Historical Software Issue 9: Statistical Software in Historical Social Research

Thaller, Manfred

software

Historical Software Section[+]

Issue 9: Statistical Software in Historical-Social Research

As announced in the last issue of this section, we will deal now with the new versions of the "big" statistical packages. More explicitly we will designate one issue to each of the following: SAS, SPSS-X, BMDP, P-STAT, OSIRIS, and GENSTAT/GLIM.(1)

Originally it was intended to cover already today the first of the packages, but during writing it became more and more clear, that one should first lay down the principles of our descriptions and evaluations. One should even explain why we publish such a series in a newsletter addressing itself to a community of researchers which will only rarely have enough influence upon the computing center of their university to change its decisions which software to buy. Additionally: the smaller programs of earlier sections dealt with problems which are so specific, that it is relatively easy for a researcher to decide if they are useful for his or her research, while the big systems we are talking about now offer so many features, that it is rather hard to imagine a research program even in the more mathematically oriented sciences, let alone in historical-social research, which will actually use all of them. So for the majority of the researchers the simple truth is, that you have to use whatever your local computing centre provides as "the" big package.

When despite all this we are going to devote so much space to detailed discussions of software products many of our readers will never have a chance to use, our reasons are these:

- the developments of the last few years have seen a large number of increases in the power of the statistical packages, which in the opinion of this reviewer open up completely new possibilities for the application of statistical computing. Much of these developments seem to have gone unnoticed. Just to give the most trivial example: SPSS features since quite some time a MULT RESPONSE procedure, which in many cases provides a pretty useful solution to old problems like "Many of the people in my dataset are supposed to have more than one occupation; what shall I do about it?" Or, less trivial, but still constrained to SPSS: using the REPORT procedure of this package, one can quite easily solve a number of problems which are typical for microanalytical studies, such as analysing data statistically, which shall at the same time be processed to select and display certain subgroups with personal names printed in uncoded form.
- some of the packages offer so large and excellently documented possibilities which are going beyond whats been considered

+ Address all communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge-Weg 11, D-3400 Göttingen.

easily available statistical computing, that their description should provide a whole number of ideas which can be turned into practical results with other programs as well.
- we think, that all the new features taken together should be reason enough to raise the question "What do I expect from the computer?" again.

To answer particularly this last question, we will devote this section to a survey of the overall trends in the developments of the big packages and to an explanation why, in our opinion, some of them should be looked at very carefully. A few of the new possibilities might actually influence the research design of many projects. As we are looking into the future, we will in general also include features, which are not distributed right now, but will be available within the next 18 months (at least with the more prominent brands of mainframes).

Major new developments have during the last years occurred in a number of areas. We will concentrate upon six questions when evaluating the individual packages: (a) Do they contain statistical techniques which are particularly suited to historical-social analysis? (b) Which didactic concept for teaching statistics do the support? (c) How much power has the user gained for transformations of his or her data? (d) How complex are the structures of files supported? (e) Do the packages have string handling and/or report writing capabilities? (f) What is provided for visual display?

a. Statistical Methods Available

Most of the new versions of the big packages offer facilities for the multivariate analysis of variables on the nominal scale. While up to now there was a kind of silent agreement, that loglinear analysis, probit analysis or the other ways of dealing with data for which the classical approaches of multivariate analysis - say factor analysis - could not be applied by someone who's computeracy was restricted to big packages, this is not true anymore. BMDP actually provides (partially) suitable programs already since a number of years.
For time series analysis similiar considerations hold true. Even SPSS - not exactly the most actively expanding package in this area - provides a BOX-JENKINS procedure since release 9; other packages provide a whole number of methods.
The important point is not, that one or the other of the new methods offered is as such of uttermost importance: what is important in our opinion is, that the old story "I would be fascinated to use advanced statistical methodology, but the methods which are available within packages designed for the non-computer-expert require data I simple don't have, while the programs implementing such approaches as my data could support I cannot handle" simply is obsolete.
While relevant new methodologies have been included in all packages, one should distinguish between certain classes of them. SPSS seems to be most definitely dedicated to successes in the business world.

