Distr.
RESTRICTED

LC/DEM/R.57
Series A, No.185
22 de agosto de 1988

ORIGINAL: ENGLISH

SMALL-AREA CENSUS DATA SERVICES BY MICROCOMPUTER
APPLICATIONS OF THE REDATAM SYSTEM IN LATIN AMERICA AND THE CARIBBEAN

ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN (ECLAC)

LATIN AMERICAN DEMOGRAPHIC CENTRE (CELADE)
Casilla 91 - Santiago de Chile
Tel: 2283206 Cable: UNATIONS

IASSIST 14th Annual Conference
Washington, D.C., 26-29 May 1988

SMALL-AREA CENSUS DATA SERVICES BY MICROCOMPUTER:

APPLICATIONS OF THE REDATAM SYSTEM IN LATIN AMERICA AND THE CARIBBEAN

Arthur Conning and Ari Silva
CELADE, Casilla 91, Santiago, Chile
and
Lawrence Finnegan
US Bureau of the Census
Washington, D.C.

Table of Contents

ABSTRACT

A study in 1983 found that census information for specific small geographical areas was often not available from the population and housing censuses in the Latin American and Caribbean countries because the political and administrative boundaries used in the census frequently do not correspond to the particular areas of interest and because most statistical offices were not able and/or willing to reprocess the large census files rapidly and at low cost on their mainframe computers. The interactive REDATAM [1] system, in English and Spanish versions, was created to solve the problem of providing small-area population and housing information by using an IBM or fully compatible microcomputer to store the microdata of an entire census on a hard disk (or laser disks for larger countries) and to permit any tabulation to be produced rapidly for any area down to city blocks or smaller.

The census (or survey) data is stored in compressed form (approximately one fourth of the original space requirements) in a database that makes it possible to access the data directly for a given small area without having to process the remainder of the data. Version 3.1 is presently available in English and Spanish with associated User and Database Generation manuals.

Facilities offered by the system include geographic selection, grouping of geographic areas, self-documented databases, interactive and batch processing, calculation of derived variables, use of weighting factors, hierarchical processing, generation of sub-databases, production of files for export to other packages and password protection.

REDATAM databases have been installed for 1980-round census data in Chile, Saint Lucia, Costa Rica, Uruguay, Dominica and Colombia and for survey data in Guyana. The processing efficiency of REDATAM makes it possible for the small Caribbean countries to use REDATAM for processing at the national level. In all the countries, hard disks have been used to store the databases, except in the case of Chile where the 16 million records are stored on optical "WORM" laser disks which permit writing data once. The use of hard disk or optical disk is essentially "transparent" to the user.

REDATAM may play an important role in the 1990 censuses in the Latin American and Caribbean countries since the system will permit the provision of timely small-area services (and at the national level in the Caribbean countries) before, as well as after, the regular data processing and publication of results are ready. This should greatly increase the use of the census data by both the governmental and private sectors, but will require some improvements in the data collection process and cartography to ensure the quality and convenience of using the small-area information.

A new two-year project will develop an extended system, to be known as REDATAM+, that will permit the cartographic display and analysis of population and other information through an interface with a Geographical Information System (GIS) and will associate multidisciplinary information describing geographical areas with multi-level population and housing data.

[1] REDATAM - REtrieval of DATa for small Areas by Microcomputer.

1. INTRODUCTION

REDATAM is a microcomputer software system for obtaining tabulations and other statistics for specific geographical areas rapidly and at low cost from large files of population or other data. Although the system may be used to process survey, vital statistics and other similar quantitative information to take advantage of its high-speed processing capabilities, REDATAM is oriented primarily to more masssive census datasets. It is designed to store all the original microdata (i.e., the values of each variable of each individual, person by person) of an entire population and housing census on an ordinary microcomputer and to allow users to obtain tables with any of the variables for any small-area within the country down to city blocks, normally within minutes and without special programming assistance.

