Distr.
RESTRICTED

LC/R. 966
9 January 1991

ORIGINAL: ENGLISH

ECLAC

Economic Commission for Latin America and the Caribbean

Interregional Workshop on Population
Data Bases and Related Topics

Jakarta, Indonesia, 4 - 19 January 1991

LATIN AMERICAN AND CARIBBEAN EXPERIENCES IN ELECTRONIC STATISTICAL DATA PROCESSING */

*/ This document was prepared by ECLAC's Division of Statistics and Projections.

Document not subjected to editorial revision.

INDEX

INTRODUCTION

This report, prepared by the Statistical and Projections Division of the Economic Commission for Latin America and the Caribbean (ECLAC), will be presented during the INTERREGIONAL WORKSHOP ON POPULATION DATABASES AND RELATED TOPICS to be held in Jakarta, Indonesia, in january 14-19 of 1991.

This paper main objective is to give the participants of this Workshop a thorough description of the state-of-the-art in statistical data processing among countries of our region. As this Workshop is mainly devoted to population databases, special considerations will be given to the 1990 census round of housing and population as well as to permanent household surveys conducted regularly in the majority of our countries.

It is hoped that the content of this paper will become a useful contribution to the Workshop sessions dedicated to analyze and discuss comparatively the different valuable experiences of the Countries and Organizations participating in this meeting. We are deeply convinced that to learn about the experiences that other participants will expose during these days will be valuable references to take into account for improving our regional technical level. Now, if our experiences at least contribute to avoid already suffered problems to other participants we will be really rewarded. On the other hand, if our regional experiences can practically contribute to enhance other participants activities in this field we will be fully satisfied and ready to share them in a bilateral cooperation fashion with anyone requesting for it.

For clarity's sake this report on regional statistics data processing activities will be organized into the following chapters : Institutional Framework, Generalized Software, Hardware Environment, 1990 Census Data Processing and Evaluations and Conclusions.

Finally two important warnings. Mention of any commercial hardware or software does not imply the endorsement of ECLAC. When this report speaks of our experiences or similar sentences it is being implied "the countries of the region".

INSTITUTIONAL FRAMEWORK

ECLAC has as one of its main and permanent duties the technical assistance in the field of statistical data processing to the regional countries. At the same time, for its own needs, specially in the Statistical Division, ECLAC has been heavily using statistical electronic computing facilities.

In order to tackle satisfactorily both aspects of these computing concerns it was decided to turn its efforts and resources towards using generalized statistical data processing Software considering the in-house usual lack of programming experts. Obviously, at the same time, a strategy of this nature was undoubtedly considered an advisable solutions for the needs of the region.

In 1985 after surveying several sources and Institutions where there could be this kind of approach to face the problems in this field, ECLAC finally decided to get in touch with the UNDP funded european project named Statistical Computing Project, from now on SCP [01], that was clearly oriented by this strategy. This contact gave very concrete results : all of software already available in this project was installed and used in ECLAC for its own applications. ECLAC took over responsibilities to disseminate in the region the methods, techniques and systems produced under that project.

In september 1986, during the 9th Inter American Conference on Statistics, ECLAC delivered a paper proposing a regional project for cooperation in this field among countries [02]. The Conference decided to request to ECLAC to present to UNDP a regional project to fund activities oriented to reach this target.

ECLAC presented the terms of the project to UNDP and it was approved to initiate its activities in july 1987. This project was named RLA/87/001 "Statistics for development of Latin America and the Caribbean" and it had the following objectives : disseminate generalized software, mainly SCP outputs, in the region; render technical assistance in installing and using those software; training of national experts and devise guidelines to support activities of cooperation among regional countries for using these techniques and systems.

This project has finished its activities last october. During its lifetime, practically three years, not less than 15 national Statistical Offices received about 20 technical missions to install and to devise applications using those software supported by the project.

