Distr.
GENERAL

WP.18
29 April 2011

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2011)
(Luxembourg, 23-25 May 2011)

Topic (iii): Innovation and related issues

Data Processing and Data Dissemination Technologies used for 2008 General Population Census of Cambodia at National Institute of Statistics, Ministry of Planning in Cambodia

Supporting Paper

Prepared by Meng Kimhor, National Institute of Statistics, Ministry of Planning (Cambodia)

I. Introduction

1. The National Institute of Statistics (NIS), which is part of the Ministry of Planning, is the focal point on statistical matters in Cambodia. The NIS compiles and consolidates statistics provided by decentralized offices and also collects primary data through household and establishment surveys, and population, agricultural and economic censuses.

2. Cambodia has a decentralized statistical structure. There are statistical bureaus and sections within Planning and Statistics departments of the various Ministries and in planning and statistical units in the provinces and districts.

3. Commencing from 1998, the Royal Government of Cambodia is committed to conducting a General Population Census every ten years in accordance with the U.N. recommendations and the Statistical Law of Cambodia. Such a census would include some aspects of housing census also. Accordingly the 2008 General Population Census of Cambodia was conducted on a de facto basis with reference to 3 March, 2008, exactly ten years after the previous census was held. This census marked the second census since Cambodia become a democratic country and the first of twenty first century.

4. Organizing the IT for the 2008 General Population Census comprised a large array of activities, ranging from pre-census tasks; such as automated mapping, GIS development and EA database design, to post-census activities; like manual coding and editing of questionnaires, data entry, data verification, computer editing, tabulation, the development of various census dissemination products (Table Retrieval System, Community Profile System, Population Database and Mapping Application based on 2008 Population Census) and maintenance of the website.

II. Location-specific information and geographic information systems

5. The work of pre-census mapping commenced in June 2006 with the development of maps at small area levels. To ensure complete coverage of the population, it was necessary to divide the entire country into small enumeration areas (EAs). The intention was to assign an EA to an enumerator so that he/she might enumerate all persons found in that area within the census enumeration period of 11 days. Proper delineation of enumeration areas in each village/town was the first most important step in census preparations. In the absence of precise and detailed maps of scale 1 to 5,000 required for a census, photographs and satellite maps were used as reference material in the cartographic field work.

6. For the pre-census cartographic work, thirty staff members from NIS were thoroughly trained both theoretically and practically in mapping. Village boundaries were checked and measurement of location of each village was taken using GPS. Enumeration areas were demarcated with an average size of about 100 households each.

III. Data Processing and Data Dissemination

7. The Data Processing for 2008 General Population Census of Cambodia was designed based on:

a. Data Capture
b. Data Coding
c. Data Editing
d. Data storage and tabulation
e. Data Dissemination

8. These are described in the following paragraphs:

A. Data Capture

9. The General Population Census of Cambodia 1998 used PC-EDIT for data capture and verification. The CSPro (Census and Survey Processing) package was used for data capture and verification of Cambodia Inter-Censal Population Survey 2004 (CIPS2004) and the General Population Census of Cambodia 2008. The data contained in the records of the 2008 General Population Census of Cambodia numbering more than 2 million (households) were entered into computer by using CSPro software.

10. The NIS took some important decisions regarding the basic concepts of data processing, e.g. it was agreed to centralize data processing and to use 100 microcomputers together with key-to-disk method for data capture. CSPro software was selected for data entry since this package was also used for the processing of the Cambodia Inter-censal Survey data in 2004 and many data entry staff were already familiar with this software. The data entry applications had checked and controlled many errors by performing a number of range and consistency checks during data entry time. In such an approach data quality was considered more importance than data entry speed.

11. Verification of data entry batches was performed in order to minimize typing errors. At the beginning of data entry full verification (100%) was in place, but as work progressed and operators gained more experience the percentage was reduced gradually. In all, some 20% of all data batches were verified.

12. For the processing of the 1998 census questionnaires an extensive productivity-based incentive scheme was established with the intention to boost the output of data entry operators. An automated Keyer's Monitoring System formed the main thrust of the scheme. It measured the performance of each individual keyboard operator and facilitated the production of monthly productivity reports. These reports were used for the computation of the incentives. With such a system in place the overall productivity of data entry operators was quit satisfactory, resulting in an average of approximately 8,000 keystrokes per hour. Hence this scheme was implemented for the processing of the 2008 census returns also.

