Census Data Processing in Kenya

In preparation for the

Census Data Processing Workshop for the contemporary technologies for census data capture and data editing, Dar Es Salaam, Tanzania, 9-13 June 2008

TABLE OF CONTENT

INTRODUCTION

The 1999 Population and Housing Census was the fourth census to be carried out since independence and the sixth since 1948. The enumeration exercise was successfully carried out between the night of 24th/25th and 31st August 1999.

The main objective of the census was to collect demographic and socio-economic data required for policy formulation, development planning and research. Being the last census to be carried out during the 2nd millennium, it was important that the data collected be comprehensive, complete and accurate in terms of coverage and quality for policy formulation and development planning at the beginning of the 3rd millennium. The success of the 1999 census was an upshot of the administrative and methodological procedures that were used during the implementation of diverse activities after taking recognizant of the experiences and lessons learnt during the 1989 Census.

DATA PROCESSING

The processing of data collected in a census constitutes one of the most important and challenging activities that have to be undertaken efficiently and expeditiously in order to justify the immense resources invested in a census. This activity entailed several processes: manual editing of the questionnaires after enumeration, data capture, data cleaning and validation, and finally tabulation.

Government's commitment to provide provisional results within six months after enumeration and final basic results within another six months greatly influenced the strategies and actions adopted at every stage of data processing in order to adhere to the commitment.

RECEIPT OF QUESTIONNAIRES FROM THE FIELD

The process of documenting the questionnaire books used to enumerate persons in conventional households and short questionnaires for special population started at EA level. Each enumerator completed a control form showing the serial numbers of the books he/she had used. This process was replicated at all levels i.e. sub-location, location, division and district. The District Census Officers utilized the above documentation to verify that all the books that were allocated to the district had been accounted for before dispatching the same to the Census Office.

The subsequent stages involved placing of questionnaire books in boxes starting with the first EA in the first division of a particular district; numbering the boxes and labelling each box indicating district, division, location and sub-location names and codes; number of books and EAs contained in a specific box. This process was systematically followed until the last EA in the last division had been sorted out.

The following was done at reception stage:

- Receive the questionnaire booklets for conventional households and the short questionnaires for vagrants, travellers, hotel lodgers, and refugees from each district.
- Record the number of books returned for each EA.
- Note any EA returning a zero population and the reasons given.
- Hand over the short questionnaires to the data entry operators for keying.
- Note any deviations from the geographical file, e.g. several EAs collapsed to form a single EA, or extra EAs created in the field during the census enumeration.
- Place the questionnaire booklets in boxes in order, label each box with the names of the district, division, location, and sub-location, the sub-location

code, number of books, and the EAs contained in the box. Number the boxes in ascending order of EA code.

- Send the boxes to the store.

The reception process should have been handled by a computer-based system but instead handled by manual process, which were of limited utility in controlling other aspects of the data capture and cleaning operations.

MANUAL EDITING

Manual editing of questionnaires was an enormous task which involved ascertaining that identification codes, numbering of households and number of individuals enumerated in each household had been correctly done. The staff were also utilized to reinforce faint codes on the questionnaires and sometimes transferring of information to new questionnaires when they could not be scanned. The number of personnel involved in this exercise ranged between 200 and 1,200 between October 1999 and December 2000.

This exercise posed a number of challenges. The working space was insufficient resulting in overcrowding and hence poor supervision. It increasingly became difficult to have an efficient tracking system as far as the movement of questionnaires from storage to editing rooms was concerned. This sometimes resulted in the mixing up of questionnaires belonging to different EAs.

The questionnaires undergo a limited number of manual checks as follows:

- Confirm the EA code on each completed questionnaire in each booklet.
- For each booklet of questionnaires confirm the population summary counts by sex on the cover.
- Edit the household numbering, household type, housing records, etc. according to the specifications that were provided by the demographers.
- Where an EA was enumerated by more than one enumerator re-serialize the households if necessary.
- Transfer the data from mutilated or dirty questionnaires to fresh ones.
- Reinforce the writing on questionnaires with faint writing.
- Liaise with the cartography section, data processing staff, and the district statistical officer on cases requiring amendment of the geography file, e.g. several EAs collapsed to form a single EA, or extra EAs created in the field during the census enumeration.
- Confirm population and household totals for each EA that were compiled in the field.

SCANNING OF QUESTIONNAIRES AND ARCHIVING

Scanning technology technically known as Intelligent Character Recognition (ICR) was adopted for the 1999 census. The idea was an upshot of an initial attempt made to process data collected in the Welfare Survey in 1996/97, the experiences gained during the processing of the 1989 census data which took about 2 years and Government's determination to release the census results in the shortest possible time. The choice of the technology influenced the design of the main census questionnaire. The exercise was planned to start in March 2000, but was delayed until June 2000. It took 3 months to accomplish. The recognition

rates varied between 95 and 98 percent, and were greatly affected by factors such as print quality and the type of pencil that was used to fill the form.

