PACKAGES FOR STATISTICAL COMPUTING AND ANALYSIS[1]

M.M. Manuel, Jr.[2]

Introduction

Applications of the computer to perform more and more complex tasks to help solve a wide range of problems are increasing rapidly. The computer has been used widely as a tool to generate secondary information materials and other tasks, to handle large volumes of information, to provide access to specific information required by a user, to make information available to many people in many locations to meet many demands and needs, to transmit information quickly, to assess user demands and needs, and to link the producer with user.

There has been a dramatic increase in the quantity of information being transferred and a significant compression of time and distance in this transfer process through technological advancement. Behind all these is a computer system that processes enormous amount of relevant raw data accumulated for specific and different purposes.

Research information for instance, to have impact on domestic and world problems, must, at most times, move from the researcher to those responsible for its timely utilization for decision-making purposes. Raw data therefore, must be processed by computer systems equipped with necessary methodologies for generation of information required by persons within the time allowed for decision-making or action to which it applies.

Since the quality, accuracy, timeliness, and relevance of information depends to a considerable extent on the methodologies available on the software systems, this paper provides some insights on some

---

[1]Paper presented at the Philippine Statistical Association Conference on Computers in Statistics, Oct. 16, 1981, Central Bank of the Philippines, Manila.

[2]Executive Director: Agricultural Resource Computer Center (ARC) at Los Baños; Coordinator; UPLB-MIS; Assistant Professor of Statistics, UPLB.

software systems used primarily for statistical computing and information generation in the country.

Software Computer Systems

Basically, software computer systems may be grouped into Programming Languages such as BASIC, FORTRAN, COBOL, PL/1, and so on, and Application (packaged) Programs such as CLUSTAN, SPSS, SAS, GPSS, MPSX, etc. In general, software computer systems used specifically for statistical analysis has the advantage of: 1) rapid analysis, once program is completed; 2) complex analysis; 3) easy handling of large data sets; 4) minimum computational error in calculation of statistics; and 5) printed results in easy readable format which becomes a permanent record.

On the other hand, software computer systems have several disadvantages: 1) analysis is sometimes expensive; 2) access to data is sometimes delayed; 3) large storage capacity is required which frequently is not readily available; 4) a high level of expertise is required to program the system; and 5) the possibility of programming error is increased if the programmer is not the person who actually supervised data collection.

The Need for Statistical Softwares

The use of the computer as a tool for data processing in general, and statistical analysis in particular, is only possible through the existence of software developed and designed specifically for this purpose. The packages developed this way, the new and those based on existing ones, are of great significance because they represent a work of the world's computer community and they reflect the needs and demands of worldwide community of users. This does not, however, in any way undermine the importance of user developed programs.

It can be emphasized that in the design and development of software packages for statistical computing and data processing, the choice of programming languages influence considerably the nature

of the final products. One therefore cannot talk about a program package without some detailed reference to the programming language used. One may therefore refer to Application Programs to include program packages (usually designed and developed by a group for commercial use over a long period of time and usually multi-purpose types and user-developed programs (are single-purpose and developed by a user over a short period of time).

Application Programs therefore can be classified according to:

A. Data Program Packages
Some of these are:

1. SAS – Statistical Analysis System is a user-oriented program package with powerful capabilities for data management, statistical analysis and report writing.
2. SPSS – Statistical Package for the Social Sciences is a user-oriented data management and statistical analysis program package.
3. CLUSTAN – A cluster analysis program (a single program) package featuring a wide variety of clustering method.
4. MPSX/370 – A program package designed to solve linear programming problems.

B Program Packages for Data Base/Data Communication System

Some of the program packages are:

1. CICS/VS – Customer Information Control System/ Virtual Storage is general purpose Data Base/Data Communication (DB/DC) system. It functions to support on-line system in the same manner as the operating system and access method do to batch processing system.
2. STAIRS/VS – Storage and Information Retrieval System/Virtual Storage – is a terminal oriented, multi-user on line dialog system for storing and retrieving data both formatted and unformatted. It offers the user a variety of resources for data base creation and maintenance besides

containing special features for data base searching and document output.

