Experiences of Applying Information Technology in the Hong Kong 2006 Population By-census

A paper prepared for
ESCAP Expert Group Meeting on Effective Use of Information Technology in Population Censuses

10-12 December 2007
Bangkok, Thailand

by
Stephen KC LEUNG
Acting Assistant Commissioner for Census and Statistics
Hong Kong, China

INTRODUCTION

1. It is an established practice from 1961 for Hong Kong to conduct a population census once every ten years and a by-census in the middle of the intercensal period. Following this practice, a population by-census was conducted in Hong Kong in the eighteen-day period from 15 July to 1 August 2006.

2. A by-census differs from a census in not having a complete headcount of the population but simply enquiring on the detailed characteristics of a large sample of the population. The size and characteristics of the entire population are inferred from the sample results. The sizeable scale of a by-census, as compared to other household sample surveys, can facilitate the provision of statistics of high precision even for population sub-groups and small geographical areas. Such information is vital to the Government for planning and policy formulation, as well as to the private sector for business and research purposes.

3. Hong Kong has been adopting Information Technology (IT) to support the conduct of population censuses/by-censuses since 1970's. In the earlier years, IT was mainly applied in processing the census/by-census data. In recent rounds of censuses/by-censuses, IT has been applied to support various aspects of the project from sampling, data collection to data dissemination. For this by-census, we have developed/enhanced a suite of computer systems to facilitate its planning and execution. This paper discusses the applications of IT in different stages of the 2006 Population By-census (06BC) in five broad areas: (i) supporting field operation with Geographic Information System (GIS)[1] technology; (ii) allowing sampled households to make appointment for enumeration at preferred time slots through an Internet application "Appointment Request Service"; (iii) providing electronic questionnaires to respondents for electronic reporting; (iv) using Intelligent Character Recognition (ICR) and Optical Mark Recognition (OMR) technologies to capture data on completed questionnaires; and (v) disseminating 06BC results through the Internet. The practice of releasing the microdata of population census / by-census will also be covered in this paper.

[1] A Geographic Information System can be seen as a system of hardware, software and procedures designed to support the capture, management, manipulation, analysis, modeling and display of spatially-referenced data for solving complex, planning and management problems. (Goodchild & Kemp, 1990)

GEOGRAPHIC INFORMATION SYSTEM TECHNOLOGY

4. GIS technology was first applied in the 2001 Population Census (01C) through a computer system named "Digital Mapping System (DMS)" which enables more efficient updating, production and maintenance of maps and effective monitoring of fieldwork progress. GIS techniques were used in the following two major aspects to support a population census in Hong Kong: (i) on field operation, the large volume, multi-format, multi-scale, tailor-made and up-to-date maps of good quality produced by the DMS allowed field operation to be conducted more efficiently; and (ii) on data dissemination, data analysis on small geographic areas could be performed more conveniently through the use of thematic maps[2]. Given adoption of GIS led to considerable cost saving, in addition to production of maps, the DMS was enhanced to support fieldwork operation of the 06BC in three stages of work: allocation of assignments, itinerary planning and monitoring, and controlling fieldwork progress.

Allocation of assignments

5. In 06BC, some 5 000 temporary field workers were needed to interview 0.7 million people in 230 000 households living in around 50 000 buildings spread around 4 000 street blocks throughout the territory. Making use of GIS techniques, the assignments were first sorted into an "assignment sequence" based on their geographical locations in order to ensure that the assignments of each enumerator would not be widely scattered. Furthermore, before allocation work started, a number of parameters (e.g. average traveling distances from field centre[3] to buildings with sampled quarters and that between buildings with sampled quarters) was determined for each unit of the sampled quarters to estimate the amount of efforts required to perform enumeration such that each enumerator's set of assignments was comparable in terms of enumeration effort. GIS techniques could tackle the above tasks scientifically and efficiently with manual intervention in the process minimized.