Equipment

Two types of scanners were procured: Kodak models 3500 and 9500. Other equipments included 110 micro-computers, servers and printers.

Software

The software AFPSPRO was provided by an Israel Company TIS. The company had successfully utilized the software to process data collected in the Turkey census of 1997. This software had ability of recognizing scanned characters, processing of the images, transferring of the data to servers, transformation of the data from images to ASCII data and archiving of the data for further editing and validation.

Stages involved in Scanning

The following stages were involved during the entire scanning process:

Guillotine

Binding of questionnaires into booklets during the census was to minimize the misplacement/mix-up that would hinder faster batching of the EAs. After the editing, the questionnaires were detached from their covers and cut/trimmed using guillotine machines. This was to ensure that they were of the same size with the questionnaire images stored in the system memory where the scanned images would be recognized and interpreted.

Although instructions and specifications were explicitly given, some problems such as wrong trimming and mixing of questionnaires from different EAs were experienced. These problems delayed the scanning process. In essence, a good number of scanned batches were rejected in the processing stage due to wrong cutting. The documents were checked for EA codes and consistencies of

household numbers before they were handed over to the teams which were charged with the responsibility of feeding the questionnaires into the scanners. The steps were as follows:

- Detaching the covers from the questionnaire booklets and placing them in the bottom of the box.
- Removing and discarding unused pages from the booklets.
- Removing the glued edge with a hand guillotine.
- Securing the now loose questionnaires for each EA by placing them in an envelope or with a rubber band.
- Returning the boxes of trimmed questionnaires to storage.

Preparation

The forms are prepared for scanning. The process was as follows:

- Check that the geography codes on each form are correct.
- Ensure that all the household numbers are in order in each EA.
- Check that all the EAs indicated on the label are in the box and make notes of any missing EAs, or EAs that have been wrongly included in the box.
- Make a note of any misplaced boxes.
- Use a movement register to forward the prepared boxes to the scanning room.

Scanning

A bit of fanning was done to make sure that all documents were separated to allow easy flow. Four staff members were involved at this stage at any given moment: one oversaw the feeding of the document at the input hopper, one observed and controlled the images on the screen, the third recorded the scanned batches in the batch distribution schedule and the last cut the labels from the printer and attached them to the scanned batches. A batch contained on average between 50 – 150 questionnaires.

Although the scanning of questionnaires was initially planned to be completed within 2 months, the process was however, slowed down due to the following problems:

- **Poor Preparation** – The mixing of EAs before the documents were passed over to scanning personnel and poor reinforcement on the pencil marks resulted in rescans. The questionnaires found illegible even after adjusting of the parameters had to be reinforced. The cutting of questionnaires was poorly done. Some questionnaires ended up with coastline-cut-edges while others were cut beyond margins. The latter had to be transferred to new questionnaires during scanning.
- **Rescanning** – Many of the physical batches returned for rescan were found not to match ASCII batches of the same in the system. A search for the correct ones had to be carried out. This verification of the batches in the system and in the cartons, further delayed scanning. The mix-up was as a result of the poor batching method of binding questionnaires with rubber bands on which the label for each batch was attached. The bands sometimes got broken and the questionnaires got mixed up.
- **Batch labelling** – the poor maintenance of the printers introduced errors when they failed to print the labels. Handwritten labels were introduced in the process. Some of the handwritten labels were incorrect. Tracing of these errors (for rescanning) had a drawback in time factor.
- **Lack of trained staff** - CBS personnel had limited knowledge of the scanning software (AFPSPRO) and maintenance of the equipment. This contributed to low production while waiting for knowledgeable people to come from Israel and maintenance engineers.

- **Flow of households during scanning –** failure to scan households in their order of serialization resulted in introduction of errors during cleaning of data. New or extra household numbers were introduced while correcting household numbers.

The above errors were observed during and after scanning process. To prevent the occurrence of such errors in future scanning, the following recommendations were made:

- All requisite equipment should be put in place before actual scanning is started.
- Questionnaires should be distinct, legible and un-bleachable or able to last more than 3 years.
- The paper that should be used for printing of the questionnaire should be predetermined by the scanner manufacturers to enable the scanning job to have quality output and easy to operate.
- During dispatch of the questionnaire to the field, a proper record of distribution by serial numbers should be maintained for future reference. In cases of translocation, this would come in assistance.
- Sooner after all the questionnaires are received from the field, a microfilming of the same should be done to retain the original data for any other reference.
- Batching of the questionnaire should be done during the preparation: each batch should be banded and thereafter put in an envelope.
- Training of supervision team on the scanning and network software is paramount.
- The fast wearing away spares should be brought in together with the rest of the equipments.
- No handwritten labelling of the batches should be allowed and, the labels should be pasted on the envelopes containing the batches.

