

# Improvements of the Census Operation of Japan by Using Information Technology

By  
Naoki Kurihara  
Statistics Bureau, Japan

## I. INTRODUCTION

1. The Population Census of Japan has been conducted by the Statistics Bureau every five years since 1920, and the coming Census is to be taken as of 1 October 2005. Since it is projected by the National Institute of Population and Social Security Research that Japan's population will reach its peak in 2006 (medium variant) and decline after that, the 2005 Population Census is expected to give essential information on Japan's population at this critical turning point. With the population of Japan over 120 million, the Census is the largest statistical undertaking in Japan. To collect and process the extremely large amount of data within a limited time, the use of information technology (IT) is essential. This paper describes how IT will be used in the Census operation of Japan, and explains a few issues to be considered for future Censuses in the author's private capacity.

## II. IMPROVEMENTS OF THE CENSUS OPERATION BY INFORMATION TECHNOLOGY

2. There are many possibilities of using IT in the Census operations. In the recent Censuses of Japan, major improvements by the use of IT have been brought about in the following areas.

- a. Mapping of enumeration districts
- b. Data capture from questionnaires to computer
- c. Dissemination of small area statistics

The so-called Internet enumeration will not be adopted in the 2005 Census because of many problems envisaged to encounter. However, IT will be utilized to the extent possible in the management of the Census operation.

### **Mapping of enumeration districts**

3. While the responsibility for planning and executing the Census rests with the national government (Statistics Bureau), the field work to enumerate the population is entrusted to the local governments, i.e. 47 prefectures and about 2,500 municipalities (cities, towns, and villages). The municipal governments employ and supervise approximately 900 thousand enumerators all over the country during the enumeration period. Because the enumeration is the crucial element of the

Census work that affects the quality of the resultant statistics, the Statistics Bureau supports the local governments by providing tools that help improve the quality and efficiency of their work.

4. One of the most important tools is the geographic information system (GIS) which is used for demarcating Enumeration Districts (EDs) and producing ED maps. The work of demarcation and mapping of EDs was computerized for the first time in the 1995 Census on a limited scale, and the coverage of the system has been extended broader in every Census.

5. For computerized mapping, digital data of the ED boundaries and digital base maps are needed. The digital data of ED boundaries owe very much to the Census Mapping System (CMS) which the Statistics Bureau developed by applying GIS in the 1990 Census. The CMS at first aimed at making the work of statistical compilation at the Statistics Bureau more efficient by maintaining digital data of all the ED boundaries. The CMS was then used for compiling small area statistics, such as grid-square statistics from statistics of EDs. As for digital base maps, the cost for them was very expensive in the market in early 1990's. As digital base maps have become increasingly popular for PCs and car navigation systems, their prices have become more affordable.

6. For the 2005 Census, the Statistics Bureau has developed a new system of computerized mapping of EDs. The system can produce ED maps by putting digital data of ED boundaries on the digital base maps. The Statistics Bureau has provided local governments (prefectures) with the system along with the digital data of the ED boundaries for the previous 2000 Census and the digital base maps so that they can produce ED maps of their own areas for the 2005 Census.

7. Owing to the computerized mapping, most municipalities can produce ED maps without cutting and pasting paper maps as they used to do before. As a result, the workload of the local governments has been considerably reduced. The coverage of such ED maps produced by computer has increased up to approximately 80% of all the EDs in Japan.

#### **Data capture from questionnaires to computer**

8. Data capture from the questionnaires is a crucial part of the census operation because of the extremely large volume of data. In Japan, data capture operation is centrally done by the Statistics Center, the organization specialized in data processing.

9. The Statistics Center used optical mark readers (OMRs) for data capture since 1965 until the 1995 Census. In the 2000 Census, optical character readers (OCRs) were used for the first time because of the following advantages over OMRs:

- ( i ) Capable of capturing numerical responses directly as well as marks;
- ( ii ) Capable of capturing image data on a full scale.