While the microcomputer-based REDATAM system is not designed to be used for obtaining tabulations for an entire census in the medium and larger countries of Latin America, it does serve this purpose for the small countries in the Caribbean. It should be noted, none the less, that frequently the real time required to receive a REDATAM tabulation for a large city of a few million persons may be much less than the real time to obtain the same information with a mainframe in a national statistical office, since the microcomputer can be left running overnight whereas the mainframe work is often delayed due to the need for a programmer and tape manipulation.

2. THE DEMAND FOR GEOGRAPHICALLY DISAGGREGATED CENSUS DATA

In 1983, the United Nations Latin American Demographic Centre, CELADE [2], carried out a study [3] in seven countries in the Latin American and Caribbean region to determine the types of requests for numerical population data which the national statistical offices receive from the public and private sectors and which the offices have difficulty in answering. The countries selected covered 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 in parentheses): Saint Lucia (125,000), Trinidad and Tobago (1.2 million), Costa Rica (2.5 million), Bolivia (6 million), Chile (11.6 million), Peru (18.7 million) and the Brazilian state of Sao Paulo (23 million).

While the findings of the study covered a wide range of topics (see Conning, 1983, for a detailed report), the major population data supply problem faced by all the national statistical offices visited was that of fulfilling

[2] CELADE is part of the system of the Economic Commission for Latin America and the Caribbean (ECLAC).

[3] Financed by the International Development Research Centre (IDRC) of Canada.

special requests for data for specific geographical areas, usually from the population and housing census. For example, the Trinidad and Tobago agency, Town and Country Planning, with responsibility for rationalizing land use in priority areas of the island, had to wait for many months for the special tabulations that it requested.

The findings of this study helped identify a major data supply problem that can be defined in the following terms:

a). The development of new projects and programmes and the improvement of social services normally requires information on, for example, the characteristics and spatial distribution of the labour supply and the population that will be benefited or otherwise affected by the action.

b). The population and housing censuses in most developing countries are the only source of existing data that has a large enough number of cases to permit useful tables to be obtained for small geographical areas, but the tabulated information is normally only available for administrative and political boundaries that frequently do not correspond to the particular areas of interest to users;

c). Statistical offices dependent on large computers and programmers for working with census data cannot reprocess census data rapidly and at low cost to obtain small-area tables, because the census data files are very large, are conventionally organized and processed and because the offices usually give higher priority to their own regular activities than to requests for special tabulations by other agencies and private organizations.

By 1983-84, it was already clear that the solution resided in the use of standard microcomputers that were just being introduced on a wide scale in the Latin America and Caribbean region.

3. THE DEVELOPMENT OF REDATAM

As there did not appear to be any low-cost commercial or other software available that was able to store large population and housing census files (usually with many millions of records), rapidly locate the data for the geographical areas of interest and efficiently process extensive quantities of data to give results for most requests in minutes or at most tens of minutes, CELADE began to develop the new system, REDATAM, in June 1975 [4]. The system is written in the "C" language.

The basic concepts underlying the operation of REDATAM may be understood by visualizing the "data matrix" for a simple census (an actual database is usually more complex). The record for each person occupies a row of the

[4] A grant was provided by the International Development Research Centre (IDRC) of Canada for the REDATAM project. Work on REDATAM and its installation in various countries also receives continuing support from the United Nations Population Fund (UNFPA) and the Canadian International Development Agency (CIDA).

matrix; for each person there is a set of variables, e.g., age, sex, education, occupation, place of birth, etc. Each variable defines a matrix column.

```
     DATA MATRIX
                            VARIABLES for each person
                          Age  Sex   Educ ... Bthplace ...
                     ----------------------------------------------
          Pers 1 |     |                    |
     PERSONS     2 |     |                    |
   (in geograph-  3 |     |                    |
    ical order)   4 |     |                    |
                 . |     |                    |
                 . |     |                    |
                 . |-----|--------------------|-----------------
                   |     | Persons in municipio, La Florinda
                 . |-----|--------------------|-----------------
                 . |     |                    |
                 . |     |                    |
         567,822 |     |                    |
    Pers 567,823 |     |                    |
```

