Historical Software Issue 4: Calculs et Analyses Sur Ordinateur Appliques aux Reconstitutions (CASOAR)

Thaller, Manfred

software

HISTORICAL SOFTWARE SECTION[+]

Historical Demography might easily qualify as that field in historical research, where the application of both, quantitative methodology and techniques supported by computers, have the strongest tradition. So it is not particularly surprising, that this subject lead to the development of some of the few program packages, which have been designed from the very beginning explicitly to support historical research. One of them - particularly outstanding for the excellent documentation, so often missing with programs originating in the research community - is the subject of this software section: CASOAR (an acronym for Calculs et Analyses Sur Ordinateur Appliques aux Reconstitutions), developed by Michael Hainsworth and Jean-Pierre Bardet at the Laboratoire de Démographie Historique in Paris.(1) CASOAR is essentially a tabulation program, that is, it should be ranged among the software available for descriptive and/or explorative statistics, not among the systems supporting analytical statistical methodology. This might be considered a shortcoming of the system: as it stands, it cannot be understood as a tool that will support the more sophisticated approaches towards historical demography, as e.g. the use of mathematical simulation (2) for the analysis of demographic developments. This is a shortcoming, indeed. But quite besides the point that CASOAR 2, implementing the very analytical tools missing right now, is announced to be on the drawing board, even as the package stands at present there are very significant virtues which can be attributed to it in compensation. Doubtless one can do tabulations with many of the existing multipurpose packages for statistics: the first great virtue of CASOAR, though, lies in the fact it can be considered as a working implementation of a consistent approach towards historical demography, as defined by the works of Louis Henry.(3)
This virtue - representing not just some tabulations one comes across by chance, but a consistent and documented approach - leads to another one, which may not be so immediately apparent. CASOAR is extremely rigid with respect to the way a tabulation can be done. There are practically no options to influence the outlook of particular results (we will discuss this from a technical point somewhat later). This unfriendly behaviour has a very welcome side effect, though: any research done with the help of the package becomes immediately and completely comparable to any other one undertaken with it. So while in a given research project one will probably always have some peculiarities which CASOAR cannot handle, and will want to get some results it is better to use some more general program for - say SPSS - using

---

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

CASOAR guarantees, that within a very short time one can acquire all the standard tabulations to compare ones "own" population with other ones, about which the usual descriptive statistics have been published. This possibility to acquire standard tabulations for a very wide range of problems is probably demonstrated best, if we enumerate some of the specialised routines CASOAR provides: "Conceptions prénuptiales et intervalle entre marriage et premiere naissance", "Mortalité maternelle", "Signatures au marriage du mari ou de la femme" or even things like the "Distribution mensuelle des premieres naissances comparées à celles des naissances suivantes".
This high degree of specialisation requires of course, that one makes very strong assumptions about the data available: indeed CASOAR as a whole is designed around a rather specialised format for the input of family reconstitutions, which in themself are not an overly common kind of source.
CASOAR is therefore a system for the specialist. So, when it is described here, we adress ourself first of all to people interested in historical demography and working with family reconstitutions. Quite besides this immediate concern, the package seems for the author to present some much more general hints how programs to be used outside of a given historical project could and should be written. But before we touch this question once more a more systematic description of CASOAR is appropriate.
As distributed by the Laboratoire de Démographie Historique, CASOAR consists of a collection of approximately 70 or 80 programs (depending on how one counts routines that facilitate the operation of the other ones at a given installation). All those programs are "main programs", i.e., to execute them you have to call them from the level of the operating system like you would call one of the big statistical packages. The system has no control language whatsoever: once you have called a program, it will do exactly its job - without giving you a chance to influence its execution.
The programs can be divided into two classes: preparatory routines and utilities on the one hand, tabulation routines on the other.
Four routines are provided for formatting: data on reconstituted families have to be entered in a mixture of freefield format and tag-content logic, which, before being processed further, is transformed into a kind of "systemfile" which contains mainly fixed format records, adding a number of indicators (e.g. number of children in the family) computed during the formatting process. The "systemfiles" can be accessed and modified by the local editor (CASOAR in itself provides no recoding capabilities).
A set of seven routines is provided, which check the data for logical errors - people having died before being born and stuff like that - and eight routines are available to select subsets of data according to user defined criteria. All these preparatory operations are supported by a number of copying routines which are of course very specific to the local computing centre and of scarce value at other installations.
Tabulation - or rather analysis - then starts with 5 routines for relatively general listing and sorting operations, which provide mainly legibly formatted copies of the original data.
Various aspects of nuptiality can be visualized by 10 routines each of which provides one type of tabulation: to restrict such a tabulation - and the following ones mentioned - to, say, couples out of a given region, you have first to call one of the selecting routines mentioned before and produce a smaller "systemfile", containing only