3. CDS/ISIS – Computerized Documentation System/Integrated Set of Information System. This is a package developed by UNESCO which is suited for storage and retrieval of textual data.

C. Special Program Packages

1. SYMAP – A program for producing on a line printer, maps with varying shades of gray.
2. SCRIPT/370 – A program designed for text formatting and editing.
3. Mailing Labels System – This system allows the user to store and to update mailing file and to produce mailing stickers.

D. Simulation Packages:

The simulation packages include:

1. GPSS – General Purpose Simulation System is a program used for simulating discrete systems on a digital computer.
2. CSMP – Continued System Modelling Program is a program designed for simulating the behavior of continuous systems.

E. Utilities

This consists of general purpose utility programs varying from SORT/MERGE to users tape/disk error connection.

Choice of Software

—An electronic data processing center is always presented with a variety of software to choose from in the market for data processing and statistical computing or data analysis. The final choice that one makes depends on what its systems is capable of maintaining to satisfy its needs and demands of its clientele. A general set of criteria to be considered are the following:

1) Ranges of Applications, and
2) Cost of Maintenance

Ranges of Applications:

In any Electronic Data Processing installation, running in batch, interactive and autobatch modes, one chooses a program package for statistical computing and analysis that offers:

1) a wide range of statistical procedures
2) extensive data management tools
3) comprehensive report-writing features

There are some popular software packages for statistical computing in the market today which are installed in some data processing centers in the country like SPSS, SAS, and BMDP to name a few. One or two have been installed and used much longer than others.

To put things in a much better perspective, let us identify some criteria for choice between and among available packages for statistical computing and data analysis. Some of these are:

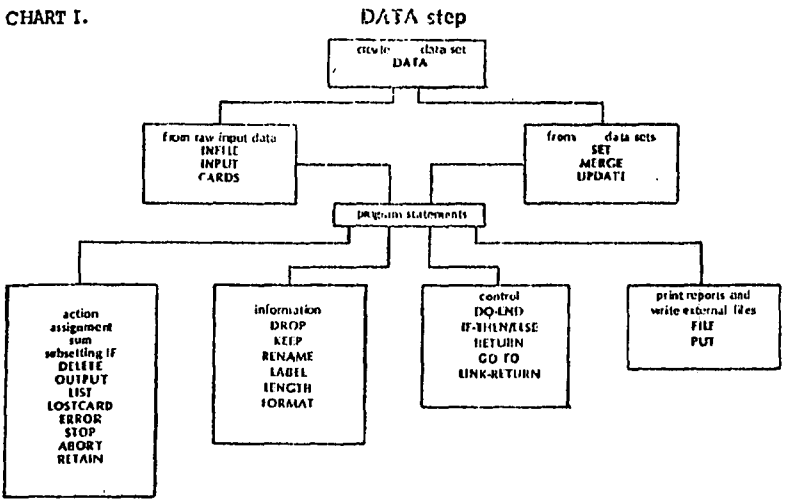
1. CAPABILITY – Is the package designed to handle as many (kinds) and as extensive (depth) statistical analyses including their attendant problems such as data tranformation, and the like?
2. PORTABILITY – Can the package be adapted to as many *makes* and sizes of computers and be used in batch, interactive or autobatch modes?
3. CALLABILITY – Within the same operating system can the package, if the need arises call on another package?
4. STORABILITY – Internally, does the use of one procedure allow the use of another for a given problem?
5. FLEXIBILITY – Can it allow use from very simple to highly sophisticated level of application?
6. OPEN-ENDEDNESS – Can the package allow for inclusion of additional procedure for local use?
7. ECONOMY – Is it cheaper to use than have the program custom-made instead?
8. EASE OF LEARNING – Will a non-programmer or non-computer-oriented individual be at ease in learning its use?

