UNITED NATIONS
ECONOMIC
AND
SOCIAL COUNCIL

5395

Distr.
LIMITED
E/CN.14/CPH/7/Add.1
Original: ENGLISH

ECONOMIC COMMISSION FOR AFRICA
Seminar on Organization and Conduct of
Censuses of Population and Housing
Addis Ababa, 17 - 29 June 1968

ELECTRONIC PROCESSING OF CENSUS DATA

Part II

Some Principles of Computer Processing of Census Data

Prepared by

the Statistical Office of the United Nations

Distr.
LIMITED

ST/STAT/33
7 May 1968

Original: ENGLISH

STATISTICAL OFFICE OF THE
UNITED NATIONS

ELECTRONIC PROCESSING OF CENSUS DATA

Part II

Some Principles of Computer Processing of Census Data

Contents

		Paragraphs
A.	INTRODUCTION	1
B.	OPTIMUM USE OF CAPACITY	2 - 8
C.	THE DETECTION AND CORRECTION OF ERRORS	9 - 21
D.	PRODUCTION BY COMPUTER OF FINAL TABLES FOR PHOTO OFFSET	22 - 24
E.	SUMMARY	25 - 26

ANNEX

Example of tabulation on a computer

A. Introduction

1. Almost all censuses in the 1970 round - and afterwards - will be processed by computers. In some cases, the input-output operations will be highly automated through the use of mark readers or optical character readers for input operations, and electronic photo-composers for out-operations. Some principles of the processing technique will be discussed here with a special view to explanations and guidelines needed for those countries that are about to enter this field for the first time.

B. Optimum use of capacity

2. Common to all computers is a central processing unit, hereafter referred to as a "CPU". The CPU contains the central memory, that is the device where the programme to be executed is stored together with the data to be manipulated (the input data, the intermediate data) and the final data resulting from the manipulations (the output data). Attached to the CPU is a configuration of input-output devices, or so-called peripherals, such as units for reading or punching cards or paper tapes, magnetic tape drives, direct access storage devices (magnetic disk, drum units, etc.), optical character readers, printers, etc.

3. Handling of data inside the central memory is much faster than reading, punching and writing of data on the peripherals. It is good programming technique to arrange, as far as possible, for the central memory to be kept busy and not idling, waiting for a record to be read in or written out. Reading and punching of cards or paper tapes, as well as reading marks or characters, is much slower than reading and writing magnetic records (on tapes, disks, etc). The general practice is, therefore, to write input-data, that have to be used more than once, on magnetic tapes or other magnetic storing media. There are even cases where a special, smaller, computer is used for such comparatively slow operations as "card to tape", "tape to card" and "tape to print", leaving the bigger computer to communicate with the "outer" world with magnetic tapes only.

4. In tabulation of census data, the standard method is to use part of the central memory for the simultaneous creation of a set of different

tables (see annex). In certain cases, areas can be reserved in the central memory for all required tables, and if so, only a single input operation for each record is required (for the tabulation) and no sorting at all of the input records is needed. If there is not space enough in the central memory for the tabulation to be made in one run, the following methods can be used, separately or in combinations:

(1) Reruns, without sorting, of the input-records, with modified programmes, until all basic tables are created. A table is said to be basic if it is not a sum of other tables.

(2) Sorting the input records before and/or between the different tabulations. In the extreme case, sorting can be used so that only a single area in the central memory is required for the tabulation. This single area is then used for the accumulation for one table-cell at a time. The extreme case has been mentioned here only to illustrate the thesis that the larger the memory, the less sorting -- the smaller the memory, the more sorting.

(3) Attaching a so-called Direct Access Storage Device, a DASD, to the central memory. On a DASD, records can be accessed directly, as in the central memory, rather than serially, as is the case with magnetic tapes. Magnetic disks, drums and even magnetic strips or cards, such as the IBM Data Cell Drive or the NCR Card Random Access Memory (CRAM) are examples of DASD.

