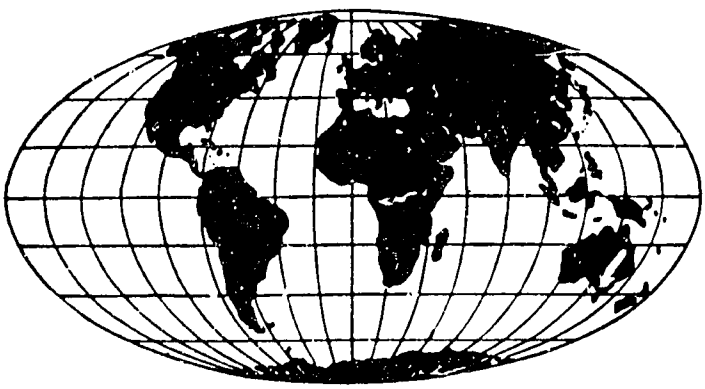
PN-AAR-202
ISN 36731

STATISTICAL SOFTWARE INVENTORY report



Prepared By

DELTA SYSTEMS CONSULTANTS, INC.

003941

under subcontract
AID/DS E 0027-SUB-1

STATISTICAL SOFTWARE INVENTORY REPORT

This report was prepared by

Delta Systems Consultants, Inc.
250 Valencia Avenue
Coral Gables, Florida 33134

Senior Editors

Julio C. Ortuzar
Joseph L. Willard

Contributors

Richard I. Buhrman
Manuel Quinteros

September 1983

The report was prepared by Delta Systems Consultants, Inc., under a sub-contract to the Data Use and Access Laboratories - DUALabs, 1515 Wilson Boulevard, Arlington, Virginia. The funding was provided by the Office of Population, Agency for International Development, Washington, D.C. 20523

PREFACE

This report was originally prepared in 1980. Funding was provided by the Office of Population, Agency for International Development, Washington, D.C. 20523. This September 1983, revision is authorized by Amendment Number 9, to CONTRACT No. AID/DSPE-C-0027.

The report is intended to be a reference document for use by statistical institutions in developing countries. It is not intended to be a procurement guide, but the authors believe that it will provide comparative information that will be useful in evaluating the various software packages prior to procurement.

The chapters on tabulation and editing software contain more detail than those on the statistical analysis packages because the tabulation and editing packages are generally les well known, and the body of information available is considerably less than that available for the widely used statistical analysis packages.

The authors are indebted to Professor Ivor Francis, who provided significant assistance in the preparation of this report. His vast experience in examination and evaluation of statistical software over the years has given him a special vantage point from which to view the capabilities and limitations of the software packages included in this document.

The comments provided by the U.S. Bureau of Census, Research Triangle Institute, and Westinghouse Applied Systems staff members were quite useful in the final preparation stages of the report.

Readers desiring information about other packages than those included in this report should obtain the text by Professor Francis. Professor Francis, formerly of Cornell University is now Professor of Mathematics at the University of Otago in Dunedin, New Zealand. His most recent compendium contains listings of more than 100 statistical software packages. The book entitled "Statistical Software: A Comparative Review" was published by Elsevier, North Holland, Inc., in 1981. The methodology and description of the rating procedure is described along with summary ratings for all of the packages included.

TABLE OF CONTENTS

Preface iii

TABULATION SECTION

INTRODUCTION 1

Major Tabulation Software Packages Included in this Inventory Report 2

PART I

Description Format 3

PART II

CENTS, COCENTS, & CENTS 4 7

TPL 12

GTS 15

LEDA 18

CENTS-AID 21

P-STAT 24

XTALLY 27

PART III

Comparison of Software for Tabulating Survey and Census Data (edited) 29

EDITING SECTION

INTRODUCTION 35

Major Editing Software Packages Included in this Inventory Report 37

PART I

Description Format 38

PART II

CONCOR 45

COBOL CONCOR 51

CAN-EDIT 56

UNEDIT 61

PART III

CONTROL 65

STATISTICAL ANALYSIS SECTION

INTRODUCTION 73

Major Statistical Analysis Software Packages Included in this Inventory Report 77

BMDP 79

EASYTRIEVE 83

MINITAB 86

OSIRIS IV 89

P-STAT 94

SAS 98

SIR 101

SPSS 103

BIBLIOGRAPHY 107

TABULATION SECTION

INTRODUCTION

This Section is divided into three parts. Part I contains an outline of the areas in which tabulation packages were analyzed along with an explanation and description of each section of the outline. In Part II, each tabulation package is discussed individually in the same format presented in Part I. Part III is an analysis of the results of a test demonstration of tabulation packages which was carried out by DUALabs/Delta Systems in conjunction with Professor Ivor Francis who was Scientific Secretary of International Association of Statistical Computing (IASC) at the time of test.

During the 1970's there were major problems in tabulating the census data in many developing countries. There were a number of reasons for this, but the principal one was the lack of generalized software available for this purpose. Many countries found it necessary to write custom computer programs or have their data sent to another country for processing. As a result, in some countries, years elapsed between the time of census enumeration and the publishing of the data.

There are now a number of generalized software packages available for producing tabulations. Existing packages are being updated constantly and new packages are being introduced each year. The packages vary widely in intended uses and in their ability to meet these intentions. There are so many packages available that sometimes it may be desirable to select more than one in a particular situation. One objective of this report is to provide users with some guidelines in determining which package is more useful under a certain set of circumstances.

This report focuses on those packages which can be used to process censuses in developing countries. There are certain conditions which exist in statistical offices in many developing countries which are quite different from the conditions in a developed country. These conditions will be an important factor when deciding which software package to use in a particular country.

With different requirements, it is unlikely that many countries will find any package that can be considered perfect. Some of the requirements are conflicting; for example, a package that is easy to use will print the tables in a predetermined format and lack flexibility for producing camera-ready copy.

Major Tabulation Software Packages
Included in this Inventory Report

As noted earlier, there are a number of tabulation software packages available for processing large data files. The packages selected for inclusion in this inventory report are shown below by name of the package, the distributor, cost, along with a brief statement of the attributes. Each package is presented separately within this section.

MAJOR TABULATION SOFTWARE PACKAGES

NAME	DISTRIBUTOR	COST	ATTRIBUTES
CENTS COCENTS CENTS 4	U.S. Bureau of the Census Washington, D.C. 20233	$300* $300* $300*	These packages are all similar in user language and capabilities. All are very efficient in terms of computer time required, and all produce camera-ready copy. CENTS is written in IBM ALC; COCENTS and CENTS 4 in COBOL.
CENTS-AID	DUALabs Suite 607 1515 Wilson Boulevard Arlington, Virginia 22209	$6,000 $3,000 for Government and Non-Profit	CENTS-AID is based on the COCENTS package with a powerful pre-processor that interprets English-like command instructions. Uses the Data Dictionary approach; written in COBOL.
GTS	U.S. Bureau of the Census Washington, D.C. 20233	$300	At present available only for UNIVAC 1100 series computers. Similar in design to COCENTS. Uses the Data Dictionary approach and provides camera-ready copy.
LEDA	National Institute of Statistics and Economic Studies (INSEE) Boulevard A. Pinard Paris, CEDEX, France	200,000 French francs	Modular package which contains data handling, editing and correction, as well as tabulation capability. In terms of computer time, LEDA is quite efficient. Written in COBOL, there are both French and English versions.
P-STAT	P-STAT, Inc. P. O. Box 285 Princeton, New Jersey 08540	$5,000 first year $2,000 renewal $1,000 university	Primarily a statistical package with a fairly good tabulation module. Written in FORTRAN, P-STAT runs on a wide range of computer systems. Now available on WICAT micro with UNIX.
TPL	U.S. Bureau of Labor Statistics 441 G Street, N.W. Washington, D.C. 20212	$300	TPL is a system designed to select, restructure, cross-tabulate and display statistical data. Written in XPL language, TPL is only available on large mainframe IBM systems with minimum 300 K-bytes of memory. Extremely powerful and easy-to-use language.
XTALLY	United Nations Statistical Office New York, N.Y. 10017	No charge to member governments	Originally written in RPG II for very small computers. Has capability for producing counts with summary totals and percentages. COBOL version now available.

**Available free to USAID-assisted countries from the U.S. Bureau of the Census.*

PART I

There are a number of desirable features that need to be considered when selecting a computer software package for processing census data. These features include but are not restricted to the following items:

A. Large volume of data

Any tabulation system processing census data must be able to handle large input files, since the populations of developing countries range from 50,000 to more than 100 million. Systems unable to process several million records in a reasonable time would not be useful in all but a few countries.

B. Various data formats

The tabulation package should also be able to process different data formats. Census data may be stored in binary, packed decimal or character format. For systems which cannot handle one or more of these formats, the data would have to be converted to the format used by the package in some situations. This is usually possible using utility routines, but can be expensive and time consuming for large census files.

C. Hierarchical file structure

Census files are hierarchical rather than flat files; they usually include both housing and population records. Some census tables require household variables to be cross tabulated with population variables, so packages should be able to handle these hierarchical files without having to preprocess the file.

D. Computing table data

Tabulation packages should be very flexible in the kinds of tables that they can produce. They should provide for the following capabilities:

1. Multi-dimensional tables

   Tables are often multi-dimensional. For example, a four-dimensional table with age and sex, by marital status, and level of education, is typical of the tabulation needed for censuses. It should be possible for all of the data to be shown on the printout page, providing there are not too many categories of each variable. Tables such as this one allow analysts to study quickly how variables are interrelated. Some packages allow for only two or three dimensional tables and would require the information above to be split into several tables printed on separate pages.

2. Large numbers of tables per run

It is too costly and time-consuming to process a large census file more than a few times, so it is important to be able to process as many tables as possible in one run. A good package that can produce twenty tables in one pass of the data is more desirable than a superfast one that can produce only two or three tables at a time.

Some packages provide for area breakdowns, that is, automatically creating country, region, and province totals. If a table such as one with age by sex is needed for every region and every province in the country, the ideal package stores only the basic age by sex table in core at one time. This will mean a substantial savings in core memory, over other packages which must hold one table for each area of the country at the same time.

3. Recoding facilities to create new variables

Packages must be able to recode input variables into different categories; for example, age is often recoded into five year intervals, 0-4, 5-9, etc. Some tables also require index variables which combine two or more variables; for example, socio-economic level might be created from the input variables of salary, occupation, and level of education.

4. Definition of universe of table

Redefinition of the universe should be possible both for an entire set of tables and for individual tables. For packages that do not have this feature, subfiles have to be created containing only the universe desired. This is costly and time consuming for large census files, and creates an additional logistical task of managing more files.

E. Weighting factors

Since complete count censuses are expensive, sometimes sample enumerations are taken, in which only a small percentage of the total population is interviewed. When sample data are processed, it is often desirable to weight each of the records in order to create tables in which data fields reflect total population. Tabulation packages should allow for a weighting factor to be set for all cases or for the factor to be obtained from each case on the data file. Ideally, this capability should provide for integer or fractional weighting.

F. Handling missing values

Census data used for tabulation are not always complete, so packages should have a feature for handling missing data. Ideally the package will be able to show missing data items as not reported.

G. Printing tables

In most cases, it is desirable that the printed output of census tabulations be created in a way that it can be used as camera-ready copy; that is, ready to be sent to the printer for publication. This shortens the time interval between data collection and publication of tables. It also keeps errors from being introduced during retyping or type setting. If the tabulation system does not provide for camera-ready output, the printout should be easy to read and to understand.

Printed titles for table data include boxheads, stub and caption lines, and footnotes. Boxheads or headings are located above the data, stubs and captions to the side of the data, and footnotes below the data. Since it is only through these titles that the table data can be interpreted, they should be understandable and complete. To be useful, as camera-ready copy, words should not be arbitrarily broken, abbreviated or truncated.

H. Generation of codebook or data dictionary

A data dictionary (D.D.) is a very useful feature since it provides for a complete description of the input data file. The D.D. should allow for the following information to be described for each variable; questionnaire ID, record type (for hierarchical files), variable name, data format, and location in record. Category names should be allowed for each value or group of values for the variable. This allows for automatic printing of headings and stubs. For recoded variables, only the new name and category names or class names should have to be coded. Another feature needed in the codebook is general descriptive information about the data file; such as file name, record length, creation date, etc.

I. Statistical measures or parameters: percent distributions, medians and means

These statistical measures are the most sought after in the census tabulations and therefore should be easy to produce. Different types of percents need to be calculated - row percents, column percents, and percents based on the total universe of the table.

J. Advanced statistical measures such as standard deviation, analysis of variance, chi-square, multiple regression, etc.

These statistical measures can be very useful, especially in determining the accuracy of the data. They are not used as often in census tabulations as percents, means and medians. They are used in the analysis of data and with sophisticated computer software they usually can be derived with aggregated or macro data much more cheaply than at tabulation time.

K. Summary comments

A brief summary of the purpose and the utility of each package will be included.

PART II

CENTS, COCENTS, AND CENTS 4

CENTS, COCENTS, and CENTS 4 were developed by the U.S. Bureau of the Census and were designed for tabulating census data. Since they are very similar in many ways, including the method of processing and in the construction of the command language, they will be discussed together in this report. The main difference is that CENTS is written in IBM assembler code (ALC), COCENTS is written in COBOL, and CENTS 4, the most recent version has both ALC and COBOL modules. CENTS 4 has other differences that will be noted in the various sections which follow.

The original CENTS software package was introduced in June 1970, and since that time, CENTS and COCENTS have been used extensively in many countries to process censuses. In 1982, the latest version of this series, called CENTS 4, was released by the U.S. Bureau of the Census.

CENTS, COCENTS, and CENTS 4 were developed expressly for producing census tabulations, so they are very powerful and at the same time flexible in the handling of input data and in producing output or tabulated results.

A. Large volume of data

CENTS, COCENTS, and CENTS 4 are capable of processing very large data files efficiently. There are no restrictions in the programs which limit the size of data files.

B. Various data formats

All three packages are capable of handling multiple data formats, even if those differing formats are on the same files. They can handle binary, packed decimal and character data, all without reformating the input data files.

C. Hierarchical file structure

CENTS, COCENTS and CENTS 4 were designed to process hierarchical data files usually found in population and housing censuses. If variables are needed for a table from different types of records, only those variables that are required need to be stored in core at any one time. This reduces the amount of core storage needed to process files with multiple record types.

D. Computing table data

The packages can produce a wide variety of tables. They may be used in producing tables for economic. agricultural, and population censuses as well as fertility, household, and other surveys.

1. Multi-dimensional tables

CENTS, COCENTS, and CENTS 4, can produce tables with as many dimensions as desired. These tables may be printed in whatever format is desired, including having more than two variables shown on the same page. The limits are the physical limits of the printer.

2. Large numbers of tables per run

All of the packages can produce multiple tables in a single run. The size of the run program grows as more tables are added, so that the limiting factor is the size of the computer processing the file.

3. Recoding facilities to create new variables

All have flexibility that allows for recoding input variables Index variables (combinations of two or more variables) may also be created. Recoded variables may be saved for use in all tables which require that new variable.

4. Definition of universe of table

There are no restrictions regarding definitions of the universe of a table. Only applicable records need to be included in any given table. The universe definition may vary from table-to-table within the same run.

E. Weighting factors

All three packages are able to handle weighted surveys. The weight factor can be set for the entire file or extracted from each record. Fractional weights are more difficult to deal with than integer weights. Both systems process table data in intergral form and if fractional weights are used, the tables must be modified during the table preparation stage. This is when tables are prepared for printing.

F. Handling missing values

The packages have some limited editing capabilities. Missing or erroneous data can be printed on a separate line of the table or be excluded from the table altogether.

G. Printing tables

CENTS, COCENTS, and CENTS 4 have separate programs for preparing the tables for printing. This table preparation phase can be as simple or as detailed as it is desired, depending on whether or not the user wants the printout to be in camera-ready format. Headings and stubs can be minimal or elaborate. Since the table data string can be saved on tape (or disk) between the data processing and table preparation phases, the tables can be reprinted with different labels without reprocessing the data file. This is especially important if the tables are from census data and errors were detected in the labeling of the original run. This also allows for the quick production of multiple printouts. The detailed work required for coding a large number of tables is probably the greatest single disadvantage with these packages.

All of the packages can produce a wide variety of tables based on many different types of data. Multi-dimensional tables up to a maximum of six dimensions can be produced. The design of the language is such that geographical coding schemes are easily manipulated so as to allows for the creation of tables of totals and special aggregations.

A table need not represent a simple crossing of two items of information. A single item may be tabulated using different aggregations or breakdowns in a single table. Similarly, multiple items may be tabulated in the same table. Multiple tables may be produced in a single run or pass of the data file. The program size grows as the user adds more tables and commands to the run; therefore, the limiting factor is the size of the computer.

Through the CENTS and COCENTS language, the user has the ability to recode input data items. This allows for grouping of values or creating index variables. Any recorded items may be saved for use in all tables which require that new variable.

There is no restriction regarding the inclusion of an observation in the universe of a table. The universe definition may vary from table to table within the run and is completely under user control.

In support of survey data files, a weight or sampling factor may be set for the entire data file or on a case-by-case basis. Fractional weights are more difficult to handle than integer values. The packages internally process data in integer form. As a result, if fractional weights are used, the tables may need modification before printing.

Among the new features provided by CENTS 4 is the ability to manipulate arrays. The user may define an array of up to 2 dimensions and may reference an array in most tabulation commands. CENTS 4 also allows for the definition of subroutines and for conditional calls to subroutines. It allows for variable table and array references. End of file processing has been simplified and new reserved identifiers have been added to provide the user with more program control.

The function of defining text for geographical areas is a separate subsystem in CENTS 4. This allows a user to create a generalized area name file which could be used in many CENTS 4 applications.

Because the internal representations of the table are saved on a storage medium between the input data processing and table presentation phases, the tables can be reprinted with different headings or stubs without reprocessing the entire input data file This is especially important if the tables are from census data and errors were found in the table labelling. This arrangement also allows for the quick production of multiple copies of the final tables.

The detailed work required for the coding of complex tables with intricate labelling was probably the greatest single disadvantage of both COCENTS and previous versions of CENTS. This job is made easier in CENTS 4 because the user interface language has been simplified. Text definition statements allow for easier coding and modification. User control over table formatting has been greatly enhanced and simplified. The user has the option of specifying table formats in great detail or taking system defaults.

Another disadvantage of previous tabulation packages is that they allowed for printing of only one decimal place for fractional items. CENTS 4 allows for three decimal places.

H. Generation of codebook or data dictionary (D.D.)

None of the packages uses a data dictionary. Description of input data and recoding of variables are all coded at the same tim as table producing commands.

I. Statistical measures or parameters: percent distributions, medians, means and ratios

These statistical measures may be produced with, CENTS, COCENTS and CENTS 4. The base of the percents can vary from table-to-tabl The only restriction is that statistical measures can only be print

with a maximum of one decimal place in CENTS and COCENTS. The algorithm for calculating medians is not as complicated as the ones used by analysis packages and therefore may not always be quite as accurate.