couples which fulfill the desired criterion, and then ask the tabulating routine to do its job on that smaller file.
30 types of tabulation are provided for various aspects of fertility, covering a wide range, as e.g. the temporal distribution of still-births, aspects of premarital conceptions, intervalls between births controlling for a number of variables and so forth.
Finally, 9 types of tabulation are available for the analysis of several aspects of mortality.
All tabulations have been designed to provide a maximum of information within the space of a printed page; this aim enforced relatively short comments, labeling information therefore appearing in abbreviated form (in French). An example is reproduced on the next page.
When we say "available" we have to say something about the practical difficulties one has to expect when intending to use the package.(4) CASOAR - originating from an IBM - is written completely in PL/1, taking some pain to avoid almost all of the refinements which can hamper the compatibility of programs in that language. As a result the system is relatively easy to implement, a tendency which is further increased by the extremely simple structure of the package. As there is absolutely no common subroutine library, one can convert the single routines as the need for them arises.
Attempts to implement the programs were undertaken at two systems: an UNIVAC of the 1100 series, where only very small peculiarities had to be smoothed over (differences in the handling of precision, changes in the declarations of files and the substitution of machine specific calls to SORT routines). This bright picture cannot be upheld for the CYBER series of CDC. As it turned out to be impossible to read the test data from tape, we did not go beyond the compilation of the programs and, as the PL/1 compiler available at CDC machines is not able to handle even some of the most simple conveniences of PL/1 (which in theory should be part of the very basic definition of the language, like the assignment of scalar values to non-scalar targets), every CASOAR routine would need at least half an hour of careful scanning for different standards, which sums up to quite a bit of time, when one talks about a package of 70 routines.
Once the package is implemented, the results come very cheap: typical tabulations performed on a set of testdata consisting of 1000 families (9960 lines of input) where performed on an UNIVAC 1100/83 in approximately 4-5 seconds of CPU usage and 25.5 seconds of I/O time (plus some requirements of the operating system). The "size needed" can scarcely be given reliably, as PL/1 uses a very large runtime library and the figures will change drastically even on machines of the same manufacturer with different organisations of that library. Anyway, with the exception of some calls to sort routines, all programs should be able to execute in about 20 K (4-byte) words.
So CASOAR is easy to implement on machines which have a reasonable PL/1 compiler and it will run rather fast. Still, when you want to introduce a package like that in the context of a given research project you may encounter quite some resistance from the side of the technical staff available. The arguments that can be raised against CASOAR can often be heard in connection with the accesibility of other programs as well, so they shall be discussed here not only because of CASOAR, but in the interest of improved conditions for an exchange of historical software in general. Here the main points:

