APPLICATION OF INFORMATION TECHNOLOGY IN THE POPULATION AND HOUSING CENSUSES IN SRI LANKA

H.R. GUNASEKERA

DIRECTOR

DEPARTMENT OF CENSUS AND STATISTICS - SRI LANKA

Paper presented at the Expert Group Meeting on Effective Use of IT in Population Censuses

ESCAP, Bangkok

December 10-12, 2007

Application of Information Technology in the Population and Housing Censuses in Sri Lanka

1. Introduction

Sri Lanka has a long history of Census taking. The first of a scientific series of decennial Censuses was conducted in 1871. Twelve regular Censuses were held thereafter up to 1981, establishing a well defined Census methodology. But the Census due in 1991 could not be held due to the disturbed conditions prevailing in the Northern and Eastern provinces of the island. The last Census in 2001 has special significance as it was the first Census to be taken in the new millennium after a lapse of 20 years from previous Census, which is the largest time gap recorded for an intercensal period in the history of Censuses in Sri Lanka. However, Census 2001 was able to carry out completely in 18 out of 25 districts due to the disturbed conditions in Northern and Eastern provinces of Sri Lanka.

This paper discusses the experience in the application of IT for the Census 2001 and strategic plans that need to be implemented for the next Census to be carried out in 2011, based on the lessons learnt from last Census.

2. Census Methodology

Census 2001 was carried out in 4 stages viz Mapping Operation, Listing Operation, Preliminary and Final Census. To avoid omissions or duplications of building units, maps were drawn manually at smallest administrative division level (Grama Niladhari Division) with clear identifiable boundaries. Such divisions were further divided into Census blocks. A Census block comprised of around 60 building units in the rural sector and around 80 units in urban sector. The Census block boundaries were also marked on the map during the mapping operation. Based on the maps prepared, all the building units were listed in a separate form during the listing operation. All building units were classified into housing units, collective living quarters, institutions and non-housing units and listed in convenient and suitable order.

During the preliminary Census, enumerators collected information pertaining to every individual who usually live in the household. Housing information were also recorded. The final Census night, between 6.00 p.m. and 12 mid night the enumerators visited all the units and verified the entries made in the schedules during the preliminary Census. Special arrangements were made to enumerate people staying outdoors on the final Census night. In Sri Lanka Censuses up to now the method of enumeration is based on 'de facto' basis.

In 2001 Census, two types of schedules were used. The Population and Housing Schedule and Disabled Schedule. Information on migration, fertility and housing were collected on a complete enumeration basis, whereas these were collected from a sample of Census blocks in 1981 and 1971 Censuses.

3. Experience of Applying IT in Census 2001

Sri Lanka has been adopting IT to Censuses since 1970's. In the earlier years, IT was mainly applied in processing the Census data. During the Census conducted in 2001, IT applications have been widened to areas such as data dissemination and GIS applications. This is in addition to improvements in technologies in the processing of Census data.

3.1 Data Capture

Data processing activities were centralized at the Data Processing Division of The Department of Census and Statistics (DCS). Processing was done using IBM S390 integrated server 3006 model B01 with several clusters of modern personal computers. The software, Integrated Micro Computer Processing System (IMPS), developed by U.S. Census Bureau was used for all the data processing activities.

Census Schedules were subjected to manual editing and coding prior to data capture. Manual editing was confined to simple checks such as validation of area identification codes and ensuring some selected questions have been coded correctly. Coding was required only in respect of three questions i.e. educational attainment,

occupation and industry. Occupation and industry coding was very laborious and time consuming, although the staff were well trained.

Data capture method used was the centralized key-to-disk procedure. Data entry program allowed for on-line range checks and produced summary statistics for the operator and batch. Dependent verification was done on 100% basis at the beginning and gradually reduced to 10% as the data entry operators get used to the job. Data entry was done under 3 record types i.e. Person, Housing unit and Household, with a length of 96 characters. This manual data entry procedures demand significant resources such as large staff with equipment and space and caused considerate delays in data processing.

3.2 Data Editing and Imputations

Computer editing was carried out by CONCOR component of IMPS. This system consisted of structural edit, range edit and a consistency edit. The error records were flagged and error print outs were sent to subject matter specialists for the necessary corrections. The corrections were sent back to computer division and fixed by the data entry operators. This was identified as a major cause of delay in releasing the timely data. Carefully selected automatic imputations were carried out to fill for missing data and /or to correct erroneous inconsistent data. Imputations were done to the edited data and summary reports were obtained for control purposes. It was felt that the imputations could have little expanded, of course with utmost care, in order to expedite the process.

3.3 Census Mapping / GIS Technologies

Enumeration area maps are vital for a successful Census enumeration. Each enumerator should have complete information about the area to be covered and the location of each housing unit to be interviewed. These area maps were developed for the first time in 2001 Census using the base maps available from Surveyor Generals Department. Two types of maps, i.e. smallest administrative area (GN/Ward) maps and Census block maps were developed. Each GN/Ward maps showed the location of all Census blocks in that area. Each Census block map showed the boundaries, streets /roads, lanes, landmarks and location of every building unit to be visited by

enumerators. Maps prepared for GN Divisions were digitized after the Census. These digital maps are being extensively used for data presentation purposes. Several Atlases were produced based on the digitized maps. Web based interactive mapping down to the GN level is now possible through the DCS web-site. However, it was found that certain maps digitized are deficient in terms of identifying the boundaries properly.

