PROCESSING SAMPLES FROM CENSUSES IN DEVELOPING COUNTRIES

Karen A. Stanecki and Susan Enea Adamchak
U.S. Bureau of the Census [1]

Abstract

Recently the question has arisen to the benefits of conducting a full scale census (i.e. complete enumeration) in countries that can ill afford to conduct them. Also, there has been increasing pressure in the Sahelian countries of Africa to produce current, reliable population estimates for policy dialogue purposes. In an effort to speed availability of data, a number of countries conducting censuses in the 1980/1990 rounds processed only samples of census data. Advantages and disadvantages of these procedures are discussed. The cases of Burkina Faso and Morocco are elaborated in detail.

Introduction

In many of the countries in Africa, processing of the census had become a long and drawn out process. The reasons for this have included lack of planning, inadequate financial resources, and loss or lack of trained staff. In the past, delays almost always resulted from the inability to process the census data using computer installations found in some Central Statistical Offices. In many cases, by the time the census has been processed and published, the data are out of date. Many uses could be made of the population data for policy planning and various ministries in these countries have begun urging the national statistical offices to release these data in a more timely fashion.

In Africa, perhaps more so than in the rest of the world, the need for timely data is particularly acute, given the short history of census taking in the region, and the consequent lack of data required for population and development planning. In response to this increasing pressure, a number of national statistical offices have opted to process samples of the census records.

A number of countries worldwide have implemented census data processing plans which give priority to quick publications of sample results. The procedures have been carried out with varying degrees of success. The People's Republic of China, with a population of more than one billion, was able to tabulate and publish results of a 10 percent sample fifteen months after the 1982 enumeration. Aware of considerations of cost and precision, Hong Kong (population 5.7 million) conducted a 100 percent headcount by age and sex, and a 20 percent sample of social, economic and demographic characteristics in the 1981 census. Data were input and edited in fifteen weeks, and publications based on the sample data released soon after. New Zealand (3.3 million) canvassed the total population in the census of 1981, and then selected a 10 percent sample of returns which were given priority in coding and data processing. This enabled initial publication of data by November 1981, eight months after the census date, while processing of the rest of the returns continued.

Not all cases of sample processing are successful. Zimbabwe, for example, published data from a 10 percent sample of all census returns in 1985, nearly 3 years after the enumeration date. Processing of the full census is still in progress, five years later. Thus, the country has realized neither a time savings, nor are the full data available.

Central Statistics Offices now planning censuses for the 1990 round stand to learn from the experiences of other countries. However, the decision to process a sample of records should not be made without careful consideration. Processing a sample of records could seriously delay the processing the entire census. Or, as occurred in Morocco, the complete census is never fully processed limiting the usefulness of the census in not having the data for small areas available.

Censuses also provide a sampling frame from which future data collection activities are based. In Chad for example, the lack of a sampling frame (the last census having been conducted in the late 60's) has stymied all attempts to collect information for regional agricultural planning and sectoral health planning.

This paper describes the experience of two countries that decided to process a sample of records 1) to reduce the costs of processing--Morocco and 2) to provide advance tabulations--Burkina Faso.

Morocco Background

In considering sampling in the census process, one normally thinks of applying it at the collection stage, gathering some data items on a 100 percent basis and others (more complex or detailed items, and subsequently those requiring extensive office coding) on a sample basis. Generally this is carried out with the use of two questionnaires, a short form with the 100 percent items only and a longer form containing both 100 percent and sample items. Two mutually exclusive and exhaustive subsets of the total population of listed living quarters (housing units) are predesignated, usually in a systematic fashion. One subset is administered the long form and the other the short form. As an example, one-fifth of the listed units may be chosen for the long form and the remaining four-fifths receive the short form. Since the long form also contains the 100 percent items, the 100 percent data information is available for the entire population.

In preparing for the 1982 Census of Population, the Moroccan central statistics office, Direction de la Statistique (DS), stated its unwillingness to collect data using the combination 100 percent and sample approach described above, on the grounds that there is a requirement to make available some of the results of the 100 percent information very quickly. They would therefore have to code and key the 100 percent items from all the questionnaires (long and short forms) to produce, for example, data on educational attainment, and then repeat the keying operation for the 100 percent items later

when the more detailed long form information is readied for processing. This then would result in a wasteful and costly duplication of effort.

Instead, DS collected the detailed data on the entire population, but processed and tabulated only a sample of it. It was pointed out that this procedure was also wasteful since a large proportion of the information on the questionnaires collected might never be used. Indeed the expense of conducting never-to-be-used detailed interviews for 75 percent of the households, or about 2.75 million households, far outweighs the expense of keying the sample households twice (about 900,000 households).

