A GENERAL DATA - PROCESSING PACKAGE

.... describing a set of general programs developed by the research staff of the Department of Economics and Geography at the University of Salford for the use of laymen in the storage, retrieval, tabulation, analysis and interpretation of data collected from repeated observations of a standardised set of conditions.

A paper prepared for the colloquium of the Computing and Mathematical section of the British Sociological Association held at the University of East Anglia in April 1968.

J F Hall Research Assistant SSRC Work and Home Survey
Dept. of Economics and Geography
University of Salford

A Set of General Programs for Data Processing

(Storage, Retrieval, Tabulation, Analysis of Social Survey Data)

J F Hall Department of Economics and Geography
(SSRC Work and Home Survey)
University of Salford April 1968

1. In 1963, the then Liberal Studies Department of the Royal College of Advanced Technology, Salford, was given a grant by the then DSIR to conduct an investigation into the economic and social effects of Salford Corporation's slum clearance and rehousing activities, with special reference to overspill. At that time the KDF9 was little used and the investigators saw the project as a chance to use a computer to process the data from the proposed household surveys. However, since the research staff engaged on the project had no computer experience at all, and the computer staff no experience of large scale data processing, both parties had to learn from scratch some of the techniques of the other, a rewarding, if lengthy process.

2. Apart from the results of the survey, the University of Salford acquired first, two competent programmers with academic backgrounds respectively in psychology, and classics and sociology, and secondly, a set of general data processing programs designed originally to specifications determined by individual sloth, but of practical use to any feckless social or other scientist with neither time nor inclination to learn the techniques essential to satisfactory tabulation and analysis by computer. It is this set of general programs which conerns the present paper.

3. It would not be amiss before dealing with the programs proper to point out that computer usage is made considerably easier if some form of precoding is used on questionnaires. Of course, this is now standard practice, but standardisation of some of the intervals used in time, distance and quantity leaves a lot to be desired at present. It is standard Salford practice to code non-response in the following manner: refusal: -3; no reply: -2; don't know: - 1; not applicable: 0; and all answers from 1 upwards. Each interview is also given a number in sequence as it is coded (from 1 up to however many interviews there are in the survey, provided that there are no more than 999 interviews) and a unique code entry to mark the end, (in this case 1000, the last interview of all being marked with -1000).

4. When all interviews are completed and precodings ringed, or open ended answers coded according to a preferably short code-list, all interviews are then transferred in code form to standard code sheets consisting of rows of data arranged in five columns. These sheets are then punched on to paper tape and verified. The processing starts with all interviews coded and punched on paper tape. At present the general programs can cope only with interviews which are all the same length and it is assumed that a location in one interview refers to the same variable in all other interviews. There is a general program which can expand data from an original paper tape to standardise the location when transferring to magnetic tape storage.

5. Once the paper tapes are ready, the general programs are used for transfer to magnetic tape storage and for all further manipulation and tabulation. All the programs are written in KDF9 Algol, and established on disc as "Kidsgrove" Algol. When the data is transferred from the paper tapes it is stored in binary arrays, each of which is given an 8-character name, on a magnetic tape which is also given an 8-character name. The tape names must be supplied by the user, but the array names are fixed by the programs, although work is in progress to modify the programs so that the user can also supply his own array names. (The reasons for preferring the latter will be evident when storage on one tape of more than one survey is discussed.)

6. The General Programs.

Storage on magnetic tape: CBL27JW -- KP4

The first general program transfers all data from paper tape to magnetic tape, referring to the unique entry of (-)1000 to separate interviews, as binary Algol arrays. Each array keeps the original code items in order and retrieval is executed simply by referring to the location in the arrays of the particular variable under consideration. No checking for errors is made at this stage as it is easier to check when the data is already on the tape. To use this program, one simply needs to decide on an array size, which it is advisable to make slightly longer than the actual interview length to accommodate mispunched interviews which may be under- or over-length. The arrays holding each interview are declared in the program as ranging from 1: n where n is a value to be supplied on the steering tape and is normally