- All work should be scanned sequentially to eliminate introduction of errors through guesswork during cleaning.

Irregular Batches

All the misplaced questionnaires, found thereafter were scanned and stored in the same parent carton with the rest of the documents belonging to that sub-location, although considered out of range.

Double Scan Observations

When two or three documents went through unnoticed and were scanned as one document, they introduced a double scan. The image observed would show the front view of the first and the rear view of the second or the third document depending on how many went in at any time. The double scan resulted to missing households in a batch though not missing physically.

Box Labels

A box contained a number of EAs comprising of many batches. The scanned batches were labelled sequentially. The first and last batch labels were fixed on the box to indicate the range of batches in that box. This was to assist in editing during completion and cleaning stages. When wrong or inconsistent geo-codes were found during scanning, the whole box was returned to preparation section for correction.

Batch Control Sheet

The batch control sheets contained columns for entries of batch, box, EA and geo codes. The records were used for reference in case of rescanning and verification during cleaning.

Rescanning

Rescanning of batches occurred mainly when batches could not be traced during subsequent processing stages or the contents could not be easily read or recognized. When the re-scans were done, new batch numbers were generated and the old ones deleted from the control stations/archives. The errors observed were caused by the following:-

- Direct rejection by the processor computers – due to poor quality of paper/poor cutting.
- Corrupted batches – blanks and asterisks – due to poor recognition of characters and corruption during transfer from AFSPSPRO to ASCII.
- Mixture of questionnaires of different EAs – observed during cleaning.
- Double Scan – scanning of two or more documents as one.
- Change of geo codes – a number of changes were done on part or whole contents of a carton. The batches with changed geo codes had to be rescanned.

COMPLETION

In this stage completion operators check and enter fields or characters that were not recognised by the character recognition software. There were three shifts with a total of 175 completion operators and 16 supervisors.

The AFPSPRO software displays the image of the field, or of the entire questionnaire, and prompts the operator to enter the highlighted fields. Range checks are included for each field but no consistency checks are carried out. Fields for which the operator cannot determine a suitable value are sent to the exception stage for handling.

EXCEPTION

In this stage operators attempt to deal with unresolved cases from the completion stage. This stage involves referring to the original documents.

All the unresolved cases from the completion stage are handled in this stage. The resolution sometimes involved referring to the original questionnaires that were held in temporary storage. One of the major problems was the length of time taken to obtain a batch from temporary storage, and in some cases the batch was never found. In such cases the supervisors used their own judgement to resolve the problem.

LESSONS LEARNED

Several lessons were learnt during the implementation of various activities pertaining to this important component of census operations. The following are the key lessons learnt:

- Having adequate storage facilities to accommodate the voluminous number of questionnaires is vital as it facilitates easy retrieval and monitoring. The lack of storage space, racks and cabinets resulted in loss of substantial time when tracking down the whereabouts of questionnaires belonging to some areas. This constraint slowed down other processes.
- Given the complexity of the processes to be undertaken, it is of paramount importance that staff involved in the implementation of various processes be fully trained. This was a major oversight across board.
- The large number of steps and the overall speed of the processes mean that computer-assisted monitoring and control systems are required. Use of manual monitoring and control systems are inefficient.
- The requisite hardware to support complex software such as the one deployed in this census has to be in place before any data processing activity is embarked on. This requires careful technical evaluation of all aspects of hardware and software.

CHALLENGES

Many challenges were experienced during the processing of the data. Most of them pertained to the complexity of the technology that was deployed. This was further compounded by inadequate number of in-house personnel trained to man various activities and delays in procurement of appropriate hardware. However, in spite of the above constraints, the quality control procedures that were put in place assisted greatly in tackling the challenges. The creation of an easily accessible census data base constitutes an important development as far as the enhancement of utilisation of the data is concerned.

Delay in procurement of hardware and software posed a big challenge. It is vital at this aspect be taken seriously and catered for in advance before undertaking the enumeration exercise. A careful technical study should be undertaken before a new technology is decided upon. Such a study should be able to bring on board the costs and other crucial burdens the use of the technology would impose.

The initial arrangements to process the data collected in the pilot census using scanning technology did not materialize due to delays in formalisation of the contract with the firm which was envisaged to provide the requisite software. However, a sample of questionnaires involving 25,710 households and 127,166 persons was manually processed. This sample was utilized to run the edit specifications and tables. This constraint of not using the scanning technology at this stage posed a big challenge later during the processing of the main census data. It was not possible to assess in advance the limitations and requirements of the software.

The reception process was handled by manual process, which was of limited utility in controlling other aspects of the data capture and cleaning operations.

No tracking system was developed to track the questionnaires, batches etc. through the process

No utility had been prepared to remove blank records.