J. Advanced statistical measures: for example, standard deviation, analysis of variance, chi-square, etc.

CENTS, COCENTS and CENTS 4 cannot produce these advanced statistical measure

K. Summary comments

The CENTS family of tabulation packages offers the most widely used, completely portable systems available for use in developing countries. The U.S. Bureau of the Census supports the packages and can provide them free or at a small cost to other governments. The cost for reproducing the magnetic tape and one set of user documentation is $300, and this may be waived to USAID - assisted countries. The address is: U.S. Bureau of the Census, Washington, D.C. 20233.

T P L

TABLE PRODUCING LANGUAGE

Table Producing Language (TPL) is a computer language system designed to select, restructure, cross-tabulate, and display data. The system was designed by the U.S. Bureau of Labor Statistics to reduce the need for special computer programs that produce cross-tabulations, expand its tablemaking capability, and broadly speaking, improve production schedules by reducing time between collecting data and viewing the results in tabular form.

There are two steps in table production when using TPL. First, the various types of data must be named and their lengths and acceptable values specified. These entries collectively are called a codebook. Codebook preparation is a one-time activity after which many TPL users can reference the file for producing tables. Data to be tabulated may be stored in a wide variety of formats including hierarchical files. The second step uses TPL statements to reference codebook variables in producing tables. A TPL request indicates which data are to be tabulated and how they are to be formatted. The system can use arithmetic operations to calculate new variables and can delete, reorder, and regroup old variable values.

The TPL system can (1) place variables side-by-side, (2) allow levels of subdivision (variables within variables), and (3) designate separate, two-dimensional grids or the repetition of the grid for additional variables.

The system also can calculate averages, medians, minima, maxima, quantiles, and relative time. It also can create new variables from existing variables, calculate additional data after tables have been compiled, deal with subsets of information in the data file, and group, delete, or reorder values of existing variables. Many tables can be produced in a single run. Further, the user can arrange output in any sequence.

A. Large volume of data

TPL can handle large input files. Although processing speeds tend to be slower when a file of more than a few million records is used. There are no restrictions in the programs which limit the size of the file.

B. Various data formats

TPL can process many different data formats, even if they are in the same file. Types of data allowed include: EBCDIC, binary, floating point, and packed decimal. All data formats are described in the data dictionary (D.D.). Record formats may be defined in different ways if a certain portion of a record represents one variable in some cases, but another variable in other cases.

C. Hierarchical file structure

TPL can process hierarchical files. If the structure of any hierarchical unit is incomplete (e.g. a housing unit with no household records), the system will not include that unit in the tabulation.

D. Computing table data

1. Multi-dimensional tables

Tables of essentially unlimited dimensions may be produced by TPL.

2. Large number of tables per run

TPL can produce multiple tables in one run.

3. Recoding facilities to create new variables

There are excellent recoding facilities in TPL, inluding the ability to create index variables. Recoding is not allowed in the D.D.

4. Definition of universe table

Redefining the universe of any table is easily done with TPL.

E. Weighting factors

There is no automatic weighting facility in TPL, but weighted tables may be created through the use of TPL's COMPUTE command.

F. Handling missing values

Missing, or erroneous data can be shown in tables in a separate category. It may also be ignored for tabulation purposes, with the error record being printed.

G. Printing tables

TPL has a wide range of commands with flexible options for printing tables in a variety of formats. Tables may be produced in camera ready format. Coding becomes more complex if the tables are complicated, but default options may be used which print the tables in a form which is quite readable. TPL is one of the few packages which can control where words are split. At the U.S. Bureau of Labor Statistics (BLS) where TPL was developed, the package interfaces with a Print Control Language which produces camera ready copy on microfilm. TPL has the ability to use both upper and lower case letters in titles and to overstrike titles for darker print if desired.

H. Generation of codebook or data dictionary (D.D.)

TPL does create a separate D.D. It includes a description of the physical characteristics of the data file as well as descriptions of file structure, record formats and variable names and locations. Variables may not be recoded in the D.D. The format TPL uses for describing the data file is very much like the DATA Division of a COBOL program. The D.D. must therefore be coded by a programmer who is familiar with the file; however, the D.D. only has to be coded once for any given file.

I. Statistical measures or parameters: percent distributions, medians and means

These measures can be produced very easily in TPL. Unlike most packages, TPL does not have to produce these statistics for every cell of the table when they are requested.

J. Advanced statistical measures, for example, standard deviation, analysis of variance, chi-square, etc.

These measures cannot be produced automatically by TPL. However, because TPL is a language, formulas for many of these statistics can often be written into the program.

K. Summary comments

TPL will only run on large IBM or IBM compatible computer systems. The system is the easiest to use of all of the systems included in this section and except for very, very large files should be efficient enough to satisfy most users.

The TPL system is available from the U.S. Bureau of Labor Statistics. The fee will range from $300-$500 depending on site and user requirements. The address is: U.S. Bureau of Labor Statistics, 441 G. Street, N.W., Washington, D.C. 20212.

G T S

Generalized Tabulation System

Developed for processing large census files at the U.S. Bureau of the Census, GTS can process large volumes of data efficiently. It has been under continuous development since 1976 and although support and training are limited, it has several very desirable features. GTS is operational only on large UNIVAC computers.

GTS is designed to tabulate, summarize and display data from large surveys and censuses. Statements are coded in free-form format using an English-like syntax. The syntax includes a data dictionary facility. The user may construct a dictionary to describe his data records and to store frequently used GTS source statements. GTS has a limited editing facility. The values of data may be changed within a record, however records may not be added or deleted.

GTS builds a COBOL program based on the source statements to perform tabulations and produce table matrices. These table matrices may be summarized based on a hierarchical structure (usually geographic) defined by the user. The user is allowed full control over the appearance of the displayed tables. During the display process, arithmetic calculations may be performed on rows and columns of the table. Alphanumeric text may be substituted for table cell contents.

A. Large volume of data

GTS was designed specifically to handle very large files of data and as such is very efficient in tabulating and displaying census results.

B. Various data formats

GTS can handle either ASCII or field data, character data or binary data. Data in EBCDIC format (the standard IBM format) must be converted prior to tabulation.

C. Hierarchical file structure

Hierarchical files are processed easily with GTS. The package permits different types of records to be described. The user may then select the type of record wanted for a given table.

D. Computing table data

1. Multi-dimensional tables

GTS can produce tables with as many dimensions as desired. These tables may be printed in whatever format is desired, including having more than two variables shown on the same page. The table limits are the physical limits of the printer.

2. Large numbers of tables per run

If a sufficient amount of core memory is available, large, numbers of tables can be produced in each run.

3. Recoding facilities to create new variables:

GTS has extensive recoding facilities, including the ability to create index variables and to recode alphanumeric variables.

4. Definition of universe of table

There is a very useful SELECT command in GTS which allows the user to determine precisely the universe of the table.

E. Weighting factors

There is an automatic weighting feature in GTS. The weighting factor may be set for all cases or obtained from each case.

F. Handling missing values

Missing data may be ignored, replaced or displayed in a separate part of a table. If it is replaced, a corrected data file can be created under the GTS EDIT mode.

G. Printing tables

Tabulations are capable of being produced in camera-ready output, but the coding required for stubs and headings can be very time consuming. Titles and labels will be very readable and understandable if the user takes time to code them well.

H. Generation of codebook or data dictionary (D.D.)

GTS uses a separate D.D., which is coded only once for a file. The information coded in the D.D. gives record layouts and names and formats of variables within each type of record. The D.D. in GTS has some limitations. Category names for variables, if coded, are essentially comments. They are not used to create headings and stubs automatically. Recoding of variables cannot be done in the D.D.; recoding commands must be included with table commands.

I. Statistical measures or parameters: percent distributions, medians and means.

Percents, medians and means can be produced fairly easily in GTS. They do not have to be produced for every cell of the table. The package is flexible and permits the user to decide which field to use for the base of the percents.

J. Advanced statistical measures; for example, standard deviation, analysis of variance, chi-square, etc.

GTS can calculate a square root of a field so that standard deviations and variances may be calculated by using parameter coding instructions.

K. Summary comments

The GTS system operates only on large UNIVAC (Series 1100) systems. The software is similar to the COCENTS system from a users standpoint and can be obtained from the U.S. Bureau of the Census by special arrangement. The address is: U.S. Bureau of the Census, Washington, D.C. 20233.

LEDA

LEDA is an all purpose package that has three separate modules. Each module has distinct uses and when used together the package can provide data handling, error detection, editing and recoding procedures in addition to the data tabulation and display capabilities. The three modules: CASTOR, POLLUX and HELENE are each COBOL program generators that use a free-format language with optional French-like or English-like syntax.

The package was developed by the French Institute of Statistics and Economic Studies (I.N.S.E.E.) in the early 1970s and has been distributed and supported by I.N.S.E.E. since 1975. It has been used for processing census data in France and in several developing countries.

A. Large volume of data

Even though its developers identify LEDA as a survey analysis package, it can process large data files very efficiently.

B. Various data formats

LEDA is capable of processing input files in multiple data formats including binary, packed decimal and character data.

C. Hierarchical file structures

LEDA was designed to process hierarchical files usually found in census data files.

D. Computing table data

1. Multi-dimensional tables

   Typical census cross tabulations can be created by the HELENE module of LEDA but there is a limit of three levels permitted in any table display. The number of cross tabulations is unlimited — only the display of the tabular data is limited.

2. Large number of tables per run

   Multiple tables may be produced in each run. The limiting factor is the core memory available in the computer being used.

3. Recoding facilities to create new variables

   The LEDA software has the full capability of creating new variables either in the data dictionary or through recodes. Additionally, fractional data can be handled through floating point arithmetic.

4. Definition of the universe of table

There is wide flexibility permitted in the HELENE module to select records for each table being processed. Further, as in the case of most packages designed for census use, the universe may vary from table-to-table within the same run.

E. Weighting factors

LEDA provides for both fractional and integral weighting of data.

F. Handling missing values

LEDA has superior facilities in both the POLLUX module and CASTOR module for data correction, rectification, imputation and data manipulation.

G. Printing tables

The table display procedure used in LEDA limits the number of levels within a table to three. This limitation results in additional pages of output when dealing with complicated multi-level tables. Within this restriction, full stub and heading information can be created as desired for camera-ready copy.

H. Generation of code book or data dictionary

The code book when properly prepared reflects the hierarchical file structure including:

- The description of the logical structure
- Name and characteristic (code or quantity) of each variable
- The values of the code variables

I. Statistical measures or parameters: percent distributions, medians and means

LLUA has the capability of producing all of these with fairly complete flexibility.

J. Advanced statistical measures: for example, standard deviation, analysis of variance, chi-square, etc.

Only the standard deviation can be obtained using the HELENE module.

K. Summary comments

The LEDA system is available from and supported by the French Institute of Statistics and may be obtained directly from them. The address is: National Institute of Statistics and Economic Studies, (INSEE), Boulevard A. Pinard, Paris, CEDEX, France. The cost for the LEDA system is 200,000 French Francs. This cost may be waived under certain circumstances.

CENTS-AID

CENTS-AID is a package developed by Data Use and Access Laboratories (DUALabs) of Arlington, Virginia. The system uses modified versions of the COCENTS tally and consolidate modules, and has added a language preprocessor and other modules which make the command language easier to use. Although some of the flexibility of COCENTS is lost, the commands are more English-like, and there are more commands.

CENTS-AID was originally developed specifically for use with U.S. Census Public Use Sample files. One of its major features is its flexibility in handling input data.

A. Large volume of data

CENTS-AID is able to process large data files. It has been used to process census files of almost 3 million records and is able to process the files efficiently.

B. Various data formats

Although the earlier version of CENTS-AID could only process data in character format, the new version, CENTS-AID 3, can process data files in binary, packed decimal and EBCDIC/BCD formats, even if the different formats are in the same file.

C. Hierarchical file structure

Hierarchical files of up to [illegible] levels can be processed. Tabulations can be performed at different levels within a single run.

D. Computing table data

1. Multi-dimensional tables

CENTS-AID can create tables of up to eight dimensions. A complete multi-dimensional table can often be printed on a single page.

2. Large numbers of tables per run

If there is sufficient core storage on the particular computer used, many tables may be processed in one run involving only one pass of the data. The number of tables allowed depends on many factors and can be altered to suit the capacity of the computer being used. CENTS-AID is able to create tables with area breakdowns automatically.

3. Recoding facilities to create new variables

There are extensive recoding facilities in CENTS-AID, including the ability to create index variables, and the ability to recode alphabetic codes.

4. Definition of universe of table

CENTS-AID has three commands for restricting the universe of tables. These commands allow flexibility in specifying different universes within a single run. Universe restrictions can be applied to data transformations as well as table definitions, resulting in substantial cost savings.

E. Weighting factors

An integral weight factor can be set for all tables or for each table individually. A weight factor may also be drawn from each record.

F. Handling missing values

Missing or erroneous data may be detected through the recoding facilities of CENTS-AID, and can be excluded through use of the SELECT IF command. CENTS-AID 3 has a MISSING VALUES command that works in conjunction with the computation of advanced statistics.

G. Printing tables

Although tables produced by CENTS-AID cannot always be produced in camera-ready copy, they are very readable and understandable. Label descriptions of categories of variables can contain up to five or six words. Titles and boxheads may be printed automatically by the system using D.D. labels, or they may be spelled out in detail if camera ready copy boxheads are desired. If printed automatically, long words will be split arbitrarily. However, options in the CENTS-AID 3 allow the user to control the splitting of words in labels, and to produce and revise table formats without processing the file. CENTS-AID has the option of printing totals above and to the left. Some packages force totals below and to the right. When printing for a table flows from one page to the next, CENTS-AID does not have an option for controlling where the break occurs. For tables divided into wafers, a break may occur in the middle of a wafer instead of between wafers as it should.

H. Generation of codebook or data dictionary (D.D.)

A D.D. approach is used in CENTS-AID. A permanent Data Base Dictionary file (DBD) can be used repeatedly to provide all necessary descriptive and labelling information. When a DBD is used, no

variable or category labels need be supplied in the table command set, except for recoded variables. DBD's are created by a separate system (LEXICOGRAPHER). If no DBD has been created, the descriptive and labelling information can be included in the table command set. These data descriptions include variable name, data format, and record location in the DATA command, and category names in the VAR LABEL command. The variable and category names are printed automatically in table headings and stubs. The DATA command uses symbols and numerics with no accompanying descriptive information and is therefore of little meaning without a manual. The new version of CENTS-AID eliminates this inconvenience with a more descriptive DATA command.

I. Statistical measures or parameters: Percent distributions, medians and means

All three of these statistical measures may be produced. Row and column percents, as well as percents based on the total cell of the table, may be produced. If percents or means are requested, they will be shown for every cell of the table.

J. Advanced statistical measures, for example, standard deviation, analysis of variance, chi-square, etc.

CENTS-AID is able to produce some advance statistical measures, such as standard deviation, variance, and chi-square. CENTS-AID-3 includes additional measures such as Pearson's correlation, sum of squares, and sum of cross products. It also allows creation of a matrix file for input to some of SPSS' statistical routines.

K. Summary comments

CENTS-AID is maintained and supported by DUALabs of Arlington, Virginia. The CENTS-AID Software is patterned after COCENTS and the user language is quite similar to SPSS. The system is available for a one-time fee of $3,000 - $6,000 depending on type of user. The address is: DUALabs, Suite 607, 1515 Wilson Blvd., Arlington, Virginia 22209.

P-STAT is a large applications system which is typically used to clean data, manipulate files and do statistical analyses such as correlation, regression, factor analysis and crosstabulation. Its principal applications have been in areas such as demography, survey analysis, medical research and education. It contains approximately 95,000 Fortran statements and is in use on a variety of different computers at both educational and commercial installations around the world. Supported systems include IBM OS, VS, DOS/VS, VM/370, UNIVAC 1108, DEC 10/20, CDC CYBER, H6000, etc. P-STAT runs well in both batch and interactive modes. Additions for interactive use include width control, a text editor for correction of simple errors and, now being tested, automatic screen holding.

P-STAT is supported and marketed by P-STAT Inc. of Princeton, New Jersey, and was originally developed by Roald Buhler at the Princeton University Computer Center in 1962.

P-STAT is included in this tabulation section because the developers of P-STAT participated in the tabulation demonstration carried out by DUALabs and Delta Systems, and the results showed that, in fact, P-STAT does have a powerful tabulation feature.

A. Large volume of data

P-STAT has no limitation on the size of the data file, but processing speed is slowed with large files.

B. Various data formats

The tabulation programs in P-STAT accept input only from P-STAT system files. P-STAT has a file building program called DATA which reads data in many formats and then creates files in a standard P-STAT format. This program can also do some pre-editing of the data. It can check for invalid codes, make sure the file is sorted and detect duplicate or missing records in a questionnaire. Since P-STAT can read only P-STAT system files all input files must be preprocessed by the DATA program.

C. Hierarchical file structure

P-STAT does not have the ability to process hierarchical files. These files have to be split by the DATA program.

D. Computing table data:

1. Multi-dimensional tables

P-STAT can produce tables up to six dimensions. The format for printing the variables used is flexible. If many variables are used, the tables do not have to be split and printed on separate pages.

2. Large numbers of tables per run

Multiple tables can be produced in one run.

3. Recoding facilities to create new variables

P-STAT has extensive recoding facilities, including the ability to create index variables. All recoding is done in the DATA program. This means that if a second set of tables are needed with recoded variables not created during the first DATA run, the DATA program would have to be run again to create a new P-STAT system file.

4. Definition of universe of table

The universe of a table may be redefined for each table created.

E. Weighting factors

Complete flexibility in weighting Data — either integrally or fractionally — is available in P-STAT.

F. Handling missing values

The DATA program has facilities for detecting missing or erroneous values. The user can decide whether or not to include records with missing data in any tables.

G. Printing tables

P-STAT cannot produce camera-ready copy tables, but the tables are very readable. Horizontal and vertical lines can be used optionally to separate cells. Labels for rows and columns do not split words unless that is desired, but labels will be truncated if they are more than two or three words long.

H. Generation of codebook or data dictionary (D.D.)

A separate D.D. is not created is P-STAT, but a D.D. approach is used. There is a codebook section where labels are assigned to the values of each variable. Every variable must be included. There is no description for each variable of the data format and record location, since this has been taken care of by the DATA program. This makes the codebook more readable, but less flexible; if another recode variable is needed, the Data program must be rerun.

I. Statistical measures or parameters: percent distributions, medians and means

These statistical measures are easily produced with P-STAT. Statistical measures, when requested, are always produced for each cell of the table.

J. Advanced statistical measures, for example, standard deviation, analysis of variance, chi-square, etc.

P-STAT can produce most advanced statistical measures, including standard deviations, correlations, and factor analysis.

K. Summary comments

P-STAT is primarily a statistical package that has a fairly good tabulation module. The system is available on a yearly lease from: P-STAT, Inc., P.O. Box 285, Princeton, New Jersey 08540. The cost varies by the type of user from $1,000 - $5,000 per year.

