UNITED NATIONS

ECONOMIC
AND
SOCIAL COUNCIL

LIMITED

ST/ECLA/Conf.43/L.4
11 July 1972

ORIGINAL: ENGLISH

ECONOMIC COMMISSION FOR LATIN AMERICA

SEMINAR ON THE PREPARATION AND USE OF POPULATION AND HOUSING CENSUS TABULATIONS

Organized by the United Nations, through the Economic Commission for Latin America, the Statistical Office of the United Nations, the United Nations Trust Fund for Population Activities with the collaboration of the Latin American Demographic Centre

Santiago, Chile, 14-19 August 1972

COMPUTER REVIEW AND EDIT OF A POPULATION AND HOUSING CENSUS

By

Howard G. Brunsman
United States Bureau of the Census

The purpose of the error detection and correction programme is to improve the accuracy of the census data. As indicated in the United Nations paper* errors may arise because the information recorded on the census schedule is not a correct representation of the facts or because of errors in the various stages of processing the data from the enumeration document to the resulting statistical tables.

The United Nations paper refers to the three types of errors that might be detected. These are:

(a) Omission of entries

(b) Inadmissable entries

(c) Inconsistent entries

All three types might occur either in the data recorded on the census schedule or in the processing of the data. An entry on the schedule might be omitted because the respondent did not know the answer to the question. The recorded data might be inadmissable. For example, the reported place of birth might be non-existent. Or the recorded data may be inconsistent with other characteristics. For example, highest grade completed may be too great for the age of the person. Reported bathing equipment may be inconsistent with water supply. Some of the errors may be corrected by a detailed examination of the schedule. A missing sex report may be supplied by an examination of given name. Inadmissable place of birth might result from relatively minor misspelling. The only way in which many errors may be corrected is by reinterviewing the respondent. Obviously this approach is not feasible on a wholesale basis for a nation-wide census.

Many of the errors result from the various processing operations. The most common errors arise from improper manual coding and in the process of converting data to machine readable form. Errors of these types can be corrected by reexamining the enumeration schedule, but it is very expensive to search out the enumeration schedule and to determine the required correction. And if the error proves to be in the recording of the data on the schedule, nothing is gained by the search. Also if the errors are relatively rare, the accuracy of the data is not seriously impaired if these are corrected by imputation.

* The Use of Computers for Error Detection and Correction in Population and Housing Census, preparaed by the United Nations Statistical Office, April 1971.

/Inconsistencies in

Inconsistencies in the data might be revealed by a case-by-case examination of each record or by the preparation of a detailed cross tabulation of the variables. For example: Inconsistencies in the relationship of grade attending to age might be detected by tabulating detailed grade by detailed age. Adjustments might then be made in the table before publication. But such adjustments might result in producing inconsistencies with the tables showing grade attending but not by age for smaller areas. Therefore it is better to locate the inconsistent cases and change them before any tabulations are prepared.

As indicated in the United Nations paper, the role of the computer can be restricted to the detection of records with errors without the automatic correction of the errors. In the days of the unit count tabulator, the tabulator could process a batch of punch cards and sort into a separate pocket all cards with a specific type of omission or inconsistency. These cards were corrected manually and the corrected records inserted in the file. In corresponding manner the computer programme can print a report showing the content of all defective records and the nature of the defect. The programme can produce duplicates of defective cards so that only the defective items need be corrected.

The fact that the computer is able to detect certain types of errors must not be used as an excuse to eliminate all verification. The recorded information may be incorrect even though it is fully consistent. The computer error detection programme might be used to identify the work units with an excessive number of errors. The presence of errors that the computer is able to identify may suggest the presence of additional errors. Work units with excessive errors might then by subjected to a more thorough verification.

In many cases the true nature of an error is not immediately apparent. An inadmissable sex code may result from the puncher skipping an item. In such a case the relationship code of "4" might appear in the sex column where only codes of "1" or "2" are acceptable. An error of this type will place incorrect data in many of the characteristics. It is better to reprocess such records rather than supply the data by allocation. Also an improper housing unit identification code on the record for one person may create an apparent error in the count of persons in the housing unit, and the appearance of several groups of persons with the same identification but no housing unit. Cases of this type also should be taken care of by reprocessing rather than allocation.