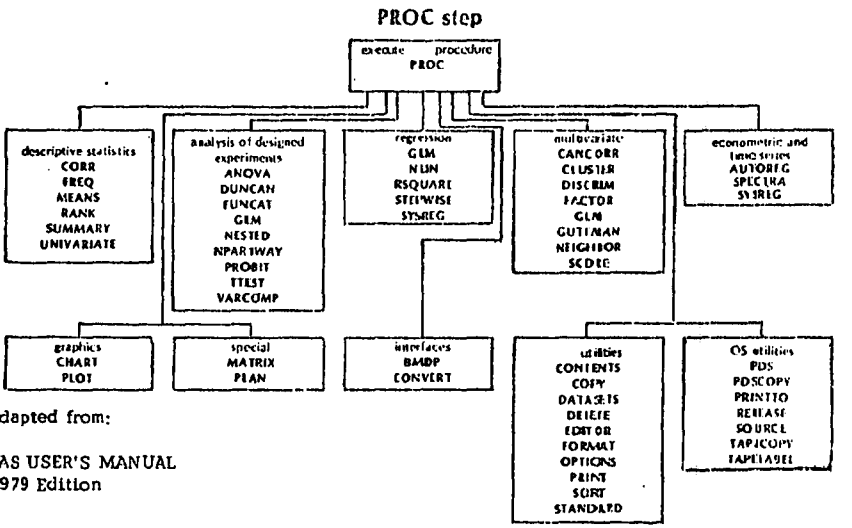
9. EASE OF USE – Is it free of computer jargons and tedious exercises when used?
10. RELIABILITY – Will it always work? If it works, can we depend on the accuracy of the result?
11. STABILITY – How long will it stay in use? Is it supported by established organization for maintenance?
12. INTEGRITY – Are the method and algorithm used the most recent and of the highest quality and made by authoritative people?
13. SHARABILITY – Internally, does the use of one procedure allow the use of results of another for a given problem?
14. ADAPTABILITY – Can the output/display be custom-made that you sort of get the feel of it?
15. LIMITATIONS – What are its limitations?
16. EFFICIENCY – How is user productivity increased?

- does the package have file-handling capabilities that carry out utility applications such as file updating, data retrieval, sorting and merging?
- does the package require a minimum number of JCL (Job Control Language) to run jobs, save data sets or write operating system (OS) files?
- does the package allow programs written in other programming languages such as COBOL or PL/1 be generated with a few package-supported (procedures) statements?
- how long does one get to learn to use the package? Is the language of instruction simple and tied to concepts that users are already familiar with? Do occasional users need to go through the learning processess each time they use the package?

One therefore looks for a package that gives the user a broad spectrum of statistical procedures for analytical work. This does not only include key statistical capabilities available but also accompanied by data manipulation tools that make it easy for the user to do all his processing in the total environment provided by the software. From a few in the software market today, one may be able to find a software that has the capabilities shown in CHART I.

CHART I.





Adapted from:

SAS USER'S MANUAL
1979 Edition

Statistical Analysis System (SAS) and Statistical Package for the Social Sciences (SPSS)

Two of the most frequently used softwares for data processing and statistical computing and analysis are the SAS and SPSS. These have stayed much longer in most computer installations here and abroad because they have been maintained well by their developers.

Using the criteria for selection identified in the preceding section on SAS and SPSS, we have the following: SAS:

1. Capability – SAS is an easy-to-use system for data analysis, and can well provide the following functions:
   - o information storage and retrieval
   - o data modification and programming
   - o report writing
   - o statistical analysis
   - o file handling

   – *information storage and retrieval – SAS can read data values from cards, disk, or tape, then organizes these* values into an SAS data set, which may be combined with other SAS data sets; it may be analyzed statistically, and reports may be produced; automatically self-documenting.
   – data modification and programming – set of SAS statements and functions is available when data value need to be modified, these include statements that create new variable, accumulate totals, check for errors and powerful tools such as DO/END and IF-THEN/ELSE statements.
   – report writing – writes data in almost any form, including preformatted reports, punched cards, and other output files.
   – statistical analysis – statistical analysis procedures in SAS ranged from simple descriptive statistics to complex multivariate techniques.
   – file handling – SAS has tools for editing, sub-setting, concatenating, merging and updating data sets. Multiple

input files can be processed simultaneously and several reports can be produced in one pass of the data.