10. Owing to the first advantage, it became possible to design a user-friendlier questionnaire in the 2000 Census. In the 1995 and earlier Censuses, the respondents were asked to write numerical responses, such as month and year of birth, in both numbers and marks. Writing marks is not only an extra burden for the respondents but also a cause of response errors. OCRs have enabled us to eliminate such mark fields from the questionnaire, and this has made the questionnaire more compact. The second advantage has made the data editing work more efficient because the editing staff can refer to the full image data of the questionnaire on the screen of their own PC. In the earlier Censuses, when the staff needed to refer to questionnaires, they had to get them physically, which required time and manpower. With OCRs, necessary information can be retrieved on the spot electronically, and such extra work has been eliminated. This change has enabled the staff to make faster and better judgment.

11. OCRs, however, are not free from problems. The main problems are with the accuracy and speed of recognition. As for accuracy, in some cases, characters cannot be recognized (non-recognition), while in other cases, the result of recognition is wrong (false recognition). In OMRs, non-recognitions or false recognitions seldom occur as long as the responses are clearly marked in the right position. But the recognition performance of OCRs depends on not only the ability of the machine but also the quality of written characters: when characters are not neatly written, non-recognitions and false recognitions will occur. The speed of recognition also varies according to such factors as the complexity of the form and the legibility of characters. For the purpose of the Census of Japan, the character set has been limited to numbers, although the available OCR models can recognize alphabets and Japanese characters as well. By limiting the character set, the recognition rate and the speed have increased.

12. To cope with the recognition problem, great care is taken in every phase of work from planning to implementation. In the planning phase, OCR models are first selected on the basis of catalog specifications. But because the performance of OCRs is affected by various factors not specified in the catalogs, the machines are tested with the real questionnaires of pilot surveys or a certain sample survey to measure the performance, i.e. accuracy and speed. According to the result of the test, the patterns of handwriting that are unrecognizable or falsely recognized are analyzed, and the recommended form of handwriting is developed for each number. They were included in the instruction to fill the questionnaire.

13. In the implementation phase, unrecognized characters are judged and entered manually, and the recognition performance is constantly monitored on the real time basis: If non-recognition occurs, the unrecognizable characters are displayed on the operator's screen, and he/she judges it and enters the correct numbers. Moreover, there is another group of operators assigned to monitor the accuracy of recognition: Sample inspection is applied to batches of questionnaires, and if a sample batch includes more errors than the permissible level, the batch is rejected and recaptured. As a result of the control, the non-recognition rate was 0.8%, and the false recognition rate was estimated

to be 0.1% at maximum for the 2000 Census. 12 units of OCRs were rented, and it took eight months to process 56 million sheets of questionnaires in the 2000 Census. The average reading speed was approximately 5,600 sheets per hour.

14. In the 2005 Census, the same OCR technology is planned to be used, but the possibility of capturing the Japanese characters was tested to introduce automated coding of hand-written responses for the first time in the Census. The test was targeted to the coding of “destination of commuting”. This item was selected for testing because the responses fall in one of the approximately 3,000 municipalities, which means that the words and characters appearing in the response are limited, and even if one or two characters are unrecognizable, there is a possibility to infer them from the context by referring to the dictionary of the municipality names.

15. The result of the test showed that the accuracy of recognition was not sufficiently high, and it was concluded that automated coding of hand-written responses will not be adopted in the 2005 Census. The Statistics Bureau will continue its study of automated coding from hand-written characters.

#### **Dissemination of small area statistics**

16. The census data have been made available in various media such as printed reports, CDs, and the Internet as for other survey data. In February 2004, a new web site named “Statistical GIS Plaza” was launched to provide small area statistics of the 2000 Census. Users can look at the census data of any area in the form of maps. There have been many accesses to this site, and favorable comments have continuously been fed back to the Statistics Bureau.

17. The small area statistics provided in the Statistical GIS Plaza are compiled on the basis of geographic units called “cho” or “aza” (address blocks). There are approximately 211,000 address blocks in Japan, each containing about 220 households on average. For each address blocks, statistics of such basic characteristics are made available as population by age group, industry and occupation.

18. The statistics for area blocks are displayed on the base map showing the boundaries of area blocks and geographical objects such as roads, railways, rivers, etc. The site provides various functions to sum up statistics of two or more address blocks. For example, the population of an area within a certain radius from an arbitrary point can be easily computed. Many of such functions are useful for business planning.