XTALLY

This package has been maintained and distributed by the United Nations Statistical Office since 1976. The system has been used by United Nations personnel in many developing countries to process census and survey data. XTALLY is capable of producing multi-dimensional cross tabulations summing one or two variables, providing sub-totals, percentages, ratios, means and totals at all levels of tabulation. It is written in RPGII and is designed to be used on small (as little as 24K bytes of main memory) in those cases where the larger more flexible tabulation packages cannot be installed.

A. Large volume of data

There is no inherent limit on the size of the input data file. Processing speeds for large files are slow on the computers for which there is currently a version of XTALLY available.

B. Various data formats

XTALLY can process only character format data. Packed decimal and binary data are not allowed. Records can be a maximum of 99 characters in length.

C. Hierarchical file structure

Hierarchical files cannot be processed.

D. Computing table data

1. Multi-dimensional tables

XTALLY can produce tables of up to seven dimensions up to three for column variables and up to four for row variables.

2. Large numbers of tables per run:

Many tables can be produced in one run with XTALLY.

3. Recoding facilities to create new variables

XTALLY can recode input variables into different categories. Each variable can be recoded into more than one set of categories. Index variables cannot be created.

4. Definition of universe of table

The definition of universe can be changed for each table. This is done by eliminating certain values of table variables.

E. Weighting factors

A weighting factor may be used for all records. Weighting factors cannot be extracted from individual records.

F. Handling missing values

Missing data can be eliminated or displayed through the recoding facility. The value of one variable cannot be checked against that of another variable for validity.

G. Printing tables

XTALLY cannot produce camera ready copy tables. The names of variables used in a table are not printed as labels for the table. Only the names associated with individual codes of variables are printed. These names may be a maximum of four characters long and are therefore hard to read and understand.

H. Generation of codebook or data dictionary (D.D.)

A data dictionary is generated in a separate step in XTALLY. Data locations are defined, variables are recoded, and associated category values are assigned.

I. Statistical measures or parameters: percent distributions, medians and means

Percents and means may be calculated in XTALLY, but not medians. The base for percents can be the total cell for the table. Row and column percents may also be calculated.

J. Advanced statistical measures, for example, standard deviation, analysis of variance, chi-square, etc.

These measures are not available in XTALLY.

K. Summary comments

The XTALLY system is recommended only when no other system is available. The restrictions of the system limit its value in census and survey tabulations. The system is available without charge to member governments from the United Nations Statistical Office, New York, N.Y. 10017.

PART III

A COMPARISON OF SOFTWARE FOR TABULATING SURVEY AND CENSUS DATA[1] (edited)

1. Introduction

In this paper we describe a demonstration of tabulation capabilities of statistical packages used to tabulate results from surveys and censuses. It is in part an extension of an earlier review (Francis et al, 1976): we have added different programs, more complexities, including hierarchical files, weighted observations, and some very large data sets.

The test has two phases, each having its own benchmark problem: Phase 1 examined programs for tabulating results of a population census while Phase 2 reviewed programs for presenting results from sample surveys. Phase 1 was conducted in cooperation with DUALabs and Delta Systems Consultants, Inc. One object of Phase 1 was to identify programs that might be used in some forty developing countries for processing the current (1980) round of censuses.

Developers undertook these tests voluntarily. We have the unpleasant task of criticizing their efforts. This inevitably includes some elements of subjectivity both in the choice of test problem and in some evaluations we make of the results. But one of the two principal aims of this exercise is to try to introduce some objectivity into the evaluation of statistical software by designing some standard problems and using measures of software performance that are at least partially objective.

2. Phase 1

2.1 The Test Problem

The test file consisted of approximately a one percent sample taken in a mythical country called Popstan. This country was created by the International Statistical Programs Center (ISPC) of the United States Bureau of the Census to provide a case study for the world-wide 1980 round of population censuses. The Popstan case study will be used by the Census Bureau in its training program for officials from developing countries.

It should be noted that this test problem for Phase 1 was developed with CENTS and COCENTS packages in mind. These packages had been developed also by the ISPC as the CENsus Tabulating System, and a CObol version, to handle this kind of problem. Therefore Phase 1 cannot be regarded as a comparison of general tabulation capabilities.

[1] Francis, I., S. P. Sherman - Cornell University & Buhrman, R. I., J.L. Willard - Delta Systems Consultants, Inc., A Comparison of Software for Tabulating Survey and Census Data, American Statistical Association, Statistical Computing Section, Aug. 1979.

Rather it seeks to discover how well several other more general-purpose packages fare in trying to perform a specific task for which these two were specifically designed.

The file was prepared in sequence by population record within housing record within Province within Region. There were a total of 277,755 records of which 74,565 were housing and 203,190 were population records. The file was sent to developers on magnetic tape in a format of their choosing.

The ten tables for this test were based on tables recommended in the Popstan case study and are consistent with international recommendations of the United Nations Statistical Office and the regional UN Economic Commissions. Since many developing countries want to be able to use computer-created tabulations as camera copy for publishing census results, the participants were asked to produce the tables in a form as close as possible to the specifications. In addition, they were asked to write a brief technical report including management, programmer, and clerical time, computer CPU and I/O time.

2.2 Meeting Specifications

The seven packages[2] for which results are included are listed in Figure 1. The developers of all seven packages had earlier claimed that their packages had tabulations as "a principal purpose" and could efficiently handle files of over 200,000 records.

To meet specifications a table had to be (a) visually attractive and (b) numerically accurate. CENTS, COCENTS, GTS and TPL produced tables that presented the information most effectively and were camera-ready as far as lay-out is concerned. The other packages' tables fell short of this standard in varying degrees.

With regard to numerical accuracy, minor coding errors resulted in several errors in the tabulated data in CENTS, COCENTS, and P-STAT. In the calculation of the medians only CENTS and COCENTS produced the correct values for all tables. TPL, CENTS-AID, and GTS, produced one or more incorrect medians through coding errors. P-STAT and LEDA did not produce medians, although P-STAT is able to produce cumulative percentiles so that the user can compute the median.

[2] This test was undertaken prior to the release of CENTS 4 in late 1982, and therefore does not include CENTS 4. The developers of XTALLY declined to participate in the test and for that reason XTALLY is not included in the test.

2.3 Efficiency

Some measures of efficiency are listed in Figure 1. The four most efficient in terms of CPU time, CENTS, COCENTS, GTS, and CENTS-AID are special-purpose tabulation packages directly related to each other through the U.S. Bureau of the Census: the first two were developed for efficient census tabulation, and the other two evolved for different purposes. LEDA and TPL are also special purpose tabulation packages, LEDA (which also has filing and editing capabilities) from the French Government's I.N.S.E.E. and TPL from the U.S. Bureau of Labor Statistics. The last, P-STAT, maintained by a private company, is the only package among the seven which is not designed primarily for tabulation. (Given the origin of the test problem it should not be surprising that CENTS and its relatives were the most CPU-time-efficient.)

The CPU times are converted from actual reported CPU times using an IBM conversion table for comparing different IBM machines. The only non-IBM machine in this test was a Univac used by GTS and so its time is approximate. It should be remembered that all these times are only approximate, since comparative ratings are based on balanced configurations, and the precise hardware used in these tests was not known in most cases.

The developers estimated their own personnel time, and no attempt was made to verify these numbers. The CENTS and COCENTS tests were done by the same person who reported that approximately 9 hours were spent on each for a total time of 18 hours for the two tests. However, since the parameter instructions are quite similar, it is likely that if two different people had prepared the two sets of instructions, more than 9 hours would have been required by each. Moreover the programmer himself pointed out that he had the advantage of the extremely powerful text editor and job submission system WYLBUR.

If a real census called for 500 tables rather than our 10, the time for writing and checking the table programs, their headings and labels, would be relatively more favorable to CENTS-AID, P-STAT, and TPL, since our tests unintentionally favored the packages which did not require the creation of a data dictionary. If there had been more tables, this one-time coding requirement would have become less significant.

On the other hand, in a real census, the differences among the CPU times would be more dramatic too. For example, if a population were 14,000,000, and the small computer to be used required 20 times more CPU time than those in this test, and if eight passes through the file were needed, than the CPU time required would be 5,000 times the amount used in this test.

2.4 Trade-offs

In general efficiency is sacrificed to achieve portability. CENTS, written in ALC, is most efficient but least portable across machines. COCENTS, written in COBOL, is less efficient, but more portable. P-STAT, designed for portability (of code and data files) and scientific computing power, was written in FORTRAN which is not efficient for file handling nor character manipulation for tabulation.

P-STAT and CENTS AID have omnibus commands to perform different analyses. Such commands may, for example, produce a table similar to one requested but fail to meet exact specifications. One ought not to be surprised that to get a table printed precisely as specified, one might have to either use a specialized tabulation package or sacrifice the high level language.

The level of user languages range from CENTS and COCENTS, which have a low-level, almost assembler-like language, to P-STAT, CENTS-AID, GTS, and TPL which have almost English-like commands. P-STAT can be interactive. The low level languages are the hardest to code correctly and debug, especially by people who do not use them every day, whereas it is easier for one person to check another's program in one of the high level languages.

CPU TIME IN SECONDS TO COMPLETE TEST

NAME OF PACKAGE	COMPUTER MAKE	CPU TIME (SECONDS)	CONVERTED CPU TIME*
CENTS-AID	IBM 3033	130	554
CENTS	IBM 370/168	39	107
COCENTS	IBM 370/168	136	374
GTS	UNIVAC 1108	324	283**
LEDA	IBM 3032	285	784
P-STAT	IBM 360/91	630	1732
TPL	IBM 3033	170	841

* All times converted to equivalent IBM 370/158 Model 3 time assuming two runs for each system.

** Times for the UNIVAC and CDC systems are approximations.

FIGURE 1

PERSONNEL HOURS TO COMPLETE TEST

NAME OF PACKAGE	TOTAL	MANAGEMENT	CLERICAL	PROGRAMMER
CENTS-AID	42	4	3	35
CENTS	9		2	7
COCENTS	9		2	7
GTS	28	2	8	18
LEDA	15			15
P-STAT	20			20
TPL	36	10	2	24

FIGURE 2

PACKAGE PORTABILITY

NAME OF PACKAGE	MINIMUM CORE (K BYTES)	LANGUAGE(s) WRITTEN IN	ESTIMATED NUMBER OF SYSTEMS INSTALLED[3] (as of 12-31-1982)
CENTS-AID	90	ANSI COBOL	4
CENTS	24	ALC	1
COCENTS	32	COBOL	14
GTS	144	COBOL	1
LEDA	320	CPL/1 and ALC	3
P-STAT	180	FORTRAN	13
TPL	300	XPL and some ALC	1

FIGURE 3

NUMBER OF PAGES PRINTOUT PRODUCED
BY EACH PACKAGE BY TABLE NUMBER

NAME OF PACKAGE	TABLE NUMBER										
	1	2	3	4	5	6	7	8	9	10	1-10
CENTS-AID	17	2	2	3	1	3	2	3	12	1	46
CENTS	3	1	1	1	1	1	1	1	7	1	18
COCENTS	3	1	1	1	1	1	1	1	7	1	18
GTS	3	1	1	1	1	1	1	1	7	1	18
LEDA	2*	2*	2	1	1	1	1	1*	20*	1	32
P-STAT	12*	1	1	4	2	4	2	2	42	1	71
TPL	3	1	1	1	1	3	1	1	10	1	23
AVERAGE NUMBER	6.1	1.3	1.3	1.7	1.1	2.0	1.3	1.4	15.0	1.0	32.3

* Did not produce desired statistic for this table

FIGURE 4

[3] This is the number of different types of computer systems, not the number of installations.

EDITING SECTION

INTRODUCTION

Computer editing of statistical data files can be a very challenging task with as many opportunities to introduce new errors as there are ways to correct errors that are found. Unlike the tabulation procedure, where control totals can be established to insure that all cases are accounted for, most errors cannot be verified in absolute terms; and in fact some errors are not detectable with the most precise computer analysis.

The principal purpose of the editing phase in a population census is to produce a data file that contains acceptable and consistent entries for all applicable data items for every housing unit and person enumerated.

The usual definition of editing[1] includes:

a. The checking of each field of every census or survey record — that is, the recorded answer to every question on the questionnaire — to ascertain whether or not there is a valid entry.

b. The checking of entries in predetermined fields or combination of fields to ascertain whether or not the entries are consistent with one another.

During the initial phases of the processing cycle (data gathering, coding, and data entry) that occur during a census, errors affect and tend to falsify the universe under study. By introducing a computer editing procedure, these errors can be detected, corrected and documented so that the users of the tabulated results are aware of the quality of the data.

This report concentrates on those editing packages that have been used to process national census files, with particular attention paid to the capabilities of the various software packages to detect, correct and document errors.

[1] Fellegi, I.P. and Holt, D. "A Systematic Approach to Automatic Edit and Inputation". Journal of the American Statistical Association, 71 (March 1976), 17.

As computers have become the universal tool for tabulating and analyzing large census data files, it seems logical to extend this usage to include a method of monitoring or overseeing the census processing activities with a computerized management system. Several countries have implemented such systems, but not until 1981 was a generalized system installed and used in a developing country. The system - called CONTROL, was designed in 1978 and subsequently enhanced and installed in two countries in Latin America.

Although not a tabulation or editing package, the CONTROL software package is designed to complement the editing process. CONTROL is a management reporting system designed to monitor various activities of the census and to provide timely information on the progress of a census or large survey. The information is provided in tabular and graphic form by major phases of the project. It shows where bottlenecks occur, and it shows the timeliness of production schedules. The system keeps track of all census geographic areas and identifies missing areas and duplicate areas. The CONTROL software is described in this section on page 65.

Major Editing Software Packages
Included in this Inventory Report

Four editing software packages that have been designed exclusively for statistical data checking and correction have been included in this section. While other statistical packages have data editing capabilities, these four packages are presented to illustrate various approaches to the specific task of data editing. The packages are shown below with the name and address of the developers, and a brief description of the attributes of each package.

MAJOR EDITING SOFTWARE PACKAGES

NAME	DISTRIBUTOR	COST	ATTRIBUTES
CONCOR A	CELADE Latin American Demographic Center Casilla 91 Santiago, Chile	Cost of reproducing materials	A package to edit and check consistency of census files, it is very efficient in terms of computer time. Requires IBM or plug compatible system with 64 K-bytes of memory. It is written in ALC. Separately maintained by the World Fertility Survey in London, U. K.
CONCOR C	U.S. Bureau of the Census, Washington, D.C. 20233	$350*	More flexible with more features than CONCOR A, but somewhat slower. Written in COBOL, it requires at least 128 K-bytes of memory. It has been adapted to run on a variety of computers.
UNEDIT	United Nations Statistical Office New York, N.Y. 10017	No charge to member Governments	UNEDIT provides minimal editing capability for small computers with RPG II compilers. Requires only 32 K-bytes of memory, and is easy to install and use.
CANEDIT	Statistics Canada R. H. Coats Building Tunney's Pasture Ottawa, Ontario Canada	Negotiable Status	CANEDIT is the most advanced system. It was designed for use by the Canadian Census Office and is written in ALC and PL/1. It is difficult to install and use in developing countries. It requires at least 300 K-bytes, large disk storage and a data base management system (e.g., RAPID, TOTAL).

PART I

In describing the methods of detecting errors in census data, it seems logical to follow the same order as the data themselves are structured; i.e., from individual variables, to records, to the full questionnaire. (See Figure 1.) There are some other errors that are out of the scope of a usual editing phase, such as missing or duplicated enumeration areas, or missing or duplicated households. These types of errors could be detected and corrected better at an earlier stage and one approach to solving problems such as these are discussed in the section on CONTROL.

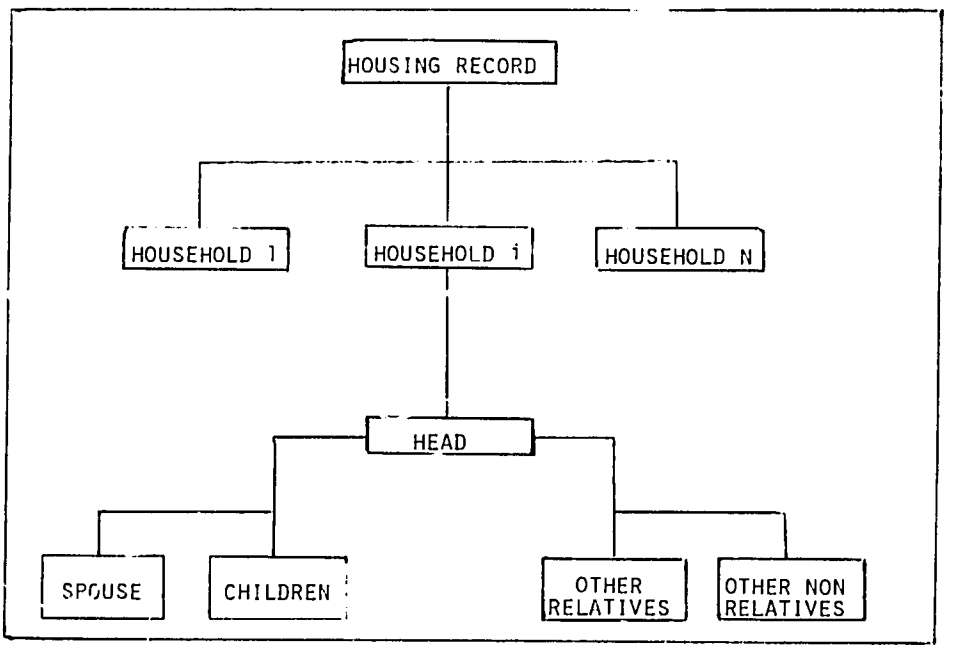


FIGURE 1: STRUCTURE OF THE INFORMATION ON A POPULATION CENSUS UNIT (QUESTIONNAIRE).

A. Types of errors to be examined

1. Intra-records checks

Examining each information element (variable) separately, a Range Test can be applied to detect invalid codes (qualitative variables) or out of range values (quantitative variables).

Taking into account the relationships among variables, two types of checks or tests can be performed to detect errors: (a) Logical Relationships Test; for example, "RELATIONSHIP TO TO THE HEAD OF THE HOUSEHOLD" equal to "SPOUSE" implies "MARITAL STATUS" equal "MARRIED"; (b) Arithmetic Relationship Test; for example, "AGE" must be greater than or equal to "YEARS OF EDUCATION" plus 5.

2. Inter-records checks

Logical and arithmetic relationships should also be applied to variables of different records either of the same type or different types. Examples of this relationship could be: "SEX of SPOUSE" is "FEMALE" if and only if "SEX" of "HEAD" is "MALE"; "AGE of HEAD" must be greater than or equal to "AGE OF OLDEST CHILD" plus 12.

It should be noted that the differentiation between 1 and 2 has been made because, in the last type of relationship, variables of any record must be available at any time. This means that in a population census the editing system should treat a household unit as the unit under study rather than the individual record being the unit.