Itinerary planning

6. After allocating the sampled quarters to the enumerators, the assignments as

[2] A thematic map shows the spatial distribution of one or more specific data themes for standard geographical areas. The map may be qualitative or quantitative in nature.

[3] One field centre was set up in each working district to facilitate data collection work during the fieldwork operation period.

shown on the assignment list were arranged in such a way that the shortest route (in terms of walking distance) to visit the assignments would be incurred. Making available a suggested itinerary to enumeration enabled the enumerators to better plan their household visits in a short span of time, and hence, led to higher productivity and efficiency. This information was particularly useful for the enumerators without local geography knowledge.

Monitoring and control of fieldwork progress

7. Thematic maps on fieldwork progress were produced based on the records of completed enumerations everyday during the operation period to facilitate the monitoring and controlling of the progress of enumerations. With a glance at the map, the management could quickly grasp an idea of the overall progress of enumerations so that special attention was paid to identify any possible problems and appropriate actions taken to speed up the completion rate.

ONLINE APPOINTMENT REQUEST SERVICE

Implementation Approach

8. An online Appointment Request Service (ARS) application was designed for sampled households to make appointment requests for enumeration over the Internet in the 06BC. It provided a new channel to respondents in addition to phoning the 06BC telephone enquiry centre (TEC). Through the ARS application, authenticated households could make, change, cancel or enquire appointment requests for enumeration online by themselves.

9. To facilitate the processing of the large number of appointment requests made through the ARS application, a daily cut-off time was set at 8:00 a.m. each day starting from two days before the commencement of the 06BC operation. After the cut-off time of the day, the appointments requested kept in the ARS application would be transferred to the relevant field centres for follow up.

10. In the field centres scattered around the territory, daily summary reports on the contact and appointment request details of households by individual field centres and enumerators were generated from the ARS system for enumerators to follow up

on the appointments. Enumerators would then confirm the exact date and time of interview with each respondent by phone.

Usage

11. The functions of ARS were found user friendly to the public and it had achieved the aim of alleviating the workload of the TEC. Among all appointment requests for enumeration made, about 17% of appointments were handled by the ARS, with the remaining handled through the TEC.

ELECTRONIC REPORTING

12. Data were collected from the households through conducting face-to-face interviews in the past censuses/by-censuses. In the 06BC, respondents were given an option to provide data through electronic means. Sampled households opting for electronic reporting could download their electronic questionnaires (e-Qs) from an Internet application and return them to the 06BC office after completion through the Internet.

Implementation Approach

13. The e-Q was in Portable Document File (PDF) format, with extension features to enable respondents to use the freely available Adobe Acrobat Reader for self-completion.

14. Sampled households/persons wishing to use the e-Q could call the enquiry centre for registration before commencement of the fieldwork. Some basic information of the household and its members was collected during the registration. Separate e-Q, protected by a unique password, was generated for each member of the household.

15. Enumerators would then deliver the password letters to the registered households/persons. Using the password for each household member, respondent was able to login at the designated website to download his/her e-Q and, after completion, submit the data portion via Secure Socket Layer (SSL) of Internet, by simply clicking a button built in the e-Q.

Usage

16. Some 5 400 households registered for using e-Q to report data. Eventually, about 4 500 households had completed the e-Q, accounting for about 2% of all successfully enumerated households. The usage rate was in line with previous expectation given the experience from a pilot survey, a relatively low publicity profile having been adopted and the complexity of the 06BC questionnaire.

Lessons learnt

17. In order to complete and submit the e-Q, respondents had to install specific version of Adobe Reader, which was 6.0 or above. Also, since the e-Q was bilingual (Chinese and English), respondents needed to install the Asian Font Pack when using an English version of Adobe Reader. Most enquiries on using the e-Q from respondents during the operation period were related to this issue. Wider publicity to caution users on this area would be launched in the coming round of the population census with a view to reducing the effort required to handle such enquiries in view of the much larger number of respondents involved. In sum, it is essential for the e-Q solution to have minimal system requirements on the respondent's computers for it to be successfully implemented.

