities (potential outliers). The data that fit the pattern found can be treated by traditional statistical methods for the analysis of mass data, while the singularities are

analysed by an individualistic approach bound to the context of the generation of the data and the problem at hand.

To ensure that the general model is not influenced by the singular data, the EDA approach requires robust methods. The deviations from the model should be easy to understand; their interpretation and evaluation in the context should not be limited by lack of expertise in complex theoretical models as the adequate interpretation determines further steps of modelling interactively; therefore the approach requires simple models, which are almost free of assumptions for the data and the real problem.

The style of EDA also requires techniques which allow separating the bulk or the centre of regular data (with small residuals from the fit, i.e., the model) from singularities (i.e., the potential outliers with large residuals from the fit); here this specific character of techniques was named as magnifying potential. EDA techniques are signified by a special magnifying potential, which enables the detection of singularities which can open ways for the interactive search for new insights. Furthermore, the techniques allow for a flexible search for covariates which may explain the model and the deviations by the context of the problem. The interactive style of EDA facilitates the search for relevant relations in the context of an initial problem. The argument to generalize (intermediate or final) results is established by the insight gained from results in the light of the context.

5.2 Comparison of EDA and classical statistics

The following comparison highlights the differences in the approaches showing clearly that the methods serve different needs. Whereas classical methods are unambiguous only if random data are available and if it is undisputed how to build the model (maybe from pilot studies or from contextual knowledge), exploratory methods can be of advantage in any situation without respecting the usual constraints.

Category	**EDA**	**Classical statistics**
Type of analysis	*multiple modelling*	*unique model optimal to criteria*
Models	emerge from repeated analysis	have to be determined prior to optimal analysis
Data	may be arbitrary	has to be from random sources
Goal	to describe a problem	to enable a decision or inference
Type of results	*many views*	*a unique result*
Essential factors for results	the interaction	the model used
Agreement on results	by knowledge of context	by optimization criteria
Real world and model	are tightly connected; to search for explanations	are strictly separated; the fit is “checked” later

5.3 Perspectives for the future of modelling and the exploratory approach

Contextual knowledge determines the repeated steps of an exploratory analysis and the models are built up interactively. In order to interpret intermediate results adequately, the techniques used and the context has to be familiar to the modeller.

To facilitate this key activity, the EDA approach requires simple concepts. Any need to refer to complex theories and non-transparent assumptions actually hinders the progress of analysis. Therefore, it is vital that results are directly comprehensible and relatively free of assumptions. By no means can EDA be automatically applied like conventional statistical methods (as was done during the early boom in the 60's) by delegating the analysis onto a computer programme though computer packages are helpful and have facilitated its approach by their visual capacity, i.e., by their ease to produce various diagrams of the data.

Exploratory techniques have become quite popular in the pilot phase of applied projects where they are used to clarify the perception of the investigated problem and help to examine the intended methods (models) for plausibility of inherent assumptions. EDA *generates models* and insight (or not) whereas classical models enable the testing of conjectures within *preset models*, which hardly can be checked for validity.

It is worthy to mention that an exploratory approach is not restricted to the simplest statistical methods as an alternative. Tukey and co-workers have elaborated on methods to parallel also higher-level methods like the analysis of variance by exploratory substitutes (see Hoaglin, Mosteller, and Tukey, 1991).

Generally, applications of mathematics enjoy a wide approval and they are accepted as an objective way to deal with a problem. This high status goes back to the optimization criteria that are applied to derive the solution within the level of mathematics (within the mathematical model). Usually it is neglected that the process of application comprises also the modelling as well as the evaluative and interpretative phase. The questions "How to derive a model for a problem situation from real world?", "How to judge the fit of models to the problem at hand?", and "How to interpret mathematical results for the problem within the context?" reflect the inherent subjectivity during these phases of applying mathematics. Within a naive paradigm of the natural sciences it is hardly achievable to address the resulting subjective features of the final results explicitly.

A popular argument in favour of any model – simplified or sophisticated – may be linked to a statement by Box who said "Essentially, all models are wrong, but some are useful." (Box and Draper, p. 424). This sounds like a cynical statement in two ways. First, models are not reality and thus they are essentially wrong so that the statement is a mere tautology. Second, if the modeller can insist on the model used, it is useful at least for him. And he might not expect objections if the model is highly complex as it would take time for an opponent to analyse in which respect the model covers relevant aspects of the real situation that is modelled. The vital question is what aspects of the problem situation are reflected by the model and its inherent assumptions and what aspects are left out. This shows simply how poor the discussion about model building usually is. The reference to Box is taken as a free-rider to use a model without such an analysis.

Exploratory data analysis may be linked to an *innovative* paradigm of modelling that opens ways to cope with the subjectivity of applications. The paradigm is already represented in applications of mathematics. The rapid progress of exploratory techniques in the model-building stages is a clear sign of a greater awareness that one has to address the questions directly. Multiple analyses are still used "silently" and the best-fitting result is presented, which is a clear indication that there will be a longer way to elaborate on the paradigm to make its ingredients better known and better received.

The proper place for EDA is in the model-building phase when a phenomenon has to be explored from diverse perspectives and context plays an essential role for judging which models might yield a suitable representation of relevant features of the situation in real world. A second part where exploratory analyses play a vital role is model-checking (independence, type of distribution, transformation of scales etc). Models derived by EDA should be used formally in a replication study rather than in the same study as this would cause serious methodological problems (multiple testing increases type I error and reduces statistical power). An improved practice of applying mathematics will be the prized reward of the endeavour to enrich modelling by the style of EDA. As Tukey (1980) already stated, "We need both exploratory and confirmatory [statistics]".

References

Barnett, V. (1982). *Comparative statistical inference*. Second edition. New York: Wiley.

Biehler, R. (1995). Probabilistic thinking, statistical reasoning, and the search for causes – Do we need a probabilistic revolution after we have taught data analysis? In Garfield, J. (ed.), *Research Papers from The Fourth International Conference on Teaching Statistics (ICoTS 4), Marrakech 1994* (18pp). Univ. of Minnesota: The International Study Group for Research on Learning Probability and Statistics.

Borovcnik, M. (1986). Zum Teilungsproblem. *Journal für Mathematik-Didaktik, 7*, 45-70.

Borovcnik, M. (1992). *Stochastik im Wechselspiel von Intuitionen und Mathematik*. Mannheim: Bibliographisches Institut.

Borovcnik, M. and Ossimitz, G. (1987). *Materialien zur Beschreibenden Statistik und Explorativen Datenanalyse*. Wien: Hölder-Pichler-Tempsky.

Box, G.E.P., and Draper, N.R., (1987). *Empirical model building and response surfaces*. New York: Wiley.

Cleveland, W. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Ehrenberg, A.S.C. (1981). *Data reduction*. New York: Wiley.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1991). *Fundamentals of exploratory analysis of variance*. New York: Wiley.

Jablonka, E. (1996). *Meta-Analyse von Zugängen zur Modellbildung und Folgerungen für den Unterricht*. Berlin: Transparent Verlag.

Tukey, W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Tukey, W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34 (1), 23-25.

Velleman, P.F. and Hoaglin, D.C. (1981). *The ABC's of EDA: Applications, basics, and computing of exploratory data analysis*. Boston: Duxbury Press.