Structural editing could also be considered within the scope of Inter-record checks. This particular type of editing refers to the completeness of the household information; that is, to the detection of missing or duplicated records of any household unit. In the specific case of a population census, this type of editing is trivial since the structure is very simple. A housing record contains a variable that shows the number of households in the housing unit and each household record contains either one or two variables that show the number of population records, or the number of population records by sex; and finally, the population records within each household are numbered.

B. Correction of errors

In a census, most of the corrections will be made either automatically or mechanically, although some of the corrections such as inserting missing records might be made manually. However, in the editing of surveys, corrections are usually made by hand mainly because the number of cases and the number of errors are usually smaller and the information is usually more accurate.

1. Manual correction

Here, the computer merely detects the error, with actual correction being left to the personnel in charge of the census or survey operation. The computer should provide all the information needed to identify the case or questionnaire containing the error, and identify the variable or variables in conflict. By referring back to the source document, the person (or persons) responsible for the editing procedure determines the changes that need to be made on each questionnaire and updates the file. Changes refer to a single variable, a set of variables, or a complete record, or the whole questionnaire.

In order to update a file, the editing system should provide a module to perform this task. This means: (a) the specifications for the update module should be understandable for subject matter people, as they are the ones who are going to determine the actual changes to be made in the basic information; (b) The system should provide the necessary features to minimize the possibility of introducing new errors rather than correcting the detected ones. An error in the identification of the questionnaire could alter a good questionnaire, leaving the wrong one in error.

2. Automatic correction

This is undoubtedly the most cumbersome phase of the editing. When the error has been detected by a range test, there is little doubt where the problem is or which variable needs to be corrected. However, when a relationship (consistency rule) involving two or more variables is not satisfied, only part of the problem has been solved. The variable(s) in error still needs to be identified.

The process of identifying the variables that need to be corrected is probably the most difficult part of the entire editing process. Therefore, a very important feature of the system would be the automatic identification of the variables in error.

Once the variables containing erroneous information have been identified, they need to be corrected either by a logical deductive procedure, by a probabilistic method, or by a combination of both.

When using the logical deductive procedure, it is known beforehand that there is one and only one code that can satisfy the set of relationships established for that particular variable (evaluative variable). In this case, use of decision tables could make the task easier, and therefore, the editing system should allow for the use of matrices. If the values in a given matrix never change (i.e., probability of assigning value at any coordinates of x and y = 1.0) the matrix is called a cold deck.

The common probabilistic method is the hot deck approach in which cases not requiring edit provide code values to update a matric while cases requiring edit are supplied codes based on their x and y coordinates in the matrix.

A fully automatic imputation feature has advantages and disadvantages. The advantages would be: (a) It reduces the amount of work left to the user; and (b), it increases the possibilities of ending up with an error free file. On the other hand, the disadvantage is that it generally requires a large computer system to run such a software package, which is not always available in developing countries.

Unfortunately, less sophisticated systems that require fewer computer resources may not provide the level of accuracy desired. For example, suppose the editing system derives the variables on which the imputation is going to be based from the consistency rules. Then the minimum requirement for the imputation takes into account at least all the dependent variables or those that show some degree of relationship with the one that is being inputed. Unfortunately, the dependent variables or related variables are not always linked to the variable being inputed by a direct relationship. For instance, the responses for the variables "MARITAL STATUS" and "CHILDREN EVER BORN" have a high correlation, although they are not linked by any relationship.

It is quite difficult to define precisely what would be desirable or ideal. Probably the ideal is some point between the full automatic imputation and a mechanical imputation in which most of the work is left to the user.

C. Documentation of errors detected and corrected

Although this is a secondary stage in the cleaning of data, it is no less important than the two preceding ones. It allows the user to make a judgment of whether or not the procedures applied were correct and produces a quantification of the changes that were made to basic information.

Documentation of errors detected and correted may be of two kinds

(1) Overall information giving a first impression of the cleaning process. For example, the marginal frequency distribution of inputed cases in values that are either absolute or relative to the total number of cases.

(2) Detailed information in a file containing all the questionnaires in error, plus the identification of the variables corrected with their corresponding values before imputation.

It is desirable to be able to control the level of information or the amount of detail to be printed out. In a census file with millions of records, it isn't possible to review all errors in detail, so that summary information may be more useful than line item detail information. The same is true for the area breakdown of error data. The practical matter of producing census results in a timely way inhibits the detail investigation of very small area error conditions in most cases.

There is a final documentation procedure that is useful. This is the automatic creation of an error file which lists all errors that have been detected by the editing system. By separating the errors and producing a new file, it will eliminate the need to search the entire file for errors at a future time and create a new resource for evaluating the census procedures. And this error file will allow users to further analyze the editing operation.

It is also important to have the option of choosing or defining the detail with which the error statistics are to be printed out.

D. Input file

When dealing with statistical data, there are a wide variety and types of applications. Some of them have a very complex data structure. Others, such as a population census file, have a less complicated data structure.

As it was stated in A.2, a good software package is one in which the inter-record checks are able to call up any variable at any time. This means that the editing system must be able to handle complex hierarchical files as well as flat files; and therefore, it should provide a way to define not only the variables contained in each record type, but also those that have multiple occurrences within a particular record type. In addition to being able to define housing record variables, household record variables and population record variables, the generalized editing system must be able to recognize, when referencing a household or population variable, which of the occurrences is being referenced. In other words, if the user is referencing the variable "SEX", the identification of this item is not complete since there are, for the same questionnaire, several items with the same label. In this particular case, one should specify in addition to the item name, the household identification, and within that particular household, the specific person or line number to which all of the population records the variable "SEX" belongs.

E. Error free output file

Since the main purpose of editing is to generate an error free file, there will always be an output file regardless of the method used in the edit. This output file could be in the same form as the input file, retaining the same structure and format, or it could be a different structure (such as a flat file) with its own format. The first case is probably the most common because it makes documentation of files easier. On the other hand, if the tabulation system that is to be used cannot handle hierarchical files, the user would need or prefer a flat file as output.

F. General aspects

When examining software packages, there are some general characteristics to take into consideration.

1. Portability

In order to have a portable system, there are at least two minimum requirements: (a) The system should be easy to install on different computer systems; (b) The documentation, (installation guide, user's manual) of the system should be sufficiently good enough to provide ample guidance so that little or no assistance would be required in its implementation and use in distant locales.

2. Reliability

The system should have been used or tested thoroughly so that users will have confidence in it.

3. Language

The language should be easy to use and, therefore, easy to learn. Sometimes, this characteristic is in conflict with flexibility and power of the system, but at a minimum the language should contain a sub-set of commands or functions that would allow inexperienced programmers or statisticians to perform simple editing.

4. Performance

The system should be efficient in terms of CPU time to permit it to be used on small-to-medium-size computers.

In summary, the most important features of such an ideal editing system can be outlined, and this outline will be the base for the analysis of each individual editing system referenced in this report. The structural editing as mentioned earlier in this

part (Page 38), usually can be better handled in a separate manner. To the extent the individual packages are able to deal with these types of error conditions, it will be noted. Specifically, the detection of missing or duplicated records and the subsequent creation or deletion of these errors are generally thought of as structural edits. The other errors are results of erroneous responses, enumerator, coder, or data entry operator errors or combinations of these. The format to be used is as follows:

A. Types of errors

1. Intra-record checks: Range test
   Arithmetic relationships
   Logical relationships

2. Inter-record checks: Arithmetic relationships
   Logical relationships

B. Correction of errors

1. Manual correction

   Ability to modify the whole questionnaire, a single record or a variable within a file.

2. Automatic correction

   (a) Automatic identification of variables in error
   (b) Manipulation of matrices
   (c) Method of inputation

C. Documentation of errors

1. Overall error statistics
2. Error messages identifying questionnaire, record and variables in error (for manual correction)
3. User control of the degree of detail in documenting errors
4. User control of area breakdown of error statistics
5. Automatic generation of error file (containing all the questionnaires found in error)

D. Input file(s)

E. Output file

F. General aspects

1. Portability
2. Reliability
3. Language characteristics
4. Performance

G. Summary comments

PART II

CONCOR

Editing and Imputation System
(Assembler version: CONCOR-A)

The CONCOR edit system was developed in 1975 to provide a simplified method of identifying and correcting errors and inconsistencies in census and survey data files. CONCOR was conceived originally to apply to census editing, but later it was expanded to be applicable to complex surveys such as the World Fertility Survey (WFS).

A. Types of errors

1. Intra-record checks

The system allows users to check relationships among variables of the same record easily. This relationship can be either logical or arithmetic and either of the one way implication (→) or two way implication (↔ if and only if). The use of compound conditions (Boolean expressions using "AND" "OR" operators) are accepted.

CONCOR-A has a range command that performs that validity code test for qualitative variables and out-of range tests for quantitative variables.

Whenever the test fails, the user has the option of either marking the variable for a later imputation, or assigning an optional code such as the "NOT REPORTED" code. In addition, the system will automatically generate an error message.

2. Inter-record checks

The system does not provide for full inter-record checking capabilities. However, all those record types that have only one occurrence (that is, a housing record) within a questionnaire, will be kept in memory during the entire processing cycle of questionnaire. Therefore, the variables defined in those record types will be available at any time, and full inter-record checks are allowed in those particular cases.

When a record type has multiple occurrences (i.e., a population record type), only the variables belonging to one occurrence are available at a time. So, if inter-record checks between records of the same type are needed, the user has to save the whole set of records or at least the variables involved in the relationship, using the same programming techniques required by any general purpose high level language such as COBOL or FORTRAN.

B. Correction of errors

1. Manual correction

The system has a module that allows users to update files by means of four different operations:

(a) Changing the codes or values of one or more variables independently of the record they belong to, by simply giving the name(s) and the new code(s) or value(s) of them.

(b) Replacing an entire record; a set of records, or the whole questionnaire, if necessary.

(c) Inserting a record or a set of records for a particular questionnaire.

(d) Deleting a record; a set of records, or the whole questionnaire.

It should be pointed out that CONCOR-A doesn't provide any failsafe method that ensures the correct spelling or keying of the questionnaire identification. In other words, if a mistake is made when writing down or key-entering the questionnaire identification, changes may be made on another completely unrelated questionnaire.

2. Automatic correction

(a) Automatic identification of variables in error

CONCOR-A does not provide this feature to identify variables in error. In order to decide which variable needs to be corrected, the user will have to code the routine using various commands.

(b) Manipulation of matrices

The system handles one or two dimensional matrices. If more than two dimensions are needed, the user has to recode variables so that only two dimensions are required.

(c) Imputation method

The hot deck approach is available in CONCOR-A. In order to use this method, the user must design his own hot deck matrices and decide on which variables the imputation is going to be based.

Documentation of errors

1. Overall error statistics

At the end of the editing execution, the system provides the following statistics:

— Absolute and relative frequency of error occurrences by variable.

— Absolute and relative frequency of errors by consistency rule.

— Absolute and relative frequency distribution of imputed codes or values by variable.

— Total number of questionnaires in error.

— Total number of questionnaires read.

— Total number of records read.

— Total number of output questionnaires.

— Total number of output records.

As an option, the user can get detailed statistics of the errors detected. This information can be used either for manual correction or more detailed analysis of the correction procedures.

2. Error messages

The documentation of error discovery (and correction)is as follows:

— Identification of the questionnaire in error.

— Identification of the variables involved in a specific error and their corresponding codes or values either before or after the correction has been made.

— Literal error messages.

— The complete questionnaire print out.

3. User control of the degree of detail

The user has the option of specifying:

— Only the overall statistics.

— Overall statistics plus identification of questionnaire and variables in error, and the error message.

— The above plus the printout of the questionnaire.

4. User control of areas breakdown of error statistics

CONCOR-A does not provide this feature for the area breakdown. The system only provides the total error statistics for all the data in the file. If the user needs to have error statistics by some specific area level, the batch will have to be prepared in such a way that the end-of-file agrees with the end of the area.

5. Automatic generation of error file

CONCOR-A automatically generates an error file containing:

— All the detailed statistics specificed above except for the error message.

— All the questionnaires in error.

D. Input files

CONCOR-A can accept hierarchical files as well as flat files. The former can be of up to 50 different record types, each of them requiring a common field for: (a) The questionnaire identification — from 1 to 9 digits — and (b) The record type identification — 1 or 2 digit field. The maximum record length is 100 characters (bytes). The system can accept three different types of input variables: Numeric, alpha-numeric and binary variables. Numeric variables can be from 1 to 9 digits, alpha-numeric variables can be from 1 to 4 characters and binary from 1 to 8 hexadecimal digits.

E. Output file

The system can generate, as an option, an "error free" output file. This file is not automatically generated by the system, therefore the user has to specify the variables wanted on the output file, in the order desired on the output record (location) and the length of each variable either in bytes or digits, depending on the type of output file (binary or external numeric). There are some advantages in having the user define the output file. For example, the output file can contain, in addition to the input variables, new variables created through the editing phase or it can transform a complex hierarchical file into a simple flat file.

F. General aspects

1. Portability

The system was written in IBM 360/370 assembler language (ALC). Therefore, CONCOR-A can only be implemented on IBM 360/370 computers, with a minimum of 64K bytes of memory, 5 megabytes of direct access storage or 4 tape drives.

The documentation of the system is fair. There is a user's manual which could be improved by inserting some examples that would help to clarify the application of each command of the language.

There is no system installation manual available. However, at the end of the user's manual, there are a few pages that offer some help to programmers in installing the package. Also, there is no programmers guide available. This means that whenever a complex editing program has to be written, assistance will usually have to be provided to show the user some of the capabilities of the system. It also means that it is going to be very difficult for someone unfamiliar with CONCOR-A to make modifications and enhancements to the programs.

2. Reliability

CONCOR-A has been extensively used by the United Nations Latin American Demographic Center (CELADE) and the International Statistical Institute World Fertility Survey Project (WFS). CELADE has applied it to the editing of various sample data files and to different types of surveys. WFS has been using it as the main tool for the editing of fertility surveys throughout the world.

The results have been diverse. Normally, users who didn't know the system well enough had problems using it, mainly because errors introduced in language specifications were sometimes not always detected by the system. It should be pointed out that CONCOR-A was originally designed to be used within CELADE rather than as an exportable software package. As a result of these experiences, CELADE and WFS have improved the system to make it more reliable. Both institutions undertook a joint effort in 1978 and added to the original system a "syntax and semantic analyzer" module that is now implemented. In spite of this, a complex system such as CONCOR-A written in assembly language will probably never be 100 percent reliable. However, it can be rated as reasonably reliable.

3. Language characteristics

CONCOR-A language is a high level editing oriented language. The language requires a semi free-format specification and contains a basic set of commands that is easy to learn and easy to use. This basic set of commands is oriented to manual correction, and in order to make automatic corrections, some skills in programming techniques are required.

The language has some special commands for the editing of statistical data, but, on the other hand, it does not have minor features that would help to make the editing an easier task.

4. Performance

One of CONCOR-A's major attributes is its efficiency in execution time. For example, a program containing approximately 900 editing rules for the Fertility Survey of Peru, (with an average of 500 variables per questionnaire) on an IBM 370/148, processed at a rate of 10,000 records per minute. (CPU time)

G. Summary Comments

There are several conclusions that can be made about CONCOR-A.

— CONCOR-A meets some of the requirements of an editing package, but is far from the ideal editing system.

— It almost certainly will make the editing operation much easier; reducing the amount of programmer time that would be required to perform the same level of editing with a general purpose high level language by perhaps as much as 80 percent in some cases.

— CONCOR-A has a serious weakness in that the inter-record check feature among records of the same type, now requires some programming skills to accomplish successfully.

— The documentation of the system has been enhanced recently, but needs some improvement, especially in the installation and programmer guides.

COBOL CONCOR

Editing and Imputation System
(COBOL Version: CONCOR-C)

This version of the CONCOR editing system was first developed by the U.S. Bureau of the Census in 1976, and has since been enhanced by it and NTS Research Corporation of Durham, North Carolina. The editing technique is similar to the CONCOR-A version discussed earlier in this report, while the user language of CONCOR-C is more powerful and more extensive than the A version.

Structural editing procedures permit some errors to be identified, and as in the case of the CONCOR-A version, there is no method available to correct these errors within the system.

A. Types of errors

CONCOR-C has special functions available for both intra-record and inter-record checks. The relationship of variables to one another may be tested in simple or complex ways. Generally, every variable of a given questionnaire may be available for examination at any time during the editing process.

1. Range tests

There is a range command that performs a validity test for quantitative variables and code tests of qualitative variables. When an out-of-range condition is detected, several options are available to the user. Unless suppressed, error conditions will generate error counts and error messages automatically. The error conditions may be flagged for correction later or may be assigned an optional code such as "Not Reported".

2. Logical and arithmetical relationships

Variables within the same record may be checked easily for errors in consistency, and checks may be made across logical records for variables within the same questionnaire.

B. Correction of errors

1. Manual correction

At present there is no method of manual correction available in CONCOR-C.

2. Automatic correction

Corrections may be made by imputations from either hot deck or cold deck arrays or through arbitrary allocations. Counts of all of the imputations are kept for later examinations and analysis.

The user must design the hot or cold deck matrices to be used and must assign the variables to be edited to the matrix from which the imputations are to be made. Matrices from two-to-five dimensions may be created in CONCOR-C

C. Documentation of errors

1. Overall error tatistics

The overall error statistics procedure provided by COBOL CONCOR is very much like that provided by CONCOR-A. At the end of the editing execution, the system will provide the following statistics:

— Absolute frequency of error occurrences by variables.
— Absolute frequency of errors by consistency rule.
— Number of times each variable was imputed (at the present time, the frequency distribution of imputed codes by variable is not working).
— Total number of questionnaires in error.
— Total number of questionnaires read.

2. Detailed error statistics

Optionally, the user may obtain detailed statistics that can be used for analysis of the correction procedures which were performed automatically. In this case, the information the user receives from the system consists of:

— Identification of the questionnaire in error.
— Identification of the variables involved in the failed consistency rule, plus their respective values.
— Literal error messages that help to identify the problem.

3. User control of the degree of details

The user may optionally obtain from the system:

— Error statistics by variable.
— Error statistics by Edit-Specification.
— Error statistics at the questionnaire level.

4. User control of area breakdown of error statistics

The user has complete control in specifying the area level the error statistics will be produced.

5. Automatic generation of error file

CONCOR-C automatically generates an error file containing all of the information needed to produce the error statistics mentioned above.

D. Input files

CONCOR-C can process fixed length hierarchical files. The file may be composed of up to 50 different record types, and the hierarchy may be defined by

(1) A set of variables that usually represents different geographical areas.

(2) A questionnaire identification that must be unique at the smallest level or area.

(3) A record type identification.

(4) The variable names within each record type.

The variables contained in the input file may be numeric, alpha numeric, halfword binary, fullword binary or doubleword binary coded data.

E. Output file

CONCOR-C produces an edited output file identical in format to the unedited input file which allows the output to be resubmitted as input to re-edit the file. This second reading will show if any new inconsistencies were introduced during the initial edit process. Optionally, the output file may be transformed into a flat file and newly created variables may be written out as part of the new file.