It is worthwhile mentioning that four technical seminars were conducted in that period. The first was held in Santiago, Chile, in the last two week of 1988, to present the packages available, to discuss and persuade to experts and specially to managers the advantages of replacing ad-hoc applications systems by applications systems implemented under generalized software. The second seminar was carried out in Rio de Janeiro, Brazil, end of may 1989 to discuss about hardware and generalized software for entering and editing statistical data. The third seminar took place in Cuernavaca, Mexico, end of November 1989, to discuss matters related to Data Bases and Electronic Dissemination of statistical data. The last seminar was conveyed in Quito, Ecuador, in the first week of 1990, where was focused in depth the hardware and software for statistical computing using microcomputers, LAN and open architectures or UNIX environments.

During these last five years, ECLAC has achieved a close cooperation with Statistics Canada, the Bureau of the Census of USA, Statistics Sweden, Statistical Office (INE) of Spain and recently has been initiated contacts with The Netherlands CBS for joint activities aimed to assist our countries for using their in-house developed BLAISE package.

This close contact with those well developed Statistical Offices and the SCP project makes sure to our region a valuable framework for keeping pace with the most up-to-date methods, techniques and efficient generalized statistical software.

Although ECLAC is not currently executing any regional project can continue its support to the region thanks to its links with those Statistical Office and the SCP project as far as ECLAC is permanently interested in this field for its own tasks and because it is not a costly duty to disseminate up-dated information on new systems and even submit them or facilitate contact with the appropriate source.

ECLAC and the region has a permanent forum to review their statistical concerns every two years during the Conference of the Statistical Directors of Americas. Statistical computing is a relevant item in the agenda of this Conference.

Finally, it is worthwhile mentioning that the national experts in statistical computing of our region are currently doing their best to get support from the Statistical Directors of the Americas to have national funds devoted to organize periodic technical conference in this field. The initial proposal is to carry out this meeting every two year and as soon as possible to hold it yearly. There is an unanimous consensus among the regional statistical computing experts that this technical conference will be the most useful and efficient activity for tracking joint developments and sharing valuable experiences.

GENERALIZED SOFTWARE

Briefly, it is convenient to remind regional strategies and policies for statistical data processing, active clearly up to mid of 80's. Centralized computers facilities, production run under batch mode, systems development in interactive mode, classical systems analysis and design, programming language were usually COBOL, FORTRAN and ASSEMBLER, data management through sequential files even on direct access devices and data dissemination on printed forms and publications.

Every application was an had-hoc task, from the very beginning of data entry step up to the target output. This style produced a rigid interrelation among the expert group in charge of the application, the programs and the data files and even with the exploitation of the system. Any fail or a momentarily lack of one of those components implied a crash. Specially in the aspects of application maintenance and enhancements. And so on it is possible to continue enumerating drawbacks and drawbacks. It is not necessary as everybody knows perfectly well what is the end of this story.

As it was stated in the INSTITUTIONAL FRAMEWORK chapter, for the weakness of working style mentioned above and considering many other aspects such those rigorously presented in [03], the Latin America and Caribbean countries are using under regional backstopping activities and bilateral cooperation the following generalized software :

DATA ENTRY

KEY ENTRY III : a commercial software for IBM PC compatible; rather powerful in checking input, manages several records layout simultaneously, verification capabilities, an easy parametric programming language and two style of production; one for bulky fast keytyping and other for standard user with record layout on screen. Basic statistics for production management.

ISSA : a proprietary system. It is a component of an integrated system for survey processing. A complete and sophisticated tool for complex input with a powerful programming language letting real-time validation and consistency intra and inter records. Several statistics can be produced through programming for control purposes.

CENTRY : a component of the IMPS system developed by the ISPC of the Bureau of Census of USA. As a well-known system among our Statistical Offices it is not necessary to point out nothing more.

A few countries are using dedicated minicomputers with their vendor data entry programming language and own Operating System (OS). These equipments are IBM, DEC and UNISYS.

DATA EDITING

CONCOR : the well-known ISPC system. It is being used in its mainframe version under different OS and computers. Lately several countries are moving to or choosing the IMPS version. By far it is the most used system for editing purposes in the region.

