Distr.
GENERAL

WP.20
6 April 2010

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 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (iv): Innovation and related issues including census systems

Modern Information and Communication Technology (ICT) application in Population & Housing Census – 2011

Prepared by [(Mr.) Migara Ransiri Fernando, Department of Census & Statistics, Sri Lanka]

I. Introduction

1. The Department of Census and Statistics (DCS), Sri Lanka is to collect, compile and disseminate relevant, reliable and up-to-date statistical information required to plan a better future for our country and the people for Sri Lanka, to monitor the progress of development and other socio-economic activities and to measure the impact of various governmental policies of the economy of our country and the living standards of the people.

2. Population and housing censuses are the most important statistical activities for any national statistical office (NSO), particularly those in the developing countries. This is because population censuses in the developing countries provide the basis for many other statistical operations, including other censuses such as agriculture and economic, and also are used as the frames for various household survey programs. In addition, population censuses are the main source of many indicators needed for measuring the achievement of the Millennium Development Goals.

3. Sri Lanka is one of only a few developing countries who have a long history and experiences in conducting population and housing censuses. The 2011 census will be the 14th decennial census carried out by the Sri Lankan Government. The senior technical staff involved in the planning and preparation of the 2011 census were also involved in the 2001 Census of Population and Housing.

4. Upgrading the capacity of the technical staff of DCS is very important to successfully carry out the Census of Population and Housing 2011. Simultaneously, application of the latest information and communication technologies as digital mapping and Geographic Information System (GIS) are very necessary for this activity. Many developing countries including Sri Lanka did the previous Census of Population and Housing in the year 2001 by using various technologies but results cannot be published up to date. Therefore,

this opportunity is very important for Sri Lanka to provide most valuable up-to-date information for decision making and policy planning by 2011 Census of Population and Housing.

5. One of the most important areas that need to be developed is census data capture. Many developing countries used Optical Mark Recognition (OMR) and / or optical Character Recognition / Intelligent Character Recognition (OCR/ICR) for their census data capture in the 2000 round of censuses.

6. The greatest benefit of this technology is acceleration of transferring the data from paper questionnaires to digital format, so that the data can be available for planning and policy decision making sooner.

7. The other important technology to be applied is in the area of digital mapping and GIS. Map preparation is also crucial to provide accurate Census Block maps to the enumerators. Availability of satellite image digital maps will make Census Block maps production much faster than using manual operations. Though, the initial cost for development is high, the long-term benefits are enormous. Once the technology and digital mapping system have been developed a great saving in time, human and financial resources will be made for any future survey, census and other statistical operation, production, and dissemination.

II. Digital mapping

8. Digital mapping is the process by which a collection of data is compiled and formatted into a virtual image. The primary function of this technology is to produce maps that give accurate representations of a particular area, detailing major road arteries and other points of interest. The technology also allows the calculation of distances from one place to another.

9. Though digital mapping can be found in various computer applications, such as Google Earth, the main use of these maps is with the Global Positioning System or GPS satellite network, used in standard automotive navigation systems.

A. Conversion of the printed mode to electronic mode

10. The roots of digital mapping lie within traditional paper maps. Paper maps provide basic landscapes similar to digitized road maps, but this is cumbersome and covers only a designated area. No specific details are shown such as road blocks. In addition, there is no way to “update” a paper map except to obtain a new version. On the other hand, digital maps, in many cases, can be updated through synchronization with updates from different servers.

11. Early digital maps had the same basic functionality as paper maps. They provided only the “virtual view” of roads. Generally, outlined by the terrain encompassing the surrounding area. However, digital maps have improved with the expansion of GPS technology in the past decade, live traffic updates, points of interest and service locations have been added to enhance digital maps user friendly. Traditional “virtual views” are now only part of digital mapping. In many cases, users can choose between virtual maps, satellite (aerial views), and hybrid (a combination of virtual map and aerial views) views with the ability to update and expand digital mapping devices.

B. GPS Navigation Systems

12. The principle use by which digital mapping has grown in the past decade has been connected to Global Positioning System (GPS) technology. GPS is the foundation of the digital mapping navigation systems.

13. The coordinates and position as well as atomic time obtained by a terrestrial GPS receiver from GPS satellites orbiting Earth interact together to provide the digital mapping programming with points of origin in addition to the destination points needed to calculate distance. This information is then analyzed and compiled to create a map that provides the easiest and most efficient way to reach a destination.

14. More technically speaking, the device operates in the following manner:

(a) GPS receivers collect data from "at least twenty-four GPS satellites" orbiting the Earth, calculating position in three dimensions.
(b) The GPS receiver then utilizes position to provide GPS coordinates, or exact points of latitudinal and longitudinal direction from GPS satellites.
(c) The points, or coordinates, output an accurate range between approximately "10-20 meters" of the actual location.
(d) The beginning point, entered via GPS coordinates, and the ending point, (address or coordinates) input by the user, are then entered into the digital map.
(e) The map outputs a real-time visual representation of the route. The map then moves along the path of the driver.
(f) If the driver drifts from the designated route, the navigation system will use the current coordinates to recalculate a route to the destination location.

