United Nations ECE/CES/2012/56

Economic and Social Council

Distr.: General
15 May 2012

English only

Economic Commission for Europe

Conference of European Statisticians

Sixtieth plenary session
Paris, 6-8 June 2012
Item 3(a) of the provisional agenda
2010 round of censuses – innovations and lessons learned

New technologies adopted in the Albanian Population and Housing Census

Note by the National Statistical Institute of Albania

Summary

The paper focuses on new technologies used by the National Statistical Institute of Albania in collection and processing of Population and Housing Censuses. As a result of introducing new scanning technologies, the efficiency of data capture has improved notably. This has reduced the costs of the census processes and will also improve timeliness of releasing the first census results to the public. Moreover, new tools for managing the field operations were utilised, such as SMS messaging to a web-based GIS Monitoring System, which turned out to be valuable for leading the census operations. The experience gained and the investment in technology will continue to be used for future censuses and other surveys in the Statistical Institute.

I. Background

1. The National Statistical Institute of Albania (INSTAT) has been using new technology to assist in all phases of population censuses for many years. Although some of these technologies have been used in the past, they have been further developed during the last Population and Housing Census.

2. Development of a well-planned management structure for the field operations of census taking had a great impact on the quality of the data collected. It also facilitated the monitoring of the enumeration progress and redaction of the related costs as it was possible (i) to plan accurately materials and staff resources needed in each area and (ii) to limit unforeseen events during the general enumeration. Since a population census is a large time-critical project, with many interlocking operations, the use of a modern Census Management System is of vital importance.

3. INSTAT has until 2010 been using manual data entry to transfer data from questionnaires to database systems. In 2009, a scanning system was implemented to support the upcoming Census of Non-Agricultural Economic Enterprises. The work under this component focused on getting the scanning system configured for the Census pilot in April 2010 and subsequently for the Census of Non-Agricultural Economic Enterprises. The system is configured to handle all surveys and censuses conducted by INSTAT. The aim of the system was to obtain a higher level of quality of the captured data as well as to shorten the time of data capturing. The activity was supported by different donors and internationals experts.

4. In Albania, this approach was used by the INSTAT for the 2011 Population Census. Its use allowed INSTAT to scan and process 20 million of pages in 5.5 months.

II. Census management and monitoring system

5. Since a population census is a large time-critical project, with many interlocking operations, the use of a modern Census Management System is of vital importance. This system was developed for printing the delivery notes and labels to boxes containing census forms, for assembling the boxes to be sent in the field, for tracking the transportation of the boxes from INSTAT to the different Census Offices and from the Census Offices to INSTAT at the end of the fieldwork, and finally, for verifying that all the materials in the box were received. The census management tool was also used to prepare the boxes and the materials for the Post Enumeration Survey. Furthermore, the census management tool and the related data were also used as a base to properly organize and manage the data processing operations.

6. An innovative approach for monitoring the field work coverage was the use of SMS messages.

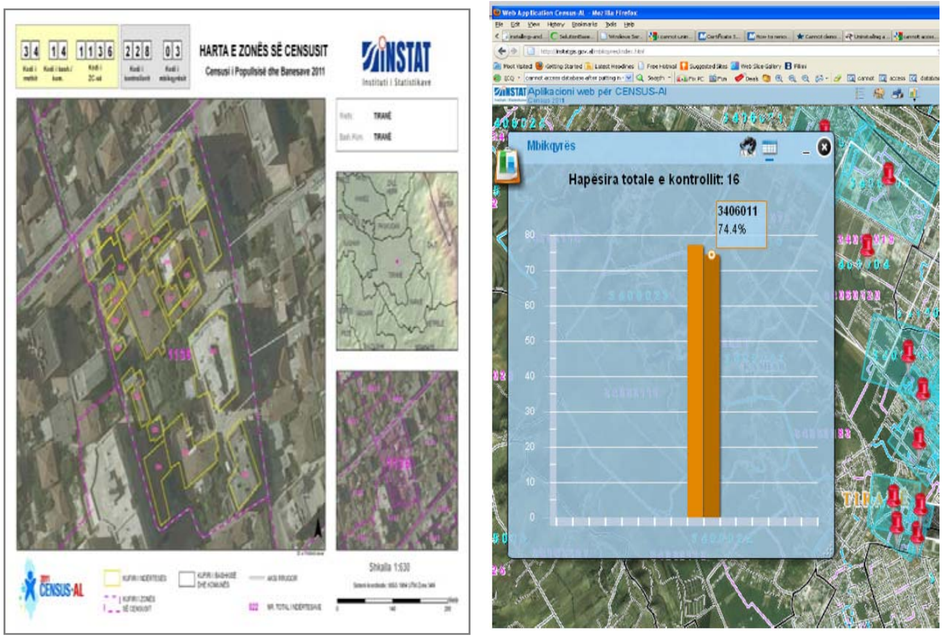
7. Everyone involved in the field work was provided with a SIM card which permitted communication free of charge among the staff. This was useful for addressing problems encountered during the enumeration process timely and efficiently as well as to properly solve them. Moreover, the enumerators and controllers send an SMS every morning to a short number to record daily figures about the total of households and persons interviewed the previous day. A web-based Geographic Information System (GIS) application was developed and updated daily with the SMS figures.