13. The enumerator’s summary statements were entered into the computer before starting main census questionnaire (Form B) and the Houselist (Form A). This information have provided the basis for the provisional census results (population by sex at national and provisional levels) released on September 3, 2008 presided by Deputy Prime Minister of Royal Government of Cambodia.

B. Data Coding

14. It is generally accepted to minimize and simplify to the highest extent possible the task of manual editing of census questionnaires. This kind of editing is better done on microcomputers using specialized software packages. In fact, as experiences in other developing countries have shown, there is a real risk of office editors introducing new errors.

Nevertheless, an instructions manual for office coding & editing needs were prepared before data entry operation. The instructions were basic and focused on checking the geographic identification of each questionnaire, ensuring all code boxes were properly filled (if not, NR was imputed) and were legible, and the skip patterns were checked..

15. The various coding schemes, required for manual coding were developed. The NIS coded Occupation on the 3-digit level and the ISCO coding scheme was used . Similarly, the response to Industry was coded on the 3-digit level using the recently revised ISIC coding scheme. To simplify and enhance the search for a particular Occupation and Industry code a set of structured coding indexes was prepared. The coding schemes for Mother Tongue, Place of Birth, Previous Residence, and Place of Work were carefully reviewed and adopted. The NIS carried out manual coding & editing using some 70 staff members. As was mentioned above, the editing component was limited to checking the geographic identification of each questionnaire and a few more items.

16. It was decided to assign all data processing staff to this task for an initial period of approximately one month. In other words, all designated coders together with all data entry operators would first build up a stock of coded/edited questionnaires ready for data entry. The added advantage of this approach was that data entry staff received training on editing rules and became thoroughly familiar with the various ranges and skip patterns. This knowledge improved the quality of data entry, more so since an intelligent data entry program was be used with an integrated number of consistency checks.

17. After the initial build-up period, some 70 designated office editors continued this activity until all questionnaires were completed. Manual coding & editing of main census questionnaires commenced in May 2008 after an initial two-week training course. The activity was completed by December 2008, i.e. the entire process took about 8 months.

Although the editing rules were kept to a minimum the combined set of rules and coding schemes proved too exhaustive and complicated for a single person to comprehend. It was decided to divide the editing and coding task amongst three teams. In such a scenario, team # 1 was assigned the task of editing the geographical identification and the first 15 questions on the questionnaire. Team # 2 was responsible for the editing of the economical characteristics and for Occupational and Industrial coding. Team # 3 was assigned to editing the Fertility roster, Housing Amenities, and the Mortality roster. The breakup of the coding and editing task amongst teams did not only improve the quality of work but also increase dproductivity.

C. Data editing

18. The main objectives of computer editing were to validate the geographical codes, the batch structure, the completeness of the batches, and the detection and correction of inter- and intra-record consistencies errors. Automatic imputations, including hot-deck techniques, was used where possible to correct the inconsistencies encountered.

19. The edit rules for computer editing were prepared prior to software development with substantive inputs from subject-matter specialists. The CSPRO package was used for computer editing. The program required very

thorough testing to ensure all imputations are carried out correctly. Testing was done on dummy files and, at a later stage, on real data sets. The output files generated by CSPRO also required careful scrutiny and examination to check the effects of the imputations.

20. The CSPRO editing program was executed three times on each data set. The first time to load appropriate seed values for the hot-decks, the second time to perform the actual imputations and the third time to ensure the resulting data file was free of errors. Six dedicated computer editors were assigned to this task. They used the Local Area Network to upload the data entry batches and then execute the editing applications.

21. Once the E.A. batches were validated they were concatenated to a higher geographical level (i.e. the District level). The actual concatenation was done through a build-in function of CSPRO. Checks against the Census frame were performed to ensure that E.A.'s are neither duplicated nor missing.

22. The main outputs of CSPRO editing were clean data files, free of errors and ready for tabulation. The package also produced a number of reports for monitoring the imputations by type and frequency. These reports were archived in folders for future reference and the computation of error rates.

