(File REDA15.PUB 28c/06/84)

GEOGRAPHICALLY DISAGGREGATED CENSUS DATA FOR PLANNING IN DEVELOPING COUNTRIES

Contributed paper
to the IUSSP 1985 General Conference, Session F.15:

Utilization of Demographic Knowledge
in Policy Formulation and Planning

Arthur M. Conning
Latin American Demographic Centre (CELADE)
Santiago, Chile

Geographically Disaggregated Census Data for Planning in the Developing Countries

Arthur M. Conning
CELADE
Santiago, Chile

I. INTRODUCTION

This paper will describe the practical difficulties that many agencies in developing countries face to obtain quantitative population data on small areas and will propose a concrete solution to facilitate their access to such data. The discussion, stimulated by a recent examination of selected Latin American and Caribbean statistical offices and their users, will focus on the utilization of geographically disaggregated population data for planning and the provision of social services; it will not consider aspects involved in the utilization of the same information for policy relevant social science research, since the needs and practical solutions are likely to be somewhat different.

A. The role of the National Statistical Offices in the supply of population data

The national statistical office is the major collector and supplier of population data in most developing countries. Although other agencies may also collect population-related information, the national statistical office is the official institution responsible for the national population and housing censuses and usually organizes and makes available vital statistics and carries out periodic household surveys and various special surveys.

National statistical offices are basically supply-oriented, in that their primary purpose for collecting, processing and publishing data is to supply quantitative data to other agencies, rather than to use the data for internal ends (Data for Development International Association, n.d.; p31-32).

In the Latin American and Caribbean region, and probably in other regions as well, considerably less use is made of the population data collected by the statistical offices than expected. For instance, in one of the few systematic studies in the region of the utilization of the results of population studies, Ortega (1980) found that none of his informants in the Dominican Republic cited the 1970 census as a source of information that directly influenced policy or program decisions, although it was a source of general statistics.

B. Identifying population data supply problems

While there are many reasons why existing population data and substantive results may not be adequately utilized by agencies that 'should' be doing so, a problem-solving approach to increasing utilization is to first identify, and then help meet, the needs of would-be users who experience difficulties obtaining the population data that they explicitly request from the national statistical office in their country.

II. FACTORS AFFECTING THE SUPPLY OF GEOGRAPHICALLY DESAGGREGATED POPULATION DATA FROM STATISTICAL OFFICES TO THEIR USERS

A. Examination of the experience in selected statistical offices

Using the approach outlined above, the Latin American Demographic Centre (CELADE), with funds provided by the International Development Research Centre (IDRC) of Canada, carried out an examination in 1983 of the difficulties encountered by national users in the retrieval and use of quantitative population data produced by Latin American and Caribbean statistical offices. Emphasis was placed on identifying key problems that would be amenable to solution through technical assistance, training and technology transfer activities of CELADE.

Information was obtained on the situation of the statistical offices and their users in seven countries. Although they cannot be considered to 'represent' the Latin American and Caribbean region since every country has its own special characteristics and needs, they do cover various different situations with respect to physical and population size of country, language, cultural background, computer facilities and experience of the statistical office, etc. The countries were (1983 estimated populations are given in parentheses): the English-speaking Caribbean island countries of Trinidad and Tobago (1.2 million) and St. Lucia (125,000); the Central American country of Costa Rica (2.5 million); and the South American countries of Bolivia (6 million), Chile (11.6 million), Peru (18.7 million) and the Brasilian state of Sao Paulo (23 million in the state).

The computer access and data processing experience of the statistical offices varied (in 1983) from St. Lucia, which has no computer and had its 1980 census processed on a computer in another country, through Bolivia, Chile, Costa Rica and Peru with varying degrees of computer power and experience available, to the state of Sao Paulo statistical agency that utilizes advanced computer technology to supply a variety of numerical and bibliographic information services (however, since it does not have the 1980 census data for the state, requests for reprocessing are referred to the national statistical office).

In each statistical office, the Director and substantive and technical staff were asked during unstructured interviews to identify user agencies which have made large or frequent requests for population data. The relevant persons in many of these were in turn interviewed, using a similar set of topics as in the statistical offices.