8. During the field operations of October, INSTAT monitored the census coverage through the SMS-based reporting system and the interactive web-based GIS Monitoring System.

9. The web-based GIS Monitoring System provided a valuable tool for the monitoring of the census progress from the side of field Supervisors. All the supervisors had their accounts to login into the GIS Monitoring System, giving them the opportunity to monitor the progress of their staff and areas. The information provided by each enumerator via SMS was automatically stored in a central database and the GIS Monitoring System updated daily the total figures accomplished by each actor.

10. On the other hand, during the enumeration period INSTAT produced a daily summary coverage report with the data received the day before by SMS (See figure 1). The reports showed cumulative data by district and by Municipality. The GIS Monitoring System showed data by small area, giving the opportunity to geographically locate areas of poor progress.

Figure 1
An image of the GIS Monitoring System


3 4
1 4
1 1 3 6
2 2 8
0 3
HARTA E ZONËS SË CENSUSIT
Censusi i Popullsisë dhe Banesave 2011
INSTAT
Instituti i Statistikave
TRANË
TRANË
Shkalla 1:530
CENSUS-AL
Web Application Census-AL - Mozilla Firefox
Aplikacioni web për CENSUS-AL
Mbikqyrës
Hapësira totale e kontrollit: 16
3406011
74.4%
80
70
60
50
40
30
20
10
0

11. The data reported by SMS and uploaded on the GIS application was compared every day against the data reported by INSTAT regional offices. The two reporting systems showed some differences during the first period of the enumeration. The differences were caused by under-reporting from areas not covered by the mobile network, reporting errors in both systems and differences in the estimated number of dwellings in the two systems. Those differences almost disappeared in the further phases of the enumeration.

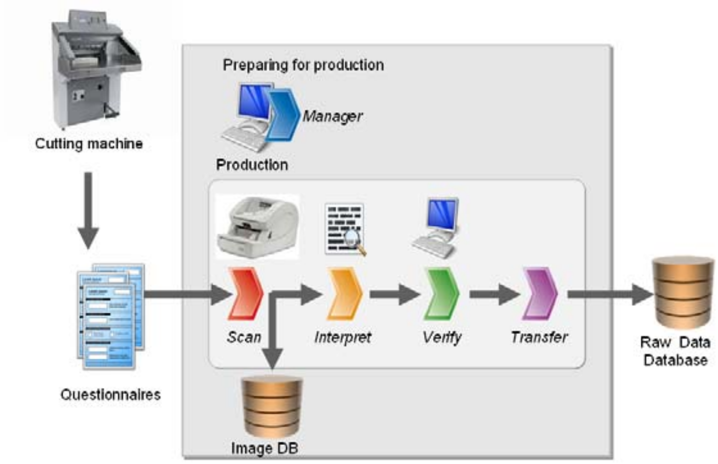
III. Scanning technology

12. In general, the Population and Housing Census is a large-scale data collection project that encompasses an entire country. To shorten the data capturing process, a selection between using a large number of data entry operators or deploying new technology for data capture had to be made.

13. The decision between using manual data entry versus automated entry was partly based on timetable requirements and the consideration between staff and hardware costs.

Other factors, such as whether it was feasible or possible to implement more sophisticated technology also was taken into consideration.

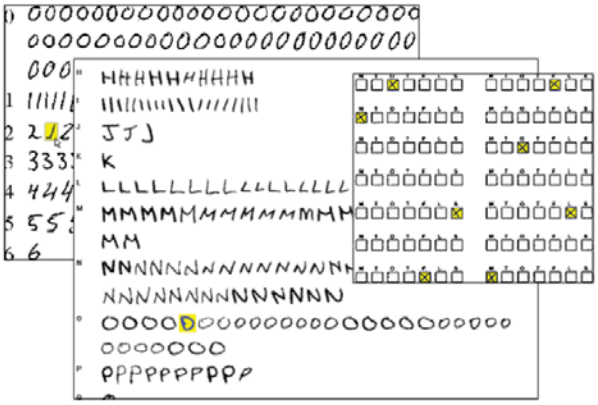
Figure 2
The data capture process using an Intelligent Character Recognition system



14. Image scanning technology is a system used to capture data from a questionnaire (form) with a limited amount of human intervention: once the forms are scanned, the images are saved and passed to an Intelligent Machines Research Corporation's system of Intelligent Character Recognition (IMR/ICR) that attempts to infer the contents of the answer. Depending on the confidence of the recognition process, the IMR/ICR system either accepts the inferred result or rejects it. (See figure 2.)

15. Even with highly sophisticated interpretation engines a certain number of Key for Image (KFI) operators are needed to correct errors and, for obvious reasons, they cannot all be highly qualified. Consequently, there are two important requirements at this stage: verification should be simple and fast, and the KFI operators must not insert additional errors. The solution adopted by INSTAT was the so called 'Mass verification': the images of all interpreted characters were presented as a group according to their value. If a character appears in the wrong group, the verifier selects it and he will be automatically taken to the field in which the character appears, to correct it. First, digits are mass verified, then letters and finally mark fields. (See Figure 3.)