19. In designing this site, special attention was paid to the privacy concern. As the area unit is small, there is a risk of disclosing an individual’s characteristics inadvertently. Therefore, the statistics provided in the Statistical GIS Plaza have been limited to the basic characteristics without

cross-classifications, and the statistics about the characteristics that are perceived as delicate have not been included in the site. The balance of wider use of statistics and protection of privacy will remain to be an important issue in disseminating small area statistics.

20. Besides the Statistical GIS Plaza, the Statistics Bureau runs the web site of the Census results, which have more than 600,000 accesses every month. People can get most of the Census tables through the site free of charge. The Statistics Bureau will make efforts to improve the web site of the Census results to become user-friendlier hearing user comments and opinions.

### **III. SOME ISSUES TO BE CONSIDERED FOR THE FUTURE CENSUSES**

21. One of the most important issues to be considered for the future Censuses is the use of IT in data collection. We recognize that the statistical offices of a number of countries provide options to the respondents to submit their answers by the Internet.

22. There are many obvious advantages with the Internet response method. One is that those who are familiar with the Internet can submit their responses easily and quickly. As more and more people are feeling uncomfortable that enumerators/interviewers see their questionnaires, the feasibility of the Internet response is being considered as option for such people. As the preference of the people at large is becoming diverse, providing wider options for response methods other than paper questionnaires is an important consideration for getting good responses. If all or nearly all people would return their responses by the Internet, the workload of enumerators will be reduced.

23. The Internet response method is also advantageous for data processing. If the respondents submit their own data by the Internet, there is no need for data capture. In the Internet response method, functions to check data consistency can be included to ensure accuracy of responses. The Internet responses will reach the processing center almost instantaneously, and will expedite the data processing work, as long as the respondents submit the responses punctually.

24. While the Internet response method is an attractive option as a means of data collection, this option will not be taken in the 2005 Census because of the following problems.

#### **(a) How to ensure accuracy**

With the Internet response, it is quite difficult to ensure punctual responses from all the respondents, unless the respondents are highly cooperative. In order to remind the respondents who do not meet the deadline, the most effective way will be the enumerator's visit. But this will not reduce the workload of the enumerators, and the advantage of the Internet response method will be partially lost.

#### **(b) How to achieve cost effectiveness**

To judge the feasibility of the Internet response method, an important factor is the percentage of the people who choose it. If the Internet response rate is at a low level, the workload and the cost of the enumerators will not be significantly saved, while the Internet response

method will just add the cost.

(c) How to maintain the security and confidentiality of data

There are also fears in the handling of personal information on the Internet. To adopt the Internet response method, the system should be carefully designed in view of protecting the confidentiality of the respondents' data.

(d) How to make the method compatible with the enumerators' work in the field

If the Internet response is to be adopted along with the conventional method, data collection by the enumerators has to be well coordinated with the Internet response so that duplicate responses and non-responses may be avoided.

As mentioned above, there are still many problems to resolve before adopting the Internet response method in the future Censuses. The Statistics Bureau will continue to work on devising a new IT-related method that can be used in the future Census.

25. Another important issue is how to meet the diverse needs of users for the Census results. There are various demands for providing more detailed Census tables to exploit the Census data to the extent possible. But, it is impossible to compile and provide every cross-classification table because the budget and time for compilation are limited. On the other hand, use of micro-data is still restricted in Japan for various reasonable reasons. Therefore, by the use of IT, the Statistics Bureau will research a mechanism of tailor-made compilation.

#### **IV. CONCLUDING REMARKS**

26. In the Population Census of Japan, new methods utilizing information technology are adopted in mapping of the EDs, data capture, and data dissemination, but not in data collection.

27. The data collection process is the most important process that determines the quality of data in terms of coverage, accuracy, and timeliness. The Internet response method has many difficulties to overcome, and will therefore not be applicable in the 2005 Census of Japan. So, the data collection process continues to depend on the conventional method because it is still reliable and affordable in Japan.

28. The 2005 Census is to be taken in about seven months from now. The Statistics Bureau will make utmost efforts to conduct the 2005 Census successfully. The experiences and lessons learned from the 2005 Census will be reflected in the planning of the 2010 and future Censuses.