3.4 Data Dissemination

Census results were released in 3 phases; preliminary, sample and final. All such information were disseminated through printed and electronic media as well as via internet. Most of the information including the data for districts were disseminated through user-friendly interactive CD-ROM's. The data bases were prepared in ACCESS and interfaces were developed using VISUAL-BASIC. These CD-ROM's contain all the tabulations contained in published reports and some information at lowest administrative division (GN) level which are not included in the printed publications. Statistical tables in CD-ROM's were classified into broad categories for easy reference under population and housing characteristics such as age-sex composition, education, economic activities, housing stock and household activities etc. Users can select the necessary tables under each such category.

4. Plans of Applying IT for Census-2011

Based on the experience and lessons learnt from Census 2001, the following application of IT are under consideration for the next Census to be carried out in 2011.

4.1 Census Data Capture

Timeliness is a major concern for releasing Census data. As pointed out in 3.1, manual data entry is one of the major cause for delays in Census data processing. In addition to delays, data entry can be a potential source of error unless carefully planned verification procedures are adopted. Application of image based form processing technology such as OCR/ICR has been a salient feature of the IT plans for

the Census 2011. Before implementation of any such technology, greater attention should be given to examine its suitability. The success depends on the extent to which the necessary conditions prevail such as the quality of the specially printed questionnaires, training given to enumerators in recording the responses clearly and of course on the selection of technology.

A pilot study is currently being carried out to examine the suitability of form processing technology. Completed schedules from a special enumeration carried out in recently cleared Eastern Province of Sri Lanka (3 districts) are being processed under this study. After interacting with various vendors and going through the demonstrations of their products on automatic form processing the work has been outsourced. The results are encouraging so far and the whole procedure has to be critically evaluated.

The possibility of decentralizing the data capture at regional level will be explored. This needs substantial allocation of resources at district level in terms of equipment and human resources.

4.2 Computer Assisted Coding

Although most of the items of the Census questionnaire are pre-coded, certain items need to be coded before computer processing begins. Even the form processing technologies such as OCR / ICR methods are used, this type of coding needs to be carried out prior to scanning. Occupation and industry are two items usually need to be coded. Past experience reveals that this is laborious, time consuming and can be an important source of error. Software can be developed to assist the coders in finding the most suitable code for particular occupation or industry classification by matching the description given through the keyboard to possible set of codes in the list. The matching has to be done for characters in local languages. This computer assisted coding system (CAC) will greatly reduce the time taken for coding and the risk of having errors.

4.3 Computer Editing and Imputations

As discussed in 3.2 the exchange of error print outs between Data Processing Division and subject matter division caused considerable delays in processing. It is decided to explore the possibility of introducing on-line editing procedures in the next Census, in order to improve the efficiency and avoid unnecessary delays. Edits to be done on-line should be carefully identified as over ambitious editing procedures may cause the so called corrected data to be significantly flawed. This on-line error correction is to be done by trained staff of the subject matter division as they have a good understanding of data.

Computer imputations should be minimized as far as possible; but lessons learnt from Census 2001 indicate that carefully designed imputations are needed. Hot-deck imputation methods can be explored. This technique uses information obtained from previously processed records with similar characteristics as the 'best fit' value in replacing missing values or values that have failed satisfying edits.

4.4 Census Mapping / GIS Technologies

Pre- Census mapping activities have already commenced for the Census 2011 and expected to be completed by mid 2010. The objective is to prepare a digital map for each Census Block with all geographical features. Satellite images are used as base whenever possible to create/ update maps. Global Positioning System (GPS) technology is effectively utilized in the mapping operation. A simple, hand-held GPS receiver gives latitude and longitude co-ordinates with reasonable accuracy of key points. This is no doubt, a significant improvement for correctly identifying the boundaries of Census blocks. Such information is also valuable in developing GIS for later use. During the mapping operation active co-operation is obtained from all relevant key stakeholders.

4.5 Data Dissemination

Currently, Census information is disseminated through printed as well as electronic media and also via world wide web. Information in the internet is limited to pre-defined sets of tabulations. This static format will be continued in future for tabulations which are in heavy demand. For the users who need more detailed data than static web pages, provision of interactive tabulation system will be explored. In addition interactive viewing of maps to enable a thematic view of information will be very useful for data users. However confidentiality of data should be maintained in order to safeguard the privacy rights of the public and a clear policy need to be established. One long term objective is to establish 'One Stop Service System' where Census data together with administrative and survey data would be aggregated for policy relevant analysis.

4.6 E-Census

Sri Lanka uses more traditional approaches to collect data in the Census. Usual practice is the enumerator to visit the household and completes the questionnaire. The past experience shows that the most affluent people in urban areas, particularly in capital city of Colombo, do not cooperate well with the enumerators in completing the questionnaire in spite of the wide publicity campaign. Therefore it is planned to collect Census information from a limited number of households with such affluent people in Colombo city. This will be accomplished through self administered questionnaire posted on internet. Such households will be selected at the time of house listing operation prior to the Census by asking their consent. Households opting for electronic reporting could download their e-questionnaires (e-Q) from an internet application and return them after on-line completion. The e-Q is supported by data encryption during transmission and whilst held on the web server an unique password is given for individual e-Q's to safeguard data confidentiality.