AERO : a SCP output developed by the Statistical Office of Hungary. It is installed and fully tested in ECLAC. Argentina is evaluating its application to the next population census combined with CONCOR. This system is available only for IBM 370 lines [01].

DIA : an INE Spain developed system. A complex package underlying Fellegi-Holt model for automatic imputation. IBGE Brazil is working with INE Spain to take advantage of some of its modules for their in-house developed editing system.

In the region we have two concrete efforts in developing generalized system for statistical data editing. They are IBGE in Brazil and INEGI in Mexico [08]. It is happen they are participating in this Workshop, so we hope to hear from them some details on their systems during the discussion panel. Both systems will be fully proved during their population census. Besides, these two countries have implemented for their census an in-house developed system for computer assistance in automatic codification. Clearly, by this time, these two countries are the only ones meeting conditions and assigning funds for these kind of activities.

DATA MANAGEMENT

A few words before listing available systems for managing statistical data bases. This paper is not intended for presenting the complex world of data bases. It is worthwhile to reference to the reader to a very valuable report [04] to take into account the several aspects involved in this subject matter.

RAPID : developed by Statistics Canada is currently one of the systems available in the SCP project. It is implemented under IBM 370 lines running OS/VS. ECLAC, with the direct assistance of Statistics Canada, has migrated this DBMS to IBM DOS/VSE environment.

BOS : developed into SCP project. Currently is not only a complement to RAPID to facilitate relational operations on those databases but also can manage files residing in different media in a relational fashion. To this extent is by this time a real free standing DBMS. It is developed to be run on IBM 370 architecture under OS/VS. ECLAC with direct assistance of Statistics Sweden adapted this system to IBM DOS/VSE environment. Lately, SCP has produced a DOS PC version [01] and [04].

REDATAM : developed by ECLAC Demographic Center. It runs on IBM PC compatible under DOS. As of mid of this year a PLUS version will be released. By the time being is the most used DBMS in PC environment in our region. For details on this package see [04] and [05].

During this year ECLAC is planning to install and or full test three DBMS which are considered very well suited for satisfying several regional requirements.

TEMPUS : developed by INE Spain [06]. It is designed to manage time series and there is a version for mainframe and other for PC under DOS, both version fully compatible make sure applications and data base portability.

URUCIB : developed under the project URU/84/006 funded jointly by UNDP and the Planning and Budgeting Office of the Uruguay Republic Presidency. It is oriented to manage time series for tracking their behavior in order to make possible timely decision making at executive level [07].

AXIS : a DBMS developed by Statistics Sweden. It is a powerful DBMS oriented to manage time series and tables and to provide several statistical computing facilities. In the first semester of this year is expected to deliver the PC version fully compatible with the mainframe version running on IBM 370 line under VS2 OS.

TABULATION AND ANALYSIS

CENTS : the well-known ISPC system. It is being used in its mainframe and PC version. Lately the version belonging to IMPS is replacing the PC version.

TPL : it is available for IBM 370 line under OS VS1 and VS2 in the version developed by the Bureau of Labor of USA. Currently it is available a commercial version for compatible PC developed by an USA soft-house, QQQ Software Inc, Arlington, Virginia.

INTERTAB : a system developed by SPC project. It is available versions for IBM 370 line under OS VS1, VS2 and CMS. The version for compatible PC is at the phase of designing.

SAS and SPSS : both systems are intensively used in their mainframe version. By the time being a few countries have introduced their compatible PC versions.

INTEGRATED SYSTEMS

These systems must be considered a well-defined group in the field of generalized systems. These systems provides in an comprehensive environment facilities powerful functions to tackle all of the steps comprised in processing surveys and in some cases census data. Usually they are oriented to real experts due to their inherent complexity in design and applicability. Lately new available systems are intending to make easier and more friendly their use for facilitating the design and production of surveys by subject matter experts.

IMPS [09] and ISSA system [10] are the utmost used systems.

The BLAISE system [11], developed by CBS of The Netherlands, is a new very attractive alternative due to its rather easy applicability in many surveys conducted by several countries in the region.