So regarding the availability of new statistical methods. SPSS will rank very low compared to almost every other package. Just leaf through the late update manuals: scarcely any example for the new procedures is taken from social science any more, most of them explain how to deal with your companies accounting. (SURVIVAL has been included for the benefit of medicine, which, as "big science" in the opinion of this reviewer is in all questions of funding much more like a business operation than a social science.)
P-STAT and SAS are interested in the business with business quite unmistakably as well; at least one of them distributed a leaflet a short time ago on "How To Increase Your Companies Profit with the Help of ...". Still reading the documentation of both packages one has more the impression that the improvement of methodology relevant for social science is still a pressing concern for them.

SAS will probably score highest on "second thoughts". While I could not find any hint at it in SAS systems literature yet, there are even rumours, that the old GENERAL INQUIRER philosophy of computer supported content analysis is going to be reimplemented with an explicit link towards SAS in mind.
BMDP, OSIRIS, and GENSTAT/GLIM are still securely based in scientific applications and contain a very large spectre of methods offered, though BMDP and GENSTAT/GLIM never were particularly intended for social science applications.

b. Learning and Teaching

All packages have features that are intended to support one in learning how to use them; there exists a large number of different opinions, though, what an appropriate strategy might be. On the one extreme we still have SPSS, where in most cases the user is simply expected to fill in the names of variables into the various slots provided for them, without ever coming to know, which kind of algorithm is actually used; the opposite extreme would be GLIM, where the statistical formulae are (with a slightly adapted notation of course) more or less the kind of command a user in expected to give.
Overstating just slightly one might say, that SPSS would be the dream of an historian with little numeracy but definitely the nightmare of a statistician; GLIM the joy of any statistician but almost being designed to frighten the ordinary beginner away from the computer (at least if he or she is an historian).
Whats the role of the computer if we want to propagate and teach quantification in historical social research? Shall it be an instrument with which we can prove, that results actually can be produced within a rather short time and without overly endulging in the details of statistical derivations? Or shall it be the means, by which one can actually understand how a statistical argument is put into a result, without being bored to death by computing it all with paper and pencil?
The answer to this question will of course have important implications for the didactic concept of any introductory course into

quantitative methods and for the selection between available software. This reviewer admits, that he himself doubts the wisdom of an overly detailed statistical training for historical-social researchers and that in his opinion a program package should take care of the statistical part of the job as much as possible, leaving the historian free to worry about the conceptual difficulties and the problem of sources which may become obscured by an overindulgence into "whats really going on". On the other hand one has to admit, that that is exactly the approach, that leads to the application of misunderstood methods - and is much more dangerous, when one talks about applying a sophisticated linear model, than when one produces the next crosstabulation.(2)
What a system designer should do, would ideally be to provide both possibilities: (a) ready made commands, where one can produce results during the production phase of one's work without bothering much over the details of the equations somewhere in the background as long as one still understands what the method is doing in principle and (b) features in the command language which allow the student in an introductory course to increase the transparency of whats going on, by giving explicit commands how to undertake a given statistical computation. In principle this is possible with all of the packages already now of course; though I doubt if it is good teaching practise to explain to a beginner how one computes just a correlation coefficient by the pre SPSS-X COMPUTE command. In practise all the packages in question seem to have advanced towards this ideal during the last years - though the differences may be largest here and will be the most difficult part of the individual reviews to write.

c. Enhanced Data Handling Capabilities

Important developments have in all the packages been made with regard to data modification, transformation or handling. Particularly the uncomfortable situation that one has to start a computation to find out about the mean of a variable (or one of the more sophisticated parameters of its distribution), print it out and then feed it in again to use it as part of the computation resulting in an appropriate index has been improved upon considerably.
This is in our opinion one of the most important fields of development for three reasons:

- all approaches advocating to put into a missing variable a number derived from the parameters of the distribution concerned will be much more easy to implement in future. (Just think about the simple putting of the mean into missing variables: when you have to write the numeric value of it into an applications program, you have to compute it again and over again every time you update, add to or correct your data - what in practical work means that you will probably never try to make this substitution before you are in the very last stages of a project. As soon as the mean is a property of the variable you can use in a data transformation statement, you can rerun it arbitrarily often without the error-prone typing of figures - and start from your very first analysis with your data in much better shape.)