Description of the problem

Based on the findings of the survey of the above-mentioned countries (see Conning, 1983 for a detailed report) a major data supply problem involving geographically disaggregated population data was identified; it may be defined in terms of the following points:

(1). Among the many factors that must be taken into account when planning, constructing and operating infrastructure projects and social services are the characteristics and spatial distribution of the local labor supply and of the population that will be benefited or otherwise affected. Physical planners involved in a concrete spatially located project normally require this data for the specific area of interest, in contrast to social science researchers who usually want information on, say, all small areas (however defined) within a city, region or a country. Local planners seek to specify while researchers seek to generalize.

(2). Summary population figures are seldom sufficient for planning a spatially located project. For example, knowing only the total number of women in the catchment area of a new hospital does not permit a careful analysis to be made of the number of maternity beds required. Rather, tables such as those showing women by marital status, age and recent fertility must be prepared for the planner.

(3). The population census in most developing countries is the only source of existing population data that has a large enough number of households and individuals to permit useful tables to be obtained for small geographical areas of a country.

(4). As the area of concern to a development project seldom corresponds to political or administrative boundaries, users want to construct the actual area by aggregating the census data of various contiguous smaller areas such as enumeration districts.

(5). It is impractical for a statistical office to publish volumnes or keep computer printouts of all the possible tables that might be requested in the future for all the small areas of an entire country (many Latin American nations have more than a hundred thousand enumeration districts).

(6). If a planner requires a set of tables for a specific area that cannot be made to correspond to one or a combination of existing tables available from the statistical office, the only present recourse is to return to the microdata for reprocessing.

(7). Most statistical offices in the Latin American and Caribbean region are not equipped to re-process census microdata *rapidly and at low cost*, in part because the data files are large and are conventionally organized and processed. Furthermore, since the statistical offices have many other important functions, the request for special small area population tabulations by another agency usually receives very low priority, resulting in long waits by users for urgently required quantitative data.

(8). Most user agencies cannot do their own re-processing of the census data since they do not have access to programmers or to computers adequate for reprocessing the large files of conventionally organized census microdata.

The words 'rapidly and at low cost' are emphasized above since, not withstanding the fact that most of the statistical offices examined can produce census tabulations (except the small Caribbean island which does not have direct access to a local computer), there are normally delays of many months and/or prohibitively high costs for most users. As these delays and costs are normally the result of having to employ scarce computer programmers and having to use large, costly and frequently over-burdened mini- or mainframe computers, any solution must eliminate dependence on both programmers and the main computer. This suggests that it would be desirable to base a solution on the direct utilization of low-cost microcomputers by intermediate users (such as library staff) and final users.

III. OUTLINE OF A MICROCOMPUTER-BASED SYSTEM FOR PROVIDING SMALL AREA CENSUS DATA IN DEVELOPING COUNTRIES

The remainder of the paper will outline the design of an approach that would greatly facilitate the rapid, low-cost provision of ad hoc census tables required for physical planning and related needs. The development of the system, to be known as REDATAM ('Retrieval of small area census DATA by Microcomputer'), is being considered by CELADE.

As the goal is to improve the access of planners and others to such data as soon as possible, efforts are being made to locate existing systems that could be adapted to the conditions in Latin America and the Caribbean. Todate, however, no systems have been found that have all the features desired (see Section III.B).

A. Definition of the REDATAM system components

The REDATAM system has two basic components, one to create and store the database of census information and the other to retrieve subsets of this information in tabular form. The components have distinct functions and different relationships to the ultimate users.

Database creation and storage component

Because the REDATAM database file of census data will be large even in the case of a small country, the database in each country will have to be specially organized to maximize retrieval speed and convenience (see Section III.C). While the approach used to store the information will have an effect on the flexibility possible during retrieval and on the rapidity of response, in all cases a careful analysis will have to be made by substantive staff, in collaboration with information specialists, on the nature of the tables that are likely to be retrieved (e.g., what combinations of variables, codes and areas will be requested) and the probable frequencies of different types of requests.

This component of the REDATAM system normally will be used only once with each dataset to create the corresponding database. The procedure will involve converting the original census data conventionally stored on magnetic tapes into a suitable database and then transferring the information to the microcomputer hard disk that will be used for retrieval. It is assumed that a programmer using a mini- or mainframe computer will carry out this 'one-time' database creation operation.