Normally, to make a tabulation of "Age by Bthplace" for all the persons, the computer reads the data of each person sequentially and uses the age and birthplace data found for each person to derive the table. The ordering of the persons in the data matrix is irrelevant. However, since REDATAM is directed towards obtaining tabulations for a user-defined geographical area, such as the municipio de La Florinda, it is convenient to order the persons by the code of the area in the REDATAM database, that is, the person records are sorted beforehand on the geographical identification so that all the persons in the municipio of La Florinda are together. REDATAM thus can ignore all the rest of the data and jump immediately to the persons in La Florinda.

The computer then could read the data for all variables on each person within La Florinda to pick out the information on Age and Birthplace to create the table of interest. But this is inefficient since time is wasted reading all the unused variables as sex, education, etc., which are of no interest in this tabulation. For this reason, the REDATAM database is also structured using an "inverted file" (data is stored by variable rather than by person record) so that only the variables of interest need be accessed for the persons in the area selected.

Since a very large number of records must be stored, REDATAM compresses the information and eliminates redundant data making the final REDATAM database around one-fourth the size of the original microdata. Depending on the number of variables in the census, it is sometimes possible to place the census data of up to a million persons on a 20 megabyte hard disk.

In practice, the REDATAM database is normally constructed with hierarchically-linked housing and population data to permit the production of tables that treat either population or housing or both. This makes it possible to study, for example, the housing characteristics of immigrants.

It is important to note that, while REDATAM is designed to be used without programmer assistance for obtaining tabulations and other results, the one-time creation of a REDATAM census database is complex and requires a proficient programmer. The census data usually is first cleaned to eliminate logical inconsistencies, must be sorted by geography and then transmitted ("downloaded") from the mainframe environment to the microcomputer. Information on each variable must be placed in the REDATAM dictionary and the entire geographical hierarchy of names and codes must be entered and related to the census data and finally the REDATAM database must be generated. While all this can be done relatively easily with the REDATAM software for survey data already on the microcomputer, the sheer magnitude of most census data files introduces its own complications. Nevertheless, once the REDATAM database is created, its use is equally simple whatever the complexity involved in its creation.

The system can employ either a hard disk for the storage of the databases from small and medium-sized countries or an optical "WORM" (Write Once Read Many times) laser disk for the data of large countries. The storage medium is transparent to the user. As there is normally no reason to change the database, the "write once" limitation of present laser disks merely improves security.

4. CHARACTERISTICS OF THE REDATAM SOFTWARE

Summary of user-oriented features

Full details on the facilities for data manipulation, statistical processing, geographic selection, etc., can be obtained by consulting the User's Manual (1987a; 1988a). Information on database generation is provided in the corresponding manuals (1987b; 1988a). The list that follows outlines the most salient features of REDATAM version 3.1 from the point of view of a user.

English language and Spanish language versions: The REDATAM software and all manuals are available in separate English- and Spanish-language versions. The tutorial and examples in the manuals refer to a small demonstration database that comes with the software. The system is designed to permit the easy inclusion of other languages for all screens and help.

Geographic selection: The user defines the universe of cases to be processed by selecting the specific geographic areas of interest, which in some databases can be created from city blocks or smaller. Maps usually must be consulted when census-defined areas without names are involved.

Self-documented database: There is a complete dictionary of variable names, descriptions of categories, etc. Similar information can also be maintained for recoded or derived variables.

Interactive: The user interacts with the system through menus and other facilities and there is extensive context-sensitive help. A "batch processing" mode also exists so that various long processes with millions of cases can be left overnight without user intervention.

Calculation of derived variables: New variables may be defined by recoding and through the utilization of arithmetic operations. The new variables may be temporary or may be incorporated into the database.

