May 24, 2001

DRAFT

TRIP REPORT

Kathleen E. Chamberlain
Chief, Information Systems Branch
Governments Division
U.S. Census Bureau

Colombo, Sri Lanka
May 14-25, 2001

Trip Purpose and Summary

At the request of USAID/Colombo and the Department of Census and Statistics (DCS), Kathleen Chamberlain traveled to Colombo, Sri Lanka, to review the data processing system for the 2001 Census of Population and Housing. The DCS plans to use the Integrated Microcomputer Processing System (IMPS), developed by the International Programs Center of the U.S. Census Bureau, for data entry, editing, and tabulation of the census data. Chamberlain worked with counterpart S.A.S. Bandulasena to test and modify several programs, combine others, and develop a menu from which operations staff can choose programs to run.

Details of the Trip

The census will be conducted on July 17, 2001, with 100,000 census officers collecting data on approximately 19 million persons in over 4 million households throughout the country. The DCS plans to conduct a *de jure* census during the last week of June and first week of July 2001, returning to the households on census day and conduct a *de facto* census. The census questionnaire consists of 25 questions per individual, 9 questions per housing unit, and 7 questions per household. Multiple households per housing unit are not uncommon in Sri Lanka; the determination being based on the presence of separate cooking facilities. During enumeration, a separate 17-question questionnaire will be completed for disabled persons.

The last census of Sri Lanka was conducted in 1981. Much of the institutional memory has been lost over the past 20 years, particularly in the area of computer processing. Information technology professionals are in great demand and many have left the DCS for higher salaries. For the next several months and throughout the census processing period, it will be important to have adequate programmer support to modify existing programs and write new programs as the need arises.

Chamberlain worked primarily with systems analyst/programmer S.A.S. Bandulasena and Statistical Officer W. D. P. De. A. Gunatilake to review the computer programs and overall system flow. In reviewing and testing the computer programs, she and Mr. Bandulasena made numerous corrections and other changes to the editing programs. (All program changes were made to separate copies of the programs so that the originals remain unchanged.) The data processing plan involves considerable manual review of questionnaires after the editing programs identify errors. Such manual review is reasonable during processing of early batches of data, but should be bypassed as soon as possible in the interest of time. Non responses should be left as non responses, and automatic imputations should be done only where obvious.

A system for tracking and control of census books should be developed. The system should allow for check-in of batches and tracking them through each phase of processing. This will ensure that all batches go through each phase of processing and no batch is processed multiple times. IMPS/CENTRACK can be used for this purpose. Chamberlain reviewed CENTRACK with Mr. Bandulasena and developed a prototype system.

Chamberlain reviewed the tabulation programs and demonstrated for programmer E.A.G. Sarath Perera a strategy for tabulation by geographic area. She and Bandulasena ran a large batch of pilot census data through the imputation program, and made the resulting file available to Perera as test data for the tabulation program. The data provided a more realistic test of the tabulation programs and uncovered several errors in the program.

Chamberlain developed a menu for computer processing of census questionnaire batches, instructing Bandulasena on how to modify the menu to include additional functionality. The menu software is generalized and can be used for other applications

Attachment B contains a description of the census data processing plan and a list of recommended modifications. The system requires further testing and streamlining, and additional menus should be developed to simplify processing and minimize error. Once questionnaires are received from the field and data entry begins, additional changes and further streamlining will be needed.

Recommendation for future visit in support of the Census

It is recommended that an additional 2-week visit by a computer specialist conversant in the IMPS software be scheduled shortly after data entry has begun. The availability of actual data will assist in further streamlining the editing programs and will present a realistic picture of the time required for computer editing and imputation. Estimated timing for this visit is October, 2001.

ATTACHMENT A

Persons Contacted

USAID/Sri Lanka

Address: 44, Galle Road
Colombo 3, Sri Lanka

Colombo - USAID -6100
Department of State, Washington, DC 20521-6100

Phone: 94-1-472–855
Fax: 94-1-472-850/60

Gary Robbins, Project Manager

Department of Census and Statistics (DCS)

Address: 15/12, Maitland Crescent, Colombo 7
P. O. Box 563, Colombo, Sri Lanka

Telephone: 94-1-695-291
Fax: 94-1-697-594

A. G. W. Nanayakkara, Director General (dcensus@lanka.ccom.lk)
(wimalnn@lanka.ccom.lk) (home phone: 883-774)
Suranjana Vidyaratne, Director

Department of Census and Statistics (DCS)

Address: 16/7, Albert Crescent, Colombo 7
(Between Australian High Commission and St Bridgets Convent)
(Across the street from Aesthetic Studies Institute and Nat'l Museum)