Data retrieval component

The REDATAM data retrieval component, which is the public face of the system, will be used with the database each time a table is requested. Hence, this component must be microcomputer-based, so that access to a large computer does not have to be obtained, and oriented toward easy operation by information suppliers and their users without dependence on programmers. One can envision the microcomputer placed, for instance, in the public library of the statistical office, and if the demand is high enough in some individual agencies, directly in them.

B. System specifications

To be useful in the conditions found Latin America and the Caribbean (see Section II.B), as well as in other developing regions of the world, it is desirable that the REDATAM system meet at a minimum the following specifications:

(1) The data retrieval component must be a user-friendly, interactive system (for the set-up of the request), and able to be operated in its normal retrieval mode without programmer assistance. Ideally, the user should be able to 'learn' to obtain the tables required with little or no assistance from programmers or from CELADE, perhaps through computer-aided instruction or video cassette ('user training at a distance').

(2) Requests for ad hoc tabulations should be possible for any small area identified in the microdata and for any variables and codes available in the

original data. It should be noted that a planner can make at least two different types of request: (a) retrieve data on a *specific* small area, which with proper database organization and retrieval software should be possible without passing through the data of other areas; and (b) retrieve data on, or identify, all areas of the country that have a given population characteristic (e.g., all districts with less than 50% of the school-age population attending school), in which case the data of the entire country must be processed.

The REDATAM system will be designed to answer the first type of request -generation of tables for *specific small areas*-, which is far more frequently required by physical planners than the second. When planners or social scientists make the second type of request for a whole country or a large region, it will be more efficient to use a large computer.

(3) As requests will normally be for tables for specific areas (or contiguous areas), the system need not have the capacity to store the entire country on its hard disk, but it should be able to make different segments of the country rapidly available on the same machine.

(4) The retrieval component should permit easy recoding or construction of any variable or combination of them and should allow the selection of subsets of the population within a geographical area (e.g., tables referring to employed males aged 12 to 64).

(5) It should be possible to construct a table for any area of interest using information aggregated from smaller (normally contiguous) areas.

(6) The response time to obtain tables should be reasonably short, i.e., a user visiting the statistical office should be able to request a set of tables, begin the retrieval and wait for the printed result.

(7) It should be possible eventually to link the system to other databases so that maps of the area of interest can be drawn and/or population data related

to other spatially located information.

(8) The REDATAM system software should be written to permit easy conversion to different microcomputers so that eventually there is at least one available from an authorized dealer providing maintenance in each Latin American and Caribbean country (or subregional grouping of countries). The microcomputers selected and the peripheral hardware should be reliable even when used without special climatization equipment.

(9) Although it is difficult to speak of hardware prices, which vary from country to country, the total cost of the hardware and any associated commercial software that might have to be purchased should not be more than around, say, US$10,000 to $12,000 (for government agencies exempt from importation duties) at the time that the system becomes available for distribution. Because prices are dropping, the system can be designed with, say, more hard disk space, than can presently be purchased for the same money.

C. Selection of the storage methods

There are at least two general storage alternatives to consider in the development of the microcomputer-based REDATAM package: aggregate data storage and microdata storage.

The aggregated data storage approach

The aggregate data storage approach would involve the one-time creation of a database of all the tables of likely present and future interest for each small area. For each such area, one may imagine a 'super-table' with all the variables and their codes of interest, from which any smaller tables can be reconstructed according to the request of the user. As this approach would store actual tables for each area, a priori decisions would have to be made on the tables, the level of detail of the codes for each variable, and the areas.

Unfortunately, however, when the usual variables of importance are included with the codes normally available, the number of cells in the super-table will be very high, as seen from the following example (the number of categories for each variable is given in parenthesis):

Rural/urban residence(4) x Sex(2) x Age(100) x Marital status(7)
x School attendance(3) x Educational level(30) x Type of activity(12)
x Occupation(100) x Industry(100)

which gives a super-table of 60 billion cells for *each* geographical area! Even using data compression techniques, such as the removal of cells with zero population, the storage required for each area will be enormous.

Thus, for this approach to be feasible, the number of variables and their codes would have to be very limited and the total number of areas small, even though this forced pre-selection would result in the loss of much information. For instance, many detailed occupational codes would have to be grouped making it impossible for a fisheries development project to determine how many fishermen were available in an area if that occupation happened to be grouped into a more general category. Furthermore, as it would be difficult to store even a small super-table for every census enumeration district, planners would not be able to build up ad hoc planning areas from these smaller areas. Consequently, although some variations of the table storage approach might be very useful for social science analyses, it is not very convenient for detailed planning purposes.