5. Before further discussion of the three methods, the preportions of a typical census case will be given here. In the united Nations "Principles and Recommendations for the 1970 Population Censuses" there are 22 tables ranked as "first priority". The total number of locations, or cells, in these tables is around 5,000. In this count, no totals are included. For instance, the requirements for Table 7 is taken as 40 and not 63, as is required when the marginal distributions are included (for explanation of "marginal distribution" see the annex). Many of the tables will appear on the lowest level in the hierarchy of geographical subdivisions. It can therefore, easily be seen that the required number of table-locations, taking

into account the geographical distribution, might run into the millions or even tens of millions. If such cases have to be dealt with, without sorting and rerunning, the only way is to use a kind of DASD. A census, where the material is collected and prepared in an order that has no relation whatsoever to the tabulation order, might be a case for a DASD. Generally speaking, a DASD offers an attractive solution where the input material as a whole has to be taken as a unit for processing and where no tabulation can be finalized before the very last input record has been prepared. In many cases, however, the way in which the census-data are collected results in a suseful presorting. This is normally the case in a de facto population census. The data for an enumeration area can be processed, independently of the data for other areas, as soon as they have been prepared. This is called batch-processing. Assuming the above-mentioned recommended tables, only some 5,000 locations are needed for a complete tabulation in one run. If an additional presorting according to, for instance, sex and age is made, the required capacity comes down to the order of a few hundreds. (The capacities quoted here are for the tables only, and are in addition to the requirements for the actual programme and a possible supervisory programme.)

6. The three methods can, as already mentioned, be combined. It should also be noted that a DASD can play a very important role, even if it is not large enough to store all required tables simultaneously.

7. Storing in a DASD is slower than storing in the central memory. It might, therefore, be economical in some cases to programme so that the most frequent cases are tabulated in the central memory and the less frequent in the DASD. The computer can be programmed to find out where the tabulation should take place in order to minimize the processing time. Input data for which no destination is reserved in the central memory or in a DASD (because no DASD is attached or its capacity is surpassed), can be "dumped" on a magnetic tape for future processing. In some cases, it might even pay its way to dump excessive data on punched cards. Inherent in the dumping technique is the possibility of reducing the volume of input data from one run to another.

8. The tabulation of an input record consists of finding, for each table, whether any data from the record should be added to any location in the table.

If data have to be added to the table, the problem is to find the serial numbers of the locations within the table that have to be "updated". When the numbers have been found, the additions take place. A serial number calculated for a table must, of course, fall within the limits of the table. One way of ensuring this would be to reserve for each table, a residual group where the records with erroneous codes could be added. This would, however, create chaotic discrepancies between tables and would not solve the problem in the case where it is not clear whether a record belongs to a table or not. The method normally chosen is to ensure "processability" through a special checking programme. This imposes a kind of precision in processing that goes far beyond what is statistically required.

C. The detection and correction of errors

9. Precision is achieved through volume control and editing. The standard method is to let the enumerator summarize - on a checklist - a few basic data from the questionnaires, such as number of persons, households, living quarters, industrial enterprises, agricultural holdings, etc. By summarizing the checklists throughout the hierarchy of geographical subdivisions, two things are achieved. Firstly, a few, provisional, statistical data are quickly made available. Secondly, a framework is created for checking completeness and uniqueness. This is the volume control and it is of fundamental importance throughout all the handling of material and data.

10. Editing consists of two phases: the detection of errors and the correction of errors. Some editing is, and must be, performed manually. However, experience has shown that manual editing is not efficient enough. Furthermore, the records must be error-free as they appear in the central memory. Errors might be introduced in punching, in mark sensing or in optical character reading. The computer can be programmed for the first phase only or for the two phases combined. In the first case, the computer will print a list of the erroneous records and the character of each error. This procedure might be combined with rejection of the erroneous input records. The corrections will, in this case, be made manually. In the second case the appearance of an error will initiate an automatic correction

procedure, which is either deterministic or stochastic. A deterministic procedure is one where a given condition defines a unique value. For instance, the rule always to set sex as female whenever a number of children born alive is reported, is a deterministic procedure. A stochastic procedure is one where a given condition defines a frequency distribution of possible values for an erroneous (or missing) code. The actual value is determined by a random process. For instance, a rule to determine a missing age by randum drawing of a number with a given distribution, is a stochastic process.

11. The correction procedures must, of course, be constructed so that no new errors or inconsistencies are introduced. A simple way to formulate the stochastic rules is to assume that missing or erroneous data are distributed as the correct data. But this is one of the more questionable methods to use. Errors sometimes appear in a very systematic way. A coder might constantly misunderstand certain answers (perhaps because of language differences) or he might have memorized certain codes incorrectly. People of one religion might be more unwilling to state their religion than people of other religions. It is a very tricky problem to formulate the rules for automatic corrections. But once it has been done, the gains are considerable. Verification of coding and punching can be skipped or reduced to sample operations.