2. Portability – SAS runs on IBM 360/370 computers and plug-compatible machines as AMDAHL, ITEL, CDC OMEGA, MAGNUSON, RYAD, etc. under OS or OS/VS. It can be used either in batch, interactive (TSO) or autobatch mode.

3. Callability – Any OS program can be called from SAS by putting its name in a PROC statement. For example, an OS programme named MYPROG can be called from an SAS job using the SAS statement:

   PROC MYPROG [CC=n] ;

   where the [CC=n] is optional and gives the maximum acceptable return code the program can have to execute the SAS job. A SASLIB DD statement should be included in the job control deck describing the OS program called by the SAS job.

   In addition, SPSS, SAS72, BMDP, DATA-TEXT, OSIRIS data sets (system files) can be used by an SAS job using the PROC CONVERT statement and will convert these to SAS data sets containing the same information as the input system files.

4. Storability – more than one procedure is allowable or can be used in a given problem, hence, storability is high for SAS.

5. Flexibility – SAS provides a wide range of application, from a very simple to highly sophisticated one as follows:

   - originally developed for statistical needs, it provides accurate and up-to-date statistical techniques.
   - data handling – changing, rearranging, and combining data values, generating new variables with no limits to the ways in which they can be combined; used by many as a data base management system.
   - report writing – presenting computer results in a form that most users can understand.

- data management maintenance – SAS includes a group of utility procedures for copying data sets, investigating tape contents, moving program libraries, etc.

6. Open-endedness – SAS allows inclusion of additional procedures, changes and additions, enhancement to existing procedures, and new system features.
7. Economy – in terms of computer running time, a custom-made or tailored-made program perhaps will always be more efficient than a software package but considering the cost of preparing the program/system designing, developing, etc., a software package like the SAS will be cheaper and maintenable in the long run, and will cover a wider range of applications.

SPSS:

1. Capability – SPSS is designed for the analysis of social science data and offers a large number of statistical routines commonly used in the social sciences. It allows a great deal of flexibility in the format of data, as well as providing a comprehensive set of procedures for data transformation and file manipulation.
2. Portability – SPSS Inc. supports its software on IBM 360s, 370s, 3030s and 4300s (and software compatible counterparts, such as the AMDAHL 470, CDC OMEGA and NAS computers) under OS, DOS, and CMS, operating systems, on DEC VAXES and DEC PDP-11s; on BURROUGHS large systems; on CDC A CYBER 70; on UNIVAC 1100 series; and on XEROX sigma series. It operates either in batch or interactive mode.
3. Callability – OSIRIS-SPSS interface enables SPSS to read a typed OSIRIS dictionary file and fixed format BCD OSIRIS data file as input for an SPSS run.
4. Storability – the use of more than one procedure for a given problem is allowed in SPSS (One-case-at-a-time philosophy of SPSS).

5. Flexibility – flexible enough to use for simple application to highly sophisticated one.
6. Open-endedness – installing or adding a new statistical procedure is possible in SPSS, though one may be intimidated by the size and apparent complexity of SPSS. There is some guidelines that outline the basic considerations in the addition of new statistical procedure.
7. Economy – There are certain factors to consider whether a software package like SPSS is more economical to use than a custom-made program or system, and these are partly dependent on the type of applications they are used for, in which case, justification of economy becomes relative and dependent.

Let us assume a hypothetical data shown in the following table:

TABLE 1. SAMPLE DATA FOR SAS AND SPSS

ID No.	*Agency*	*Sex*	*Job Exp (Yrs.)*	*Monthly Salary*
1	Govt.	M	2	4,000.00
2	Priv.	M	6	8,000.00
3	Priv.	F	15	12,000.00
4	Govt.	M	20	9,000.00
5	Govt.	M	25	10,000.00
6	Priv.	F	5	9,000.00
7	Govt.	F	10	8,000.00
8	Priv.	M	12	10,000.00
9	Govt.	M	8	6,000.00
10	Govt.	F	9	7,000.00

The simple statistical analyses conducted on the sample data using the two packages are:

A. Frequency distribution tables for Agency, Sex and cross-tabulation between Agency and Sex, and
B. Descriptive Statistics such as means, standard deviations for Job Experience and Monthly Salary.