Figure 3
Scanning the manually filled census forms with Key for Image operators



16. In addition to the mass verification phase, a parallel process was in place for quality assurance: a selected group of operators has checked for a second time all the individual questionnaires flagged with at least four inconsistences. A specific application was designed to show the questionnaire images and the value stored in the different fields, highlighting the values that were generating an inconsistence. Here, the objective was not to correct enumerator's mistakes but to be sure that the inconsistences were not caused by a misinterpretation of the IMR/ICR system.

17. Based on INSTAT requirements, a company called “ReadSoft” developed a Census Control Package for monitoring and managing large scale scanning operations. The main task for several INSTAT employees from April onwards was to test, troubleshoot and communicate any problems to the ReadSoft developers. Despite some initial problems, when the scanning of the Population and Housing Census summary questionnaire started in November 2011, the Census Control Package proved stable enough to be used.

18. A well-designed form accomplishes two things: it makes it easier for the respondents to fill in their answers, and it leads to a minimum amount of manual work when capturing the results. Thus, the design work is best done in cooperation with experts of the scanning system, since they know the usability issues and technical aspects that need to be considered. Census questionnaires were designed with a clear layout and adequate space for answering with a large font size and appropriate page breaks. Filter questions targeted to different subgroups and skips were highlighted in a different color.

19. A serial number was printed in all pages of the questionnaire. Identifying each form uniquely helped during the data capture process especially if the form was mistakenly scanned more than once.

20. A routine was put in place during the printing of the questionnaires, ensuring that a random sample of every days printing was checked with the scanning system. Even though every questionnaire passed this pragmatic test, at the end of the operation some misalignments were discovered in a limited amount of printed questionnaires. A work-around was implemented in the scanning system to cope with these issues. A total of about 1,200,000 booklets (6 individual questionnaires) were printed.

21. When imaging technology is used, adequate training of enumerators on how to properly fill in the forms is crucial to give the ICR the best chance to recognize the handwritten characters correctly. INSTAT had a specific session during the enumerators training on how to write characters into the response areas of the census forms, what kind

of a pen to use, et cetera. Time and effort was invested to ensure that the census forms were completed as accurately as possible and returned in the best condition.

22. In September 2011, a new Data Processing Centre was established with 60 personal computers (PC), 10 advanced PCs and 7 optical scanners. The capturing center was organized in five sub-systems. Each system consisted of one scanner, twelve verification computers connected to one database so that in case of problems the loss of production capacity could be kept at a fifth of the maximum. This turned out to be an adequate approach: once a system had a problem, the other sub-systems were working normally.

23. The success of a data capture process depends, to a certain extent, on the quality of the personnel involved in the process. INSTAT organized training for the personnel to improve their awareness of their duties and of the importance of the accuracy of their work. Special sessions were devoted to train the scanner operators, making sure they understand how to handle and maintain the scanners and the scanning software, to the KFI operators, training them in the use of mass verification and other relevant parts of the software, to the data verifiers, making them familiar with the highlighting of inconsistencies and the general validation strategy.

IV. First assessment of data quality

24. By the end of April, INSTAT finished the data capturing from the Census forms and the work for the data cleaning has started. The data cleaning procedure consists of: i) localizing the errors in the collected data, ii) applying deterministic rules implemented in *ad hoc* programs for correcting the systematic errors, iii) imputation to correct for the remaining errors and missing values.

25. As a result of this approach, an analysis of the impact of the edit and imputation procedures by comparing initial and final distributions will be performed.

26. At this stage the cleaning procedure was developed and tested on a random sample of 51,000 individuals (13,000 households, about 2% of the total). A deterministic step has been developed, consisting of 19 deterministic rules.

27. The probabilistic step has been based on the definition of 131 explicit logical edits that have produced a complete set (explicit plus implicit) of 439 edits. The application of this set of edits to the random sample has determined the following result: i) 27,306 exact records (53.02%), ii) 24,189 erroneous recors (46.97%).

28. The results seem to confirm the general good quality of the census data and of the data processing as a whole.

V. Conclusions

29. The Albanian experiences with the Population and Housing Census highlights the fact that scanning is a relatively cheap way to considerably speed up data processing. Looking at a population of some 2,850,000 people, we can expect the full set of cleaned data to be available months earlier than via manual data entry, employing a modest number of personnel. This situation will lead to a much faster availability of census outputs, with final census tables expected to be available 9 or 10 months after completion of the census enumerations.

30. Also the experience with the SMS and the web based GIS Monitoring System was highly valuable and very useful to identifying issues related to the coverage in a timely manner.

31. Finally, the inital investment in terms of hardware/software and knowlgment will be reused for future censuses and other surveys. INSTAT is indeed planning to conduct the Agricultural Census in October 2012 using the same technologies adopted with the Population and Housing Census.