F. Faulbaum, R. Haux und K.-H. Jöckel (Hrsg.) (1990). SoftStat '89
Fortschritte der Statistik-Software 2. Stuttgart: Gustav Fischer, 575 - 583

Software Tools for Statistical Data Analysis in Education and Teacher Training?

R. Biehler and W. Rach

Summary

This paper presents results and methods of a project situated in the frame of research into perspectives for an innovative education in probability and statistics at the secondary level of schools providing general education. Statistical software tools including PC–ISP, S, STATVIEW and DATADESK are compared and analyzed. Constitutive elements of design are reconceptualised in order to develop a model of data analysis tools which could guide the development and application of statistics tools in the classroom.

1. The background of the study

The background of the project forms a potential line of innovation which is to integrate statistical graphics, multivariate real data and simulation into teaching. This integration is to be considered with various general objectives in mind. New modes of working and knowing belonging to statistics are to be made accessible for teaching. The new technologies are to be used to encourage an interactive–flexible mode of working with data and methods, and a systemic holistic approach to problems instead of applying isolated individual methods. Besides, the new technological achievements are to be used to attain a better understanding of statistics and a richer, more motivating and application-oriented teaching: for instance graphics and simulation can be used to illustrate and visualize, multivariate real data shall be used to link theory and practice and to treat more interesting 'real life' applications. With regard to existing curricular subject matter, this would imply, among other things, to develop descriptive statistics further towards Exploratory Data Analysis (EDA), that is to place more emphasis on an exploration of data from various points of view, and on statistical graphics as well as on the interpretation of data. A second potential line of growth is to transform inference statistics towards a more comprehensive understanding of statistical modelling which includes the developing and checking of models with data and to interpret and discuss results in the broader context of the data. By developing competencies and attitudes concerning the use of graphics, concerning data bases, simulation, and modelling, an innovative stochastics education could make a more important contribution to the educational goals in schools providing general education.

Two projects realized at the Institut für Didaktik der Mathematik (IDM) in close cooperation are oriented towards elaborating a perspective of innovation. The project GRAPHDAS (Graphische Darstellungen zur Datenanalyse im Stochastikunterricht) studies, in cooperation

with a group of mathematics teachers, how ideas and techniques of Exploratory Data Analysis may be used to enrich mathematics instruction at the secondary levels I and II (cf. Biehler 1988, Biehler/Steinbring 1989). Along with this project, statistics software is used to allow for complex data analyses with elementary means of representation and operation. The fairly simple representations used are boxplots, scatter plots, histograms, stem-and-leaf - displays, enhanced scatter plots with numerical summaries such as lines or median curves. These are applied together with operations like identifying and localizing points and subsets, splitting the data into groups, reduction to subsets, zooming and overlaying graphics, definition of new variables.

In a second project SOMA (Software im Mathematikunterricht), we analyze and compare mathematical and statistical software tools. The SOMA project was more concerned with basic research into the design and use of software in various applicational contexts, and rather less concerned with direct recommendations of certain data analysis tools for the classroom. Our objective was to reconstruct constitutive elements and concepts of design and to reconceptualize them in order to develop a model of data analysis tools which could guide the development and application of statistics tools in the classroom.

Taking in account our intended educational applications, we placed the emphasis on tools which support an interactive graphical data analysis, and which, at first sight, seemed to represent certain design conceptions in a prototypical way. Within the "MacIntosh line", these were mainly the systems STATVIEW and DATADESK, besides MACSPIN and ELASTIC. For the systems aligned to command language resp. on interactive languages of statistics, we focussed on PC−ISP and S, but made allowance for GAUSS as well.

2. Models and dimensions for evaluating statistics software

2.1. Tools for data analysis in the conflict between user orientation and intended application

Statistics tools are influenced, among other things, by two factors: by the way the user is conceived of, and by the conception of the intended applications which are represented in the statistical methods available, in the problem types to be treated, and in the conception of statistics and of statistical activity. There is, as a rule, a relationship of conflict between an alignment to the user, and an alignment to applications.

From the point of view of intended applications, systems with well−designed extensible command languages (e.g. PC−ISP, S, GAUSS) are preferrable. To quote some of the advantages of these systems: locally and globally, they offer greater freedom of activity, a fact which makes for a more supple adaptation to different methodologies or individual styles of statistical work. After a protracted stage of learning, they convey an impression of economy and efficiency within the paths of familiar use. Without further ado, they permit extensions of language which can be used by the developer of methods, but also during exploratory data analyses for a redefinition, modification, refinement of methods. This usage conforms to the philosophy of EDA and to pedagogical concerns with regard to

developing(!) statistical methods in the classroom. Another advantage are the possibilities for "record keeping" ("script" or "monitor files" in ISP, "Audit files" in S). "Monitoring" one's work or that of others through using the system to record user commands (and software responses) is important in educational settings but, of course, with indubitable profit to any kind work pertaining to data analysis, in particular for EDA.