There are various systems that provide aggregate local area data, such as the Small Area Statistics Package (SASPAC) which makes selected information on to the enumeration areas available for various British censuses (LAMSAC, 1983), and the Kentucky Economic Information System (Renfro, 1980). These and others, such as that based on geocoding developed by Statistics Canada (1982), work on mainframe computers or large minicomputers and are far too costly to consider for use in developing countries. The Urban Data Management Software (UDMS) developed by the United Nations Centre for Human Settlements (HABITAT, 1983), on the other hand, works on a small microcomputer that stores simple aggregate summary data for areas within, say a city. While it has many interesting features, for the purposes here, it is very limited and cannot provide detailed tables for individual areas.

The microdata approach

This approach, which would store microdata, i.e., the information on each variable of each of the individual units of observation, has the huge advantage that tables at any level of detail can be produced for any area that is coded in the microdata or for any combination of such areas. The storage required would be affected primarily by the number of records (inhabitants) in the country.

If a census record with, say, 60 characters can be compressed to 30 without loss of information, then, in this case, a 10 megabyte (MB) hard disk could store information on around 300,000 persons (approximately 10,000,000/30) or a 50 MB disk on around 1.5 million persons, which is enough for many small countries of the Caribbean. If a fast method of loading the microcomputer hard disk is used (such as streaming tape), a larger country could be divided into segments, each of which could be made available quickly whenever needed.

The far greater flexibility of this alternative makes it the preferred choice for the REDATAM system. The crucial question will center around the efficiency with which the database can be used with a microcomputer and associated software to produce tables for given small areas (which can have substantial populations in large cities). The next section provides suggestions on how to organize the data for this purpose.

D. Organization of the microdata database for efficient table production

If the records of microdata are organized properly and stored on a hard disk during processing it should be possible to jump directly to the geographical area of interest to produce tables for that area, without passing through all the other previous data. With suitable software, this requires only prior sorting on the codes for geographical area during the creation of the REDATAM database. Furthermore, if prior substantive analysis of likely needs shows that a high percentage of tables are requested for particular subgroups, as the school age population and the economically active population, then these categories can be subgrouped by sorting within each geographical area, allowing a jump directly to the subgroup of interest within the geographical area.

In addition, it may be efficient to use a transposed file so that only the variables of interest for each person need be read, rather than reading all the variables for each person as in a conventional file. In a transposed file there is a record for each variable with the value (code) for each person on that variable; in contrast, in a conventional census or survey file there is a record for each person with the values for each variable for that person). Only the variables of interest need be read in the transposed file, while in a conventional file the entire record of a person has to be read to obtain the information on the variables of interest.

IV. IMPLEMENTATION IN THE DEVELOPING COUNTRIES AND IMPLICATIONS FOR THE 1990 ROUND OF CENSUSES

Although appropriate technology will help alleviate difficulties of providing small area data in many developing countries, improved technology is only a starting point. There still will be many substantive problems with the data and with the actual application of the system in the environments found in specific countries and their statistical offices. As many of the difficulties that will arise have their roots in the way that the previous censuses were organized and carried out, the more that REDATAM (or other similar system) is used to facilitate the supply of small area data, the more evident will be the deficiencies and defects in the previous censuses, which in turn, it is to be hoped, will stimulate efforts to reduce these defects in the 1990 round of censuses.

First of all, as noted above, the efficiency of the system for producing tables will depend on the organization of the database in each specific situation, which will vary from country to county, depending on physical size, population, the intended use of the data, the variables included in the census, their definitions, the data quality, etc. These are primarily substantive matters that will have to be decided in each case and the corresponding decisions taken with respect to the organization of the database.

Second, for very small areas, the 1980 and prior census data are likely to have magnified errors due to poor control over the original data gathering and data entry. While an attempt has been made by all the Latin American and Caribbean countries with 1980 censuses to check and correct the logical consistency within individual records and, to some extent, within households, much less effort has been devoted to securing consistency for enumeration districts. As these are small and normally refer to the work of a single enumerator, the accidental entry of a package of questionnaires in the wrong area may greatly increase the population in one area and greatly reduce it in the other, even though for larger areas the errors will disappear. Of course, similar gross errors can result from packages doubly entered in the same area or simply lost. If the use of small area data is stimulated, the problems in its substantive utilization may lead to improved procedures in the 1990 censuses.

