12711

PSD.1/INF.6
14 March 1980
Original: ENGLISH

ECONOMIC COMMISSION FOR AFRICA

First session of the Joint Conference of African Planners, Statisticians and Demographers

Addis Ababa, 24 March - 2 April 1980

COMPUTER PROCESSING OF SURVEY DATA IN DEVELOPING COUNTRIES

(Prepared by Overseas Development Administration, United Kingdom)

Introduction

1. One of the major difficulties which survey programmes are likely to run into is that of computer processing. In some countries the absence of hardware is a problem, although with the cost of hardware dropping rapidly, this is not perhaps as severe in the long run as the high cost of software and the difficulty of programming.

2. Typically, the time required for programs to be written to process the data is large in relation to the overall time-scale of the particular survey; the tabulation requirements must be specified well in advance and cannot later be changed without considerable delay; and further analysis, suggested by the initial results, is greatly inhibited by the programming constraint. In these circumstances considerable processing delays which significantly reduce the value of the survey are not uncommon; worse, there are instances in developing countries of data collected at considerable expense never being fully tabulated.

3. There are two aspects to this problem which require attention. The first is the need for statistical software which effectively by-passes or at least greatly reduces the need for programming. Ideally the statistician should be able to communicate his requirements directly to the computer in a form which is both convenient to the statistician and as flexible as possible. This leads to the second aspect, which is that statisticians must be prepared to get involved in computing and to understand the limitations as well as the power of computer processing so that they can exploit the new technology to the best advantage.

Involvement of Statisticians in Computing

4. Too often, since computers with their powerful processing capabilities have taken over from hand-tabulation methods, statisticians have tended to abdicate responsibility for data processing. The systems analysis and programming is left entirely to generalist computer personnel who are expected to produce results to correspond precisely to the statistician's requirements, although these are often imprecisely or ambiguously expressed as well as being unnecessarily demanding. Moreover since computers are known for their super-fast processing capability, the job is expected to be done within hours or days rather than weeks or months.

5. Delays and frustrations follow when the desired results are not produced within the required timescale. The reason is not that the computer itself is slow, but that the time required to write, compile, test, correct, re-compile and retest the necessary programmes (several may be needed) is usually very substantial. Once written, they will work (with great speed) over and over again: unfortunately this attribute is valueless in the case of a one-time survey, and which 'continuing' survey is not materially changed from year to year?

6. How is this problem - the programming problem - to be overcome? The first essential is that statisticians recognise the problem and educate themselves in computing. For only if they understand how a computer works, and what is necessary to make it work correctly, will they be able to explain precisely what is required, and to appreciate what aspects of the processing are fundamental and what are mere frills which can be dispensed within the interests of speed and efficiency.

7. For example, in processing a survey the fundamental power of the computer lies in its ability to add a number to a particular cell in each table. This is because this procedure is repeated over and over again, and the more it is repeated the more cost-effective it is. But when a table has been completed, it must be printed out; and this is a far more complex procedure than the formation of the table itself, and one which is only executed once by the machine. Thus it is most important that the form in which a table is printed should be the simplest possible, if the results are required quickly, unless easy-to-use software is available for table manipulation.

Software Requirements

8. There is a wide and bewildering choice of statistical software but the packages do vary in their facilities and field of application as well as ease of use, etc. Fortunately the publication by the International Association for Statistical Computing entitled 'A comparative review of statistical software' (Francis, 1979) has contributed greatly to the ability of a statistician to identify packages which meet his particular requirements.

9. What is required in this context is a programme for processing survey data, as opposed to ad-hoc enquiries from an existing data base or statistical analysis such as regression, etc. In particular the requirements may be listed as follows:

(a) card reading including the interpretation of invalid punchings such as imbedded blank spaces, etc. (blank spaces to be distinguished from zeros);

(b) error reporting including detection of invalid codes and logical inconsistences;

(c) automatic error correction including the assignment to a 'not stated' category of invalid codes and other 'hot deck' imputations;

(d) ability to handle hierarchical data structures;

(e) grouping and derivation of variates for tabulation;

(f) flexible and clear table printing suitable for reproduction in publications; and

(g) ability to apply grossing up factors including calculation of sampling errors.

10. While there are other packages with similar capabilities, RGSP (Rothamsted General Survey Program) is the only one described in the 'Comparative Review' as a package for survey analysis. The strengths and weaknesses of this programme are discussed below.

RGSP

11. RGSP is a tabulating package for survey analysis which meets all the requirements listed above. It has been developed over many years by Frank Yates, one of the leading survey statisticians of his generation. A copy of the prospectus, from which an idea of the structure and capabilities of the package may be obtained, is attached.

12. A unique feature of the package is the separation into two parts of the data validation and formation of tables on the one hand and the manipulation and printing of the tables on the other. The first part produces a file of tables, separate from the original data, which allows tables to be extensively manipulated and printed in a variety of forms without retabulating the original data. It is often difficult to decide precisely, in advance, what groupings of income or similar attribute of a household or smallholding should be adopted. With RGSP the user may define a grouping of relatively small intervals which can be combined later into wider intervals after examination of the results. This ability to produce summaries of otherwise unwieldy tables is of tremendous value. Such files may also be updated with ease - useful, for example, in incorporating late returns.

13. RGSP makes no attempt to provide facilities for statistical analysis, but it does provide the means of converting the table files into files acceptable to other packages such as GLIM and GENSTAT.

14. A major criticism of the package in its present form lies in its dependence on a user-written FORTRAN execution programme (particularly for error detection and reporting) in Part 1. Steps have already been taken to reduce this dependence and such a programme is no longer required to form tabulations in relatively straightforward applications, including those with a simple household/person data structure. If one is prepared to take advantage of it, the FORTRAN programme does allow almost unlimited flexibility in the processing strategy, while relieving the programmer of the more tedious aspects of table formation through the use of the RGSP subroutines, parameters for many of which are specified in the Part 1 Instructions. Thus the definition of the actual tabulations required is quite independent of the FORTRAN programme and can be varied up until the last moment prior to compilation and running.