However, in case students use statistical working environments, these

- have to be simple enough and clearly arranged as a model and mental representation of the working environment;
- should remain transparent and consistent both mathematically and as a model of the working environment;
- should be easily learnt and offer itself to be handled intuitively according to the paradigm of the incidental user.

These requirements can be more easily met by visual languages than by command languages. Implementations based on the MacIntosh user interface like DATADESK, ELASTIC, STATVIEW show that it is this kind of user orientation which makes computer-aided flexible interactive data analysis accessible for school. With regard to application orientation, however, they are inferior to the above command language systems.

2.2. Reference model for data analysis tools

For a comparative analysis and evaluation of software, our starting point is the following model of reference, within we distinguish between "low level" and "high level" functions of statistical software. The high level funtctions are intimately related to the design of the user interface.

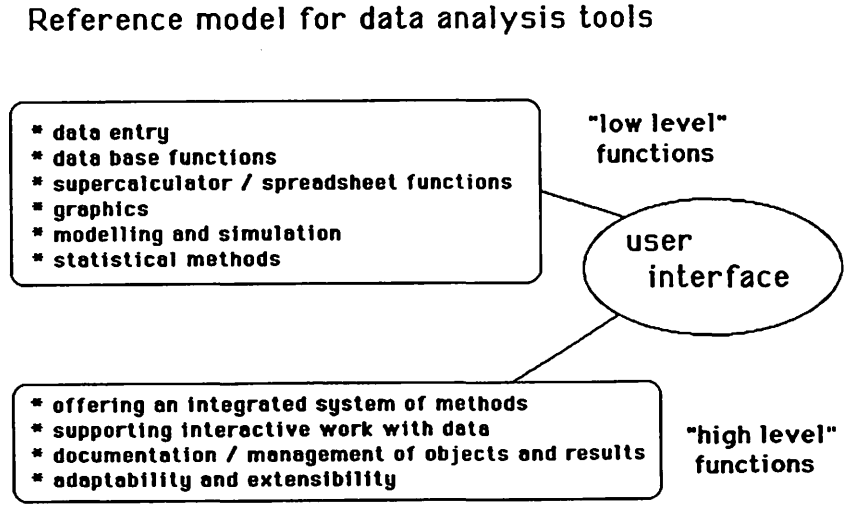

Fig. 1

We assume an intuitive understanding of these functions and should like to begin by discussing two functions which are both didactically relevant and related to our perspective of innovation. More detailed descriptions and comparative evaluations for the other software functions can be found in Biehler/Rach (1989).

The functional domain **"modelling and simulation"** designates the possibility of generating and processing random numbers for the most various probability distributions in order to simulate probability models, as well as the possibilities of handling deterministic and mixed models in relative independence of particular statistical methods. This, for instance, subsumes possibilities of defining mathematical functions and of comparing them to data without requiring that these emerge from a particular statistical procedure, e.g. the possiblity of overlaying straight lines by rule of thumb or eye into a set of data, or the possibility of graphically representing and comparing various binomial distributions. Such opportunities are important for developing new methods, for theoretical studies in statistics, and for those applications of data analysis, where the intention is to compare models and data in an unconventional way. In this area, there are many points of contact with traditional subject matter of mathematics education. Here, the opportunities offered by systems like DATA-DESK, STATVIEW and ELASTIC are limited by comparison.

By **"supercalculator/spreadsheet function"**, we understand the possibilities of carrying out general mathematical operations with data (vectors), for instance applying basic arithmetical operations to pairs of data vectors, and also the applicability of mathematical functions to transformation of data, up to the possibility of calculating with matrices. Besides transforming data and defining new variables (on the basis of old ones), this functional domain permits to execute procedures which are not yet contained in the ready made library of "statistical methods".

While professional statistics software is mainly judged according to the extent of its **library of statistical methods**, the requirements for secondary education are somewhat different. A great variety in the field of **elementary** statistical methods is required, enriched by opportunities to extend, modify and elaborate these methods, and a skilled integration of these statistical methods into a graphical, algebraic-numerical and data base environment. These featues are important for increasing the efficiency of a statistical working environment for elementary applications.

Now we shall treat conceptions of statistical graphics in more detail.