Statistical results: As REDATAM is directed to normal census processing, three basic statistics can be produced: frequencies, cross-tabulations and averages, the latter two with up to four variables in a given table. Results can be requested for sub-areas within the zone of interest, as well as for the entire zone. Decimal values can be used if non-integer weights are employed, for example, to expand a census sample or a survey.

Hierarchical processing: The system works with two levels of variable (housing and population) and results can be obtained for each of the levels separately or combined. For example, the number of foreign-born persons by sex and age within households can be calculated and cross-tabulated with a housing quality indicator.

Generation of sub-databases: A REDATAM sub-database for a specific area and/or for a specific sub-set of variables can be downloaded for utilization on another machine (which may have less storage).

Data files for export: When the basic statistical operations in REDATAM are not sufficient, an appropriately formatted parameter file and the dataset for the variables of interest for the selected area can be created for immediate processing by SPSS or SL-MICRO.

Storage and printing of results: The statistical results can be kept for later printing or inclusion in a document, as well as for the production of reformatted output using other packages such as WordPerfect, Wordstar, Lotus, Mathplan, etc.

Protection of information: Since it takes a user only a minute or two to define the area and obtain tables on the persons living on a city block or smaller, REDATAM permits the assigning of various levels of protection to each database via passwords to prevent unauthorized use at very disaggregated levels or when the number of cases is very small in a single table.

Equipment required

The following equipment, easily available in the Latin American and Caribbean countries, is required for operating REDATAM:

1. IBM PC with a hard disk, an XT, AT or 386 or a fully compatible microcomputer.
2. 640K main memory.
3. At least one floppy disk.
4. Monochromatic or colour monitor.
5. Printer with at least 80 columns.
6. Operating system PC-DOS version 2.0 or higher.

7. A hard disk with approximately 1.4 megabytes (Mb) available is required for the REDATAM system and the demonstration database that comes with it. The total amount of hard disk (or laser optical disk) space for storing the actual census or survey of interest will depend on the number of persons enumerated and the number and size of the variables on the questionnaire.

Availability of the software

Demonstration copies of the REDATAM software, version 3.1, and the accompanying manuals and test database may be obtained in English or Spanish by writing to CELADE, Casilla 91, Santiago, Chile.

5. EXPERIENCE WITH REDATAM IN THE LATIN AMERICAN AND CARIBBEAN REGION

The initial REDATAM version 2.0 was operationally tested on the 1980-round population and housing census data in the national statistical offices of Chile and Saint Lucia, respectively, for around a year in each and then made available for general distribution around mid-1987. In addition to the unforeseen demonstration effect that stimulated interest in other countries before the system was publicized, the tests helped to determine enhancements and modifications that have been incorporated into the present version 3.1 (released for general distribution in May 1988) and to identify the major extensions that will be developed over the next two years (see below).

Table 1 lists the census databases installed to date. The database for the approximately 16 million records of Chile are stored on three "WORM" laser disks of around 115mb each, corresponding to all regions north of Santiago, the metropolitan area of Santiago, and all regions south of Santiago. Since a laser disk drive was not available in Colombia and a sufficiently large hard disk was not accessible when the databases were created (March 1988), a separate database was generated for each of the six regions of the country and for Bogota. In the case of Uruguay, the REDATAM database was made (August 1987) for a 15 percent sample of the census to permit working with the information while the data entry is being completed for the entire census. Costa Rica has also created databases for household surveys and for birth and death vital statistics.

Census databases have been generated for the Caribbean countries of Grenada, the British Virgin Islands and St.Vincent, but have not yet been installed in the countries. A REDATAM database has also been created for the state of Rondonia, Brazil, to illustrate REDATAM to the national statistical office (IBGE) for possible use in the 1990 Brazilian census. The Brazilian authorities are considering the utilization of REDATAM to obtain rapid results from the 1990 census pilot tests and are looking into the mass distribution of public-use 1990 census data in the form of REDATAM databases that would be shipped along with a copy of REDATAM to permit users to work on their own microcomputers.