15. The disadvantage of the FORTRAN programme is that it makes the package less straightforward for the non-programmer; more important perhaps, it presupposes the ability of the local computer installation to provide an advisory service on FORTRAN which is not always possible, particularly where COBOL is the main language in use. However, it is intended that an enhanced version of the package will be developed which makes a FORTRAN programme redundant for any but the most complex application.

16. Another possible criticism of the package is that the variates are referred to by numbers, rather than names, which tends to diminish the self-documenting properties of a set of instructions. On the other hand it certainly makes the instructions extremely concise (50 tables can typically be specified in fewer lines) and the search for suitable acronyms for similar variates is avoided. It is not suggested that this should be changed.

Example of the use of RGSP

17. Considerable success was achieved using RGSP to process the Seychelles Census in 1977. Carried out in August 1977, the enumeration covered 62,000 people living in just under 12,000 households. Two punched cards were produced for each household and one for each of the 42,000 persons aged 12 or more (66,000 cards in total).

18. The processing was carried out in UK at Rothamsted using five RGSP Part 1 programmes, 2 for validation (inter-card and intra-card respectively) and 3 for tabulation (2 for households and 1 for persons). Several Part 2 programmes were subsequently used to manipulate and print the tables. One of the Part 1 programmes also produced a list of households with certain characteristics which was subsequently used as a sampling frame for household surveys.

19. The RGSP instructions and FORTRAN programmes were all written by a statistician in Seychelles who visited Rothamsted for the initial processing when a set of tables covering individuals was taken back to Seychelles. Thereafter a number of Part 2 programmes were written in Seychelles and sent to Rothamsted for running and the final versions of the tables, which were received ten weeks later, were reproduced (after some cutting and sticking) in the final report.

20. This exercise demonstrated the ability of a very small statistical office to process a relatively large survey within a very short time period, without recourse to intermediary programmers and with the added disadvantage of extreme remoteness from the computer installation. While the expertise of the personnel involved both in Seychelles and at Rothamsted undoubtedly played its part in the success of the exercise, the package itself and its flexibility were crucial factors.

Portability and Size

21. RGSP is written in FORTRAN and is therefore relatively straightforward to implement on most machines. Versions currently exist for the following machine ranges:-

ICL System 4

ICL 1900 and 2900 series

IBM 360/370

CDC 6400/6500 and CYBER 174

A version is planned for the NCR Criterion range.

22. The package will fit in a machine with 128K bytes of core store or more, but consideration is being given to the feasibility of producing a 'small' version which would run on a 64K machine. However, given the rapidly falling cost of storage as the technology advnaces, it will not be worthwhile devoting a great deal of effort to this matter, even if it proves to be feasible.

Comments on other possible packages

23. There are a number of alternative packages which merit consideration. The first is COCENTS which has been widely used for census processing in LDCs and by WFS. This programme however falls far short of RGSP in terms of ease of use to the extent that WFS had to develop a programme to generate COCENTS instructions. Nor does it contain any validation facilities. Its main merit at present lies in its small size. CENTS-AID is another contender (? used in Kenya) but again lacks validation facilities. It is not clear whether it is as flexible as RGSP; certainly it does not appear to have the extensive table manipulation facilities of the latter. Similarly the widely-distributed SPSS has serious limitations, being unable to handle hierarchical data or validation and having rather messy print layouts.

24. The French package, LEDA, looks to be the closest competitor to RGSP and has the 'advantage' of generating COBOL programmes. However it is not itself written in COBOL and has been implemented only on the IBM 360/370 range apart from the CII-HB IRIS 80. It also appears to require considerable core store.

25. The pair of programmes developed at the UN Statistical Office also seem attractive particularly for small computers (32K). These are UNEDIT and XTALLY but are at present only for use in UN supported projects, not having been made generally available. They are written in RPG-2, and have somewhat limited application in terms of acceptable data structures. The tabulation process looks rather slow, because presumably tables are not formed in core, but by a sorting method.

Conclusion

26. On balance RGSP appears to be the best of all the currently available packages which could be used for the analysis of non-trivial surveys especially household surveys, such as are commonly carried out by government statistical offices. It should be made clear however that there is a limitation on the number of cells for any one table and that the package is not suitable for tabulating trade statistics, for example, for which a sorting method is essential. However tables of up to 200 rows by 10 columns for example are quite acceptable and larger ones can be produced by splitting them up into appropriate pieces.

27. The programme has not been used within the UK Government Statistical Service and it is appropriate to ask why not. The GSS has not developed a package of its own for tabulation purposes nor, yet, adopted any other as standard, although SPSS is quite widely used. An underlying reason for this lies in the decentralised nature of the statistical service and hence reliance for processing on departmental computer services over which the GSS or individual statistical offices often have no direct control. Statistical computing in each department must compete with other, usually more pressing, demands (such as payrolls) and the staff of the computer unit are often not attuned to the particular problems involved.

28. The tendency has been for each department to seek its own solution, often by developing software in a limited way itself. More recently a package for interactive statistical analysis and manipulation, 'Package X', has been developed centrally but this is not suitable for tabulating. There are plans to obtain a tabulating package for general use throughout the service, but because there is a requirement for it to be capable of modification by the computer departments it must be written in COBOL, which is the only language which is universally known by government programmers.

29. It is believed that any statistical office which can obtain access to and master the facilities of RGSP will be considerably better off than many statisticians who are responsible for analysing survey data in UK.