12. Automatic corrections must be under constant statistical control. The editing programme must report the number of errors by type by coder by punch operator for each batch, and from time to time the prerequisites for the stochastic rules must be checked.

13. Not all errors can be corrected automatically. For instance, errors resulting from misplacing of questionnaires are of this nature.

14. Editing is not always limited to the individual record per se, its codes and the interrelations between its codes. Editing might cover the interrelations between records of the same group, such as records for all members of a household. Editing is sometimes combined with the creation of summary records as well as with transfer of data from master records to detail records or transfer between matching returns in different censuses.

15. Editing is a prerequisite for the functioning of the tabulation-programme. This does not mean that editing only before the tabulation starts is a sufficient

procedure to ensure acceptable tables, nor does it mean, for practical reasons, that all editing must be done before the tabulation is started. The following two examples will clarify the two points.

16. Example 1. In a census of population, a question about the kind of diploma of vocational training was asked. The answer "Driver's license" was given the code for "Other diploma". In the individual record per se the code for "other diploma" was acceptable. In the tabulation, the high frequency for "Other diploma" raised doubt and resulted in the detection of the error.

17. Example 1 shows that a study of the proportions within a table can reveal errors that cannot be, or have not been, detected in the individual record per se. The computer can be programmed to analyze the tables and to signal cases to be investigated.

18. Example 2. In a census of population the tabulation plan might have a table, for the country as a whole, with complete cross-classification of industry by occupation. The capacity might very well be sufficient for testing the industrial and occupational codes independently of each other on the batch level but insufficient for a practical and economical testing of the combination of them on the same level. By excluding the cross-classification from the original testing and tabulation, the batch-processing might be undertaken with a fair degree of accuracy. When the batch-processing is over, the cross-classifications can be sorted and tested against a file with permissible combinations. After corrections, the countrywide table can be produced.

19. Example 2 raises the question of whether the errors detected in the second round of editing will affect the tables made directly after the first round of editing. If the end of the tabulation has been reached, the rigorous requirements set by the computer are no longer at hand and the problem is reduced to one of statistical significance; if any possible decision based upon the tables can be reasonably suspected to be influenced by an additional round of corrections, the corrections have to be made; otherwise they need not be made.

20. Corrections can be done systematically by running the whole tabulation programme with the erroneous input records as negative components and their possible replacements as positive components. The tables, consisting only of the balances between the two components are then used for updating of the table file. It is thus not necessary to rerun all input records in order to make the corrections.

21. The basic tables, that is, the tables created directly from the input records for the individual cases, are in their turn added together to new tables in one or more hierarchies. The editing of tables is not limited to the basic round; it has to be made after each step in the aggregation. The editing of the tables is not only a hunt for errors, but it is, as a matter of fact, an advanced analysis and interpretation of the statistical material.

D. Production by computer of final tables for photo-offset

22. It is well known that the computer can be used for printing tables with headings, stubs, page numbers, etc., in a form that can be used for reproduction by the offset method. There are, for certain computers, character sets available for simultaneous printing with different alphabets, such as Latin-Greek, Latin-Arabic, Latin-Hebrew etc. This means that bi-alphabetic tables can be printed in one run.

23. Perhaps less known is a method called electronic photo-composition. In one such system, now in use, a computer is programmed to direct a beam of light through a selected image of a typographical character, thus projecting the character on a photographic film. In another system, now being developed, the computer is programmed to direct the beam in a TV-tube to compose a page on the screen. The screen is photographed and the film is, as in the other system, used for offset reproduction. A big variety of type-fonts are available and the results can hardly be distinguished from ordinary printing. The advantages with the automatic production of the tables are: higher speed, lower cost (at least in developed countries) and higher accuracy than with conventional methods. It has previously been said that editing is a prerequisite for the tabulation programme. Editing is also a prerequisite for the automatic production of tables.