Third, a related problem results from the use of field sampling in various censuses for the application of many questions. In areas with very small populations, sampling error (as well as response error) will be high. REDATAM should be able to 'refuse' to supply very small populations in such cases.

Fourth, while the database can be created and tables for specific areas generated without consulting maps if the codes of areas of interest are known, most users requiring information on ad hoc planning areas will have to define the area of interest by grouping smaller areas located on a map. Consequently, the maps employed for the census data collection will be fundamental for such users. The existence and quality of these maps will vary among areas within countries and among countries. To the extent that small area census data is made more easily available, the procedures and the emphasis placed on coordinating the maps with the data may change during the 1990 censuses. In addition, some countries may be stimulated to consider the use of geocoding in their next census; this in turn has its dangers in some situations because it may further complicate the difficult and massive census-taking operation, which could in some cases reduce the overall quality of the census or lead to delays in producing the national results.

Finally, as small area data becomes more easily available, new groups of regular users may appear from both government agencies and private companies, increasing the overall utilization of the census data and perhaps creating additional pressures for the improvements suggested above as well as for new or different questions in the census. Thus, the use of new information technology could play not only an important part in the integration of population variables into the planning process, but may lead to improvements and increased richness of information in the censuses and increases in their utilization.

BIBLIOGRAPHY

Conning, Arthur M., 1983. Report to IDRC on the REDATA Pre-Project Mission, 6-24 June 1983: An examination of problems encountered by national users in the retrieval of quantitative population data produced by Latin American and Caribbean statistical offices. CELADE, Santiago, Chile.

Data for Development International Association, n.d. Reproduced by: National Technical Information Service, U.S. Department of Commerce, Springfield, VA. PB 274 079.

HABITAT, 1983. Urban Data Management Software (UDMS): Users's Manual. United Nations Centre for Human Settlements (Habitat), Nairobi. HS/23/83.E

LAMSAC, 1983. Brochure: '1981 Census Small Area Statistics/SASPAC'. Local Authorities Management and Computer Committee. London.

Ortega, Manuel M., 1980. Utilización de investigaciones en Republica Dominicana: El caso de la Encuesta Nacional de Fecundidad de 1975. Santo Domingo: Instituto Technologico de Santo Domingo.

Renfro, Charles, 1980. An online information system for aggregate state and local economic data. Journal of the American Society for Information Science. September 1980.

Statistics Canada, 1982. Geography and the 1981 Census of Canada (Geography Series). No. 2 - GEO 82.

GEOGRAPHICALLY DISAGGREGATED CENSUS DATA FOR PLANNING IN DEVELOPING COUNTRIES

Summary

A practical approach to increasing the utilization of the population data collected by the statistical offices in developing countries is to identify and to seek to solve the particular problems that these offices have when trying to satisfy actual requests for population data. Using this approach, a 1983 study found that Latin American and Caribbean statistical offices are not equipped to meet requests rapidly and at low cost from planning and social service agencies for detailed small area census tables.

To solve this problem the United Nations Latin American Demographic Centre (CELADE) is concerned with the development of a system known as REDATAM (REtrieval of small area census DATA by Microcomputer), that will enable users to obtain ad hoc census tables for specific small areas without programmer assistance or the use of large computers. Census microdata, stored on a hard disk of a microcomputer rather than preselected tables, will be used to allow the creation of any tables from the original variables and codes for any geographical area identified in the microdata, or for any larger planning area that can be built up from continguous smaller areas. As the system will be designed to make tables for specific small areas, suitable retrieval software and database organization will allow rapid processing by jumping directly to the area of interest without passing through the rest.

It is likely that greater availability of small area census data in developing countries will uncover many substantive problems and deficiencies with the existing census data due to the way the previous censuses were organized and controlled. These difficulties, and the probable appearance of new users from government agencies and private companies who discover the value of the small area data to them, may stimulate pressures to reduce the deficiencies and make other changes in the 1990 round of censuses.