F. General aspects

1. Portability

The system is written in COBOL and requires a minimum of 128K bytes of main memory and at least four million characters of random access disk storage. By 1983, there were versions of CONCOR-C available on IBM, Honeywell 66, ICL 2900, UNIVAC 1100, NEC 500, WANG VS and Perkin-Elmer 3200 Systems.

The implementation of the system is rather difficult because the number of programs (19 modules) that make up the system, and the large number of files these programs handle.

2. Reliability

The current version of CONCOR-C has been installed and used in a number of countries. Extensive testing by the distributor and others has been carried out and all known error conditions have been corrected. The output of CONCOR-C should be carefully reviewed by users to insure that the results are consistent with the intended editing procedures desired.

3. Language characteristics

CONCOR-C is a high level language oriented to the editing of statistical data. As such, it contains some commands similar to commands one is likely to find in other general purpose high level languages, plus the specific commands that make CONCOR-C an editing oriented language.

The CONCOR-C language is very close to the English language so it is not too difficult to learn, but at the same time, it is not so easy to use that any subject matter person could write an edit program. As with CONCOR-A, it contains a basic set of commands that could be easily used by demographers and other social scientists when used to detect errors. When automatic correction is desired, more sophisticated commands and techniques are required to properly apply CONCOR-C to the editing procedure.

The language has numerous special features to help make the editing task easier. For instance, there are some internal CONCOR counters that are automatically defined by the system and available to the user. Examples of these are:

(a) Record type counter, which is in fact, a counter for each different record type giving the number of occurrences.

(b) Invalid record type counter, which gives the number of invalid record types in a particular questionnaire.

(c) Questionnaire counter, which gives at any time, the number of questionnaires that has been read in.

4. Performance

The run time required to process a national census file will be dependent on three principal factors.

(a) The size of the file.

(b) The processing power of the computer to be used.

(c) The complexity of the editing procedures to be used.

Within these constraints CONCOR-C should be efficient enough in most instances to process a national census of population. The 1983 version is significantly faster than the February 1980 version of CONCOR-C, and appears to be adequate for use in many countries.

G. Summary comments

CONCOR-C shows promise of being useful in census and survey projects in developing countries. The user language is very powerful, but the coding required to edit complex relationships can be rather complicated. For example, in a census of population and housing that uses multiple "hot decks", several thousand lines of coding may be necessary to use CONCOR-C. This can have two distinct disadvantages:

1. The coding skill needed to perform accurate editing procedures requires very competent programmers.

2. The more instructions that are required, the less efficient the system will be in terms of computer line.

These disadvantages may have minimal impact on sample surveys where "hot decks" may not be used and where the size of the file may only be in the hundreds or thousands.

As noted elsewhere in this report, all of the other editing packages have similar constraints or disadvantages.

CAN-EDIT does not have the capability of performing structural editing, and does not attempt to address this type of editing.

A. Types of errors

This system does not make any distinction between intra-record checks and inter-record checks, treating the questionnaire as if it were composed of a single large record. This restriction, along with some other features of CAN-EDIT, severely limits the system and this will be addressed later in this report under Input/Output constraints.

1. Range Test

When the user defines the Data Dictionary, he also defines the valid set of codes for each variable, specifying a category name and its corresponding code. The association between a category name and its corresponding code will allow the user later, in the edit specifications, to refer to a particular event by the category name rather than the code itself. (Suppose for example, that code 2 of the variable "Relation-to-head" means "Spouse"; then, the user can refer to the event, "Relation-to-head = Spouse").

As it can be seen, this method of defining the range test is appropriate for qualitative variables but it is not appropriate for quantitative variables. In fact, CAN-EDIT cannot handle quantitative variables, and if quantitative variables are encountered, the user has to stratify them in groups (by means of a recoding operation) so that it ends up looking like any other qualitative variable. Whenever an invalid code is detected, the variable in error will be flagged for a later correction.

2. Arithmetic relationships

Since quantitative variables cannot be handled by the system, arithmetic relationships are not allowed.

3. Logical relationships

This is undoubtedly the strongest of CAN-EDIT's features. A statement composed of a single set of conditions is entered which defines an impossible combination of codes in different

variables which, in the CAN-EDIT vocabulary is called a "conflict rule". Whenever that set of conditions is met, an error has been detected and the conflict rule is flagged for later identification of the error. The system contains a module called "Edit Rules Analyzer" whose functions are:

(a) Identification of possible edit rules that are in contradiction with each other.

(b) Identification of redundant edit rules that will not help to detect errors other than those that are detected by the remaining edit rules.

(c) The derivation of implied edits which are error conditions not specified by the user but which are implied through logical deductions based on the edit statements the user provides.

The functions provided by the Edit Rules Analyzer can be extremely helpful when dealing with complicated surveys where there are likely to be a large number of edit rules and complex inter-relationships.

B. Correction of errors

1. Manual correction of errors

The system has no provision for manual correction.

2. Automatic correction

(a) Automatic identification of variables in error

Once the system has applied all the user edit rules to a particular questionnaire, the system uses an algorithm to identify what is called, "the minimum set of fields for imputation" which means, the identification of the minimum number of variables that, once corrected will allow the questionnaire to pass all the edit rules. This procedure is completely transparent to the user, and behind it, is a complete statistical theory developed by Fellegi and Holt.[1]

(b) Manipulation of matrices

In the case of CAN-EDIT there is no need for matrix manipulation since the identification of variables in error as well as imputation is fully automatic.

[1] I.P. Fellegi and D. Holt, "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association, Mar. 1976, Vol. 71, 353-17.

(c) Method of Imputation

Having determined the fields to be changed, the system provides a fully automatic imputation method based on the hot-deck approach using one of the following two alternatives:

(i) First Preference — "Primary Imputation" in CAN-EDIT. All selected variables are imputed simultaneously by transcribing codes to the variables in error from a single donor record.

(ii) Second Preference — "Secondary Imputation" in CAN-EDIT — Imputations are performed at a predefined sequence proceeding one variable at a time, using a series of donor records. The The variables that define the characteristics of the donor record(s) are automatically derived from the user edit rules specification. In addition the user can explicitly specify auxiliary constraints for variables that are correlated but not linked by a logical relationship.

When no donor record is found that satisfies the selection conditions then the selection criteria is replaced by removal of the auxiliary constraints. If no donor record is found after relocation of the auxiliary constraints, the system passes to secondary imputation. If a suitable donor cannot be found in the secondary imputation alternative, the system will arbitrarily assign a possible solution. Auxiliary constraints are not supported in the second imputation alternative.

C. Documentation of errors

Overall error statistics — CAN-EDIT can produce the following overall statistics:

(i) Number of records that failed at least one edit rule, and number of records that passed all edit rules.

(ii) Number of times each variable was imputed because it failed a range test, and the same for those that failed an edit rule.

(iii) Number of records that were corrected using the primary imputation approach, and number of records that were corrected using the secondary imputation method.

(iv) The same as (iii) but by variable instead of by record; the number of variables corrected by the secondary imputation method, is broken down in:

— Number of times the variable had one and only one possible solution.

— Number of times the variable was imputed using the hot deck approach.

— Number of times the variable was arbitrarily imputed.

CAN-EDIT does not produce a frequency distribution of imputed codes by variable.

D. Input file

One of CAN-EDIT's limitations is the inability to handle hierarchical files. Since most of the statistical data - both census or survey data - are structured as hierarchical files, the user needs to convert them into flat files. This operation even if it is not too difficult, could be time consuming and expensive in terms of computer time. In the case of a national census, the user will probably need to split the input file into several files in such a way that it processes first the one person household as one file, then the two person household as another file, etc.

The other severe limitation in the system is that it only handles numeric variables. Alpha-numeric variables are not accepted.

The input file must reside on disk and it can only be accessed by means of a custom designed data base management system called RAPID that has been developed by Statistics Canada.

E. Output file

Since the input file resides on a disk, changes can be made only on the same file. Therefore, the input file ends up being the error free output file.

F. General Aspects

1. Portability

The system can run on IBM or IBM compatible computers with 250 K bytes of memory and direct access storage. The documentation of the system is good and there are a number of papers that relate to CAN-EDIT's methodological basis or to the system itself. Nevertheless, CAN-EDIT was not conceived as an exportable software, but rather to meet the requirements of Statistics Canada, and using of their significant computer resources which are far greater than the resources that are likely to be found in LDC's. Furthermore, CAN-EDIT is an element within what in Statistic Canada is known as, the "Generalized Survey Processing System", which is an intergration of several sub-system working under a Data Base environment. Under these circumstances, very few installations in LDC's, if any, can make use of the system.

2. Reliability

CAN-EDIT has been successfully used in different applications such as the 1976 Population Census of Canada. Therefore, CAN-EDIT can be considered a reliable system.

3. Language characteristics

CAN-EDIT is a generalized system for the editing of qualitative data. It provides a so-called "set language" for the user in which he specifies the editing rules. This language is easy to learn by subject matter people as well as by programmers. In general, there is very little work for the user to do, and it mainly consists of the specifications of the Data Dictionary and the proper edit rules. From this point of view, CAN-EDIT is undoubtedly the most sophisticated and powerful available editing system: a simple language, little work left to the user, and an error-free output file.

4. Performance

Part of the price the user has to pay for having a system that does most of the work for him, is a very low speed in executing the editing specifications. Unfortunately, the speed is so slow that is is almost impossible to implement CAN-EDIT on medium-size computers, much less on small computers. As an example, in the processing of Canada's 1976 population census, using an AMDAHL 470 computer (about 30 percent faster than the IBM 370 - 168), it took about 250 hours of CPU time to process 23 million records that contained a very short questionnaire (nine questions).

G. Summary Comments

From a theoretical point of view, CAN-EDIT is very close to the ideal editing system. It has several features that make it the most desirable of the existing software packages in this field. Although there are some serious limitations that make it unusable in developing countries, the great merit of CAN-EDIT is the development of its methodological basis which is a unique and important contribution to the state of the art.

UNEDIT

Generalized Editing Software Package For Census and Survey Data Processing

The UNEDIT system is a portable generalized software package developed to meet common editing needs in census and survey data processing. The UNEDIT package is written in RPG II language and can be run on small computers with a minimum internal memory of 32K bytes.

UNEDIT provides a limited capability to perform structural editing. Under some conditions, it is possible to detect missing records. However, when there is a duplicated record and a corresponding missing record, the total number of records will match with the total shown on a higher level record, and therefore, the error will not be detected.

There is no way to detect duplicated records, nor is there a way to create new records or delete duplicated records.

A. Types of errors

1. Intra-record checks

UNEDIT has an easy way to sepcify the set of valid codes or acceptable ranges. When an invalid code or value is detected, the system will automatically print out an error message. In addition, the user has the capability of assigning a specific code or value to the variable in error.

Logical as well as arithmetic relationships can be checked, although some restrictions should be noted. (a) A relationship in UNEDIT is composed of a set of conditions which define an unacceptable combination of codes or values in two or more data fields in the input record. Each condition consists of a variable name, an arithmetic relationship operator (such as greater than, less than, etc.) and a code or value of the variable. Therefore, a condition cannot involve two variables. (b) Each condition is implicitly joined to the next condition (if it exists) by an "AND" logical operation. "OR" and "NOT" logical operators cannot be used.

2. Inter-record checks

The system does not provide the full inter-record checks feature in the "edit specification part" of the program. However, there is a portion of the program called "pre-edit" where the user can define variables of different record types as new

variables, and then be allowed to use the newly created variables in the edit specification. Even though it is somewhat tedious and time consuming, the user does have the ability to perform a full inter-record check.

B. Correction of errors

1. Mannual correction of errors

The system has no provision for manual correction.

2. Automatic correction

(a) Automatic identification of variables in error:

UNEDIT does not provide this feature.

(b) Manipulation of matrices:

Not allowed.

(c) Method of imputation:

The system doesn't have any method for imputation other than assigning a specific code or value to variables.

C. Documentation of errors

1. Overall error statistics

UNEDIT provides, at the end of each area, the following statistics:

(a) Total number of input records

(b) Total number of rejected records

(c) Number of rejected records having invalid codes

(d) Number of records having inconsistent codes.

In addition, at the end of file, it provides the number of invalid codes by variable, and the number of inconsistent codes by "inconsistency-condition-number".

2. Error messages

UNEDIT will print out, for each individual rejected record the following:

(a) Invalid code messages giving the variable name, the field position and the invalid code

(b) Inconsistency messages, giving the inconsistency-condition-number.

3. User control of the degree of detail in documenting error

Not available.

4. User control of area breakdown

When defining the areas under which the household units are numbered, the user defines also the area breakdown for error statistics. The user has the choice of defining from 1 up to 3 different areas.

5. Automatic generation of error file

Not available.

D. Input files

UNEDIT can process hierarchical files of up to 9 different record types. Each variable is identified by: (a) A variable name from 1 up to 4 characters. (b) A record type identification if the input file is a multi-record type file. (c) Location of the field in the input record - 1 to 99 -. (d) Length of the field - 1 to 4 positions - if there are multiple occurrences of one specific record type, an occurrence identifier needs to be specified.

The hierarchy is based on a questionnaire identification - - 1 to 3 area codes and a household number - and a record type identification.

Two different types of input variables are accepted: Numeric variables (zoned decimal) and Alpha-numeric. Packed decimal or binary variables are not accepted.

E. Output file

UNEDIT can automatically generate, as an option, an output file which will have the same structure as the input file.

F. General aspects

1. Portability

UNEDIT was written in RPG II and was designed to run on small minicomputers with a minimum of 32K bytes of memory, and can be installed on many small computers that have RPG II compilers. The documentation of the package is limited and it could be considerably improved with some complete examples of the coding along with some brief explanations of some of the instructions. Given the simplicity of the package, it probably does not need a programmer's guide and installation manual.

2. Reliability

According to the developers of UNEDIT, it has been successfully applied to census files in several countries and fully tested in the office. Under these circumstances, they rate the system as very reliable.

3. Language characteristics

The way the user presents the edit specification to the system is by means of a list of positional parameters coded on fixed fields. This presents some advantages and some disadvantages. Most of the edit specifications can be easily written by subject matter people, with little knowledge of computer programming. However, once the program is written, the specifications are not self-evident. Probably the most important feature of UNEDIT is the analysis the system makes of edit specifications in order to detect redundancies and contradictions. Notwithstanding these capabilities, the package contains virtually no special editing features other than the basic editing commands.

4. Performance

At the present time there are no available estimates on which to base the expected run time for UNEDIT. The original implementation was accomplished on the IBM System 34 and this system is not likely to be used except in very small countries, so that the system efficiency will not be a major consideration in its use.

G. Summary comments about UNEDIT

The UNEDIT features and capabilities seem to give the impression that the system needs to be further enhanced to be useful for a national census. It does provide the features for detecting errors in the basic information, and a reasonable documentation of the errors detected. It does not provide, however, a way to correct those errors either by hand or automatically, unless the user wants to always assign a specific code to a variable in error.

The package is very slow when actually executing, therefore it may not be suitable for processing large census files. On the other hand, when dealing with small files and small computers, it could considerably reduce the amount of programmers time needed to write a custom edit program.

PART III

CONTROL

CENSUS MANAGEMENT SYSTEM

The duration of a national census activity may take 2 or 3 years from the time of the enumeration until all processing is completed. Even in small countries, there will be millions of records to be processed through various phases until the final tables are published and analysis begun. In past censuses, most countries established a manual system to keep records of this very important aspect of census processing. In recent years, efforts have been made to automate the record keeping with the use of computers.

By 1982, a generalized management reporting system called CONTROL had been installed and used to monitor the processing activities in two Latin American countries. The CONTROL system was designed by Delta Systems Consultants, Inc. in 1978, and first implemented in early 1981.

CONTROL is designed to complement a data editing package such as CONCOR so that certain types of errors that cannot be uncovered by one system can be found and corrected by the other.

A. Methodology

CONTROL is dependent upon a good cartographic system that accurately shows various geographic and administrative boundaries and structures. Then, a hierarchical coding system needs to be developed which will be incorporated into a master file or master data base. The data base can be considered as a tree whose root is the country and whose final level (nodes) are the smallest geographic areas, the enumeration area. Between the root and the last level nodes there will be as many levels as there are geographic subdivisions within a given country, and as many nodes as elements that each subdivision contains. Each tree node represents a record of the data base information of all its children or branches. In particular, the root will contain summary information at the country level.

In order to facilitate the explanations, assume that a given country is divided in regions; each region in turn is subdivided in provinces and each province in districts; finally the districts are subdivided into enumeration areas.

Starting from the base that the geographic/administrative structure of the country has previously been defined, and assuming there will be no alterations, part of the data base that remains to be defined is the last level of the tree or the enumeration areas for each district of the country. This information can be given by the cartographic section in anticipation of the census event.

Usually, the statistical office has an estimate of the number of housing units for each enumeration area and sometimes these estimations will include the expected population by enumeration area. With these information elements, one could proceed to create a data base that in the future will be the nucleus of the control system.

As can be seen, the control unit or follow-up unit of the entire operation will be the enumeration area. Its size can vary, from between ten to one hundred units, depending on the type of census and the characteristics of a specific country. Obviously, the smaller the enumeration area, the more precise the control will be; but at the same time, the greater the volume of work required by the system. Once the data base has been created, the system will be ready to perform the following tasks:

(1) To calculate for each of the intermediate and higher levels (districts, provinces, regions and country), the expected totals of the number of housing units and persons.

(2) To print-out forms and other types of materials having the geographic identification and bearing a name to keep separate each enumeration area unit. These forms will make it easier for the user to communicate with the system, to be able to update the data base.

(3) To facilitate the distribution and control of materials such as questionnaires and other items.

(4) When the information for the enumeration areas is returned to the central office along with the data gathered from the field, the data base can be updated with the following important information elements:

   (a) The pre-printed form for each enumeration area will be entered into the system to show the units that have been returned from the field.

   (b) Into the same form can be added the actual size of the unit giving the number of housing units and population counted. Such information can be written down by the same interviewer and verified by the chief of each different section.

(5) Based on the above information, the system will be able to produce two very important reports. These are:

(a) The coverage report, showing at the different levels, district, province, and region, in absolute and relative numbers and the total number of enumeration areas returned from the field. It will also be possible to obtain for each district, the identification of the missing areas.

(b) The report on the preliminary figures derived from the manual count. This report will show the number of housing units, population by sex, average number of persons per household, and the masculinity index, at the district, province, region, and country level.

(6) Later, when the information gathered from the field has been entered into magnetic media or punch card, CONTROL will create the master file of the census. The system will automatically update the data base, entering each enumeration area accepted and added to the master file. Before the unit is accepted by the system, three conditions will have to be satisfied. These are:

(a) The geographic identification codes as well as the number of the enumeration area must be contained in the data base.

(b) The number of housing units and the population counted by the system at the time of inputing the census data should not differ from those previously entered to the system for the hand count in a percentage greater than given to the system by the user at the time of the system implementation.

(c) The flag for the enumeration area added to the system should be turned off showing that a particular enumeration area has not been previously added to the data base or the master file.