HARDWARE ENVIRONMENT

It is convenient to point out that our region comprise a diversity of countries considering their population and extension. It is Brazil with more than 130 million people and around 8.5 million sq. kms but it is also Saint Lucia with about 130 thousand peoples and 616 sq kms.

Into this range we have big, medium and small size countries. Consequently, from the very beginning era of the electronic processing of statistical data among our countries, the computing facilities incorporated reflect this diversity. There are, in the region, huge computer and also small microcomputer.

Briefly, we can say, the regional computing resources are :

- Huge mainframe with hundred terminals and dozens of gigabytes in random access with remote communication facilities. We are speaking of IBM-3090-like computers.

- Big and medium size mainframe with dozen of terminals and around of 5 megabytes in random access. They are IBM-43xx-like computers.

- Small mainframe and minis. Usually with around 20 terminals and less than 1 gigabyte in random access. We can mention Texas Inst equipments, Wangs 100, IBM 3x, etc.

- Compatible IBM PC under DOS, XT-type model with 256/512K RAM, 20 Mbytes in HD and DS/DD FDD 5 1/4" with dot printer. Free-standing use.

- Compatible IBM PC under DOS, 286 with 640K RAM, 40 Mbytes HD and high density FDD 5 1/4" with usually NLQ dot printers. All of them used free-standing.

- Compatible IBM PC under DOS, 386 with 1 Mbyte RAM, more than 80 Mbytes HD and FDD 5 1/4" and 3 1/2" dual densities, very often with laser printer. All of them also used free-standing.

It can be found, since only recently, several supermicros under DOS and some minis or supermicros under UNIX.

In the field of data transmission and remote connection, two countries are using their own satellites, several countries have PSDN facilities covering almost all the country and linked to international network and most of the other countries hardly can make use of rather noisy public telephones lines.

1990 CENSUS DATA PROCESSING

The region has dedicated a considerable amount of efforts to prepare and to tune all of aspects involved in a housing and population census. To give backstopping to all needed activities ECLAC executed an UNFPA funded regional project, RLA/88/P08 "Apoyo a la ronda de censos del 90". This project, during years 89/90, carried out eight seminars covering all aspects considered critical and relevant in a census. These seminars were held in eight different countries of the region. It can be said all these seminars reached very satisfactorily their objectives according to evaluations and concrete experiences made by participants national experts.

Two of those seminars were devoted to census data processing. One was held in Santiago, Chile, 12-15 september 1989 on "Users access to census data" and the other was carried out in Caracas, Venezuela, on 28-31 may 1990 to discuss about and evaluate "Computing systems for census data processing".

These two seminar complemented with those mentioned in INSTITUTIONAL FRAMEWORK submit to the region a thorough and comprehensive information and know-how to face the electronic data processing involved in the censuses.

As results of these seminars and the own past experiences of the countries the following different methods and systems have been adopted by the countries for this round of census :

<u>Data Entry</u>

1. Minicomputers dedicated for bulky data input with low level of checking during the operation of key typing.

2. Stand alone compatible PC's for medium volume data input but incorporating high level of checking and even consistencies during the key typing operation. Centry, ISSA and Key Entry

III are the set of systems to be used.

3. OMR devices. According to evaluations and consequent decision already taken will be used devices able of reading around of 7,500 docs/hour connected to a microcomputer to monitor the reading process and to store data to be later transmitted usually to mainframes for editing and in a few cases to go into data editing in another PC.

Data Editing

1. CONCOR in its mainframe version will be used in those countries where mainframe or minis are going to be used. When the census will be processed using only PC's, the IMPS component will be used. One of these two option seem to be going used by many of the countries.

2. In-house system. Brazil and Mexico have developed fairly sophisticated and powerful generalized programs.

Data Management

1. For mainframe environment RAPID and BOS. For microcomputers field REDATAM.

Tabulation and analysis

1. CENTS and TPL will be the standard tool for census tabulation plan. Brazil has its own powerful generalized package.

2. SAS will be used for tabulation analysis and other statistical applications at the majority of mainframes sites. SPSS PC+ will be the tool at PC sites.