The REDATAM databases created to date are all installed in the national statistical offices of the respective countries. The statistical offices use the system for their own purposes and each also provides tabulations for specific areas on request to other governmental agencies, universities and to the private sector. The Colombian national statistical office, DANE, has indicated that as part of its policy to decentralize information, it will provide each of its regional offices with its corresponding REDATAM database of the 1985 census. The use of passwords provided by the system will help protect the confidentiality of the data in these situations.

Table 1. REDATAM population and housing census databases created to date (April 1988)

Country and date of the census	Dwellings (Thousands)	Persons (Thousands)	REDATAM space (Megabytes)	Storage Method
Chile (1982)	4 000	12 000	300.0	Laser disk
Colombia (1985)(Basic questionnaire a/)	5 800	27 800	70.0	Hard disk
Colombia (Enlarged questionnaire a/)	600	2 800	60.0	Hard disk
Costa Rica (1984)	500	2 500	60.0	Hard disk
Uruguay (1985) (sample)	147	450	15.0	Hard disk
Saint Lucia (1980)	30	125	3.0	Hard disk
Dominica (1981)	17	74	2.5	Hard disk
Guyana (household survey)	8	42	1.0	Hard disk

a/ The full questionnaire in the 1985 Colombian census was applied to 10 percent of the population. The basic questionnaire with a reduced number of variables was applied to the entire population.

Although the design of REDATAM is explicitly directed to the problem of rapidly obtaining tables for specific areas selected from a database of a population and housing census database of countries with various millions of inhabitants, the processing efficiency of the system has also proven very to be useful for routinely tabulating the entire censuses of Saint Lucia and Dominica with 125,000 and 74,000 population, respectively. Since these countries do not have mainframe or minicomputer facilities for census processing, their 1980-round census data were elaborated in a regional center in Barbados, along with the data of most of the other smaller Caribbean countries. Until REDATAM was installed these countries had to rely on the pre-conceived printed tabulations made for them of the whole country and of administrative zones which often do not correspond to the area of interest for individual users.

The experience gained to date in the Latin American and Caribbean countries shows that the REDATAM software normally is first used by the national statistical office, which after creating the census database, provides small-area data services (see CELADE, 1987c). Hence, most users of census data in the countries have not, themselves, utilized REDATAM, but request the tables of interest for areas defined by them from their national statistical office.

However, since REDATAM permits "downloading" a sub-REDATAM database for an area of interest allowing the user to work intensively with the data on his or her own microcomputer (if the statistical office is willing to give out the data), it is likely that there will be some diffusion of the system to other agencies. In addition, some institutions have begun to use REDATAM with their own survey or other data when, for example, processing speed is important and the features of a much more complete, but slower, system such as SPSS, are not required. As REDATAM can export the data and parameter cards to SPSS, once the REDATAM database is created, SPSS can be utilized when required.

Discussions have also been held with United Nations authorities in New York for arranging the distribution of REDATAM to countries in Africa, Asia and the Arab world. It is not known whether there has been any actual use of the system outside Latin America and the Caribbean except that the demonstration REDATAM software has been installed without modification in the national statistical office in China on a Great Wall microcomputer.

6. REDATAM AND THE 1990 CENSUSES

Most governmental and private organizations concerned with the development of policy or the formulation of plans and projects need timely data along with geographical disaggregation. But in the past, including the 1980-round of censuses, many Latin American and Caribbean countries have been unable to publish their volumes of census tables until years after the date of collection, making the very expensively collected census data more historical than timely.

However, once their 1990 census data are entered into the computer (all the data or a sample), the national statistical offices can use REDATAM to provide small-area data services on request (and national services in the case of the Caribbean countries), long before the census publications are produced or ready for distribution. The country may prefer to make these services available to the general public only after the data have been edited to remove logical inconsistencies, but this still will be months or more before the tabulations are produced and published. Statistical office staff from a number of countries have suggested that REDATAM may permit them to publish fewer tables.