1. "Programs like the ones CASOAR consists of are very simple; im-

BXGT CASOAR.HE53

INTERVALLES ENTRE VEUVAGE ET REMARIAGE DES - HOMMES

DUREE DU VEUVAGE	AGE AU VEUVAGE											
	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-98	IND.	TOTAL
0-2M	0	0	0	0	1	0	0	1	0	0	0	2
3-5M	0	0	0	3	0	0	1	0	0	0	0	4
6-11M	0	0	0	1	0	1	0	0	0	0	0	2
1 AN	0	1	0	1	1	0	0	0	0	0	1	4
2 ANS	0	0	0	0	1	1	0	0	0	0	0	2
3 ANS	0	0	0	0	0	0	0	0	0	0	0	0
4 ANS	0	0	0	0	0	0	0	0	0	0	0	0
5 ANS	0	0	0	0	0	0	0	0	0	0	0	0
6 ANS	0	1	0	0	0	0	0	0	0	0	0	1
7 ANS	0	0	0	0	0	0	0	0	0	0	0	0
8 ANS	0	0	0	0	0	0	0	0	0	0	0	0
9 ANS	0	0	0	0	0	0	0	0	0	0	0	0
10 & +	0	0	0	0	0	0	0	0	1	0	0	1
TOTAUX	0	2	0	5	3	2	1	1	1	0	1	16
NONREM	0	0	3	5	2	8	8	8	8	22	6	68
DESIND	0	0	0	0	0	0	0	0	0	0	0	0
INTIND	0	0	0	0	0	0	0	0	0	0	0	0

INTERVALLES ENTRE VEUVAGE ET REMARIAGE DES - FEMMES

DUREE DU VEUVAGE	AGE AU VEUVAGE											
	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-98	IND.	TOTAL
0-2M	0	0	0	0	0	0	0	0	0	0	0	0
3-5M	0	0	0	0	0	0	0	0	0	0	0	0
6-11M	0	0	0	0	0	0	0	0	0	0	0	0
1 AN	0	0	0	0	0	0	0	0	0	0	0	0
2 ANS	0	0	0	0	0	0	0	0	0	0	0	0
3 ANS	0	0	0	0	0	0	0	0	0	0	0	0
4 ANS	0	0	0	0	0	0	0	0	0	0	0	0
5 ANS	0	0	0	0	0	0	0	0	0	0	0	0
6 ANS	0	0	0	0	0	0	0	0	0	0	0	0
7 ANS	0	0	0	0	0	0	0	0	0	0	0	0
8 ANS	0	0	0	0	0	0	0	0	0	0	0	0
9 ANS	0	0	0	0	0	0	0	0	0	0	0	0
10 & +	0	0	0	0	0	0	0	0	0	0	0	0
TOTAUX	0	0	0	0	0	0	0	0	0	0	0	0
NONREM	0	1	3	2	1	0	1	0	1	0	2	11
DESIND	0	0	0	0	0	0	0	0	0	0	0	0
INTIND	0	0	0	0	0	0	0	0	0	0	0	0

plementing a program written for another project will need almost as much effort as writing it anew; when you write a new program you have a much better chance of getting exactly what you want; lets create our own system."
You may get most of the calculations performed by CASOAR by cleverly playing around with the data modification cards and some of the options of SPSS, nothing to say about higher programming languages. Just: you may write your first program in less time than you need to convert the first CASOAR routine. But when you have found out what the local difficulties are the remainder will come incomparably quicker. And: while it is easy to write "just a small program of 100 lines" (and the majority of CASOAR's routines are such) it may mean quite an effort to write 70 of them, designing at the same time data representations that are sensible for all routines concerned, take care of possible logical errors of your input data and so on. Furthermore CASOAR is definitely beyond the documentations of other packages that have grown out of particular projects: a set of testdata is available together with the routines, which in most cases will reproduce the tabulations given in the manual while in the remainder it is at least sufficient to allow a quick estimation, if your implementation was successful - and how long it may take to make sure that a particular program written from scratch produces always the correct results, is a story usually not covered by the "2 hours to write such a small program".

2. "Packages like CASOAR are so inflexible that to make the slightest changes you have to dig deep into them, understanding exactly a strange program structure."
CASOAR is a very specialised system: it is enough for your data to contain events before the thirty years war or in the early twentieth century to require changes in the PL/1 source code of some of the routines. Still, in such cases the package will support very well everybody who understands what is meant by the virtue of getting 90 percent of the results by 10 percent of the effort. If, during a project using family reconstitutions, you have a system like this available, you are able from the very beginning to analyze your data as they become machine readable - while the EDP expert in your project can all the time develop those special routines needed for your application. By that you avoid the blindness hampering all to many projects where everybody enters data for months without the slightest idea what will come out, when the local wizzard finally has finished the development of an ad hoc system.

