CENSUS DATA EDITING

1. INTRODUCTION

One of the major challenges facing National Statistical Offices (NSOs) is the provision of good quality data – reliable, timely and relevant and this could be achieved by having effective and efficient Data Editing or Data Quality Control measures in place throughout the different phases of a census.

There is increased demand for evidence-based information collected through well-organised censuses and the challenge is providing these evidence-based data for the users to make relevant and reliable inference of the data.

Data editing is the process of detecting errors during data collection and processing and rectifying and correcting these errors to ensure the data are accurate and reliable. Data editing evolves around standard data quality control procedures that are normally followed, whether it be during questionnaire design, pilot testing, field enumeration, office editing or the actual data processing phase to ensure good quality statistical information are collected and captured.

There are two types of errors normally encountered during a census:

A. Coverage Error
B. Content Error

2. COVERAGE ERROR

a) These are typical errors that occur in the field during the enumeration phase and some common errors include omissions of persons and households, incomplete questionnaires, duplications of persons or households, wrong categorisation of persons, loss of information/questionnaires and so forth.

b) Since they are coverage errors, they are usually resolved in the field where enumerators or supervisors will have to revisit the problem areas to correct the errors.

c) One tool used to minimise these errors is the undertaking of a Pilot Census where the different phases of census is practically tested out like the testing of logistic and administration operations, questionnaire design, intensive enumerator/supervisor training and field operations, data processing and editing.

d) Doing a Post Enumeration Survey (PES) is another way to evaluate and measure the extent of these coverage errors.

3. CONTENT ERROR

Content errors arises from poor reporting and recording of information by the enumerators during interview, poor communication between enumerator and respondent, poor questionnaire design, incorrect coding, poor data entry design and erroneous in editing and tabulation processing. This paper will focus in details on the standard quality control and data editing procedures that are used to resolve these content errors.

A. Questionnaire Design (main questionnaire / listing)

1) In terms of data quality control, the design of the questionnaire largely determines the quality of the data being collected. How the questions are worded, formatted and structured ensures collection of good reliable and relevant information.

2) All instructions for each question should be clear and understandable to assist the enumerators while they are communicating the questions with the respondents and ensuring the relevant information required for each question is recorded. Having the questions translated in the local language helps the respondents fully understand the questions.

3) All skips and filtering questions need to be clearly defined to allow for the information to be accurately captured. Most questions are structured so that there is consistent flow of information and any inconsistency in the skips tends to allow for relevant information not to be captured during the interview or data entry.

4) Questions should be relevant and adheres to international standards, definitions and classifications. A lot of countries are still following the broad standard classification and this need to be modified into country-specific standards to accurately define the correct classification.

5) Pre-coded questions are recommended in questionnaires as this helps the enumerators to accurately determine the correct codes for the answers from the respondent.

6) It is highly suggested to allow the involvement of Data Processing (DP) programmers in questionnaire design. Since the DP programmers will be designing the data entry

screen, they should have a good understanding of the skip questions, the inter-relationship and links between questions as well as the determination of the indicators from the information that needs to be collected. Other information on the data structure (dictionary), data formats and data codes are relevant information required by the DP specialist to ensure the information is accurately captured during data entry.

7) Coupled with DP programmers, it is also suggested to involve the GIS people, especially in the listing stage, as they have a good understanding of the geographical fields to be included and in what format these fields should be. Having well-defined geographical fields on the questionnaire allows correct and accurate mapping of data to maps, whether it be at EA, Household or small-area locality levels. Previous experiences have shown that a lot of time is spent on editing and verification of these geographical fields against the data and maps and these could be minimised with the assistance of the GIS people.

8) Consider minimising response-burden where the number of questions had to be weighed against the respondent's willingness in giving correct information. Correct information tend to be lost when the respondents feel tired and burdened with the amount of questions as they tend to answer anyhow so as to quickly complete the interview.

B. Coding

1) Having a code list available for the enumerators as well for the coders ensures information are correctly categorised in their relevant codes. Codes needs to be well defined with clear descriptions.

2) As mentioned in points 5 above, pre-coded questions are always recommended to assist the enumerators to correctly determine the correct answers for each response.

3) Classification codes (ISCO, ISIC, Education etc.) had to be well defined based on country specifics and requirements. Some countries uses the standard classification (with broad categories) directly and redefining these broad categories to detailed country specifics will give a more relevant and reliable indicator of the specific categories.

4) Common mistakes arises in the miss-coding of classifications like occupation and industry where the descriptions recorded are not clear, hence, coding of these variables to ‘Others’ or ‘NEC’ category or they are just lazy to go through the code list.

5) Unfortunately, CSPro is not so powerful enough for automatic coding but could be done for simpler coding during data entry where a list is shown by typing a common word and the correct code is selected from the list.