between 5 and 10 greater than the actual number of elements occupied by the interview coding. The steering tape contains the tape name in string quotes, the array size, and the coded interviews with - 1000 as the last entry of the last interview. The other general program for transfer of data from punched paper tape to magnetic tape is for use in standardising data locations by inserting zero values to expand interviews of a shorter length to the length dictated by that of the largest quantity for a particular variable. For example, the number of persons in a household may vary from one to twelve, and details of persons will consequently occupy twelve times the space for the latter as for the former, unless one were laboriously to punch zeros for eleven non-existent persons. Likewise details of each job a person has had, or of each address a family has lived in, will occupy different amounts of space. CBL01JH00AP4 allows only the actual data to be punched and automatically inserts zeros into the arrays for non existent items on the original interview. In this way data locations are standardized as before. The program operates again on the end-marker (-) 1000 to separate interviews and going through a loop containing a procedure in order to expand interviews. It is not necessary to define the number of expansions necessary for any particular run as an array is written to magnetic tape once the end-marker has been located. Extra data must be supplied on the steering-tape in the form of three integers for each expansion referring respectively to the highest value of the criterion variable (e.g no. of persons), the number of code classifications for each criterion variable (e.g. age, sex, identity, marital status) and the location in the final array of the element

containing the value of the criterion variable. Needless to say the final array organization must be known before the steering tape is punched. To illustrate the program's use with an actual example, the Salford household survey contained information about persons, jobs of male tenants, and addresses for each household: information was also collected for each person (age, sex, identity, marital status, length of marriage) for each job of the male tenant (nature of work, location, means of travel, comparative costs, time, hours, pay relative to previous job, duration and reason for leaving) for each address (nature of tenure, comparative size, condition, rent, duration of residence and reason for leaving), for each married child (frequency seen by husband, wife) and for other kin seen at least once a month (identity, frequency seen by husband, wife). i.e. 6 criterion variables with 5,8,6,2,2,3 dependent variables. An array size of 295 was allocated, the end-marker being element 289. Data fed in as below produced the desired result.

```
SSRC DATA   ;    295;
12; 5; 2;
12; 8;63;
10; 6;160;
9;  2;221;
6;  2;240;
17: 3;253;
1;  1;287;
```

Any failure of a run will be due to punching errors (e.g. a minus sign after a number, or 1000 missed so that the number of entries between two successive 1000's exceeds the array size). A successful run will be terminated when the computer detects the

unique -1000 entry, and prints out a message of the form "RUN SUCCESSFUL." In all subsequent operations the tape name in string quotes and the array size must be remembered as they are an essential part of the data necessary for defining which information is wanted.

7. Correction and tracing of errors.

Various programs are available to check that each interview is the correct size and that the values of the elements fall between the lower and upper limits specified in the coding list, and for correction by replacement of individual elements within interviews or by deletion or replacement of whole interviews. For instance the most accurate programe CBL28JW -- KP4 performs the first operation described above, but demands a longer steering tape than most programs, since it is necessary to provide a list of the lower and upper limits of each code item in the interview except for the interview number and the end marker. Output is of numbers of interviews containing errors, of the numbers of erroneous elements and the erroneous value of such elements. Two other programs perform the other operations, CBL29JW--KP4 and CBL32JW--KP4. A useful short check for length is provided by a Whetstone program, CR0001800APU, which prints out the interview numbers and the end markers of all interviews on a tape. This also gives a check on duplicate entry of numbers or of sequence errors.

8. Manipulation

Many programs have been written to transfer data from one magnetic tape to another, but these have all had the limitation that the array names were not variables that could be declared as data and

and consequently all information from different surveys has had to be kept on separate tapes. Sometimes it is desirable to select certain items of information for various purposes such as regression or component analysis or simply as a check on contents. Such was increasingly the case at Salford and consultation with the computer staff produced a procedure from the Kidsgrove library known as "instring", and two associated procedures for reading and writing arrays named in these strings which have enabled the construction of two very fast programs for the transfer of arrays from one tape to another where the array names as well as the tape names are supplied as data, and for the transfer of a specified number of individual elements from each array of a certain name to paper tape and to a line printer output, the latter being particularly useful as a preparation for multiple regression. The program for transfer of named arrays can be used either to transfer arrays from a master tape to a worktape or, more importantly, to build up a master tape from a set of worktapes. This saves tapes and avoids complaints from the computer staff. Later it is intended to incorporate these procedures in the tabulation programs. At some point in the future a system will have to be developed for allocating array names to individuals in order to avoid the possibility of a catastrophic duplication of names, just as there is a system for allocating program names.

9. Tabulation

Anyone who has ever tried to extract information in tabulated form from a computer storage of any kind will appreciate what a blessing

it is not to have to define each table in a separate program, but simply to supply certain minimal information and let the computer do all the work. We have at Salford two such general tabulation programs each with different historical development, one evolving only partially from the other.