/The automatic

The automatic correction of errors may consist of treating inadmissable and inconsistent entries the same as omitted entries. For many characteristics this is the best procedure. This is especially true for characteristics with many categories, such as occupation or industry or country of birth. The inclusion of a "Not Reported" category, forces the user to exercise judgement in the interpretation of the data. For example, in some age groups, the number of persons not reporting school attendance might be greater than the number not attending school. The interpretation of such data would be quite different if it is assumed that the "Not Reported" groups is attending school, is not attending, or is distributed in the same proportion as those reporting on attendance. Thus the inclusion of a "Not Reported" category warns the user to exercise caution in the use of the data. This is a awarning that might be lacking if the computer programme (or, more realistically, the programmer) had decided how these cases should be classified.

For certain characteristics it is preferable to have the programme assign a code in all cases where errors are detected. The programme is able to assign values to missing variables by taking account of other characteristics of the record. It is preferable to have the programme assign a code for sex in all such cases in order to avoid increasing the cells in each table by one third to provide a "Not Reported" category as well as Total, Male and Female. The programme can take account of whether the person is reported as wife, or is a head with wife present. Also whether fertility is reported or the activity is shown as housewife. Even when these clues are not available, it is preferable to assign sex taking account of relationship to head and marital status.

The allocation of age avoids the difficulty decision of whether to include or exclude persons with age not reported from tables that are restricted to persons 15 years old and over, 5 years old and over, 5 to 29 years old, etc.

The programme is also able to impute a missing age on the basis of characteristics of the person or the presence and characteristics of related persons. Economic activity is not reported for children. School attendance and education are not reported for persons under 5. There is usually a relatively close relationship between the age of the wife of head and the head and between the wife of head and age of oldest child. There is also a close relationship between age and grade attending for a person attending school. In corresponding manner, it is better to assign a value to marital status and relationship when they are not reported.

/School attendance

School attendance might be shown for persons 6 to 29 years of age. The programme might be able to take account of educational attainment and age for persons not reporting attendance. The group not reporting might be concentrated in persons 25 to 29 years of age who are seldom attending school. Thus the values assigned by the computer may be more reasonable than those assumed by the user.

The analysts are far more likely to accept the computer assignment if they are given a report showing the characteristics that have been changed and the nature of the changes. This report can be a diary summary. It can also include a case-by-case report of all changes that have vbeen made. The analyst can review this report and submit revisions for those records for which the action of the computer is not acceptable.

Within the past year, I have worked with the Executive Office of the Census of Nicaragua in the preparation of a computer edit programme for their 1971 Census of Population and Housing. This programme is being run on an IBM 360/25 with 32K. The input of the programme is in the form of punch cards with a separate card for each housing unit and a separate card for each person. The card for the housing unit is followed by the cards for the persons in the unit. Persons are grouped by household within the housing unit. The output is a binary tape record with each variable expressed as a one or two byte binary number. The output also contains an indication of which variables have been changed by the edit process. Thus it is possible to prepare tabulations showing the number of cases where the variable has been changed and the distribution of the variable for cases where no change has been and the distribution of cases supplied by the edit process.

The programme prints a report for each work unit showing:

1. Total number of records.

2. Number of records with one or more errors, and as a percent of total records.

3. Total number of errors and as a percent of total records.

4. Total number of housing records.

5. Number of housing records with one or more errors.

/6. The

6. The total number of housing errors and the distribution of housing errors by type.

7. Similar numbers and distribution of population records.

In addition to the diary by work units the programme is able to produce a record of the content of the input record before edit and an indication of which variables have been changed and their value after the change. This record-by-record report may be shown for all records, for the records, for a housing unit and its occupants when any allocation has been made for the group. Or it can suppress the report for the separate records.