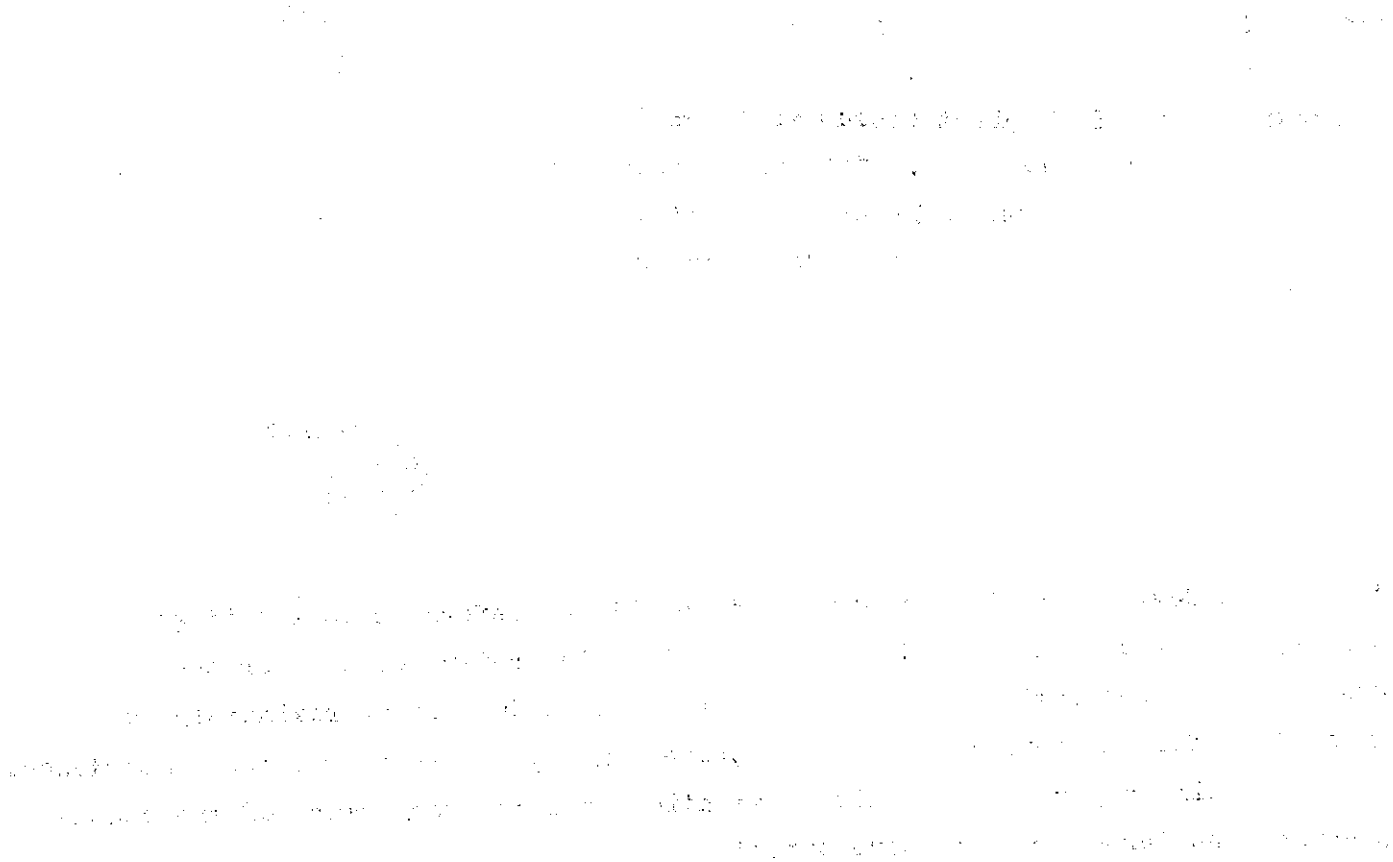
24. The edited results of the tabulation programme are normally not completely reproduced. The reasons are many. For convenience in processing, some tables are created for the smallest geographical sub-divisions but are meaningful only higher up in the hierarchy. Some errors can only be found but not corrected, resulting in scrapped tables. Some tables might contain all possible entries, including those for which no amounts are reported and those entries are not to appear in the printed tables. The selection and modifications of the tables to be printed are to a certain extent governed by the results of the tabulation and can therefore not be fully planned in advance.

E. Summary

25. At the center of the processing of census data stands the tabulation. In tabulation, the computer is used as an extremely fast accounting machine with a large number of registers. Verification of coding and possible punching can to a considerable extent be taken over by the computer. The computer can read not only punched cards or papertapes but also marks and characters (typed or handwritten). The computer can perform the first phase of the editing: the detection and reporting of errors. To a certain extent, the computer can also perform the second phase of the editing: the corrections. Every second step in the processing is an editing step, where the results of the previous step are analyzed and the way for the next following step is cleared. Interwoven is a volume control to ensure that no material is mistakenly processed more or fewer times than required. The computer can be programmed to prepare the source for offset reproduction of the tables.