10. CBL22JW--KP4 is a program whose Algol text covers nine pages of the widest computer stationery and which will tabulate in unlimited numbers of tables of one, two or three dimensions using the simplest of data. To tabulate from the interviews on a magnetic tape, all one needs to supply on the steering tape are the tape name, the array size, the number of dimensions of the first batch of tables, the number of tables in the first batch (limited to 60x1, 30x2, 20x3) and the location of the code items to be tabulated for each table. The program operates by scanning the whole tape to find the lower and upper limits of the variables for each tabulation, declaring an array to fit these limits, and then rescanning to fill the elements of this array from the interviews on the tape. There are, however, certain limitations to this program in the amount of storage taken up by the program itself, in the limit of three dimensions, and in its lack of combinatorial procedures (i.e. if details are required of the age and sex of people and there is a maximum household size of 12 catered for in the coding, then there will be 12 separate tables to add by hand, an infuriatingly laborious task). But for most users the program is perfectly adequate.

11. GR00036--KP4 represents an attempt to remedy some of the short comings of CBL22JW--KP4, especially in the provision of combinatory

facilities and a reduction in the amount of computer storage required. It also caters for tabulation in up to five dimensions. Whereas CBL22JW declares the maximum number of arrays even if it is only using one or two of them, GR00036 declares only three, a one-dimensional array holding the interview, a two-dimensional array holding the limits of each parameter and a six-dimensional array for tabulation, in which five of the six are used for actual tabulation and the sixth as a counter to hold tables separate. The Algol text runs to only three pages of short, double-spaced lines. Declaration of the number of dimensions for CBL22JW requires a separate declaration of a set of arrays for each dimension up to three, and a separate procedure declaration and list of procedure calls, and it is this three-fold repetition which takes up all the space. The original declaration of a six-dimensional array in GR00036 avoids all this repetition and the problem of dormant dimensions is solved by leaving their range at 1:1. Data requirements are almost the same as for CBL22JW except for the addition of the overal number of tables required. The only difference in operation is that a specification 3;12; followed by twelve groups of three integers specifying the locations of items to be tabulated, will result in the production of 12 separate tables of 3 dimensions by CBL22JW, but a single 3-dimensional table which is the sum of 12 tables by GR00036. Both programs identify tables by referring to the locations of the items tabulated, giving the upper and lower limits of each.

e.g. TABLE 4

(CBL22) ITEM 2 AND 63 AND 160

LIMITS OF ITEM 2 ARE 0 AND 3

LIMITS OF ITEM 63 ARE -3 AND 10

LIMITS OF ITEM 160 ARE -2 AND 10

(GROOO36)

TABLE	ITEM	LOWER	UPPER
4	2	0	3
	63	-3	10
	160	-2	10

It is assumed that the user is sufficiently prudent to retain at least one copy of his original coding list!

12. Statistical analysis

There are also a number of general programs for the classical statistical analyses and tests, but the calculations involved are familiar and present little difficulty in programming. Most frequently employed is the χ^2 test, but we also have programs for Kendall tau, Spearman rho, Pearson correlation, Aitken regression and some programs from Liverpool for factor analysis, principal components by quartimax and varimax rotation, and multiple regression.

13. With the programs described above it is possible for anyone to process, analyse, tabulate or test any information gained from repeated observations of a standardised set of conditions, whether or not he is familiar with the niceties of programming, and this applies not only to data collected by social survey, but also to information collected in a serious of scientific experiments. Since their inauguration they have been used to process several

surveys and literally thousands of tables have been produced and tested. The research team has no experience of punched card or disc-file processing, and cannot comment on the superiority or otherwise of magnetic tape, but one would suspect that it is superior to both in that all the information is kept in one place and location is easier, and in that there are fewer restrictions in coding or tabulation.

14. Promising future developments would seem to be, for Salford, a progress towards the "dynamisation" of the general programs leaving the user to supply definitions at present inflexibly incorporated in the programs, and for researchers and survey practitioners as a whole to agree on standardisation of intervals for parametric variables in order to allow direct comparison of one set of results with another, and on conventions for certain types of coding procedure referred to in paragraph 3. Is it too much to hope for a properly produced questionnaire resulting from a pooling of information or based on the results of highly developed factor - and principal component analysis? At any rate the emergence of a separate mathematical and computing section of the B S A is a welcome and overdue development.

J F Hall

April 1968