15. GPS is made up of three segments: Space, Control and User. The Space Segment is composed of 24 to 32 satellites in Medium Earth Orbit and also includes the boosters required to launch them into orbit. The Control Segment is composed of a Master Control Station, an Alternate Master Control Station, and a host of dedicated and shared Ground Antennas and Monitor Stations. The User Segment is composed of hundreds of thousands of users. GPS receivers use to provide three-dimensional location (latitude, longitude, and altitude) plus precise time.

16. GPS has become a widely used aid to navigation worldwide, and a useful tool for map-making, land surveying, commerce, scientific uses, tracking and surveillance, and hobbies such as geo-caching and way-marking. The precise time reference is also used in many applications including the scientific study of earthquakes and as a time synchronization source for cellular network protocols.

III. e-Census

17. Every ten years the Department of Census & Statistics is faced with the daunting task of counting the number and key characteristics of each person in Sri Lanka, including those living in remote areas. To do this, the DCS employs a large number of temporary enumerators and collect information from each and every household to the printed census schedule at the final census night.

18. With the introduction of e-Census as part of the 2011 Census of Population and Housing, the task has become somewhat easier for both the DCS and, potentially, all Srilankans.

19. e-Census is a tool that provides an accessible online form. It also provides all Srilankans with a robust, secure and easy alternative to completing the paper census form, and makes it easier for the DCS to count people living in isolated places.

A. Features

20. When collecting personal information, it is imperative to be able to ensure the privacy of that information. Accordingly, security for e-Census was ‘gold-plated’, with strong protection against malicious attacks from the internet, as well as an innovative solution to encrypting data within the data centre to ensure that only authorized DCS officers can view respondent data.

21. Respondents had to be able to save and exit their form, return later and view data they had previously entered. This meant the system had to be able to decrypt their data to present back to them, without the system administrators in the data centre data being able to access decrypted data. Data from fully completed forms was transmitted to the DCS in encrypted form, where it was decrypted for processing. Equipment and software that met Defense Signals Directorate security standards were used to meet these requirements.

22. As an application, it had to be simple, and easy to use for people who don't regularly use the internet. It also had to perform well over slow dial-up connections. DCS expect to test the application extensively to ensure it was fully compatible with a wide range of commercial and open source web browsers.

23. A key integration challenge was letting collectors know which households had used e-Census. This is solved by at the household listing process.

24. DCS expect approximately 60,000, of Sri Lankan households will use e-Census during the 2011 Census of Population and Housing. This is the first time an online census will be offered in Sri Lanka and there are no comparable experiences. Another issue is the peak load expected during the census period.

B. Advantages of e-census

25. From the DCS perspective, introducing e-Census has created various efficiencies in collecting and processing data. e-Census forms are processed immediately, without waiting for bulk deliveries of materials after census night.

26. Electronically capturing e-Census returns also means that data is processed more efficiently, making census results available more quickly. Also, as census enumerators no longer have to collect information from every household, in the future and DCS can reduce the number of enumerators.

27. e-Census gives the DCS scope to re-engineer the census process and improve the overall quality of the census. Further, e-Census will enable enumerators to focus on the quality of the coverage in their area and resolve exception cases, which will lead to improved census data. There will also be less printed material, ultimately reducing the cost and environmental impact of the census.

IV. Computer Assisted Manual Coding for Industrial Classification and Occupation Classification

28. Some census/survey data are classified into pre-specified categories during a process known as coding. If the computer assigns codes without human interaction, then this is called automated coding. Automated coding never assigns codes to 100% of the cases, so some type of manual coding is required. However, the use of automated coding can dramatically decrease the amount of manual coding required to finish the job. The decrease in the number of cases requiring manual coding saves time and money, and it increases the consistency in the codes assigned.

29. Manual coding operations increasingly make use of computers to aid the coders during the coding process. This enhanced process is known as computer-assisted manual coding. The major elements in computer-assisted manual coding are classifications, training, test, and validation data, software development and error and quality control methods.

A. Classifications

30. Statistical offices use classifications to categorize and classify some types of survey data. Typical examples of classifications are the Industrial Classification and the Standard Occupational Classification. For a Computer Assisted Manual (CAM) coder to work properly, careful construction and maintenance of the relevant classifications are required. A CAM coder cannot accurately classify survey responses if classifications on which it depends are poorly constructed. Good design of the relevant classifications can greatly improve the performance of an automated coder. Examples are removing overlap between concepts or reorganizing concepts. Other strategies also exist. Classifications also change over time. A CAM coder must be revised whenever the relevant classifications are. Otherwise, the CAM coder becomes increasingly irrelevant, and all development costs are lost.

B. Training, Test and Validation Data

31. Training is the term used by researchers in the machine learning community for using data to "teach" the software which categories to assign to cases. So, the software developers use the training set for building the system. The proper codes are attached to each case so the developers know to which category each case belongs.

32. The test data are used periodically by the software developers to independently verify that the training exercise is working. This is known as a "feedback loop". Feedback loops are often run daily, and a well-designed process will provide excellent guidance for the further development of the software. Care must be taken to ensure proper interpretation of testing during the development phase. It is best if the test set is drawn independently of the training set.