DS was reluctant to consider collecting data on a sample basis, although they were aware of the time and cost to be realized. DS was uncertain of its ability to convince government officials that sample data collection would yield a valid census; instead they wanted to hold their options open by collecting data on a 100 percent basis. They reasoned that if the data were collected for everyone, the entire census could be tabulated later if complaints were voiced that the sample tabulation was invalid or otherwise unsatisfactory.

The U.S. Bureau of the Census was requested to provide technical assistance to DS at all stages of census operations in Morocco. The U.S. Census Bureau's first recommendation was to reconsider the decision not to use sampling in data collection if it was likely that only a sample of the census returns would be processed. Both time and money would be saved if sampling was done at the field collection phase, rather than at the data processing phase. Given the DS commitment to proceed with their preferred plan, the Bureau of the Census next made recommendations for the sampling plan to be used to process the census returns.

Methodology

The methodology for processing a sample of census records called for dividing the census forms into random subsamples each of which would represent the entire population. The main purpose of dividing the households into distinct random subsamples was to make it possible to obtain preliminary national estimates in a timely fashion by initially processing one or more subsamples. Also this provides readily available standard error estimates if more than two or three subsamples are processed. Additional subsamples could be processed in order to produce reliable estimates at the provincial and communal levels. That is, the number of subsamples to be processed increased as the level of geographic tabulation became more detailed.

A sampling rate of 1 in 20 was recommended for national level estimates. Households were divided into distinct groups, each consisting of a 5 percent probability sample representing the nation. A 25 percent (5 random subsamples) sampling rate would be used for the provincial level statistics, including urban-rural splits.

The design recommended by the Bureau of the Census required that each subsample be uniquely identified; i.e., it was necessary to code the household census forms serially within each enumeration area (EA). It was recommended that serial codes from 00 to 19 should be assigned systematically to the forms, with a random starting number for each EA.

Although this coding procedure could have been carried out by the supervisor in the field after all the census forms for an EA had been completed, it was recommended that it be carried out during the office check-in procedures in order to ensure more quality control.

The processing plan called for the selection of a random number between 00 and 19 to determine the first subsample which would be processed to produce the preliminary national level estimates. As soon as the first census forms were checked in and assigned subsample codes, processing could begin. The additional subsamples would be selected at random without replacement. That is, a random integer between 00 and 19 excluding the first random number selected for the national estimates, would determine the next subsample. The remaining subsamples were to be selected in the same way excluding at each step the subsample codes previously selected, until the desired number of additional subsamples was obtained.

In reality, DS implemented a different sample selection procedure. Instead of selecting the households after the enumeration during the office check-in procedures, the "cahiers du district", the enumeration booklets, were preprinted with an "X" by every fourth line. Among those lines identified with an "X", every fifth appeared with an additional preprinted "X". In this way, after the census enumeration was complete, the households entered on the lines identified with "XX" constituted the 5 percent sample for the entire country, and the households with at least one "X" constituted the 25 percent sample. One quarter of the "cahiers du district" started with an "X" on the first line, another quarter with an "X" on the second line, another quarter with an "X" on the third line and the remaining quarter with an "X" on the fourth line. Within each of these four groups of booklets, there were five different starts for "XX", so there were a total of 20 different types of booklets.

There are a number of problems associated with these procedures. One problem concerns the preprinting of the "X"'s in the booklets, to identify the households in the 5 percent and 25 percent samples. If the enumerators and field supervisors knew that the questionnaires for the indicated households were going to be reviewed and processed in the office, they may have tended to treat them differently from the other questionnaires, thus introducing a bias. Although the enumerators were not told the purpose of the "X"'s, it is quite likely that they were curious about these marks and found out why they were included in the booklets or assumed that there was something special about them. Another disadvantage of the preprinted "X"'s is that whenever a line was scratched out or otherwise skipped (i.e. lines missed or left blank), the systematic selection determined by the "X"'s would have been biased.

The only information processed on a 100 percent basis was the "population legale". This information was contained in the district book which summarized the number of persons in each household into two categories: Moroccans and foreigners. This booklet, filled out by the

enumerator, had one line per household with columns for the housing unit and household identification numbers and the number of persons in each category.

Here is an example of considerable waste in field enumeration since only a sample of the household records were ever processed. Not even basic demographic information by household was ever processed. There remains the question as to whether or not the sample was biased due to the preprinting of the forms. Although the "population legale" was eventually used for a master sampling frame from which various national surveys were selected by DS after the census, details on smaller areas from the census were never realized since the complete census was not processed.