This way, the system ensures that the information gathered in the field agrees or matches with the cartographic system; or if not, the error is detected and identified. In addition, it assures that the CONTROL unit is added to the census master file once and only once. Otherwise, the error is detected and identified.

B. CONTROL as a management information system

Another important system objective of the system is to keep the director of the statistical office and the senior members of the staff that have the responsibility for the census operation, up-to-date regarding the status of the census operation. Various section chiefs will have access to some of the reports that will allow them to regularly monitor the census and to see the advances in the actual processing of the census data. That is, observing the data processing through all of the different phases which the data passes beginning when the questionnaires are stored in the main central office until they are added to the census master file. In order for the system to achieve this objective, it is necessary to ensure that the data base is permanently updated following the progression of the various enumeration areas. This can be achieved by establishing an orderly and rigid system.

(1) It is necessary to define a fixed number of phases through which the data gathered in the field should pass before being added to the permanent census master file. The first and last phase should always be "the reception of data in the main office", and "the addition of data to the master file" respectively. Between these two phases the user can define up to five different phases. For instance, "Coding, Manual Editing", "Data Entry", "Data Verification", and "Conversion and Consolidation of data".

(2) When the data base file containing all of the geographic areas with the appropriate codes has been established and sorted so that the hierarchical structure is correct, the system will create a Master Transmittal Procedure form. This form will be attached to the folio (or book or binder) that will hold all of the questionnaires for the enumeration area. The form will contain the geographic identification, the enumeration area number, the phase, and the code that identifies the particular phase. Everytime a unit passes through one phase to the next one, the form is updated with the appropriate information and submitted to the data entry to be entered to the system updating the data base.

(3) The cycle will be closed when the reports are produced by the system either periodically or at the special request by the user. The reports produced by CONTROL can be classified in two groups depending on the objectives:

(a) Summary reports designed to keep the director up-to-date regarding the census operation are the first type. The system will produce four reports that will keep the management up-to-date regarding the census operation. These reports also show the global productivity of each section in such a way that possible bottlenecks can be identified and corrected in the various phases. Following is a brief description of the reports:

(1) The number of enumeration areas already processed by geographic area and processing phase.

The user can choose the publication level (region, province, district) and within the chosen level, select specific geographic areas. The report will show for each of the selected areas, the absolute and relative number of enumeration areas that have been processed and the date of the report for each of the defined phases. The report will include the total units already processed at the country level for each phase, along with a graphic display showing the progress of the census operation.

(2) The number of enumeration areas and population already processed by month and phase.

This report measures the overall productivity of each section or phase of data processing for each calendar month following the census. The productivity is measured in number of enumeration areas processed by each section, and the corresponding population in these areas. The report will show the information flow from one section to another, and as a result, it will forecast the workload of each section that will produce a normal flow in the remaining sections.

(3) The global daily productivity report

This report is similar to one described in (3)(a)(2) except that instead of classifying the enumeration areas by geographic area, they

are classified by day in which they were processed in each section. The report will be produced occasionally on special request from the user, whenever the detail of the global productivity of each section is desired. It will be more useful at the beginning of the processing phase or when the previously defined reports identify that some type of problem exists in one or more of the various sections of phases of the census operation.

(4) The preliminary census figures by geographic area.

Even though this report will be produced only once (in theory), it has been included in this section because the figures produced will be used as a basis for the entire operation of CONTROL as well as for information directed to the census management. The report with the figures obtained from the hand counts at each enumeration area level, and transmitted to the data base at the time of first entry will at the same time show that the unit has been received in the central office.

(b) The reports designed for the systems analysts and programmers in charge of the census data processing. These reports are in greater detail than the ones previously mentioned and generally they will be used as a base for manual verification or as consultation documents.

(1) The number of enumeration areas by geographic area.

Once the data base has been created, the system will produce a listing containing all names, codes and number of enumeration areas by geographic area. The main objective of the listing is to allow a manual verification of the data base and to serve as a reference document.

(2) Missing enumeration areas by phase and geographic area.

A report will be printed to show the identification of all enumeration areas that have not yet passed through a particular phase of the processing of the data.

This report will only be produced when a geographic area has been finished (for example, district) and the report mentioned in (3)(a) shows that 100 percent of the enumeration areas have not been processed for that specific geographic area.

(3) The daily productivity report.

This report will show for each phase the number of enumeration areas processed by date.

This report will be more useful in the very beginning of the census operation, when defining workloads for each section and for establishing production rates to meet the normal flow of information during the different parts of the phases.

(4) The enumeration areas added to the census master file.

The system will produce a report every week containing the complete identification of enumeration areas and the number of housing and population records that have been added to the census master file. At the same time, it will show all of those enumeration areas that were rejected along with the proper error messages identifying the various errors.

(5) Other utility reports.

The system can produce a variety of other utility listings such as a data base dump by enumeration area, etc.

C. Other benefits of the data base

As the data base grows and becomes fully loaded with all the census data, the resulting product can be a rich source of information for other purposes. Some of the most important uses are:

1. Sample frame and sample design

The data base contains information elements such as number of housing units and population by sex. At the users choice, the population could be stored by groups of age at the lowest level such as the enumeration area. In turn each higher level (district, province, region and country), will contain the same information elements, summarized. Under the same circumstances, with the easy access to any given element of the data base, it will greatly facilitate the sampling design for future studies.

In time, the data base can be made more valuable by adding new elements of information at the enumeration area level such as, charateristics of the area (urban, semi-urban, rural), socio-economic indexes, etc.

2. The data base as an information source for urban planning, commerce, industry and services

Along with the development of the data base with additional information, the potential interest for different uses and applications will grow.

Obviously, the scope of the data base will only be limited by the capacity of the disk storage available in a given country

Having the information summarized at the enumeration area level allows:

(a) A considerable reduction in the time required for th data processing will make possible research studies that would otherwise be too expensive to finance without acces to the census master file.

(b) The re-grouping of enumeration areas, creating a sub universe that duplicates an administrative or geographic area.

Finally, when the areas of interest have been defined, th data base will allow the user to access the file for thos and only those areas of interest desired, representing a considerable saving in time and computer costs.

STATISTICAL ANALYSIS SECTION

INTRODUCTION

For more than ten years, there have been organized efforts to classify and evaluate statistical software packages. In 1977, 44 software developers participated in the first in a series of "self-evaluation" exercises. These results were cataloged and documented by Professor Ivor Francis at Cornell University and presented at the founding meeting of the International Association for Statistical Computing in New Delhi, India, in December, 1977. By refining the techniques used, subsequent evaluations were carried out in 1979 and 1981.

Other interested groups have also performed evaluation exercises to compare the capabilities of various software packages. However, the prospect of a truly independently-conducted research project that would evaluate the many packages now available is unlikely. The cost and time required to measure the different attributes of the packages probably would outweigh the value of the report.

In several instances, the computer availability is the principal factor in selection of software, because not all packages will run on all computers, or if they run, may not have all of the capabilities that are available on other versions.

None of the packages is static: all are being updated and improved regularly, so that later versions of the software may address shortcomings that were present in 1982/83 when this report was prepared. The pattern of new releases varies by developer and in some cases by the make and model of computer systems.

Developers of statistical software usually identify their products in one of five different categories that are based on the following tasks:

1. Data management and file building
2. Editing: error detection, correction and imputation
3. Data description, tabulation and plotting
4. Estimation of finite population parameters for complex surveys
5. Statistical analysis and model building

Not all packages fit precisely into these groupings, but most do. Those that perform several or all of these tasks generally can be identified as having a predominant capability that determines the grouping.

In addition to being capable of performing one or more of these five tasks, the usefulness of any software product can be measured by other factors. These other factors include:

1. Portability
2. Ease of learning and using
3. Reliability
4. Cost

These factors will be addressed in this report. The format used will be to examine the capabilities of each package; list the requirements of the operational environment; and to provide information about the cost of acquiring the system.

Ratings have been developed and are included as part of the Report Supplement. Ivor Francis[1] and his colleague have gathered the raw data on which the ratings have been based. The ratings were obtained by Professor Francis from three sources:

1. The developer of the software
2. Users of the software recommended by the developer
3. Users selected independently

All of those included were asked to rate the characteristics of the various packages on a four-point ordinal scale representing general usefulness of the program as far as each particular characteristic was concerned. The coding scheme was "0" through "3", with "3" being high and "0" being low.

In his compedium, Francis has presented the detailed scores of the three sets of ratings for twenty major characteristics in seven categories. For this report, the ratings have been shown for five categories, and the numerical ratings have been converted to descriptive rankings which are:

very good
good
modest
poor

[1]Francis, Ivor, Statistical Software: A Comparative Review, Elsevier North Holland, Inc., N.Y.

The five categories illustrated in this report (See Figure 2 on page 78) and in the supplement to the report are:

filing	accessibility
tabulation	usability
statistical analysis	

These are based on the following capabilities:

FILING

Data set size, flexibility of data input, complex data structures, missing data, storage and retrieval, file manipulation, flexibility of output.

TABULATION

Computing tables, printing tables.

STATISTICAL ANALYSIS

Multiple regression, analysis of variance, linear models, multivariate analysis, non-parametric methods, non-linear methods.

ACCESSIBILITY

Availability, number of installations, number of computer makes, core requirements, batch or interactive, maintenance.

USABILITY

Training needed in statistics and computing, language simplicity, documentation, user conveniences.

The ratings shown in the ATTRIBUTES section of the report are based on the responses to specific questionnaire items that have been grouped into five categories. The category of usability, for example actually contained five separate items in questionnaire, so that the rating provided is based on three separate groups of raters (developers, users suggested by the developers, and independent users) and on responses to five distinct items.

It should be noted that the users score can be viewed as the usefulness of this attribute of a program from the user's perspective, and that may be different from the developers intention.

It should also be noted that there may be a bias on the part of some of the developers to rate higher than warranted. This should be expected because of the following situations:

1. Even though the program may indeed be perfect when used by the developer for the purpose for which it was designed, nevertheless when the program is transplanted to a different environment under different conditions — and for the most part in this book we are looking at programs which are being used by people other than the developer — the program may not be as useful to others as it is to the developer.

2. Because of his complete familiarity with his program and his knowledge of its undocumented eccentricities, he may not be aware of its shortcomings as others see it and therefore unintentionally over-rate its usefulness to others.

3. A developer, possibly with some commercial motives, may intentionally over-rate his program.

Notwithstanding all of these conditions, the ratings provided may enable potential users as well as current users to identify those software packages that are or may be useful in terms of statistical and portability characteristics and to compare various packages using these very broadly defined capabilities.

A wide variety of sources was used to gather information presented in this report including the user documentation for all of the packages included. Other reports and publications are listed in the bibliography.

Major Statistical Analysis Software Packages Included in this Inventory Report

Of the many statistical software packages now available, eight have been included in this report. These packages are widely used and accepted; time tested; and well documented.

The eight packages are listed alphabetically by name, showing the distributor and the cost of each package. (See Figure 1 below):

NAME	DISTRIBUTOR	COST*
BMDP	BMDP Statistical Software 1964 Westwood Blvd., #202 Los Angeles, California 90025	$2,250 33-1/3% discount government 66-2/3% discount academic
EASYTRIEVE	Pansophic Systems, Inc. 709 Enterprise Drive Oak Brook, Illinois 60521	One time: DOS $14,500 OS $19,000
MINITAB	MINITAB Project 215 Pond Laboratory University Park, Pa. 16802	$1,000 per year 50% discount for academic installations
OSIRIS IV	Institute for Social Research Survey Research Center University of Michigan P. O. Box 1248 Ann Arbor, Michigan 48106	$2,400 first year $1,800 renewal year 25% discount for government and non-profit; 50% discount for academic users
P-STAT	P-STAT, Inc. P. O. Box 285 Princeton, New Jersey 08540	$8,000 first year $3,000 renewal year $1,000 university
SAS	SAS Institute, Inc. Box 8000, SAS Circle Cary, North Carolina 27511	Commercial: $7,500 first year $2,800 renewal Government: $4,000 first year $2,000 renewal Academic: $1,500 first year $ 750 renewal
SIR	SIR, Inc. P. O. Box 1484 Evanston, Illinois 60204	Commercial: $18,000 first year $ 5,000 renewal Government: $12,000 first year $ 5,000 renewal Academic: $ 4,000 first year $ 2,300 renewal
SPSS	SPSS, Inc. 444 N. Michigan Avenue Chicago, Illinois 60611	Commercial: $7,000 first year $4,000 renewal Government: $5,500 first year $3,500 renewal Academic: $1,500 first year $1,250 renewal

Costs may vary in some countries. **FIGURE 1**

As a way of showing comparisons with the other packages included, a chart has been prepared showing the ratings for five separate categories. (See Figure 2 below). The INTRODUCTION Section described each of these categories and how the ratings were derived. Each of the packages is discussed separately in this section.

NAME	ATTRIBUTES					COMMENTS
	Filing	Tabulation	Statistical Analysis	Accessibility	Usability	The ratings, which are discussed in the report, are based on developers' and users' experiences mainly in developed countries. These are measured by Professor Ivor Francis' detailed study "Statistical Software, A Comparative Review," Elsevier North Holland, Inc. Publishers, 1981. The ratings are based on summary indices that reflect the following capabilities and characteristics:
BMDP	Modest	Poor	Very Good	Very Good	Good	
EASYTRIEVE	Very Good	–	–	Good	Good	
MINITAB	Modest	Modest	Good	Very Good	Very Good	
OSIRIS IV	Very Good	Modest	Good	Modest	Modest	**Filing** Data set size, flexibility of data input, complex data structures, missing data, storage and retrieval, file manipulation, flexibility of output.
P-STAT	Very Good	Good	Good	Very Good	Good	**Tabulation** Computing tables, printing tables.
SAS	Very Good	Very Good	Very Good	Good	Very Good	**Statistical Analysis** Multiple regression, analysis of variance, linear models, multivariate analysis, non-parametric methods, non-linear methods.
SIR	Very Good	–	–	Very Good	Very Good	**Accessibility** Availability, number of installations, number of computer makes, core requirements, batch or interactive, maintenance.
SPSS	Good	Good	Good	Very Good	Good	**Usability** Training needed in statistics and computing, language simplicity, documentation, user conveniences.

FIGURE 2

A. Introduction

BMDP is probably the oldest of the major statistical software packages. The initial research was begun in 1957 and a large number of modifications and enhancements have been offered over the years. The original users of the older BMD series were intended to be researchers in the health sciences. However, the programs are now used for a wide variety of statistical applications.

Univariate descriptive statistics have always been an integral part of each BMD and BMDP program. The older programs generally restrict themselves to means and standard deviations. The BMDP programs also include such things as coefficients of variation, minimum and maximum values, skewness and kurtosis. The most complete program for univariate statistics is BMDP2D, which includes for each variable the mean, standard deviation, maximum, minimum, range, half interquartile range, median, mode, standard error of mean, frequency of each value found, relative frequencies, cumulative frequency, skewness and kurtosis and their standard errors, three robust location estimates, and a histogram of sufficient size to visually assess the shape of the distribution and detect outliers.

Over the years, many programs have been superseded by newer programs. The 1961 manual included three programs (no longer distributed) for various stages of factor analysis. These programs were replaced in 1964 by BMD03D, which was soon superseded in 1965 by BMDX72. The most recent program for factor analysis is BMDP4M, which was first distributed in 1973. In addition, BMDP4M contains numerous notes, which appear as part of the computer output, to help the user interpret his results. The description of the use of BMDP4M is supplemented by an extensively annotated computer output.

B. Capabilities

The BMDP series is a collection of individual programs. The programs communicate with each other by way of BMDP Save Files. Separate job control language is used for each program.

1. Command language

The BMDP control language consists of sentences such as:

VARIABLES ARE 15

which states that there are 15 variables. These sentences are grouped into paragraphs, each of which has a name and is separated from other paragraphs by a slash mark (/). A collection of paragraphs is called a batch.

The control language is neither card nor column oriented. The paragraphs may come in any order except for the END paragraph. Specification of input parameters to the BMDP programs is facilitated by assumed options. These assumptions are chosen to provide the user with a reasonable first pass of data through a program.

For those users familiar with the older BMD programs, the input instructions for BMDP are similar. The language for the BMDP series is the same for all programs, and diagnostic messages are improved. The recent BMDP (1979) manual contains many examples of the use of the control instructions and may easily be used as a teaching aid or for self instruction.

2. Data Sets

The BMDP series allows input and output of BMDP Save Files, which include names for variables, missing value indicators, and a description of the file including whether it contains raw data, corrections, etc. No variable format is specified when BMDP Save Files are used as input since the number of variables is retained in the file itself and since the files are binary. When the correlation matrix is saved in the BMDP series, the means and variances are also saved. If the correlation matrix has been computed with special "missing data" formulas, as in BMDP8D, this fact is also recorded on the BMDP Save File and is reported whenever the matrix is used as input to a BMDP program.

Subsets of data may be listed in BMDP1D and BMDP7D; sorted data can also be listed in BMDP7D.

Frequently analyses are to be performed using subsets of data. The old BMD programs have little capacity for subsets selection. The BMDP series programs all allow subset selection by means of the "USE" transformation statement, for example,

USE = SPECIES EQ 7.

specifies that only cases whose species is seven will be used. One could also state

USE = AGE GT 25.

Many programs allow computation of statistics automatically for each of several groups. Subsets may also be selected randomly in any BMDP program.

The control language transformations are supplemented by Fortran statements via the BIMEDT procedure (as available for IBM 360 and 370 computers or by analogous methods for other computers). Any Fortran statements may be used including subroutine calls, read and write statements, etc. The BIMEDT procedure inserts the Fortran transformation statements into a subroutine which is called by the BMDP program. The BIMEDT procedure is also used to increase the amount of computer memory for arrays so there are no intrinsic limits on the number of variables that can be analyzed.

Missing data can be represented by blanks and/or a missing value code, which can be distinct for each variable. In addition, minimum and maximum limits for each variable can be defined. The treatment of cases with missing values and data beyond limits varies with the analysis: univariate procedures generally treat each variable separately, but multivariate procedures require that data be complete for all variables that are used.

Individual cases or groups of cases can be deleted permanently or temporarily. Categories can be combined by transforming codes or by associating the same category name with two or more category codes.

Operational Environment

The BMDP (November 1982 release) package is written in FORTRAN (level H) principally and can run in 160K-bytes of memory on the IBM 360/370 series with ample room for six input and output buffers when overlays and the H compiler are used. The BMDP programs allow 15,000 words for storage of arrays, which allows enough room for 100 variables in factor analysis using double precision. The amount of memory for arrays can easily be increased by the BIMEDT procedure for especially large problems. Most of the BMDP programs can also be installed with only 5,000 words for arrays and with smaller buffers so as to run in 110K bytes or less. The number of variables on a BMDP Save File is essentially unlimited; for example, the Save File could contain 400 variables and 50 of these could be selected for factor analysis without special effort.