3. "Packages like CASOAR are so specialised that you have to rely completely upon their design, being scarcely able to add additional features or using the services offered by other programs."
The package can be used in both directions. CASOAR's routines are connected just by them being arranged round a file with a particular design which is relatively easy to understand. Once you have mastered this file structure you can work upon those files with any additional routine you can think of. (Due to the simple structure of the package those routines can be written in any programming language available - what may be important if you are frustrated by some old number cruncher that has not yet the ability to bind programs together, which were formulated in different programming languages.) On the other hand, all the routines have clearly defined and distinguished functions, so if you are really building a large scale system, providing its own utilities and filetypes, you can easily convert CASOAR into a kind of

demographic subroutine library - an important point for all projects which have data only remotely similiar to the type of family reconstitutions CASOAR requires, but are interested in the services it offers.(5)

So CASOAR:

- offers a wide range of demographic standard tabulations,
- can be implemented with moderate effort at all machines which support PL/1,
- can be used in pilot studies before decisions about own investments into program development are made,
- is reasonably fast and small,
- can be used as a nucleus of programs around which to build your own specialised routines,
- can be used as a subroutine library for historical demographers.

On the other hand CASOAR:

- operates on one and only one input format,
- has no subroutines, so any systematic change has to be repeated in all programs concerned,
- is rather inflexible if the data do not contain exactly the variables provided for,
- is rather clumsy if it comes to the selection of subsets of data for processing.

Described in this way CASOAR is of interest mainly for the community of historical demographers which is large, but still not all embracing. On the other hand, there are a number of distinctive features of this package that make it a good example to show how the exchange of programs between research projects could and should function. I admit, that being only related to historical demographers, while not one of them, I consider these characteristics of the system more interesting for computer-using historians in general, than the actual demographic calculations done. Lets list the virtues the package has in this respect, first from the documentation point of view.
The programs have a description that is not hidden in some obscure technical appendix of a paper presented at a conference, but written in its own right.
CASOAR as such is a consistent implementation of one particular approach towards a class of problems, which has been documented elsewhere. The emphasis of the documentation is on the kind of results produced; we are spared all details on how the data where (in a specific local environment) prepared for input and stored intermediately.
Even more important is the technical side:
All programs are built around one consistent type of file - no "smart" tricks for some exotic cases, but exactly one type of file, that avoids all need for periodic reformatting, when one wants to use a program that has been written a couple of months later than the last one used.
The programs are completely modular. You can use any of them without knowing anything about the other ones.
There are clearly defined logical checks that can - and should be - performed, before the data are used for analysis. No discoveries of

"dirty" data two weeks before the results are published. There are many historical projects which produce programs for special types of historical sources; many of them are of potential value to other projects. Resources for research are getting scarce - to share them, to make effort invested at one university reusable for somebody else, should be a serious concern for all of us. We might be better off, if there would be a number of packages, which are so inherently simple but well designed and documented as CASOAR, in other specialised fields as well.

FOOTNOTES

1 Michael Hainsworth and Jean-Pierre Bardet: Logiciel CASOAR, Paris: Société de démographie historique, 1981 (= 1er Cahier des Annales de démographie historique).

2 On such approaches - not necessarily restricted to demography in a restricted sense - see e.g. Kenneth W. Wachter et al.: Statistical Studies of Historical Social Structure, New York etc.: Academic Press, 1978.

3 In the manual of the system frequent references are made to his "Manuel de démographie historique", Paris, 1970 and "Techniques d'analyse en démographie historique", Paris, 1980. The later book is actually quoted with many of the separate programs for an explanation of the technique/methodology represented by a particular type of tabulation.

4 A distribution tape, containing the source code of all routines plus the JCL necessary to run the system on IBM installations plus a set of testdata, is available from the Laboratoire de Démographie Historique, 54 boulevard Raspail, 75006 Paris. The author of this paper has been entitled to distribute CASOAR within the German speaking countries; so a version being prepared to run on UNIVAC's is available as option as well. In both cases the aquisition is free of cost.

5 This is a description of whats currently being done in the context of the authors own system CLIO. Setting a frame of filehandling routines around CASOAR will obviously vastly increase the power of the system for the handling of other data structures.