- data preparation should become considerably easier.
- it becomes extremely easy to use more complicated indicators than that far, which are standardised by the parameters of one or more distributions.

The ability to handle such statistical properties of a variable is important but not all encompassing. Another development in data transformation which seems as important for the medium range, those features allowing transformations on data objects which are beyond the simple "a case out of variables" logic. Such developments have provided the power to trade information between cases (a simple, old and relatively well known example: the LAG command of SPSS) and - with a very large potential for the future - also the possibility to work with matrices derived from ones original data, sometimes giving the user at his hands the whole power of matrix algebra.

d. More Complex File Handling

The that far most oftenly lamented shortcoming of classical statistical packages has been the necessity to put data into a rectangular file. How this restriction can be overcome will probably be the shibboleth for all statistical packages that are discussed by historical-social researchers in future; actually it has been removed (but to a very different extent and with very different approaches) by all the major packages - at least in the releases announced to be available shortly. That we put this development only at the fourth place is intentionally nevertheless: the points we raised that far are of immediate relevancy for all research projects, this point and the following ones more probably only for projects which are still in their design stages.

To overcome the rectangular file and its inherent restrictions there are in principle three ways.

- to enter particular routines (possibly to be hidden from the user) which take care of file designs allowing for (a) variable length records, (b) files consisting of a hierarchy of record types, and (c) files containing networks of records of more than one type.
- to provide for variables which work as "pointers", i.e. which can be used to tell an analytical routine to select as the next data portion all or a part of the variables of a "case" which is not the consecutively next one in the file.
- to cling still to rectangular files, but allow the user to process a more or less arbitrary number of them concurrently and direct the process by which relationships between the records in them are connected.

It should be obvious that out of these basic solutions there are very many ways how to create a more complicated data design and that there are many differences between the effort required of the user in putting such a solution to practical work.

Quite besides the question how efficient a system is at allowing more complicated file structures, there remains the other one, how a set of files can be interrelated ("merged", "linked" or whatever). Let us not be mistaken: when we speak of "record linkage" in sta-

tistical programming systems, we are discussing a completely different problem than the one known as "nominative record linkage" in the discussions of software developed specifically for historical analysis. When used in the connection with social science software it is usually assumed, that there exist variables which by a finite number of code values allow a clear deterministic relationship between records in one file and those in another one. In many cases there exists the additional restriction, that every case in file "a" has to be linked to exactly one case in file "b".
Taking these and other restrictions into account, however, some of the new versions of the packages quoted contain quite a few refinements indeed: P-STAT allows e.g. a merging of two files of different sizes with a partially overlapping set of variables, where:
- two cases are considered for merging if the variables having the same name in both files are identical,
- the new file consists of cases containing all the variables appearing in any of the two original files,
- cases in "a" where no case in "b" to be used for linking could be found, get the additional variables all set to missing.

e. String Handling in Statistical Packages

We mentioned already, that a number of the packages we discuss here compete quite clearly for the software market in business applications. As one of the earliest "business only" programming languages is known as "Report Program Generator" (RPG) it is less than surprising, that a package seeking customers from business has to provide for Report Generation. Or, as this term may need explanation for someone who was that far exposed to statistical computing only: such a program takes care that
- certain entities contained in a file can be listed easily in a well readable way,
- their identifications (names of individuals or companies or cities or ...) are listed in plain text,
- if more than one record belongs to a given entity, their contents are suitably aggregated before being printed (only the name of the head of a given household is printed, plus the number of children, plus their mean age, plus the age of the oldest child, plus the age of the youngest one, plus the income of the boarder who has the highest among all boarders in this family, plus ...)
- after certain subsets of entities summary statistics are printed (at the end of each "street" the number of families and people plus the mean income of families listed for that street are displayed).