33. The validation set is used after the development is complete, and the CAM coder is ready for production use. The validation set is used for the final test before production; the data for the set are usually taken from the initial survey data collection. Again, the independence of the validation set from the training and test sets increases the accuracy of the results.

C. Software for Computer Assisted Manual coding

34. There are many algorithms in use for CAM coding. The choice depends on many factors, but often the most important factor is the complexity of the problem. Usually text responses are coded in survey operations. Some responses are very simple, and others are more complex. Usually, the more text fields in a response, the more categories in a classification, the more classifications, or the more dependencies between multiple classifications for a coding application, the more difficult it is to automate. These observations are meant as guidelines, as there are exceptions to them. .

35. Simple problems can often be solved using exact matching against a list of responses with their codes. New responses not found in the list are added after they are manually coded. Organizations find it useful to develop their own system. DCS has developed application-specific and general-purpose software for CAM Coding for Industrial Classification and Occupation Classification.

D. Error and Quality Control

36. It is not sufficient to estimate the errors for the test or validation data overall, but one must estimate the errors for each code category separately. An error control algorithm is used to control the estimated errors a CAM coder makes.

37. The decision to employ CAM coding in survey processing is not simple. There are many options and expenses to consider. Each of the options has an impact on the quality of the CAM coding operation. Native language responses are a major problem in CAM coding. In Sri Lanka, more than 95% of responses are from native languages (Sinhala and Tamil).

V. Automatic Identification and Data Capture from Census Questionnaire

38. Automatic Identification and Data Capture (AIDC) refers to the methods of automatically identifying objects, collecting data about them, and entering that data directly into computer system (i.e. without human involvement). Technologies typically considered as part of AIDC include Bar Codes, Radio Frequency Identification (RFID), Biometrics, Magnetic Strips, Optical Marks Recognition (OMR), Optical Character Recognition (OCR), Intelligent Character Recognition (ICR), Smart Card, and Voice Recognition. AIDC is also commonly referred to as “Automatic Identification,” “Auto-ID,” and "Automatic Data Capture."

39. AIDC is the process or means of obtaining external data, particularly through analysis of images, sounds or videos. To capture data, a transducer is employed which converts the actual image or a sound into a digital file. The file is then stored and at a later time it can be analyzed by a computer, or compared with other files in a database to verify identity or to provide authorization to enter a secured system. Capturing of data can be done in various ways; the best method depends on application.

A. Capturing data from printed documents

40. One of the most useful application tasks of data capture is collecting information from paper documents and saving it into databases. There are several types of basic technologies can be used for data capture according to the data type:

OCR – for printed text recognition
ICR – for hand-printed text recognition
OMR – for marks recognition
OBR – for barcodes recognition
BCR – for business cards recognition

41. These basic technologies allow extracting information from paper documents for further processing it in the enterprise information systems. The documents for data capture can be divided into 3 groups: structured, semi-structured and unstructured.

Structured documents: Questionnaires, tests, insurance forms, tax returns, ballots, etc. have completely the same structure and appearance. It is the easiest type for data capture, because each data field is located at the same place for all documents.

Semi-structured documents: Invoices, purchase orders, waybills, etc. have the same structure but their appearance depends on number of items and other parameters. Capturing data from these documents are complex, but solvable task.

Unstructured documents: (letters, contracts, articles, etc.) could be flexible with structure and appearance

Optical Marks Recognition (OMR): Form, with registration marks and drop-out colors, designed to be scanned by dedicated OMR device. Many traditional OMR devices work with a dedicated scanner device that shines a beam of light onto the form paper. The contrasting reflectivity at predetermined positions on a page is then utilized to detect the marked areas because they reflect less light than the blank areas of the paper.

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text. It is used to convert paper books and documents into electronic files, for instance, to computerize an old record-keeping system in an office. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. OCR term has now been broadened to include digital image processing as well.

Intelligent Character Recognition (**ICR**) is an advanced optical character (OCR) or rather more specific – Handwriting recognition system that allows fonts and different styles of handwriting to be learned by a computer during processing to improve accuracy and recognition levels.

42. Most ICR software has a self-learning system referred to as a neural network, which automatically updates the recognition database for new handwriting patterns. It extends the usefulness of scanning devices for the purpose of document processing, from printed character recognition (a function of OCR) to hand-written matter recognition.

43. An important development of ICR was three stage process of capturing the image of the form to be processed by ICR and preparing it to enable the ICR engine to give best results, then capturing the information

using the ICR engine and finally processing the results to automatically validate the output from the ICR engine.

B. Disadvantages and limitations

44. There are also some disadvantages and limitations to OMR. If the user wants to gather large amounts of text, then OMR complicates the data collection. There is also the possibility of missing data in the scanning process, and incorrectly or un-numbered pages can lead to their being scanned in the wrong order. Also, unless safeguards are in place, a page could be re-scanned providing duplicate data and skewing the data. For the most part OMR provides a fast, accurate way to collect and input data; however, it is not suited for everyone's needs. Another problem in data capture system is bad handwriting.