Telephone: 94-1-697-595
Fax: 94-1-697-595

Yasantha Fernando, Deputy Director, Sample Surveys Division
H. R. Gunasekara, Deputy Director, Census Operations
W. D. P. De. A. Gunatilake, Statistical Officer
Jaya Nagendran, Statistical Officer
S.A.S. Bandulasena, Systems Analyst/Programmer
E.A.G. Sarath Perera, Systems Analyst/Programmer

ATTACHMENT B

Census Processing Plan and Recommendations

Questionnaire/Batch Identification

Enumerators will visit households with questionnaire books, each book consisting of 40 pages. Pages are printed on both sides, each side accommodating 6 persons in the household. For households larger than 6 persons, and for collective quarters and institutions, the reverse of the page and additional pages will be used. A new book will be started for each census block. In urban areas, a block may require multiple books. Books are identified by the following codes:

1. Province and District (2 digits) -- 25
2. Polling Division (3 digits) – n/a (information only)
3. DS (Divisional Secretary) Division (2 digits) - 316, numbered within District
4. GN (Grama Niladhari) Division (3 digits) – 14,113, numbered within DS
5. Sector – Urban, Rural, Estate (1 digit indicator) - 3
6. MC/UC/PS – Municipal Council, Urban Council, Rural Council – 1 digit indicator
7. Ward/Village/Estate code – 3 digits, numbered within MC/UC/PS
8. Census block number – 2 digits, numbered within Ward/Village/Estate
9. Book number – 4 digits, numbered within DS Division

Within a book, a household is defined uniquely by
1. Housing Unit number, 3 digits, numbered within block
2. Type of Unit (housing, collective, institution, nonhousing, homeless, outdoor)
3. Household number, within housing unit (1-9)

Preliminary Counts

Immediately following the census, population counts by geographic area will be manually tallied so that a preliminary count can be released within a few days of the census.

I addition, summary sheets will be completed in the field offices for the purpose of obtaining more detailed manual counts. The summary sheets contain more information than is common for manual counts and will require considerable time (3-4 months) and resources to tally. For each GN Division, the number of males, females, total, age groups, 6 religion categories, and 9 ethnic categories are recorded, for a total of 20 categories. The number of categories makes the manual tabulation process highly error-prone, and will likely delay the computer processing of the census forms.
Recommendation: The number of variables for the preliminary census counts should be reduced to total persons, males, females, and 2 age groups, particularly if the inclusion of the other variables will delay the delivery of forms to the central office, and thus delay computer processing of the census..

Batching, Review and Coding of Questionnaires

Questionnaire books will be organized by District, DS Division, GN Division, Sector, and MC/UC/PS, and will be processed in this order-- that is, one district at a time, and within a

district, by DS Division and GN Division. Questionnaires will be manually reviewed for serious errors then coded. (Variables such as educational attainment, occupation and industry are entered as text and require conversion to codes.)
Recommendation: IMPS/CENTRACK should be used for tracking of questionnaires through the various operations back at the central office. CENTRACK produces forms that can be used to keep track of the operations that have been completed on a batch of census questionnaires. Information from these forms is entered into the CENTRACK database. The database can then be queried to determine the status of processing. The system also notifies of duplicate and missing batches. Chamberlain worked with Bandulasena to set up a CENTRACK prototype. The CENTRACK database must be initialized with geographic information before being operational.

Computer Hardware

For data entry, there are currently 45 PC workstations , clustered in groups of 15 each. An additional 45 stations will be purchased for data entry, for a total of 90. The current plan is to archive the census data on cartridges (810mb each)that can be read on the IBM mainframe computer.

Recommendation: CD-ROM writers should be purchased and CD's used for both archive and backup purposes. CD read-write drives are now standard equipment on many new PC's but also can be purchased as external units. CD's are inexpensive and reliable. Reliance on the mainframe computer should be reduced and eventually eliminated.

For editing and tabulation, 3-4 computers are available.
Recommendation: If 45 new computers are purchased for data entry, it is recommended that as many as affordable be top-of-the-line computers, and the current older computers being used by programmers become data entry computers while the new ones be used for the more complicated operations.

Data Entry

Data entry is expected to begin in late September or early October. Currently there are 45 work stations, clustered in groups of 15 each. An additional 45 stations will be purchased in the near future. Data entry will take place in 2 buildings at the central office. Three shifts per day, 7 days per week, are planned for the data entry operation. Data entry is expected to be complete within 9 months after start. For data entry the IMPS/CENTRY software will be used. The screens were tested during the pilot census and are satisfactory. During data entry, automatic checking is done to ensure entries are within the valid range of codes. In addition, f edits are automatically run to ensure that the questionnaire has all its required components.
Recommendation: It is recommended that this edit be expanded to include all essential "structure" edits, so that the questionnaire does not require any further manual review after data entry.