Information tabulated comparing SPSS and SAS are shown in Table 2.

TABLE 2. COMPARING SAS AND SPSS FOR PROCESSING A AND B ACCORDING TO CPU TIME, CLOCK TIME, STORAGE REQUIREMENTS, AND TOTAL COST FOR PROCESSING

VARIABLES	*SAS*	*SPSS*
CPU Time	A – 7.00 sec. B – 5.95 sec.	6.92 sec. 5.66 sec.
Clock Time	A – 49.00 sec. B – 37.00 sec.	1 min. 8 sec. 1 min.
Storage Requirements	A – 148 K B – 150 K	81.92 K 81.92 K
Cost of Job	A – ₱1.55 B – ₱1.32	₱1.54 ₱1.26

Some of the additional technical information on the two software packages are:

- Limitations – Because these packages used dynamic allocation, problem size is limited only by the amount of storage available. SAS is capable of handling 1024 variables at any one time while SPSS has a maximum capability of 500 variables.
- Mode of Use – Both software systems run in batch and interactive modes. SAS has the CMS/SAS Conversational Monitoring System and SPSS has the SCSS Conversational System.
- Reliability – SAS has been in productive use since 1968 at installations around the world, and is currently installed at over 400 locations. SPSS which was in productive use in 1970, is currently installed at over 600 locations around the world.

  Major new releases which have undergone testing for many months at both commercial and university sites are now available. Response to user problems have been quick; and maintenance has been thorough.
- Efficiency – Increasing user productivity has been the primary goal considered in the design and development of both

software systems. SAS, however, provides for free format language input option and powerful manipulation tool. Furthermore, user need not learn programming. Both SAS and SPSS are very flexible – they are end-user oriented.

Machine efficiency is at the same time improved since SAS and SPSS programs compile directly into machine code. Thus, execution speeds compare favorably with those higher level programming languages.

Of importance also, are the program statements needed to run the job which are divided into two groups.
These are:

- Statements which define or describe and create data into the systems.
- Statements which perform or execute the procedure and additional information for the procedure.

Specifically, the above statements refer to the Job Control Language (JCL) which serves as a means of communication between the software and the computer systems. Table 3 shows the appropriate JCL for the sample runs on the hypothetical data using SAS and SPSS software systems.

Table 3. Appropriate (Job Control Language) for the Sample Data Run on SAS and SPSS:

SAS JCL:

```
//        JOB CARD
//        EXEC ARCSAS
//SYSIN  DD *
        DATA SAMPLE;
        INPUT ID NO AGENCY $ SEX $ JOB --EXP
                MO __SLRY;
        CARDS;
                INSERT DATA CARDS HERE
        PROC FREQ;
        TABLES AGENCY SEX AGENCY*SEX;
        TITLE SAMPLE ANALYSIS USING SAS;
        PROC MEANS; VAR JOB. EXP MO__ SLRY;
/*
```

(Table 3 Continued)

SPSS JCL:

```
//      JOB CARD
//      EXEC ARCSPSS
//GO. SYSIN DD*
FILENAME SAMPLE
VARIABLE LIST ID NO AGENCY SEX JOB__EXP MO__SLRY
INPUT MEDIUM          CARD
INPUT FORMAT          FIXED (F2.0 1X,A4,2X,A1,1XF2.0,1X,F5.0)
N OF CASES            10
PRINT FORMATS         AGENCY, SEX (A)/
TASK NAME             ONE__WAY TABLE
FREQUENCIES           GENERAL=AGENCY, SEX
READ INPUT DATA

                      INSERT DATA CARDS HERE

TASK NAME             TWO-WAY TABLE
CROSSTABS             TABLES=AGENCY BY SEX
TASK NAME             DESCRIPTIVE STATISTICS
CONDESCRIPTIVE        JOB__EXP, SALARY
STATISTICS            ALL
FINISH
/*
//
```

SAS:

SAS INSTITUTE INC.
BOX 8000
CARY, NC 27511

Correspondence through: Sugi – SAS USERS GROUP INTERNATIONAL SAS INSTITUTE

Proponents:

ANTHONY J. BARR
JAMES H. GOODNIGHT
JOHN P. SALL
WILLIAM H. BLAIR
DANIEL M. CHILKO

(Table 3 Continued)

SPSS:

NATIONAL OPINION RESEARCH CENTER
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60611

Correspondence through: SPSS INC.
SUITE 3300
444 N Michigan Ave.
Chicago, IL 60611

KEYWORDS: PUBLICATION
FOR THE USERS OF SPSS INC
SOFTWARE PRODUCTS
SPSS INC.