C. Data Entry

1) A very efficient and intelligent data entry screen is designed based on a well-designed questionnaire where all questions are well structured and skips patterns and filtering questions are clearly defined.

2) Screens are designed with range checks incorporated to prevent invalid entries, as well as consistency and logic checks to ensure accurate entries in specific fields. In whatever capturing mode is being used, whether it is manual, CAPI or scanning, in-build edits are used to ensure correct entries are entered and captured.

3) In CAPI, the data entry screen is designed to allow for all necessary checks and edits to be incorporated in the screen to assist the interviewers in their verification processes with the respondents while doing the interviews. Compared with manual data entry using paper-based questionnaires, the amount of consistency checks to be done has to be weighed against the time needed to complete data entry – the more checks the slower the data entry will be as the data entry operators will be spending more time on resolving the errors.

4) External datasets could be linked to the data capturing application where data captured could be checked against these external datafiles for verification and consistency.

D. Double Entry

1) Double entry is a form of verification process where the questionnaire are punched twice by different data entry operators to ensure whatever is recorded in the questionnaire are correctly captured also in the data.

2) With double entry, it takes more time to complete data entry process, hence, minimised edit checks are incorporate in the system to quicken the process.

3) A balance needs to be established on the time factor as well as the quality of information being captured, therefore, the choices is to either do away with double entry and put in more consistency checks or do double entry with less checks and spend more time in doing batch editing later.

E. Edit Specifications

1) Having the edit specifications prepared well in advance allows for the DP programmers to design the data entry screen and the batch editing programs efficiently.

2) These specifications has to be prepared and provided by the subject matter people who are well versed with the definitions, classifications and the different combinations of the responses needed to determining a required indicator. Current trend with some countries is that they tend to assume that the DP programmers know these specifications but it is highly recommended that these specifications are provided by the countries.

3) Specifications has to be well defined on what variables to be checked, the skip and filtering questions and the consistency and logic checks that is required for good quality data to be processed. The use of flowcharts gives a clearer visual of the checks and edits to be done for each question.

F. Batch Editing

1) This is a crucial step in the editing of census or survey data as this involve editing at a more detailed level based on the edit specifications provided like checking and rectifying the inter-relationships between variables as well as invalid and inconsistent responses.

2) This step also involves the editing of missing and not-applicable answers where data are imputed for these ‘unknown’ information.

3) It is recommended that imputations be done at a minimum level during the editing process as there is a tendency to over-impute and causing the data to be irrelevant and unreliable. This is the reason why some detailed checks are done during data entry as they could be verified with supervisors and subject-matter people.

4) This batch editing process always take longer to complete as some edits tend to introduce new errors, more imputations, recoding and a lot of time spent on referring

back to questionnaires if need be so the onus is to have a thorough and effective batch editing application based on sound and well defined specifications.

G. Table Specifications / Tabulations

1) Similar to the edit specifications, a tabulation plan has to be provided by the subject matter people on the required output needed after the data has been cleaned and finalised. The structure and format of the tables and the variables to be used for cross-tabulations are part of the table specification requirements.

2) The specification has to clearly define the variables for cross tabulations, the variables used for universe filtering and the relevant formulas to be used to compute required indicators. Labor force and Education tables are good examples of these variables.

3) Generation of tables tends to be an effective way of doing quality check control as it will show any inconsistency in the data when cross-tabulating variables and this is then rectified in the data before the finalised tables are generated.

4) Tabulations also identify missing and not-applicable fields and how they are treated in tabulation totals as well in ratio and percentage calculations.

5) Macro-edit is the process of checking aggregates in the tables and this is essential as this will confirm the range checks, consistency and logic checks are effective during the data entry and batch editing stage. Once tables are generated, it is always advisable that subject-matter personnel thoroughly check the tables and make sure that generated figures in the different cells are consistent with other tables as well as other sources, making sure the information generated correctly indicates the current or expected trend.

H. Documentation

1) This is a common disease throughout the region where there is poor documentation of what was done throughout the whole census undertakings. Some countries still experience the same issues and problems encountered in the previous census as there is no documentation of these and how things were addressed.

2) The IHSN documentation tool is very useful tool as it allows you to document everything from the beginning of the survey right to the reporting stage. The big advantage is that it does not only document the administration, field operations, data processing and Report writing aspects, it also allows for metadata documentation where

the data could be viewed and summarised by providing simple statistical measures like frequencies, mean, standard deviation and so forth.

3) Each country has to make every effort in taking ownership of this documentation activity, as it is a good source of information for development and improvement work in the future.

4. SUMMARY / CONCLUSION

1) The focus of every census undertaking is that at the end of the day, we have GOOD QUALITY DATA with reliable and relevant information for use by policy makers, planners, decision makers, researchers and other users. The demand for evidence-based information will always be a challenge and this can only be achieved by having good effective data quality controls and editing processes in place throughout the census phase.