2.3. Conceptions of statistical graphics

We distinguish between the following aspects:
- (exploratory) working graphics
- multiple representation
- direct interaction

- multiple, linked representation
- presentation graphics

If graphs are to be used as a means of insight, the graphical functions of statistics software must meet several requirements. The basic idea is to obtain a working graph as directly as possible which then can be modified and enriched further.

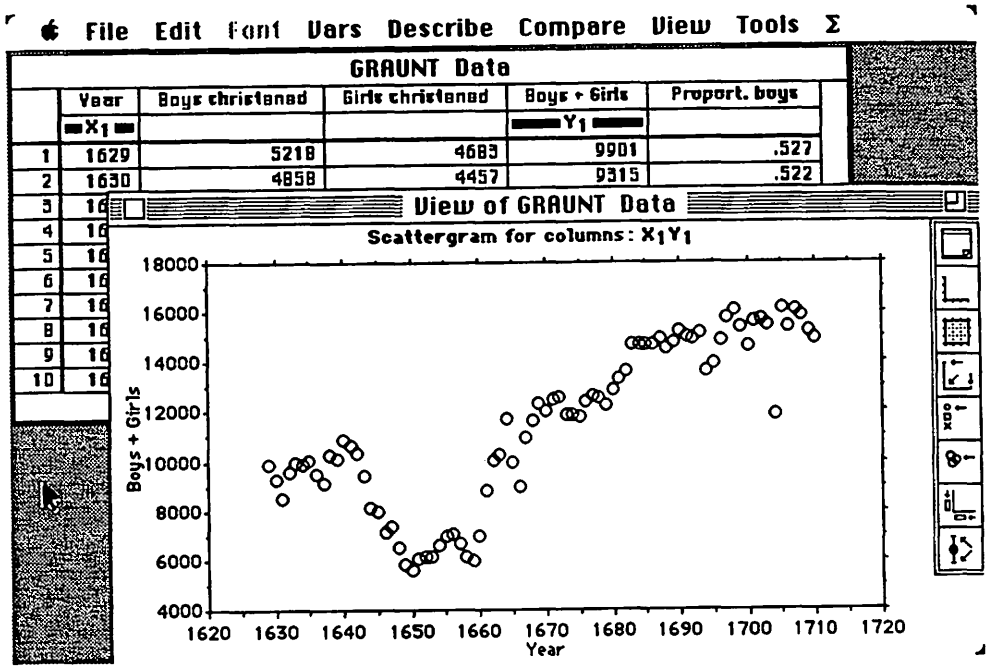

Fig. 2

Figure 2 shows a screendump from STATVIEW 512+. The working graph shown is created by selecting two variables and choosing the command *scatterplot*. A plot menu represented by icons offers various possibilities to vary and enrich the graph, for instance by rescaling, different plot symbols, new labeling, plotting "error bars" around the individual points, etc. The graph can also be quickly varied by selecting different variables from the table, or by confining the presentation to a certain subset of data. These modifications are immediately actualized in the graphics window. Other tools such as PC-ISP or S offer far broader opportunities for subsequent statistical analysis: plots already generated for instance, can be overlaid (with automatical adaption of the coordinate systems). It is also possible to add straight lines or curves to a plot, mark subsets with a special plot symbol, or use as plot symbols, instead of a simple point or asterisk, symbols with complex information (e.g. representation of further variables). Such a conception of **enrichment** or **overlay** can be helpful for elementary applications in school.

Multiple representations on the screen are very useful in educational and statistical respects. It is an important means of Exploratory Data Analysis which intends to explore a set of data from different points of view. The possibility of not only immediately actualizing graphics, but to put them into intermediate storage devices and to recall them for comparison with others is essential to counteract the volatility of screen graphics. Besides, there is the aspect that learning new representations is enhanced by relating them to familiar ones, as in relating a scatterplot of two variables to the respective histograms. Tables should be considered as semi-graphical means of representation. If tables can be flexibly manipulated, they have a considerable exploratory function. The possibility of multiple representation also includes that of presenting graphs and tables as well as texts containing numerical output (e.g. results of statistical tests) side by side on the screen and to relate them to one another. From an educational point of view it is important that students are able to handle tabular and graphical representation and their relationships competently and that graphics should be integrated into the treatment of other statistical methods.