Data dissemination

1. For on-line access by internal and external users, in most cases very restricted due to legal regulation related to confidentiality, RAPID and REDATAM will be used as they both have very powerful and easy use interface with SAS, SPSS, TPL and even CENTS. Still remain to be decided in those countries where is technical feasible the remote on-line access to the census database.

2. Microdata dissemination will be done using magnetic tapes and diskettes. Always keeping aspects of privacy and confidentiality.

3. Census tables planning will be published through camera-ready outputs or Desk Top Publishing techniques and systems.

Summary

In few statements we can say there will be two main line of working :

1. Processing restricted to PC facilities will use IMPS for the census planned processing and SPSS with REDATAM will constitute the analysis tools.

2. Mixed facilities will imply data entry through PC's, minis or OMR. Data editing in the central site to load later census data under RAPID for census tabulation plan using CENTS or TPL. Analysis will be carried out using the RAPID database interfaced with SAS or SPSS. Now, when PC will be also used for analysis on selected geography, REDATAM and its interface with SPSS and export capability in ASCII flat format will be the tool.

EVALUATIONS AND CONCLUSIONS

It is worthwhile, now, to intend a complete appraisal of the several activities and know-how reached during the intensive work done in the period covered by this paper in order to draw some conclusions useful for eventual discussion in the context of this meeting.

Firstly, it can be said about the organization of the regional activities something rather obvious but very important to emphasize at this time. Clearly, the commercial sources of software suited for statistical data processing, for the time being, are not interested in this field, specially in those aspects related to statistical data gathering and production. Hence the strategies adopted to keep close contact with developed Institutions and valuable international projects is perhaps the only way to benefit from the continuous developing state-of-the-art in this specific field [12]. It seems the best choice to share experiences and to continue bilateral cooperations not only at the level of our region but also with others to have a clear defined permanent technical seminars and workshops. What remains to be done in our region is the national support for carrying out a periodic Latin America and Caribbean Regional Conference on Statistical Data Processing.

Secondly, in the field of software and methodologies [01] [13] [14], the situation is rather stimulating and satisfactory. The only area really weak is that of data editing; we need new more powerful and statistically strong generalized packages [15] [16]. The other field where we see rather urgent decisions is that of electronic disseminations; clearly it is time to take advantage of new technologies and methods, namely, CD-ROM and GIS and mainly to strengthen network for data transmission and remote connection at national and international-linked level [17].

Thirdly, in the field of hardware, the trend is also very encouraging. What is really urgent is to speed up the integration of stand-alone PC's in LAN's and these ones linked eventually to main sites. What remain to be study and carefully evaluated is the new attractive world of LAN's or UNIX open architecture equipments versus mainframes [18] [19].

Finally, related to 1990 census data processing, we will have in the near future a very valuable set of experiences that deserves to be evaluated in an interregional workshop. What remains on the part of the countries is to implement several conclusions related to census data dissemination, specially those related to on-line access of microdata securing at the same time confidentiality on the information. The most ambitious target several countries are evaluating is the use of GIS for not only dissemination but also and mainly for planning and diagnostics at different administrative and geographical level.

REFERENCES

[01] Statistical Computing Project. A multinational european project funded by UNDP and executed by ECE. Terms and objectives of the project, technical papers and available outputs can be requested from : Director, Statistical Division, ECE, Geneva, Switzerland.

[02] COMPUTING SYSTEMS FOR STATISTICAL TASKS. Possibilities of interregional cooperation. ECLAC Statistics and Quantitative Analysis Division. 9th Inter American Conference on Statistics. Rio de Janeiro, Brazil, 15-18 September 1986.

[03] Statistical Data Processing in Developing Countries : Problems and Prospects. George Sadosky, UN consultant. Interregional Workshop on Statistical Data Processing and Data Bases. Geneva, Switzerland, 30 May - 3 June 1988.