INTELLIGENT CHARACTER RECOGNITION AND OPTICAL MARK RECOGNITION TECHNOLOGIES

18. On top of electronic reporting mentioned above, data were mainly collected using a long form (LF) paper questionnaire via the interviewer method in the 06BC. Besides, a Self-administered Questionnaire (SAQ) was left for completion by those households who could not be contacted during visits made by field staff during field operation period. There were around 284 000 sets of LF (1 set of LF refers to an A4 size booklet with 12 pages) and 12 000 sets of SAQ input form (1 set of SAQ input form refers to a 1-page A4 size form) to be processed within two months after the 06BC field operation. Given the tight schedule and the large volume of data involved, automation of data capturing by the ICR, OMR and bar-code reading technologies was adopted in the 06BC.

19. In order to achieve a high accuracy rate, only numeric fields were recognized using ICR, whereas data with multiple choices were captured using OMR technology and the images of textual fields were clipped as images and passed to another computer application for computer-assisted online coding.

20. The actual accuracy rates were: 90.4% for ICR before the operators amended the ICR data fields and 100% after the operators amended the ICR data fields, nearly 100% for OMR and exactly 100% for barcode recognition.

21. With the advancement of scanning, ICR and OMR technologies, only 3 scanners together with 20 computers were deployed to complete the data capturing work in 34 working days. On average, 8 339 LF questionnaires (equivalent to 100 068 sheets of A4 papers) were processed everyday.

DISSEMINATION OF 06BC RESULTS

22. Results of the 06BC were released in phases starting from February 2007. All statistical reports and tables in softcopy format were made available for downloading free of charge by users through the website of the Census & Statistics Department (C&SD) (http://www.censtatd.gov.hk) and the 06BC thematic website (http://www.bycensus2006.gov.hk).

23. Moreover, an online Interactive Data Dissemination System (IDDS) will be launched in early 2008 on C&SD's website as a new e-delivery service provided to the public. The IDDS will enable users to prepare statistical tables according to their own needs with 06BC macro-data files and also support data presentation using maps and charts.

MICRODATA

24. Microdata sets of population censuses / by-censuses starting from 1981 are made available to academic users at a charge. There are two microdata sets for each population census / by-census which comprised respectively 1% and 5% samples randomly selected from all occupied quarter records (except for 1981 and 1986 population census / by-census for which only 1% microdata set was available). The

microdata sets are provided in softcopy (text file) format on CD-ROM or delivered via email.

25. In addition to enabling access to the above microdata sets for off-site processing, a self-help tabulation service (STS) is also provided to government bureaux / departments and academics. Users of the service could make use of the Dynamic Reporting Module (DRM) of the 06BC Statistics Dissemination Sub-system or run customised SAS programs on the full data set of population census / by-census to produce the required tabulations in the premises of C&SD. The DRM is a menu-driven application which offers a wide range of tabulation facilities with just the clicking of buttons to produce tabulations in different formats. Therefore, users (regardless of whether they have programming knowledge or not) could tailor-make their required statistical tabulations easily and efficiently.

26. To ensure the proper use of microdata and protect the confidentiality of information in regard to individual households and persons, all record identifiers (including as detailed address of quarters) will be removed and other information suitably coded before release to ensure that particulars of individual households and persons cannot be identified. Besides, the microdata sets and STS must be used for research purposes only.

CONCLUDING REMARKS

27. Results from the 06BC are being released in a series of publications and products starting from February 2007. The planning work for the next round of population census in 2011 has just started recently. One of the most important issues to be considered is the more extensive use of IT in data collection, processing and dissemination.

28. The challenges ahead for us, and probably also other Government Statistical Offices, lie in how to meet users' demand for more statistical information, in-depth analysis, with faster turnaround, without unduly increasing respondents' burden. The experiences and lessons learned from the 06BC will definitely be reflected in the planning of the coming round of population census.