Proponents:

NORMAN H. NIE
C. HDLAI HULL
JEAN G. JENKINS
KARIN STEINBRENNER
DALE H. BERT

Cost of Software Maintenance

Technological breakthroughs in design and manufacturing processes have improved hardware performance significantly and have continuously reduced hardware cost and size over the last twenty years making them generally affordable and portable. On the other hand, since most software design and development are made in developed countries where labor of trained personnel is expensive, software (development) production which is labor intensive is becoming more and more expensive.

The following tables will provide us with some insights on the cost of the two software products:

TABLE 4. AVERAGE COST* OF SPSS BATCH SYSTEM

Year	*Release*	*Average Annual Cost Over 4 Years*	*Number of Procedures*
1971	3	$ 833	13
1973	5	1107	21
1977	7	1393	21
1979	8	1906	23
1981	9	2119	26

Source: KEYWORDS: For Users of SPSS Inc.
Software Products.
SPSS INC. Chicago, IL 60611
*Adjusted to 1967 Consumer Price Index

TABLE 5. INSTALLATION AND MAINTENANCE COSTS OF SAS AND SPSS AT THE AGRICULTURAL RESOURCE COMPUTER CENTER (ARC) IN LOS BAÑOS, LAGUNA

	SAS	*SPSS*
Releases (including updates, New Versions, etc.)	All	All
Installation Cost	$750	
Maintenance Cost		
1977 – 80	300	900
1980 – 81	500	500
Number of Procedures		
1979	50	23
1981	81	26

Average software costs for statistical packages differ between EDP centers because license and renewal fees are substantially different for academic, non-profit, and commercial customers. The first year rate for one-year service agreement for commercial customers with SAS Inc. is $2500; continuation agreements for subsequent years are $1000 per year. For degree-granting institutions and non-profit customers, the first year rate is $750; succeeding years are $300 per year.

Differences in installations and maintenance costs of SAS and SPSS are shown in Table 5.

The final consideration that one must take into account in making the choice among software products – is its cost effectiveness. Installations and maintenance costs therefore, must be spread over a wider range of applications and users.

Conclusion

In an Electronic Data Processing Installation, operating in batch, interactive, and autobatch modes, management selects among a number of packages for statistical computing and analysis that offers:

1) a wide range of statistical procedures
2) extensive data management loads, and
3) comprehensive report-writing features

Criteria for choice such as capability, flexibility and others guarantee that the package considered provides statistical procedures which meet all standard statistical needs. Further, the software should give the user a broad spectrum of statistical procedures for analytical work. Not only statistical capabilities must be available, but they must be accompanied by data manipulation tools that make it easy for the user to do all his processing in the total environment provided.

Finally, EDP management, must consider the cost-effectiveness of the software package obtained. Average cost of installation and maintenance must be spread over a wide range of applications and clientele.

REFERENCES

ALVIAR, S. M. "Statistical Packages for Statistical Computing." Paper presented at the Second National Convention on Statistics, PICC, December 2-3, 1980.

A USER'S GUIDE TO SAS: 1980 Edition. SAS Institute, Inc. P.O. Box 8000, Cary, North Carolina, 27511.

SAS/ETS USER'S GUIDE: ECONOMETRIC AND TIME SERIES LIBRARY; 1980.

Proceedings of the Sixth Annual SAS Users Group International Conference: Orlando, Florida, February 8-11, 1981.

SPSS: 2nd Edition: SPSS Inc. Suit 3300, 444N Michigan Avenue, Chicago, IL 60611.

KEYWORDS: Publication for the Users of SPSS, Inc. Software Products. SPSS, Inc.