Because the older BMD programs use fixed column input of parameters and have fewer options and less extensive output than the BMDP programs, the older programs are likely to be cheaper, assuming a correct input of parameters. However, the BMDP programs

have an English language input of parameters, so the final cost of running a BMDP program is likely to be less than any analog it may have in the BMD series.

The BMDP software has been converted to every major mainframe computer system including Japanese, German and British Systems as well. By late 1982 conversions have been completed on desktop micro-computers which have the Motorola 68,000 chip with the UNIX operating system.

D. Acquisition Information

The BMDP System is available on an annual lease from BMDP Statistical Software, 1964 Westwood Boulevard #202, Los Angeles, California 90025. The cost varies from $1,000 - $2,950 depending on type of user.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

A. Introduction

EASYTRIEVE is a widely-used and popular report preparation and data management system that can be easily learned and used. The simplified English-like command system permits the user to store and retrieve data for preformatted management and control reports. While offering a flexible range of facilities for report formatting, statistical capabilities are limited.

B. Capabilities

EASYTRIEVE can be operated in a batch mode or with the Interactive Prompting Facility (IPF) in a conversational mode on line. An unlimited number of input files may be supported, but only two files at a time are permitted at a time in the standard system.

The major functions of EASYTRIEVE deal with file input, record selection for editing and update, internal operations pertaining to data handling, and to the disposition of the output material. Data sets may be numeric, alphanumeric, packed-signed decimal, packed-unsigned decimal, or binary. SAM, ISAM or VSAM files are supported by EASYTRIEVE. The system can randomly and dynamically retrieve records from indexed data sets and data bases.

The record selection facilities of EASYTRIEVE allow users to search files and to perform logical data selection based on input statements or on calculated results. The basic system provides a full range of edit and update commands for both the batch and on-line environment. Condition Sets can be structured. Condition Sets are made up of one or more EASYTRIEVE statements that can select and process records before output. EASYTRIEVE provides the ability to delete, add, and reformat data while matching fields from a secondary file. Index searching and array processing are supported, and users can access variable location fields concurrently.

Actual operations invoke an entirely different set of actions. Extensive computational facilities are provided through user logic. Standard arithmetic operations, percentage calculations, average derivation, and almost any other user-specified calculations are included. The system-supported sort routine allows users to designate up to 10 keys for the sort program to break on. In addition, up to 9 control break levels can be established for subtotals and grand totals. All totals are calculated internally. EASYTRIEVE

can also perform special tests such as end-of-file, checking for blanks, interrogating sort status, a variety of numeric and alphabetic tests, and test-under-mask operations. Condition testing and selection functions can be specified optionally via the command language.

EASYTRIEVE output can be printed, punched, or written to a file. With the IPF option, some output can be displayed on a terminal. Any number of reports can be prepared from a single pass of the file, and with USIO, an unlimited number of input files can be processed in a single EASYTRIEVE job. The formatting facilities for data placement on reports are extensive and allow for left justification, line spacing definition, and many other controls. Automatic or alternative column headings are supported. The output processing capability can be used to provide mailing labels of any size or format through the use of a single command. Other special forms are handled as easily. Summary reports or summary output files can be generated, along with audit confirmations, pre-printed form completions, etc. Special commands allow users to produce W-2 forms[1] from the EASYTRIEVE data library. Hexadecimal prints can be defined, and users can output reports directly to a microfiche output tape. Page sizes, header information, and footing information within a report can be modified dynamically.

EASYTRIEVE also provides users with a series of utility-type operations. These operations include card-to-tape conversion and data verification; file-to-file reblocking; card, tape, and disk data set duplication; automatically formatted tape and disk prints; and subfile creation with selected records from a master file. All error messages generated by EASYTRIEVE are printed for quick analysis. EASYTRIEVE checks syntax and notifies users of any data exceptions. EASYTRIEVE consolidates all error information into a report format.

C. Operational Environment

The EASYTRIEVE System interfaces with a number of other commercial database software packages, including PANVALET, TOTAL, IDMS, IMS and DBS. The system is operational only on IBM 370 Systems and IBM plug compatible computers running under DOS/VSE, OS/VS1, OS/VS2, VM370.

[1] W-2 Forms are annual employee earnings statements

D. Acquisition Information

The EASYTRIEVE System is available from Pansophic Systems, Inc., 709 Enterprise Drive, Oak Brook, Illinois 60521. There is a one-time fee of $14,500 for the DOS version and $19,000 for the OS version.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

MINITAB II

Minitab Statistical Computing System

A. Introduction

MINITAB is an easy-to-use general purpose statistical system especially designed for students and researchers who have no experience with computers. It was designed for small to moderate size data sets which can be stored in the main momory of the computer. The MINITAB system can be useful for exploring data in the early phases of analysis, for plotting, and for regression analysis. The system may be interactive or batch and either system may be learned in a short time.

B. Capabilities: processing and displaying data

Data sets are stored in a rectangular worksheet (rows by columns) in main memory. The size of the worksheet varies with the installation, from just a few thousand numbers to several million. Several data sets can reside in the worksheet at the same time. Data sets can be subsetted (by case or variable), sorted, and concatenated; variables can be transformed, added, and deleted; values can be corrected, and recoded. Missing data are handled automatically.

MINITAB can display data in histograms, and in plots, including multiple plots on the same axes, plots with letters indicating group membership, time series plots, and probability plots.

The TABLE command forms and prints multiway tables (from 1 to 11 factors). Compact output can be produced by nesting several factors on the rows and/or columns. The cells of the table may contain one or more of the following: counts and percents, statistics such as means, medians, maxima, standard deviations, and the original data. Printing of marginal statistics, information on missing values, and the use of case weights are all options available to the user.

MINITAB II can produce statistics such as means, medians, standard deviations, and correlations; t-tests, confidence intervals, chisquare tests, and non-parametric procedures.

Regression analysis, including weighted least squares, transformations, diagnostics for multicollinearity, outliers, and high leverage points can be performed. Residuals, fitted values, and coefficients can be stored for further analysis. The following

procedures may be performed using MINITAB: Linear regression, Time series analysis, including autocorrelation, partial autocorrelation, and crosscorrelation plots, arima (Box-Jenkins) models, and regression on lagged variables. Commands for EDA, including stem-and-leaf displays, boxplots, median polish, resistant line fits, and robust smoothing of time series are available in the system.

Data may be stored in a matrix and then simple operations such as add, multiply, transpose, invert, find eigenvalues and vectors can be done. Any set of MINITAB commands can be stored for repeated execution (macros). A simple looping capability is provided. Output width can be varied. Random data can be generated from a variety of distributions. An on-line HELP facility provides general information on MINITAB as well as details of a specific MINITAB command. The MINITAB source code is very modular, therefore additional features may be added by the user.

C. Operational Environment

MINITAB II is written in ANSI FORTRAN and is currently operational on a wide variety of computer systems. This includes the following:

Amdahl (all)
Burroughs 4000, 5000, 6000, 7000
CDC (all)
Data General Eclipse under ADS
DEC 10, 20
Harris (all)
Hewlett Packard, HP 3000
Honeywell 6000, under Multics, GCOS, DTSS
IBM 360, models 40 and up
IBM 370, models 115 and up
IBM 4300
IBM 3030
} under DOS/VS, OS, VS, VM, TSO, CMS
Itel
PDP-11 11/03 and up under RSTS, RSXII, RTII, IAS
PRIME 150, 250, 350 and up
Tandem 16
Texas Instrument TI 990
UNIVAC 70, 80, 90 and 1100
VAX under VMS and UNIX
XEROX Sigma

D. Acquisition Information

The MINITAB System is available on an annual lease from MINITAB Project, 215 Pond Laboratory, University Park, PA 16802. The cost is $1,000 per year, with a 50 percent discount offered to academic institutions.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

A. Introduction

OSIRIS is an Organized Set of Integrated Routines for Investigation with Statistics designed for analysis of survey data at the University of Michigan. Recent developments that permit reading of non-OSIRIS hierarchical census files make it useful in analyzing large census-type data sets. The system features data handling capabilities for editing and arithmetic transformations and a range of multivariate and non-parametric analyses. The output of OSIRIS can be input into later stages, or it can interface with other statistical systems.

B. Capabilities

OSIRIS IV was designed to serve a broad community of users and has facilities for handling data collected for a wide range of purposes. In addition to the usual basic statistics and functions, such as cross-tabulations and classical regression and correlation analysis, several special techniques are available for handling nominal- and ordinal- scale data and for calculating sampling errors for complex designs. OSIRIS IV also has a full range of well integrated data management facilities; of special interest are the ability to handle weighted data, matrix input and output, and hierarchical datasets with variable length records. Virtually any mode of data can be used directly in OSIRIS IV. Among its other capabilities are facilities for:

- Interactive setup interpretation
- Storing, retrieving, and modifying information about the structure of a dataset
- Displaying data
- Editing and correcting data
- Copying and subsetting data
- Transforming data values, through arithmetic and logical operations both within and across records
- Generating univariate and bivariate frequency distributions and related statistics
- Producing scatter plots
- Performing multiple regression analysis
- Performing univariate and multivariate analysis of variables
- Searching among predictors for the greatest variance explanatory power
- Factor-analyzing data
- Performing cluster analysis
- Conducting analyses with multiple nominal- or ordinal-scale predictor variables

Additional commands and documentation to supplement and enhance the data structures capabilities of OSIRIS IV, include: extended capabilities for adding new and derived measures to a data structure, and procedures for subsetting the data structure; a capabilitiy for establishing what elements are present in the structure and in what proportion; and special documentation describing the generation, modification, and use of hierarchical data structures.

1. Command Language

Standard and consistent parameter keywords, make OSIRIS IV easy to learn and use, and minimize the size and complexity of the required documentation. To use OSIRIS IV, the user supplies the following items, as appropriate:

- OSIRIS IV commands indicating which functions are desired and providing instructions to the OSIRIS IV monitor
- Data formatted either as an OSIRIS IV dataset or matrix
- Recode statements creating new variables or transforming existing ones
- Entry definitions indicating how groups of variables are to be assembled from a structured file
- Control statements specifying variables and parameters, optionally defining a subset of the data to be processed.

Every variable in an OSIRIS IV dataset has a number and a fixed set of attributes associated with it: attributes such as the location of the variable within each record of the data file, the variable width, type, number of decimal places and the values to be treated as missing-data. This information is stored in a dictionary file, one record per variable. Once this information has been recorded in the dictionary, it need not be respecified by the user. Variables are then referenced in OSIRIS IV by their associated variable numbers. Dictionaries may easily be created or revised with the &DICT command.

Variables may be stored in a variety of modes:

alphabetic
character numeric
floating-point binary
integer binary
packed decimal
zoned decimal

The storage mode of the variables need not be of concern except when first entering data into the OSIRIS IV system, as OSIRIS IV data management commands can handle data in any of

The modes, and analysis commands can process all modes except alphabetic. OSIRIS IV commands which create new datasets will use the most efficient mode; however, &TRANS may be used to alter the storage mode of any or all nonalphabetic variables in a dataset.

2. Data Sets

OSIRIS IV datasets have two possible configurations:

a. rectangular: all variables for one data case are stored in one record and each variable occupies the same relative location within each record:

	V1	V2	V3	V4	...
CASE 1					
2					
3					
4					
.					

b. structured: variables are collected into "groups" each with its own record length:

GROUP 1	V1	V2	V3	V4	V5	
GROUP 2	V6	V7	V8			
GROUP 3	V9	V10				
GROUP 4	V11	V12	V13	V14	V15	V16

Selected variables from different groups may be joined together to create a rectangular record for a given run.

The first step in the creation of an OSIRIS IV rectangular dataset is to build an OSIRIS IV dictionary file, usually via the &DICT command. Once the dictionary has been created the data may be read directly by OSIRIS IV without any special "file building".

A structured dataset is built from individual rectangular files via the &SBUILD command. This type of dataset is used where there is not only a relationship between variables within a rectangular dataset, but also a relationship, usually hierarchical, between the various datasets. Each rectangular dataset becomes one or more groups in the structured dataset. A simple example is one rectangular dataset containing only household data and another containing additional data for individual members of each household merged by &SBUILD into a single structured dataset.

When a structured dataset is used in OSIRIS IV, the user gives instructions via the &ENTRY command as to how the groups are to be rearranged to create temporary rectangular records called "entries." This restructuring of the groups permits analysis to be performed on a wider range of entries than is possible with simple rectangular records and can thereby save

numerous data management steps. The dictionary for the structured dataset may contain a default entry definition which is used to restructure the dataset when no other instructions are given via the &ENTRY command.

The advantage of a structured dataset is that for many data files, a more efficient storage mode is achieved in terms of space and cost of processing. This storage technique is more flexible and powerful than other data storage techniques, and permits larger datasets to be analyzed than in other systems.

OSIRIS IV is an open system; it is relatively easy in most instances to read data which are stored in character or binary form directly into an OSIRIS IV command. Another facet of the system's openness is the ability to take any OSIRIS IV dataset and reformat it for use by other software. Thus, it is relatively easy to move outside the system; the data are not locked into OSIRIS IV. Finally, the user may add programs which use OSIRIS IV subroutines and hence use the common control statement language and OSIRIS IV datasets. Thus the software may be augmented to meet the user's special needs.

In addition to the basic user manual "OSIRIS IV: Statistical Analysis and Data Management Software System," there are many related publications of the Institute for Social Research that can serve as useful supplements. A list of these is attached.

C. Operational Environment

The hardware requirements for OSIRIS IV are an IBM 360 or 370 computer, or an IBM-compatible machine such as an AMDAHL 470 V/6, with at least 150K bytes of main storage, the equivalent of 1000 to 3000 tracks, 7294 characters each, of disk work space, and sufficient peripheral devices for user input and output files. The computer must be operated under MTS, the OS/360 or MVS operating system, or equivalent.

D. Acquisition Information

The OSIRIS IV software package is available on an annual lease from $1,800 - $2,400 depending on type of user. OSIRIS IV is available from and supported by the Institute for Social Research, Survey Research Center, University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

A. Introduction

P-STAT is a large conversational system featuring flexible file maintenance and data display capabilities and numerous statistical procedures. The cross tabulation capabilities are discussed at length in the Tabulation Section of this report. Data may be displayed in listings, plots, frequency distribution and histograms. Statistical procedures include stepwise regression, factor analysis, discriminate analysis, product moment correlations, multivariate analysis of variance (MANOVA) in addition to more routine procedures such as chi-square, F-tests, T-tests and standard deviations. The P-STAT software was developed at Princeton University and is now distributed by P-STAT, Inc.

B. Capabilities

P-STAT Data Program is the primary way of creating a P-STAT System file. It is designed to detect and report any obvious errors found in the input data. There can be many files active during a single P-STAT run. Typically the job run will be a flow from one file to the next, joining, subsetting, collating, sorting, with an occassional correlation, regression, and even cross tabulations.

P-STAT files are rectangular and are referenced by name. Usually the columns are variables and the rows are subjects. Files, variables and rows all have 16 character names or labels, which may consist of letters, numbers or decimal points. If variable or row labels are not provided, they are generated by the system.

1. Command Language

P-STAT provides a number of commands for the statistical analysis of the data. Some of the commonest statistics for social science data are provided in the FREQ and TABLES commands. The CORRELATE command does Pearson product moment correlations, either symmetric or asymmetric, on complete data or data with missing values and may be optionally in single or double precision. In addition to the correlation matrix, other optional output files include the covariance matrix, the N matrix and the crossproduct matrix. Since these output matrices are also P-STAT system files, it is very easy to save them for use in future runs. There are also commands to do biserial and tetrachoric correlations.

REGRESSION, the stepwise regression program, uses the correlation matrix as input. A number of options allow the user to control the stepping if the default values are not appropriate for his data. If the raw input data is also supplied, a file of residuals can be produced. This file, another P-STAT system file, can be easily joined with the input data for further analyses.

FACTOR can be used to do either a principle components or an iterative factor analysis. Possible output files include the roots, the output vectors, the output factors and the output significant factors. ROTATE can then be used on the significant factors to do quartimax, verimax or equamax rotations.

DISCRIM is the P-STAT command for a multiple group discriminant analysis. The discriminant functions can be saved as a P-STAT system file. This permits cross validation of a discriminant analysis or classifying new cases using previously computed weights. The DISCRIM command as it is in P-STAT version 3.07 is not stepwise, but it has already been revised for P-STAT 78 with backwards stepping and options for interactive usage which allow the user to control the deletion of variables.

Other statistical procedures include manova (multivariate analyses of variance) written by Eliot Cramer, matrix commands such as element add, and commands to produce T-tests and oblique rotations. P-STAT also contains interfaces which allow the user to read or write BMDP or SPSS savefiles during a P-STAT run. This means that users who need statistical procedures which are not available in P-STAT can easily take advantage of other systems. There is also a facility for adding user links to P-STAT which contain your own Fortran programs.

2. Data Sets

New master files may be created by updating a file with transaction records. Rows may be added, deleted or replaced. If a case of data contains multiple records, case identifiers and sequence numbers may be checked to make sure that all the

records of a case belong together. When these records are not in order, the error is reported in detail, the case is rejected and the program continues with the next case.

A case may optionally be incomplete. For example, in a file with records on mothers and their children, it is not necessary to provide dummy records so that each mother record is followed by an equal number of children records. One option rectangularizes the file, providing missing values for the missing records. Another option permits undefined record types to be ignored. This allows the mother data to be placed in one file and, in a second step, the children data to be placed in a second file.

Three different missing values are allowed for each variable. These are converted to internal missing values which are recognized by P-STAT as MISSING VALUE 1, MISSING VALUE 2 and MISSING VALUE 3. The user does not have to remember that 9 was missing on variable SEX and 99 was missing on OCCUPATION. Most of the P-STAT commands expect that there will be missing data and handle it sensibly. Missing data can later be reconverted to real values and real values can be converted to one of the missing values in the flexible recoding language.

The data modification language can also be used to generate new variables or to delete cases which do not measure up to minimum standards. Many modification clauses can be applied in a single command. If the modifications are done using the NOOP command and creating a new P-STAT file, they are permanent and the new file may be used throughout the rest of the run and saved to be used in future runs. Temporary modifications may be done any time a P-STAT file is input to any P-STAT command. Temporary modifications are only in effect while the current command is being executed.

C. Operational Environment

P-STAT is written entirely in FORTRAN except for about half a dozen small machine language subroutines for character manipulation. P-STAT's home machine is an IBM 370, but it is written in such a way that the source cannot be directly compiled even on the 370. Before the source is compiled a pass through a pre-processor is needed. This pre-processor first reads a control

card defining the characteristics of the FORTRAN on the target machine, and then generates a compileable source. Some of the characteristics that can be defined are word size, availability of encode/decode, use of END=, alphabetic variables in real or integer mode, single or double precision, and number of allowable scratch units.

As a result source programs may be generated for P-STAT which will compile without errors on almost any computer supporting a FORTRAN IV compiler. However, P-STAT is very large and loading it is often more difficult than compiling the source. To load and run P-STAT successfully, a target machine must have either a fairly large memory and a reasonable overlay facility or be a paging system. Even so, P-STAT cannot be implemented in full in a system with only a 16-bit addressing space.