Many novel ideas of ***direct interaction*** with statistical graphs as an important means to achieve insight, in particular in case of multivariate data, have emerged in connection with the PRIM concept, as it is implemented as in MACSPIN, PC-ISP and DATADESK 2.0. Besides interacting with rotating point clouds as in the PRIM conception, direct interaction is also used for other spatial transformations of data, and for brushing scatter plot matrices. Direct interaction with data graphs, similar to paper and pencil work avoids an intermediate complex symbolics. This feature constitutes an important advance for the use of data analysis tools in the classroom, and this not only for the analysis of multivariate data. Specific educational uses of direct interaction would be helpful, for instance, to create opportunities of adding summarizing lines to a scatter plot which then would not only serve to "embellish" the picture, but which permits a next step in data analysis, for instance, calculating residuals from the added line. If direct interaction with graphics is supplemented by ***multiple, linked representation,*** the chances of orientation in multivariate situations are considerable improved. Besides, multiple linked representations permit, from a educational point of view, new ways of supporting students in their flexible use of various modes of representation of mathematics and statistics (cf. Kaput 1986). Under the very requirement that data shall not remain abstract signs on the monitor for the students, it is important to be able to see which points in various diagrams (e.g. boxplot, table, histogram) "belong" to the same objects.

Presentation graphics as that form of graphics which serves to transmit the information at the analyzers disposal as poignantly as possible is a necessary supplement to (exploratory) working graphics. For presentation graphics, there are manifold software solutions: interaction on the pixel level combined with a set of geometrical primitives like in painting and drawing programs, configuration of graphics by parameters, icon-based graphics construction sets or special graphics programming languages. With regard to presentation graphics, there is justified criticism of the general level of quality, but there are hardly any consistent conceptions which might systematically guide the design of computer tools (cf., however, Cleveland 1985). A pertinent learning objective for school would be, for instance, to be able to reconstruct, criticize or improve graphics present in mass media, for instance

by means of a "graphics construction set", and to acquire elements of good graphical representation by application.

There are interesting approaches in the data analysis tools examined, but no convincing conception. DATADESK is the closest approximation to the requirement of an extended working graphics with multiple representation and direct interaction. STATVIEW offers much more possibilities for presentation and overlay graphics. Against that, the possibilities of modification, extension, graphics programming and documentation are far more developed in PC-ISP, and even more in S. It seems to be a difficult task to link working and presentation graphics.

2.4. High-level functions in user interfaces

Hitherto applying methods in isolation is practiced in statistical education. This reductionistic treatment is also reflected in the widely available, merely additive collections of statistical algorithms. In view of this, it is important to stress the value of a software tool which ***integrates*** the various aspects of statistical activity and statistical methods *into a unified system.* It is essential for multiple and iterative analyses that the output of one procedure can be used again as input for others – be it graphical or numerical methods. This aspect has been most consistently realized in PC-ISP and S. This is attractive for classroom application, for the very reason that its openess permits combinations and local "micro-worlds" of methods which are useful under the specific conditions of learning in school, but do not necessarily correspond to the conventions and requirements of statistical practice. But with systems like S and PC-ISP, there is still the problem that the user himself must (mentally) organize such modules of statistical activity, an activity for which the students may fall back on prestructurizations provided by the teacher, and orientate themselves according to these.

Menu-supported systems, by contrast, reflect structures and sequences of possible statistical activities. A sensible requirement of menu systems is to not only give access to a collection of procedures, but to use them to structure the statistical activity and the "image of statistics". This seems to be most successfully solved in DATADESK. Nevertheless, for classroom applications a possibility for configuration and designing specific working environments for students with regard to certain statistical fields of application and users' previous knowledge were most desirable. Specifical educational software tools like ELASTIC provide a closed working environment that was designed from an eduational point of view, but, however, they are apt to be criticized if their basic pedagogical assumptions and points of emphasis are not accepted.

Statistics tools offer different ways of ***documenting and administrating*** (tentative) results (plots, textual-numerical summaries, transformed data), objects (new variables and functions, macros) and processes of data analysis as the sequence of the commands given to the system (scriptfiles in PC-ISP, Audit files in S). This is crucial for organizing work with a system if interactive work is being done. This is all the more true if different persons are to work on the same problem and in teamwork. Such is the situation in a classroom where the teacher

prepares certain aspects of the data and of the graphics, and where certain problems are evolved across several lessons and under participation of different groups of learners. An essential basic idea for that seems to be a flexible conception of the *workspace* as implemented in slighly different versions in DATADESK, PC-ISP and S. The workspace possibilities in STATVIEW are rather limited, as only data tables can be stored as objects internal to the system.