Recommendation: Verification, the double-keying of data for accuracy, should be done initially on 100% of the forms, then as keying errors drop off, on smaller samples of forms (i.e. 70%,

50%, 25%, 10%). error rates should be monitored closely, and verification sample rates raised when needed. Verification should be done by a different operator than the one who originally keyed the batch.

File naming conventions

After much discussion, it was agreed that for tracking purposes, files should be named according to their combination of DS Division code, GN Division code, Sector, and MC/UC/PS code, for a total of 8 digits. The files should be further organized by District. Since book number is part of the batch-id, where multiple books make up a GN Division, the DS/GN/Sector/MC would become a temporary directory and book numbers would become the filenames. After all books from a DS/GN/Sector/MC are entered, they would be concatenated into one file and the temporary directory removed. For the main file server there would be 25 District folders (directories), each containing sub-folders for the various phases of processing (original data entry, merged to DS Division, edited, imputed, etc.) All 12 identifying codes listed above would be carried with each record on the file. See File naming example in Attachment C.

Error Detection and Correction

The current plan is to edit data by DS Division following data entry. Each DS file would pass through

1 - an edit program to check for valid codes (Range check)
2 - an edit program to perform consistency edits at the household level
3 – an edit program to perform consistency edits at the housing unit level
4 – an imputation program to perform automatic imputations
5 – an edit program to check occupation code and change invalid ones to not reported
6 – an edit program to check industry code and change invalid ones to not reported

Programs 1, 2, and 3 will produce reports showing the number of each type of error and the questionnaires containing errors. The plan is to examine the original questionnaire to resolve the errors. Programs 4-6 make automatic changes to the data, where probable responses can be derived from other responses. All 6 programs are written in CONCOR, the editing component of IMPS.

Recommendations:

1. All edit programs should use the same data dictionary so that code can more easily be moved from one program to another. Note: this was done by Chamberlain and Bandulasena.

2. It should not be necessary to re-visit the actual questionnaire to resolve errors during this phase of editing. Serious errors should have been detected during manual review, coding, and data entry. The data entry program should incorporate edits for serious errors such as questionnaire structure errors.

3. Program 1 is likely to detect only missing values (blank). This program should be replaced by a program to produce frequency distributions for important variables. The frequency program can distinguish between missing and invalid values, thus providing the same information in a

more meaningful format than the range edit. Blank values should be automatically changed to the code for "not reported" in the imputation program.

4. Inconsistencies found by programs 2 and 3 should be corrected by the imputation program.

5. Programs 5 and 6 should be combined with program 4 if possible. Note: this was done by Chamberlain and Bandulasena, but assumes the occupation and industry code lists are complete. If this is not the case, and the program becomes too large, programs 5 and 6 should at least be combined into one program.

6. All programs should be better documented internally. Comments and meaningful variable names should be used, enabling persons other than the programmer to more easily read the code.

7. Extraneous code (code that is repeated for no reason) should be removed. Chamberlain and Bandulasena did this to some extent, but time ran out.

8. After the edit programs have been revised, they should be tested with pilot census data, then again with raw keyed data from the census before being considered final. After recognizable error patterns are detected, the programs may require further streamlining.

ATTACHMENT C

Suggested File Naming Convention - example

For data entry:

Kalutara (folder) -the district currently being keyed
- DDGGGSMM.BCH - file, where DD is DS Div, GGG is GN, S is sector, and MM is MC/UC/PS
- 03074123.BCH – DS=03, GN=074, Sector=1, UC=23
- 03075121 -directory needed temporarily because of multiple books
  - book0441.bch -file
  - book0442.bch -file
  - .
  - .
  - .
  - book0450.bch -file

Main File Server:

Kalutara (folder)
- Keyed (subfolder)
  - DS03 (subfolder, first DS Division)
    - 03074123.bch - keyed batch
    - 03075121.bch - keyed batch
    - … etc. for all GN Divisions in DS03
  - DS04 (subfolder, second DS Division)
    - 04005121.bch -keyed batch
    - etc for each GN batch
- MergtoDS
  - DS03.bch
  - DS04.bch
  - etc, for each DS in Kalutara
- Edited
  - DS03ed.bch
  - DS04ed.bch
  - etc.
- Imputed
  - DS03imp.bch
  - DS04imp.bch
  - etc.