Implementing P-STAT on IBM 360/370's (or their equivalents) under OS, VS, DOS/VS or VM/CMS is very easy because load modules are provided. Implementing P-STAT on the other supported machines requires that the source be compiled and loaded at each installation. Because of the way the source is maintained, new releases of P-STAT for the supported computers can be prepared at the same time that new IBM versions are released. Then, because the generated source does not need further changes and because an appropriate loader deck (overlay, tree, segment, etc.) is supplied for each loader, implementation is usually quite simple.

D. Acquisition Information

The P-STAT System is available on an annual lease from P-STAT, Inc., P.O. Box 285, Princeton, New Jersey 08540. The cost varies from $3,000 - $8,000 depending on type of user.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

SAS

Statistical Analysis System

A. Introduction

SAS was introduced in 1966 as the Statistical Analysis System at the Institute of Statistics of the North Carolina State University, Raleigh, North Carolina. As in the case of other widely used packages, the system is now maintained and distributed by a private organization. SAS is a computer tool for data management, data analysis and report writing. The user language is made up of free format English-like commands that permit the user to custom code reports or to allow the system to produce automatic report format.

Data may be introduced into the system in any format from any device. Files may be created, stored or retrieved, and may contain variable length, mixed record types, and hierarchical files. More than 75 ready-to-use procedures are available to enter, process and analyze data. The procedures may be utilized by non-programmers with a minimum amount of training. Programmers are able to develop virtually any procedure using the powerful SAS language.

B. Capabilities

Procedures available include multiple (linear) or non-linear regression; analysis of variance or covariance; multivariate; analysis of variance; correlations; discriminate analysis; factor analysis; Guttman scaling; frequency tables and cross tabulations; categorical data analysis; spectral analysis; auto-regression; two-and three-stage least squares; t-test; and matrix manipulation.

The system produces graphics with pie charts, star and block charts, vertical and horizontal bar charts, scatter plots and overprinting options.

1. Input

Input to SAS consists of control statements, program statements, procedure statements and the data file. SAS may be executed in batch or interactively using TSO. Neither utility programs nor additional JCL is required when performing file management operations. File handling tools

include subsetting, concatenating, sorting, merging, interleaving, updating and deleting. For storage and retrieval, data can be put on disk in compressed form as an SAS data set and retained for later use.

Character data (up to 200 characters in length) and numeric data are acceptable input for SAS. Data may be free-formatted and data observations are automatically counted. Multiple data set and reports may be created in a single pass of the data file. All computations and storage of data are done in double precision.

2. Data Sets

Stored data sets solve the problem of data documentation since the following information is self contained as part of each data set:

- Job name and the SAS statements used to create the file
- Creation date
- Variable names — their labels and formats
- Number of records
- Disk space required

This information can be printed for any stored SAS data set. Additionally, SAS provides read-protect and write-protect password features that establish a significant file security capability.

3. Tabulation and graphical display

For report writing, SAS automatically generates formatted data listings, summary statistics, frequency distributions and cross tabulations. It is possible to add up to ten title lines to these pre-programmed reports. It is also possible to control the line width and page length of any report and to direct reports to various output media.

By using the powerful PUT statment, reports may be designed with variable information in the heading, user selected line width and page size, random access to any line and column on the current page and flexible formatting of output data. The PUT statement may also be used to write OS files to disk or tape or punch output cards.

The PLOT procedure provides an easy-to-use flexible capability for producing two-dimensional scatter plots. Variables to be plotted may be either numeric or character; and logarithmic plots may be produced.

The SAS package is self-contained and may be used without other software. However, the system does include a BMDP interface procedure and a procedure to covert BMDP, OSIRIS and SPSS files to SAS Data Sets.

C. Operational environment

SAS is written in PL/1 and IBM Assembler Language Coding and therefore is restricted to IBM and IBM plug compatible computers. This incudes the IBM 360/370, Amdahl, Itel AS/series, Nas, Magnuson, Two Pi, Hitachi and Nanodata. The system requires a user region of at least 150K bytes of memory.

The same version that operates in a batch mode also operates interactively under TSO. There are several TSO features. These include:

1. An editor for file updating interactively
2. Automatic adjustment of SAS to terminal line length
3. Free format data entry.

D. Acquisition Information

The SAS System is available on an annual lease from SAS Institute, Inc., Box 8000, SAS Circle, Cary, North Carolina 27511. The cost varies from $1,500 - $7,500 depending on type of user.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

A. Introduction

The SIR software is comprised of various modules including SIR/DBMS, SIR/HOST, SIR/FORMS, SIR/SQL and others. The heart of the system is the SIR/DBMS which is a comprehensive data base management system designed for both scientific researchers and business managers who may need statistical analysis of data. The system permits a wide range of record types and file structures and contains interfaces to many other statistical software products.

B. Capabilities

Version 2.1.1 of SIR/DBMS permits the creation of case-oriented hierarchical files which may be updated over time. Data may be edited, updated, retrieved, tabulated and modified as required. Built-in security procedures can limit or restrict access to data files. The SIR user may add, replace, modify or delete data at any time.

A SIR data base is defined using SPSS-like data definition (schema) commands. These commands allow for multiple record types, the definition of hierarchical and network relationships, data editing and checking (valid values, ranges and data consistency), data security at the item and record levels and multiple data types (integer, real, alpha, data, time, scaled, categorical, etc.)

The SIR retrieval language is structured and fully integrated with the rest of SIR. The retrieval language has full arithmetic and logical operations. Retrieved information can be subjected to simple statistical analysis (descriptive statistics, frequency distributions, printer plots), used in reports, used to create SPSS, BMDP, SAS, or P-STAT SAVE FILES, or a new SIR data base or written to a formatted, external data set.

The interactive subsystem includes a text editor, storage of user-written procedures and an interactive retrieval processor. The CALL (macro) facility enables the creation of generalized procedures.

Other features include various utilities for restructuring, subsetting, merging and listing, as well as automatic creation of journal files when data items are added to or changed in the data base.

The TABLE procedure produces camera-ready tabulations which may include means, medians, standard deviations, quantiles and percentages: The procedure offers complete flexibility in choosing percent base cells for various combinations of wafer, row and column percentages. The language required to produce crosstabulations is patterned after the TPL language. (The TPL package is described earlier in this report).

C. Operational environment

SIR is written in SIRTRAN which is a macro preprocessor that generates FORTRAN and assembler code. The software package is presently operational on the following computer systems:

Machines: CDC/Cyber series
DEC VAX series
Honeywell CP6
IBM and plug compatible mainframes, operating under TSO, CMS
ICL (all operating systems)
Perkin Elmer
Prime
Siemens
UNIVAC 1100, 90 series
Gould (SEL) Concept/32

D. Acquisition information

The SIR System is available on an annual lease from SIR, Inc., P.O. Box 1484, Evanston, Illinois 60204. The cost varies from $4,000 - $18,000 depending on type of user.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

SPSS

Statistical Package For The Social Sciences

A. Introduction

Although initial work on SPSS began in 1965, subsequent to the development of other statistical packages, SPSS in many ways is recognized as the leading statistical software package. Originally named Statistical Package for the Social Sciences, most know it as SPSS. It is now the most widely distributed package in the field and together with the other products of the distributor it is used at more than 4,000 sites in more than 60 countries.

Several factors have contributed to the widespread success of SPSS. The user language is natural and English-like so that the control statements are easy to learn and use. More importantly, the early user texts were fairly well written and presented.

B. Capabilities

Now in its tenth full release (version SPSS-X), the software contains a unified system of programs with a common set of conventions regarding the way the user interacts with the programs. As with most generalized package, once the user enters the data, the control instructions may be submitted in a way to carry out a variety of related tasks in any sequence required by the particular circumstance.

The system has the capability to accept input from any medium in any format. Files and sub files may be created, updated and retrieved permanently or temporarily with a complete dictionary of labels, print formats, and missing data indicators. There may be correlation, covariance and factor matrix input and output.

Statistical measures available using SPSS include procedures to produce:

- Pearson, Spearman and Kendall correlations
- Partial and canonical correlations
- Analysis of variance
- Stepwise discriminate analysis
- Multiple regression and output of residuals
- Analysis of time series

- Manova
- Paired and independent t-tests
- Factor analysis
- Guttman scale analysis
- Survival analysis

1. Data Sets

Data manipulation in SPSS can be accomplished in groups called files. A system file contains not only the data items themselves but a variety of optional information. This includes:

- Extended labels for any and all variables
- Labels for the values specified
- Indicators for missing data

Data transformations may be accomplished permanently or temporarily using arithmetic operators, logical operators and FORTRAN-like functions. There is a repeat facility for performing the same transformations using multiple variables or constants. Other capabilities include:

- Missing data assignment based on variables involved in transformation equations
- Automatic facility for creating indices based on existing variables
- Generation of uniform and normal random distributions
- Permanent or temporary selection based on logical conditions
- Permanent or temporary variable weighting
- Random data sampling
- Processing of subfiles separately or in combination

2. Tabulation and report writing

Two-way or multi-way cross tabulations are easy to obtain. The instruction set is easy-to-use and is straight forward, usually requiring two or three short commands. The CROSSTABS instruction will produce tables with up to eight-way classifications and calculates 14 different measures of association to indicate the relationship of the variables to one another. If more than one response to the same question can occur the command MULTI RESPONSE automatically solves the special problems of multiple response items.

Sub-groups may be compared and examined within tables by using the command BREAKDOWN. This procedure includes an optional analysis of variance to test whether or not the variation between groups could be due to chance. The CROSS BREAK command presents two-way to n-way analyses in cross-tabular form.

Report writing has been greatly improved with the REPORT command. This facility allows the user to override the programmed specifications to set margins, spacing, headings, labels and page lengths to improve the appearance of the reports produced by SPSS. The user need only specify the format, the variables, the breakdown and the summary statistics desired, and the customized report is printed automatically.

The SCATTERGRAM command will produce a two-dimensional graphic display of the relationship of variables to each other. In addition to the plot showing the pattern of dispersion, the command also provides an array of statistics with the plot.

Color graphics including pie charts, and bar charts which are device independent are also available under SPSS.

Other products from SPSS, Inc. include the SCSS conversational data analysis system; IDA - Interactive Data Analysis and Forecasting System; and SPSS Graphics.

C. Operational environment

The SPSS Statistical Software is written primarily in FORTRAN (95 percent) and the remainder is machine language (5 percent). The load modules are provided so that it is not necessary to have a FORTRAN compiler to use SPSS. The system is available on more than 25 computer and operating systems combinations including:

IBM 360, 370, 4300, OS, DOS, CMS and all IBM compatibles
Burroughs medium and large systems
CDC CYBER & 6000 Series
Data General Eclipse and Nova
DEC Systems 10, 20, VAX, PDP-11
HARRIS 4, 7
HEWLETT-PACKARD 3000
Honeywell 60

ICL 2900 Series
Perkin-Elmer
Prime 500-750
Siemens BS 2000
UNIVAC 70, 90, 1100

D. Acquisition information

The SPSS System is available on an annual lease from SPSS, Inc. 444 N. Michigan Avenue, Chicago, Illinois 60611. The cost varies from $1,500 - $7,000 depending on type of user.

E. Evaluation

See the supplemental information in Figure 2 on page 78, for comparative evaluation results.

BIBLIOGRAPHY

Anderson, R.E., Technical Standards and Performance Standards in the Pursuit of Quality, Bhvr. Res. Mthds. & Instmntn. V. 8, N. 2 (1976).

Auerbach Computer Technology Reports, 1983, Auerbach Publishers, Inc., Pennsauken, N.J 08109.

Beresford, J.C., Criteria and Considerations for Evaluation of a Survey Package, Proc. Comp. Sci. & Stat.: 9th Symp. on Intfc., 112-115 (1976).

Berk, K. and I. Francis, A Review of the Manuals for SPSS, and BMDP, Jnl. Amer. Stat. Asso., (1978).

Blashfield, R. K., A Survey of Major Cluster Analysis Software Programs, Proc. Stat. Comptg. Sec., (1976), 31-38.

Bryce, G. R., I. Francis, and R. M. Heiberger, Statistical Software Evaluation in the United States, Proc. of COMPSTAT, Berlin (1976), 327-334.

Buhler, R., Designing Statistical Software for Use in Both Batch and Interactive Environments, Proc. of COMPSTAT, (1976), 335-342.

Buhler, R. and S. Buhler, The Future of the P-STAT Statistical Package, Invited Paper 9.1, International Statistical Institute, Buenos Aires, Argentina, (Nov 30 - Dec. 11), 1981.

Buhler, S. and R. Buhler, Differences Between P-STAT, and SPSS, as perceived by the authors of P-STAT, Proc. Comp. Sci. & Stat.: 10th Symp. on Intfc., (1977).

Campbell, N.A. and T. L. Woodings, Improved Diagnostic Output from Statistical Packages, Invited Paper 18.1 International Statistical Institute, Buenos Aires, Argentina (Nov 30- Dec 11), 1981.

Chambers, J. M., Computational Methods for Data Analysis, New York: John Wiley & Sons, Inc. (1977).

Conference of European Statisticians, Report on the Use and Future Need for Data Base Management in National Statistical Services, Prepared by the Group of Rapporteurs on DBM, United Nations Economic and Social Council, CES/WP.9/150 (1977).

Data Decisions Software Reports, 1983, Ziff-Davis Publishing Co., Cherry Hill, N.J. 08034.

Data Sources Guide to Software Products, 1983, Ziff-Davis Publishing Co., Cherry Hill, N.J. 08034.

Datapro Directory of Software Report, 1983, Datapro Research Corp., Delran, N.J. 08075.

Davis, H. T., General Criteria and Considerations for the Evaluation of Time Series Program Packages and Libraries, Proc. Comp. Sci. & Stat.: 10th Symp. on Intfc. (1977).

Fellegi, I. P., and D. Holt, A Systematic Approach to Automatic Edit and Imputation, Jnl., Amer-Stat. Asso., 71, 353 (1976), 17-35.

Fellegi, I. P., and W. M. Podehl, Statistics Canada's Appraoch Toward Software for Survey Data Processing, Proc. Comp. Sci. & Stat.: 9th Symp. on Intfc., (1976), 125-129.

Francis, I., Design for Evaluation of Statistical Program Packages, Bull. Intl. Stat. Inst., 46, 2 (1975) 573-584.

Francis, I., Questionnaire: Features of A Statistical Program or Package, Technical Report No. 1077/014, Dept. of Econ. & Soc. Statistics, ILR School, Cornell University, Ithaca, N.Y. 14853 (1977).

Francis, I., A Comparative Review of Statistical Software, Exhibition of Statistical Program Packages, International Statistical Institute 41st Session, New Delhi, 1977.

Francis, I., Statistical Software: A Comparative Review, Elsevier North Holland, Inc., N.Y., 1981.

Francis, I., and R. M. Heiberger, The Evaluation of Statistical Program Packages, Proc. Comp. Sci. & Stat: 8th Symp. on Intfc., (1975), 106-109.

Francis, I., and J. Sedransk, Comparing Software for Processing and Analyzing Survey Data, Invited Paper 16.2 International Statistical Institute, Manila, Philippines, 1979.

Francis, I., and J. Sedransk, Software Requirements for the Analysis of Surveys, Proc. 9th Intl. Biomtc. Conf. (1976), 228-253.

Francis, I., S. P. Sherman - Cornell University & Buhrman, R. I. J. L. Willard - Delta Systems Consultants, Inc., A Comparison of Software for Tabulating Survey and Census Data, American Statistical Association, Statistical Computing Section, Aug. 1979.

Francis, I., S. P. Sherman, and R. M. Heiberger, A Look at Languages and Programs for Tabulating Data from Surveys, Proc. Comp. Sci. & Stat.: 9th Symp. on Intfc., (1976), 129-140.

Francis, I., R. M. Heiberger, and P. F. Velleman, Criteria and Considerations in the Evaluation of Statistical Program Packages, Amer. Statisn., 29, 1 (1975), 52-56.

Frankel, M., Software for Surveys: Are Existing Packages Adequate for Validating Statistical Inference, Proc. Comp. Sci.& Stat.: 9th Symp. on Intfc., (1976), 115-118.

Gentle, J.E., Comparisons of Statistical Packages by Users Having Some Familiarity with Computing and Statistics, Proc. Stat. Comptg. Sec., (1975), 114-117.

Gepner, A. L, Software Package Ratings, Datamation, (1976), 108-146.

Hardy, K. A., D. R. Kniefel, W.C. Reynolds, A Comparision of Some Existing Statistical Packages, Proc., Comp. Sci. & Stat.: 10th Symp. on Intfc., (1977).

Hill, G., Problems in Comparing Survey Packages (A Case Example), Proc. Comp. Sci. & Stat: 9th Symp. on Intfc., (1976), 242-244.

Kohm, R. F., T. R. Ryan and P. F. Velleman, Index of Available Statistical Software, Proc. Stat. Comptg. Sec., (1977).

Lauro, N. and G. Serio, Criteria for Evaluating and Comparing Statistical Software: A Multidimensional Data Analysis Approach, Statistical Software Newsletter, Vol. 8 No. 3 (Dec. 1982), 102-119.

Muller, M. E., Portability Standards for Software, Proc. Comp. Sci., & Stat.: 8th Symp. on Intfc. (1975), 173-76.

Nelder, J. A., A User's Guide to the Evaluation of Statistical Packages and Systems, Intl. Stat. Revw., 42, 3 (1974), 291-298.

Nordbotten, Joan C. & S. Nordbotten, System Analysis and Design Applied to Statistical Systems, Invited Paper 1.3 International Statistical Institute. Buenos Aires, Argentina (Nov 30 Dec 11), 1981.

Rowe, B. C., and M. Scheer, Computer Software for Social Science Data, Social Science Research Council, (British) (1976).

Ryan, T. A., Jr. and J. M. Lefkowitz, Minitab on Microcomputers: Use and Implementation, Invited Paper 17.1 International Statistical Institute, Buenos Aires, Argentina (Nov. 30 - Dec 11), 1981.

Thisted, R. A., A Comparison of Packages Approach, Proc. Stat. Comptg. Sec., (1977).

United Nations Department of Technical Co-operation for Development & Statistical Office, Annex I, Review of Software Packages for Survey Data Processing, National Household Survey Capability Programme, New York, 1982, 117-146.

Velleman, P. F., J. R. Seaman, and I. E. Allen, Evaluating Package Regression Routines, Proc. Stat. Comptq. Sec.,(1977).

Vellman, P. F. and R. E. Welsch, Some Evaluation Criteria for Interactive Statistical Program Packages, Comm. in Stats., B (1977). also Proc. Stat. Comp. Sec., (1975), 10-12.

Wilkinson, G. N., Language Requirements and Design to Aid Data Analysis and Computing, Bull. Intl. Stat. Inst., 47 (1977).

Wyllys, R. E., Instructional Use of Statistical Program Packages: Experience with BMD, IMP, OMNITAB II, and SPSS, Proc. Comp. Sci. & Stat.: 10th Symp. on Intfc., (1977).