Possibilities of "record keeping" play an important role in the frame of professional discussions about data analysis tools (cf. Huber/Huber–Buser 1988). Scriptfiles as in PC-ISP can take on different functions in a classroom context: preparing demonstrations by the teacher or by learner groups as well as documenting problem solving in order to reflect on the activity or to diagnose learning difficulties. Nevertheless, it must be stressed that this basic idea has not been realized in the present systems in a way which would make them easy to handle for teachers and students.

Possibilities of ***adaptation and extension*** have been frequently mentioned. We should like to summarize their significance within a classroom context in that they support the

- opportunity to configurate student working environments
- developing new methods in the classroom
- use of a statistics tool as a means of visualisation

Beyond that, we consider configurating and extending given systems according to individual needs to be an important learning experience with regard to new technologies. Adaptibility and extensibility is, as a rule, available only at a high cost in terms of user qualification. In the case of the command-oriented interactive languages, we consider opportunities for user-defined functions as realized in the new S-system an essential simplification. Until adaptibility and extensibility, however, can be actually used by average students, some important advances into the direction of iconic programming are still to be made.

3. Perspectives

The perspective of innovation noted at the outset requires a reorientation of teacher education and training in the field of probability and statistics. The educational offers for teachers must include experience with statistical data analysis tools, both with interactive command languages like PC-ISP, S, GAUSS, and with data analysis tools based on visual languages like DATADESK and STATVIEW.

The latter would seem to be preferable for secondary education as their user interfaces are better adapted to a student user. The "MacIntosh"–like approach to user interfaces seems to be flexible enough to be able to integrate conceptions from other data analysis tools which have been proved fruitful, as e.g. DATADESK shows in its successive versions. Classroom tests still have to be waited for before comparative statements can be made about the uses of DATADESK, STATVIEW or ELASTIC in school.

Such pilot tests in school can serve to initiate the necessary research into the effectivity of conceptions for data analysis tools which have been recognized to be educationally pertinent. Generally speaking, however, an independent development of a data analysis tool better suited to the user groups and application concerns in school must be taken into consideration if professional data analysis tools fail to meet these requirements still better.

4. References

Becker, R.A./Chambers, J.M./Wilks, A.R. (1988). The New S Language. Pacific Grove, CAL: Wadsworth & Brooks.

Biehler, R. (1988). Changing Conceptions of Statistics: A Problem Area for Teacher Training. IDM Occational Paper 114. Bielefeld. To be published in: Proceedings of the ISI Round Table Conference, Törökbalint, Ungarn, July.

Biehler, R./Rach, W. (1989). Softwaretools zur Statistik und Datenanalyse aus softwareergonomischer und didaktischer Sicht. Projektabschlußbericht. Bielefeld/Soest. To be published in: Landesinstitut für Schule und Weiterbildung (Hrsg.): Neue Medien im Unterricht: Mathematik 1989/90. Soest: Soester Verlagskontor.

Biehler, R./Steinbring, H. (1989). Graphische Darstellungen zur Datenanalyse im Stochastikunterricht. Materialien aus dem GRAPHDAS–Projekt. Universität Bielefeld: IDM.

Cleveland, W.S. (1985). The Elements of Graphing Data. Monterey, CAL: Wadsworth.

Huber, P.J./Huber–Buser, E.H. (1988). ISP: Why a Command Language? In: Faulbaul, F./Mehlinger, H.–M. (Hrsg.): Fortschritte der Statistik–Software 1. Stuttgart – New York: Gustav Fischer, S. 349 – 360.

Kaput, J. (1986). Information Technology and Mathematics: Opening New Representational Windows. In: Journal of Mathematical Behavior, 5, p. 187 – 207.

5. Software

DATADESK: 1986. Data Description, Inc., P.O. Box 4555, Ithaca, N.Y. 14852.

DATADESK Professional 2.0: 1988. Odesta Corporation, Inc., 4084 Commercial Avenue, Northbrook, Ill. 60062.

ELASTIC: 1988. Pre–release version. BBN Laboratories, 10 Moulton Street, Cambridge, MA 02238.

GAUSS: 1984, 1985, 1986, 1987. Aptech Systems, Inc., P.O. Box 6487, Kent, WA 98064.

MACSPIN: 1986. D 2 Software, Inc., 3001 North Lamar Boulevard, Suite 110. Austin TX 78705.

PC–ISP: 1988. Artemis Systems, Inc., 125 Berry Corner Lane, Arlisle, MA 01741.

S (see Becker/Chambers/Wilks 1988)

STATVIEW 512+: 1986. BrainPower, Inc., 24009 Ventura Boulevard, Calabasas, CA 91302.

STATVIEW SE+ Graphics: 1988. Abacus Concepts, Inc., 1984 Bonita Avenue, Berkeley, CA 94704.