D. Data storage and tabulation

23. The census data were stored in CD-ROM, external hard disk, hard disk and Server. After checking the census data were free of errors, the tabulation stage was started.

24. The first step in tabulating the census results was the finalization of a tabulation plan which clearly indicated the priority tables and at what geographical level they were to be produced. Table layouts of these priority tables had to be prepared and presented to the main data users for their comments and approval.

25. Actual tabulations were performed with the CENTS of IMPS package. The resulting tables required carefully scrutiny both by data processing staff and subject-matter specialists to ensure the contents are in order.

26. National census tables were produced by April 2009, after all district and provincial data sets were edited and tabulated. A detailed report on the final census results was prepared and released in July 2009 by the Prime Minister of Cambodia.

E. Dissemination systems

27. The findings of the 1998 Census data were successfully disseminated both through electronic products and printed reports and by way of dissemination seminars and workshops at the national and provincial levels. The CDs released, each for priority tables, aggregated commune database, POPMAP applications, and WinR+ Population Database were well received and used by the line Ministries, International Agencies, NGOs, planning offices in the provinces and districts, the universities, individual and institutional scholars and researchers, teachers and students, and other data users. Another important dissemination product developed was a web site with census background information, key census results, and a request page such that distant data user can demand for further detailed census information. Census tables were also supplied as demanded by data users from time to time. DUSC has also been servicing data users on an on-going basis. All these measures promoted large scale utilization of census data by line Ministries and even government departments like Provincial Planning Offices. Training workshops were conducted at national and provincial levels on retrieving the census data from electronic dissemination products and data utilization. This was continued in 2008 Census also.

a) 2008 Population Census Dissemination by print media

28. The reports mentioned under the analysis plan are being printed and published. Also published were handy data sheets and brochures containing important indicators as derived from the census analysis with suitable and attractive illustrations. Wall maps/charts and census thematic atlases were also produced.

29. Apart from reports and maps, **c**ensus priority tables on each topic would also be published. In the 1998 Census, Tables at National and Provisional levels were published. In the context of growing literacy and educational levels among the people of Cambodia, it may be useful to make available select abridged tables also at district/commune levels. This may be useful for local planning and for those who may not have access to computer facilities and consequently may not be able to avail of the census electronic products.

b) The 2008 Census Electronic Dissemination products

30. The fast-growing uses of computer and its networking call for wider and deeper electronic dissemination products with user-friendly interface, and efficient retrieval and manipulation functionality. The electronic dissemination is classified into two main categories, 1. Off-line electronic dissemination products and 2 On-line electronic dissemination products.

b.1) Off-line electronic dissemination products

31. Off-line electronic dissemination products are mainly in the form of CD-ROM. The project plans to produce a variety of electronic dissemination products based on CD ROMs. These include: a Table Retrieval System, a Community Profile System, a population database built on census micro data, a thematic mapping application and Cam Info updates. About 1000 CD-ROMs may be produced initially.

b.1.1) Table Retrieval System (TRS)

32. The Table Retrieval System stores the large number of census tables onto a single CD ROM and facilitates easy retrieval of selected tables. A user-friendly interface will be available to select multiple tables for multiple geographical areas. The application will also include a table viewer that allows exporting the tables or table cells to Excel spreadsheets for further manipulation and analysis.

b.1.2) Community Profile System (CPS)

33. This dissemination product is based on an indicator database consisting of aggregated counts, rates and ratios for all possible geographical levels, i.e. the Country (total, urban and rural), Provinces (total, urban and rural), Districts, Communes, and Villages. A variety of indicators may be considered, such as: distribution by age groups and sex, household types, median age, dependency ratio, singulate mean age at marriage, literacy rate, employment rate, educational attainment, proportions of migrants, etc. The application will allow aggregation of selected areas and will include functionality to present some of the key indicators in a graphical format and to export the profile to Excel format.

b.1.3) Population database

34. Population databases are highly recommended as they greatly expand the usability and enhance the dissemination of census data. Databases on micro-data (individual records) permit retrieval of data at any level of detail. They are ideal tools to produce small-area statistics. However, the issue of confidentiality will have to be considered and may require re-coding of some of the variables to a higher level.