26. The degree of automation a country should apply must be judged from case to case. Generally speaking, the computer must have capacity enough to keep pace with the coding and other preparations of the input records. The clerical operations and the coding system must be under continuous control by the computer through a kind of "early warning system". The returns might show unexpected characteristics that call for changes in the coding system. All of this means that the computer and the programmes must be in operating shape when the census is taken. This, according to experience, means that the preparations for the processing must start between 12 to 18 months before

the census day. If a pilot census is taken it is preferable to let it go through the processing phase. This means that the reference day for the start of the preparations will be the pilot census day, or approximately so. Ample time for the development of the processing system is of greatest importance. The processing aspect must be given attention from the very start of the census planning and throughout its development.



Example of Tabulation on a Computer

In the central memory of the computer, 400 words (locations, areas, cells) numbered 4200 thru 4599 are (in this example) available for tabulation. Each of these words can be used as a counter. In this example, the available part of the central memory will be used for the creation of:

Table 1: Population by single year of age and sex

Table 2: Population by marital status, age and sex

As can be seen below, there are 102 age-codes, 2 sex-codes and 5 codes for marital status. Thus, 204 words are needed for Table 1, and consequently the words 4200 thru 4403 have been reserved for Table 1. For Table 2, the ages form 15 age-groups. The required number of words is therefore 150 and the words 4404 thru 4553 have been set aside for Table 2. Observe that words have been reserved for the basic parts of the tables only and not for any total or so-called marginal distribution. This concept is illustrated in the following version of Table 2. The shaded areas represent the marginal distributions.

Age group	Male						Female						Total					
	Marital status						Marital status						Marital status					
	1	2	3	4	9	TOTAL	1	2	3	4	9	TOTAL	1	2	3	4	9	TOTAL
less than 15																		
15-19																		
20-24																		
....																		
75 and over																		
Not stated																		
TOTAL																		

The words 4554 thru 4599 are not used. For each input record, a "one" has to be added to the proper word in the area reserved for Table 1 and to the proper word in the area reserved for Table 2. The number of the proper word, that is the address, is a simple function of the codes for sex, age and marital status, but not of the external version of the codes but of an internal version. The external version is the one used in coding and punching the questionnaires. The internal version is calculated by the computer. The internal version may be the same as the external and the internal version may vary from table to table, as illustrated below. In this case the two functions are:

Address for Table 1 = (sex - 1). 102 + (age - 1) + 4200

Address for Table 2 = (sex - 1). 75 + (age - 1) .5 + (marital status - 1)+44

The different versions of the codes are as follows:

External code for age	Internal code for age, Table 1	Internal code for age, Table
00 for age below 1 year	001 for age below 1 year	01 for age below 15
01-99 for ages 01-99	002-100 for ages 01-99	02 for ages 15-19
XX for age 100 and over	101 for age 100 and over	03 for ages 20-24
YY for age not stated	102 for age not stated	.....
		14 for age 75 and over
		15 for age not stated

External code for sex = Internal code for sex

1 for male
2 for female

External code

External code for marital status	Internal code for marital status, Table 2
1 for single	1 for single
2 for married	2 for married
3 for widowed	3 for widowed
4 for divorced	4 for divorced
9 for not stated	5 for not stated

Sample computation of addresses:

	External codes			Internal codes and addresses						
				Table 1			Table 2			
	Age	Sex	Mar.status	Age	Sex	Address	Age	Sex	Mar.status	Address
Record 1	21	1	2	22	1	4221	3	1	2	4415
Record 2	YY	2	9	102	2	4403	15	2	5	4554
Record 3	75	2	4	76	2	4377	14	2	4	4547

On the following page, the available part of the central memory is illustrated. The words where a "one" has to be added in the tabulation of the above three records contain the symbol "plus 1".

	0	1	2	3	4	5	6	7	8	9
420	START OF TABLE 1									
421										
422		PLUS 1								
423										
424										
425										
426										
427										
428										
429										
430										
431										
432										
433										
434										
435										
436										
437								PLUS 1		
438										
439										
440				PLUS 1	START OF TABLE 2					
441						PLUS 1				
442										
443										
444										
445										
446										
447										
448										
449										
450										
451										
452										
453								PLUS 1		
454				PLUS 1	START OF UNUSED AREA					
455										
456										
457										
458										
459										