Burkina Faso Background

The "Insitut Nationale de Statistique et Demographie, INSD, decided to process the Burkina Faso 1985 population census on an ensemble of 25 microcomputers and related equipment. This decision was based on the availability of external funds with which to purchase computer equipment and the uses to which the equipment could be put after the processing of the census.

Initial manual counts of the population after the December 1985 enumeration indicated a resident population of 7.9 million, rather than the 7.0 million that had been projected from the 1975 census. There was considerable interest in the Burkina government in confirming the size and geographic distribution of the population and analyzing the implications of a population 15 percent larger than anticipated. It was estimated that data entry of the complete enumeration would take 18-24 months. As a result INSD decided to take a sample of the census records in order to generate preliminary tabulations of the census, following which the remainder of the census results would be entered and tabulated. It was predicted that initial sample results would be produced by December 1986 and the complete census would be processed by December 1987.

Methodology

There were three factors concerning the enumeration booklets that placed constraints on designing the sample of census records for processing. First was the variability in the size of the enumeration booklets. There were two types of booklets, one with 30 pages that included up to 60 households and the other with 60 pages that included a maximum of 120 households. The second factor was the variability in the number of people included in the booklets. Each page in a booklet was generally a separate household except when large households continued on a second page. Some pages only had a few people listed on them whereas others were nearly completely filled in (10 lines per page). There were booklets with only a few households listed in them and others that were completely filled in. The third factor was that information was required on concessions which are groups of households. Concessions were not always completely listed in the same booklet. The concession may have begun in one booklet and finished in the following.

Other constraints to designing the sample were the geographic levels for which the data were required and how the census forms were organized. The data were requested for the nation as a whole and for each of the 30 provinces. The most important table to be processed at each of those levels was the age-by-sex distribution.

The administrative organization of Burkina-Faso is as follows: province, departments, communes, sectors, villages. Villages vary in size from 100 to 1,000 people. It was necessary to create zones for the census that were less variable in size to control the enumerators' work load. Villages were either divided or grouped together depending on their size to form zones de control (ZC) of 4,000 to 6,000 people. A ZC was enumerated by one team of enumerators. Zones de denombrement (ZD) were formed within ZCs of about 800 to 1,000 people and were enumerated by one enumerator. There were a total of 7,696 ZDs in the country.

Since the enumeration booklets often broke up concessions, the next available unit that could be used as a primary sampling unit was the zone de denombrement.

Ideally, since the ZD's were large, a sample of ZD's would have been selected with subsampling done within them. This would have spread the sample geographically and would have reduced the sampling variance. However, if subsampling was to have be done within the sampled ZD's, a relisting by individual concession would be required within the selected ZD's enumeration booklets. Due to the size of that task and the lack of time to accomplish it, the decision was made to include the entire selected ZD in the sample.

The primary factor in deciding the sample size was the need to have a sample for the smaller provinces sufficient to provide reliable estimates for the age/sex distribution. The age distribution was broken into 5-year age groups from 0-4 to 75+. The smallest province had an estimated population from the manual counts of 105,515. By looking at the expected number of sample population for each of the age categories, it was estimated that a 10 percent sample would provide a sufficient sample to produce reliable estimates for the older age categories.

INSD believed that the urban areas were more heterogeneous than the rural areas and suggested that the sample be increased to provide reliable estimates for the tables on occupation. The urban areas contained 12 percent of the population based on the initial manual tallies. An increase to 25 percent of the urban ZDs increased the total sample size from approximately 797,600 to 938,000 which INSD felt that could be handled within the time frame allocated for data processing.

The sample was selected independently by province and by urban and rural areas within province. There are a total of 30 provinces in Burkina Faso of which five provinces have a major city and eight other provinces have an urbanized area as defined by INSD. Seventeen provinces are completely rural. A sample of 1000 ZDs out of a total of 7,696 ZDs was selected systematically, probability proportionate to the resident population size of the ZDs.

The microcomputers arrived and were installed early August 1986. By early September, data entry personnel were hired and trained and data entry began mid-September. By mid-December, the 10 percent sample of census records had been

entered. The processing of the tables however, was held up not by the data entry but by the delay in table programming. Data entry continued on the 90 percent of the records and was completed by the end of June, 1987.

In less than a year, the data entry was complete. The programming of the tables may have been completed sooner had special processing for producing the variances for the sample had not been required. Although tested software was used to produce the tables, separate programming was required to produce the estimates to calculate the standard errors. Past experience showed it to be more efficient to produce the tables and their accompanying standard errors at the same time.