The record-by-record report may be reviewed manually. In most cases it is found that the computer revisions are satisfactory. When they are not satisfactory a supplementary set of cards may be prepared for these housing units and their occupants. These records may be inserted in the basic record in place of the defective records.

Unfortunately, we have not been able to construct an edit programme that is actuated by parameter cards in the same manner as the CENTS programme. We have taken account of the desirability of facilitating the process of adapting the programme to other countries. We have made very liberal use of subroutines and macros. The variables are referenced with mnemonic lables. Age is referenced as AGE, (or should it be EDAD?). Only minor changes are required to adjust for the fact that the variable is in a different location in the input or output record. The Computer Methods Laboratory of the U.S. Census Bureau would be able to adapt the programme to the census of some other country in a small fraction of the time required to prepare the Nicaragua programme.

The edit process is performed by reading the record for the housing unit and for the members of the first household in the unit, and process this first household. Then it reads records for the next household and processes it. The record for the housing unit is edited after a new housing unit is encountered. Then the record for the housing unit and all of its occupants are written on tape. By this method we are able to check characteristics of related persons for consistency and to use relationships between persons in the Hot Deck process. For example, when a wife of head is present, the missing age of head is supplied by assuming that his age bears the same relationship to that of his wife as that of the preceding head with wife present. Occasionally this approach leads to inconsistent results and must be rejected. For example let us assume

/that a

that a 20 year old head with a 50 year old wife is followed by a head with unknown age and a 25 year old wife. The 20 year old is 30 years younger than his 50 year old wife, but the second head cannot be 30 years younger than his 25 year old wife.

I should like to outline some of the edits that are performed by the programme.

Sex

Sex is obtained by allocation when it is not reported using the following rules,

Assign as female if reporting one or more children ever born or if reported as wife of head. Assign Head as male if wife is present in household. Otherwise assign sex to head from Hot Deck taking account of presence of own children in the household, and of marital status if no children are present.

Assign sex to child if head from Hot Deck. When sex of children is not shown it is likely to be missing for more than one child in a household. Therefore the Hot Deck for this item contains the sex of the last five children. These values are rotated to avoid strings with the same sex.

Otherwise, assign sex from Hot Deck taking account of relationship to head, age and marital status.

Age

The review and edit of age is among the most complex in the programme. A reported age of less than 10 years for head or wife of head is canceled. In corresponding manner a child of the head may be only 15 to 49 years younger than the head and the parent of the head must be at least 15 years older than the head.

Assign age of wife from age of head and difference between age of previous head and wife. If this fails to yield valid age, assign age of wife from age of oldest child and difference between age of previous oldest child and wife. If this fails, assign from previous wife.

Assign age of head from age of wife and difference between age of previous wife and head. If no wife, assign from age of oldest child and difference between age of previous child and head. If no child, assign from previous head by sex and marital status.

/For others

For others attending school, assign age from grade attending and difference between age and grade of previous person attending school.

For children of head not attending school, assign age from age of wife (or head by sex if no wife) and difference between age of child and previous wife or head. As in the case of sex of child, these differences are retained and rotated for five cases to avoid strings of repetitions.

For parent or grandchild, assign age from age of head and difference between previous parent or grandchild and head by relationship. For all others and when the previous procedures fail to yield a consistent age, assign from previous person by relationship and assumed age.

Literacy

If literacy is not reported, assign as literate if education is 3 years or more; as illiterate if not.

The subjects covered by the housing edit include the following:

Occupancy status

Change occupancy status to be consistent with presence (or absence) of person records for the unit if inconsistent.

Persons in unit

Change number of persons in unit to agree with number of persons records when inconsistent.

Rooms and bedrooms

Hot deck number of rooms from previous unit by type when neither number of rooms nor number of bedrooms is reported.

Hot deck number of bedrooms from number of rooms and difference between rooms in a previous unit.

Plumbing

Hot deck water supply by type of structure when no plumbing items are reported.

Hot deck water supply from previous unit with toilet or bath for exclusive use, shared toilet or bath, or none when not reported or inconsistent.

Hot deck toilet facilities and bathing equipment from previous unit by water supply.