As the 1990 census data will be timely and can be tailored to each user's request, REDATAM will make it possible for the first time for the national statistical offices in Latin America and the Caribbean to provide the private sector as well as governmental agencies with easy access to small-area census data for their own purposes (see ECLAC, 1987, for the 1990 census- and REDATAM-related discussions of the Directors of the Statistical Offices of the Americas). Needless to say, the availability of more disaggregated information will require better primary control over the data collection process than was exercised in the 1980 censuses and improved cartography (see CELADE, 1987d; Silva, 1986).

7. REDATAM-PLUS: CARTOGRAPHIC DISPLAY WITH A MULTIDISCIPLINARY DATABASE

CELADE is now beginning development of an extended version of the system, REDATAM+, to enhance the usefulness of the system in a variety of circumstances and to keep it abreast of changing technology and increasing user

The specific extensions of the REDATAM+ software will allow it to:

a) Retrieve multidisciplinary information describing geographical areas in association with various levels of population microdata;

b) Display population, housing and other data with cartographic information and carry out spatial analysis through an interface with a full-feature geographic information system (GIS);

c) Operate within a network to permit more than one user to work with the same database; and

d) Produce camera-ready REDATAM tables for publication.

The conversion of the present REDATAM database into a multidisciplinary planning database containing timely 1990 census data and the interface with a powerful GIS to permit geographical analysis, are likely to lead many planners and other users to insist on their utilizing REDATAM themselves, with REDATAM sub-databases for the specific regions, cities or other areas of interest. The REDATAM+ system, thus, will be one means of increasing and extending the usefulness and life of the 1990 census data that will be collected at such high cost in the Latin American and Caribbean countries. Of course, the utilization of the GIS system with REDATAM+ will be possible only in situations where the geographical base file containing the cartographic description of boundaries and other geographical information is available for the areas of interest in a given country.

The REDATAM+ software should be ready for general distribution at the end of 1989 (until then, REDATAM Version 3.1, described above, will be available).

BIBLIOGRAPHY

CELADE, 1987a. "REDATAM Version 2.00 User's Manual". CELADE, Santiago. English LC/DEM/G50 (24 June 1987). [Also available in Spanish].

______, 1987b. "REDATAM: Database generation manual". Santiago, CELADE. LC/DEM/G.53 (October 1987) [Also in Spanish]

______, 1987c. "Considerations for implementing REDATAM data services". Santiago, CELADE. LC/DEM/R.49 (September 1987). [In Spanish, in a slightly different version: "Consideraciones para implementar un servicio de datos con el sistema REDATAM". Ref. document no. 7. Meeting of the Directors of Statistics of the Americas. ECLAC, Santiago, 23-25 September 1987.]

______, 1987d. "The relevance of the REDATAM system for the 1990 censuses" Santiago, CELADE. LC/DEM/48 (September 1987). [In Spanish, in a slightly different version "REDATAM: Relevancia para los censos de 1990". Ref. document no. 17. Meeting of the Directors of Statistics of the Americas. ECLAC, Santiago, 23-25 September 1987.]

______, 1987e. "REDATAM: A summary". Santiago, CELADE. LC/DEM/R.50 (Sept.87). [In Spanish, in a slightly different version: "REDATAM: Un resumen". Ref. document no. 18. Meeting of the Directors of Statistics of the Americas. ECLAC, Santiago, 23-25 September 1987.]

CELADE, 1988a. "Supplementary Manual for REDATAM Version 3.1: Supplement to the User Manual and the Database generation manual". Santiago, CELADE. Serie A-181 (March 1988). [Also in Spanish.]

Conning, Arthur, 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.

ECLAC, 1987. Report of the Meeting of Directors of Statistics of the Americas, Santiago, 23-25 September 1987. ECLAC, Santiago. LC/G.1482 (24/11/87).

Silva, Ari, 1986. El procesamiento de los censos de población de América Latina en la década de 1990: un vistazo al futuro. Notas de Población. Año XIV: No. 41 (Agosto 1986). Pp. 9-24.