Questions on editing specifications also delayed processing of the tables. The edit rules were debated in great length as to their implications for the data. These same discussions would have arisen with the processing of the complete census.

If the processing has started as soon as the data entry of the sample records was finished, the sample would have produced results 6 months in advance of the complete census. The data entry of the sample was complete by the end of December 1986. The data entry of the entire census was expected to be complete by the end of June 1987 and was actually finished the end of July. (Money for the data entry clerks ran out and new funding had to be found.) Processing of the tables did not begin until March and were not complete until the end of April. In actuality the sample saved at most three months, the difference between the end of April and the end of July. The tables will be rerun for the complete count.

Although the ZDs as primary sampling units were designed to be about 800 to 1,000 in population, they varied a great deal in size from a population as low as 200 to a population greater than 3,000. This resulted in some very high sampling errors for the initial tables produced by province. Ouagadougou, the capitol city, in particular had large variability in ZD size and as a result the estimates for Ouagadougou had large sampling errors associated with them. Had the plan been to process a sample of the census records at the beginning of the census planning phase, the enumeration booklets could have been designed differently to allow for a more efficient sample design.

The complete set of detailed tables that INSD wanted to produce based on the sample data should not be produced due to the large sampling errors particularly, for the urban areas. These tables will not be of much use as the population estimates are not reliable. The general summary tables are the only ones which will provide any estimates with reasonable sampling errors. INSD will have to wait until the complete census is processed to produce the detailed tables within provinces.

Conclusion

Sampling may be considered in the tabulations of census data for two reasons: to prepare advance tabulations, and to reduce data processing costs.

There is always a demand to obtain information from a census as soon as possible. For the population count, the census offices in many countries release preliminary figures based on simple hand tabulations. For population characteristics immediate demands of users can sometimes be met by tabulating a sample of the returns. A decision to provide advance estimates from a sample should not be made without careful consideration, since this effort can seriously delay completion of the planned series of complete census tabulations.

If a country uses computers and tested software to produce tables in final form, the need for advance tabulations diminishes rapidly because sampling will save very little time. This in particularly true in Sub-Saharan Africa where the populations are not that large; 35 out of the 47 countries have populations less than 10 million. As was the Burkina Faso experience with a population of 7.9 million, coding and data entry of the census records went very quickly. It was the programming of the tables that required more time. Sampling census records did not significantly save time.

If the census planners determine that the advance tabulations are needed, they should probably limit them to a few items of vital importance to the country. The sampling errors from the sample of records from the Burkina Faso population census limited the use of the data. Only general summary tables on immigration, fertility and mortality will provide reliable estimates.

Aside from advance tabulations, another reason for processing and tabulating a sample of the results is to save time or reduce costs. Savings can be appreciable for complex subjects requiring much manual editing and coding of the questionnaires. In general, it is preferable to collect the items requiring considerable manual editing and coding on a sample basis whenever sample results are acceptable. If information is collected on a 100 percent basis, it seems wasteful to use only part of it. Although the Moroccans conducted a complete enumeration using a detailed census form, most of that information will never be exploited since the complete census was never processed. The costs savings in processing only a sample of the census records are negated by the cost of conducting a full enumeration in the field. The use of computers, where there is a large fixed cost regardless of the amount of data, will reduce the gains from tabulating only a sample of returns.

In those cases where it has taken 5 years to process census data, advanced planning in terms of staff and equipment would have made more of an impact on timeliness of processing than the use of sampling procedures for processing. The availability of microcomputers at costs far lower than mainframe computers makes for an ideal option, either to process the entire census on microcomputers or just for data entry purposes as was done for the 1985 Population Census in Colombia. After data entry, the data was transferred to mainframe where the tabulations were made and the results were published one year after the census date.

It is imperative to stress the importance of adequate planning for all stages of the census process. In many cases sampling is undertaken in an effort to compensate for unanticipated shortfalls in staff, equipment or resources by speeding up processing. Unless a sampling plan

is carefully designed, and its full implications realized from the outset, however, it is unlikely to produce the high quality and useful data required for national planning.

REFERENCES

"Censuses of Population and Housing in Africa" Ben Kiregyera, Statistics Sweden, Vol. 2 No. 4, 1986.

"Annotated Bibliography on the Sources of Demographic Data" Vol. 1, Africa-Near East Organization for Economic Cooperation and Development.

"Census of Asia and the Pacific - 1990 Round" East-West Population Institute, East-West Center, Edited by Lee Jay Cho and Robert L. Hearn.

[1] This paper reports the general results of the research undertaken by Census Bureau staff. The views expressed are attributable to the author(s) and do not necessarily reflect those of the Census Bureau.