The ability to arrange easily for reports like that, could become a quite interesting feature for all micro studies and, local history being as popular among younger historians as it is right now, therefore a quite important feature for all projects which deal not with either a micro study or the statistical analysis of an aggregate data set, but try to combine both approaches within one research project. (As a forthcoming conference of philosophers on the impact of theories on historical research already thinks it

necessary to devote one session to the question "if there is any inbuilt animosity against theory in the new local history?", such studies should be a very recommendable thing to do, by the way.) While being interesting on its own, report writing is restricted in importance to a relatively small number of research interests. There exists, however, a windfall profit of its introduction into the new versions of statistical packages which in the long run might become more important than the reason for its occurrence: as report writing necessarily requires certain string handling capabilities, some of the recent developments incorporate basic string handling into statistical packages. While it would be absurd to think, that in the foreseeable future these introductions might make the packages powerful enough to process textoriented data, they should allow for many applications, where a small number of textual variables shall be analysed in the context of a large amount of numerical codes and other figures.

f. Improved Display Facilities

One of the main developments in EDP during the last years has been the spreading of plotting devices; it is not very surprising therefore, that many statistical packages try to provide for the plotting of distributions of and relationships between variables or derivates of variables (such as the distribution of a set of cases in a n-dimensional space defined by the first n factors from a previous factor analysis).
Indeed the relevant features of particularly the more business oriented packages provide surprisingly beautiful plots gained by really easy to use commands. As the practicability of such programs is dependent of the hardware available at a given computing centre, this seems to the reviewer a very interesting development, but one where the full impact is going to be felt only after a couple of years. (Even if your computing centre has "ample" plotting devices: in some cases already the introduction of a plotting package like DISSPLA, where one has still to have at least a rudimentary knowledge of programming languages, let skyrocket the plotting done to ten times the previous amount within less than two months. My guess is, that as soon as really sophisticated plotting options become available within the command language of a statistical package already well known, at most computing centres the plotting devices will be so hopelessly overburdened that the single user can reap the fruits only after the time it takes his or her computing centre to acquire the funds for additional hardware).
Of more immediate concern is the question, how far the routines for the display of data contain line printer options as well, i.e., how far the packages are able to produce a (necessarily less pretty) copy of a visualisation on the usual printing devices which are available right now.
The final point brings us back to where we started: among the methodical developments of the last ten years was the introduction of Exploratory Data Analysis (EDA), as a set of techniques and methods how one can display variables and their distributions in a

a way as to enable a user to get a better understanding of their properties, than by looking at the not always particularly expressive numeric parameters. This reviewer still has a certain scepticism about what may be seen as another fashion in science and still thinks, that ultimately a close scrutiny of more conventional statistical results provides a better insight into a dataset than even the most fancily baptized diagrams. But, this granted, one can scarcely doubt, that the various tools for the display of data developed by EDA provide indeed a very valuable arsenal for the early stages of research when there are many variables about which is not known very much and one has reason to suspect, that they are distributed in any way but normal - as this is just a description of what data in historical-social research are usually like, the availability of the various EDA techniques in the context of some of the big packages should be a major improvement of the software situation of our field.

FOOTNOTES

1 SPSS-X and P-STAT version 8 are described in manuals which are not yet available. As some of the other packages are to be updated soon as well, this software section does not yet contain any references to any manual. The following descriptions are based upon such manuals as are available now, plus the announcements made by the developing institutions in their various newsletters and sales representations plus information as available from statistical software newsletters.

2 What I would like to have for teaching quantification would be a system, where you could declare properties of a variable like the level of scaling the same way you do right now with MISSING VALUES, get automatically only those statistical coefficients which are applicable and at least a warning when you ask for something like the mean of a variable on a nominal scale.