[04] Establishing Population Databases : Part I and II.
Arij Dekker, UN consultant. Interregional Workshop on Population Databases and related topics. Voorburg, The Netherlands,
6-10 November 1989.

[05] Arquitectura y Filosofía de Bases de Datos : El modelo REDATAM-PLUS. Latin American Demographic Center (CELADE). Seminar on "Bases de Datos y Difusión Computacional". Cuernavaca, México, 27 November - 1st December 1989.

[06] Banco de Datos Estadísticos TEMPUS. Dirección de Información Estadística, INE España. Seminar on "Bases de Datos y Difusión Computacional". Cuernavaca, México, end November 1989.

[07] URUCIB : un sistema de información ejecutivo. Papers # 12 and 21. Seminar on "Bases de Datos y Difusión Computacional". Cuernavaca, México, end November 1989.

[08] Metodología de Vectores Teóricos. Statistical Office (INEGI) of México. Seminar on "Entrada, Detección y Corrección de errores en datos estadísticos". Rio de Janeiro, Brazil, May 29 - June 2 of 1989.

[09] Integrated Microcomputer Processing System for Censuses and Surveys. ISPC of the Bureau of Census of USA. Paper # 22. Seminar on "Entrada, Detección y Corrección de errores en datos estadísticos". Río de Janeiro, Brazil, May 29 - June 2 of 1989.

[10] ISSA : un sistema integrado para procesamiento de encuestas. Developed under a joint funded project by Westinghouse and AID. Currently is funded by IRD Macro Systems Co and AID. Paper # 2. Seminar on "Técnicas y Sistemas Generalizados para procesamiento de datos estadísticos". Santiago, Chile, June 20 - July 1st of 1988.

[11] The BLAISE system for integrated survey processing. CBS of The Netherlands. Paper # 4. Seminar on "Sistemas microcomputacionales para procesamiento de datos estadísticos". Quito, Ecuador, 3-7 September 1990.
The BLAISE system for computer assisted survey processing.
J Bethlehem, CBS of The Netherlands. Interregional Workshop on Population Databases and Related Topics. Voorburg, The Netherlands, 6 - 10 November 1989.

[12] The Impact of "New Technology" on Statistical Offices, etc.
A Westlake, consultant to ISIRC. Interregional Workshop on Statistical Data Processing etc. Geneva, Switzerland, 30 May - 3 June 1988.

[13] Generalized Systems for Surveys Processing at Statistics Canada. Wilder Boucaud, Informatics Branch. Seminar on "Bases de Datos y Difusión Computacional". Cuernavaca, Mexico, 27 November - 1st December 1989.

[14] An Integrated Solution to Distributed Information Systems. R J Bateman, SSD, B of Census USA. Seminar on "Bases de Datos y Difusión Computacional". Cuernavaca, Mexico, 27 November - 1st December 1989.

[15] Depuración de Datos Estadísticos. E García e I Villán. INE Spain.
Seminar on "Entrada, Detección y Corrección de errores en datos estadísticos". Río de Janeiro, Brazil, 29 May - 2 June 1989.

[16] Entering and Editing Data at Statistics Canada. M Jeays, Informatics Branch. Seminar on "Entrada, Detección y Corrección de errores en datos estadísticos". Río de Janeiro, Brazil, 27 November- 2 June 1989.

[17] Redes de comunicación para diseminación de información. ECLAC Computer Center. Seminar on "Bases de Datos y Difusión Computacional". Cuernavaca, México, 27 November - 1st December 1989.

[18] The Impact of Microcomputers on surveys processing at The Netherlands. CBS, The Netherlands. Seminar on "Sistemas microcomputacionales para procesamiento de datos estadísticos". Quito, Ecuador, 3 - 7 September 1990.

[19] Metodología y Procesamiento de Evaluación y Selección de Sistemas Computacionales Departamentales para el INE. INE Chile. Seminar on "Sistemas microcomputacionales para procesamiento de datos estadísticos". Quito, Ecuador, 3 - 7 September 1990.