35. It has to be pointed out that Cambodia produced even as early as 2000, a CD containing the WinR+ Population Database for the 1998 Census. The database consisted of micro-data of the 1998 census (all data records of the individual and households). This enabled the data user to produce any cross-tabulation for any user-defined geographical level. For Cambodia application the lowest selectable geographical level was the village. The CD released was called CD#4 with the following description: CD#4 Win R+ Population Database. It consists of the micro-data of the census. In all there are 2,188,663 housing records and 11,437,656 person records in the database. These records are stored in compressed (binary) format and are accessible only through the WinR+ database engine.

36. The interface for the population database in 2008 is the REDATAM+SP [REDATAM stands for Retrieval of Data for small Area by Microcomputer] package. This package enables data users to easily derive information from the database, including new variables, tabulations and other outputs. All this can be achieved via graphical windows and without the assistance of a programmer. The software also facilitates the processing of external databases in one of the common formats such as dBase and Excel. A Data Dictionary, describing in detail the structure of the database, will be included with the product.

b.1.4) Mapping application

37. Mapping and graphing databases also greatly improve the effectiveness of census dissemination as trends and patterns of the larger area, and distinct boundary and characteristics are more easily detected when displayed on maps. The project will make available map layers for the Country, Provinces, Districts, Commune and, if possible, Villages. Statistical databases consisting of aggregated count, ratios and rates for all geographical levels will complement these map layers. The layers will be in ESRI shape file and MapInfo table file formats. Data users are expected to acquire a copy of their preferred GIS software.

b.1.5) CamInfo Updates

38. CamInfo is the national adaptation of DevInfo software, a global initiative funded by the United Nations. CamInfo is Cambodia's Socio-Economic and demographic Indicator Database, providing a one-stop user-friendly computer program for storage, retrieval, comparison and dissemination of a wide range of indicators from different sources, including national surveys, censuses and administrative systems. CamInfo also allows the user to retrieve and compare indicator data values across multiple time periods, geographic levels, and other sub-group dis-aggregations. Data presentation is possible with tables, graphs and maps.

39. The National Institute of Statistics has selected indicators from the priority tables of the 2008 Census based on the perceived needs of users, for incorporation into the CamInfo regional updates. The online user interface is the same as the CamInfo updates CD-ROM. This will allow users both inside and outside of the country to access data using the Internet without installing the CD. A DevInfo workshop was held in February 2010 when CamInfo of 2008 Census was released.

40. The feasibility of using the recently introduced census dissemination tool called UN Census Info would also be explored.

b2) On-line census dissemination

41. The National Institute of Statistics maintains a web site with information on population censuses, the results of various types of surveys, periodical publication, etc. The URL is www.nis.gov.kh. Most of the information available on this web site is in a static format. For the on-line dissemination of the results of the General Population Census of Cambodia 2008 a more dynamic approach is envisaged. Priority tables and analysis outputs are suggested to be available on-line for wider and distant accesses. The possibilities of SQL database querying and on-line mapping will be explored. To abide by the statistics law on keeping confidentiality of respondents, however, security, authentication, recoding and aggregation shall be closely observed.

b3) Other Electronic products

42. As there is a vast scope to expand this type of dissemination in modern times, more electronic products may also be utilized.

c) Dissemination Workshops

43. Seminars for the presentation of census results and workshops to train planners in the line Ministries and other data users are conducted in the course of 2009-10. Such seminars/workshops are held in every province/district so as to benefit participants down to the village level. They were held in Phnom Penh and

provinces once in 2009 closely following the release of final results in September 2009 and again in 2010 at the national, provincial and district levels after the provincial reports are released.

44. The purpose of the dissemination workshops is to provide census data to planners, administrators and researchers at province, district and commune levels. Such direct interactions between the producers of census data and its users would enable the latter to understand the impact of population growth on welfare measures undertaken by the Government and also help monitor the progress made in the various development programmes of Cambodia.

V. Conclusion

45. According to experience from the General Population Census of Cambodia in 1998, the technologies used for 2008 census of data processing and census data dissemination are the same 1998 census but 2008 census has added Community Profile System (CPS) and Dynamic website in the online.

46. However, other modern methods will be considered according to actual requirements and real situation in Cambodia for using new technology for data processing and dissemination in the next population census.