NASA

NASA

USL / DBMS NASA / PC R&D

WORKING PAPER SERIES

Report Number

DBMS.NASA/PC R&D-5

The USL/DBMS NASA/PC R&D Working Paper Series contains a collection of formal and informal reports representing results of PC-based research and development activities being conducted by the Computer Science Department of the University of Southwestern Louisiana pursuant to the specifications of National Aeronautics and Space Administration Contract Number NASW-3846.

For more information, contact:

Wayne D. Dominick

Editor
USL/DBMS NASA/PC R&D Working Paper Series
Computer Science Department
University of Southwestern Louisiana
P. O. Box 44330
Lafayette, Louisiana 70504
(318) 231-6308

GENERAL SPECIFICATIONS FOR THE DEVELOPMENT OF A USL NASA PC R&D STATISTICAL ANALYSIS SUPPORT PACKAGE

Jinous Bassari

Spiros Triantafyllopoulos

The University of Southwestern Louisiana
Computer Science Department
Lafayette, Louisiana

August 2, 1984

GENERAL SPECIFICATIONS FOR THE DEVELOPMENT OF A USL NASA PC R&D STATISTICAL ANALYSIS SUPPORT PACKAGE

ABSTRACT

This is a three-level package designed to allow statistical analysis for a variety of applications within the USL DBMS NASA/RECON project. Designed with flexibility and uniformity as the main considerations, it is expected to provide computational capabilities for a variety of user needs, beginner to expert, in three different forms: a library package, an interactive package and a batch-processing package.

TABLE OF CONTENTS

GENERAL SPECIFICATIONS
FOR THE DEVELOPMENT
OF A USL NASA PC R&D
STATISTICAL ANALYSIS SUPPORT PACKAGE

I. INTRODUCTION

This is a proposal for the design, development and implementation of a general-purpose statistical package for the USL DBMS NASA/RECON project.

Statistical Packages offer to the user the power and flexibility they need, without having to write complicated programs. In addition, the user can be assured of the accuracy of the results. Many statistical packages have been developed so far, for all types and sizes of computers.

There are three major types of statistical packages available for the user:

i. Statistical Program Libraries.
Statistical Libraries are collections of programs that are bound together in one collection. The user can call them from his/her application programs, supplying the appropriate arguments and obtaining the results in a similar way.

ii. Interactive Statistical Packages.

Interactive Statistical Packages allow the user to interact with the computer. The user is put on at a "command level", where he/she issues commands and enters data. The program processes the data and returns the results to the user on the terminal screen.

iii. Batch Statistical Programs.

Batch programs allow the user to collect all his/her commands to the program in one group, code them in a particular language, and then process the entire batch. The user does not interact with the execution at all.

This research and development proposal intends to implement all three packages under a unified interface. The result is expected to be a flexible and powerful package with common characteristics between its three forms. It is also intented to be completely transportable among any computer that can support the "C" programming language. This includes 3 of the 4 large computer systems available at USL, namely the DEC UNIX VAX-11/780, DEC VMS VAX-11/780, the Pyramid Technologies 90x, and, of course, the IBM PC/XT.

II. OBJECTIVES OF THE PROJECT

The generic objectives of the project are as follows:

i. To develop a powerful, flexible, easy-to-use and transportable statistical package.

ii. To improve our knowledge in the fields of statistics and numerical computation.

iii. To obtain further experience on the design, implementation, testing and maintenance of a major software product.

The specific objectives of the statistical package design are as follows:

i. Computational power: the objective is a design that can satisfy most user needs in terms of available functions and options. The package should offer a full range of commands for most applied statistical computations.

ii. Design flexibility: The design must be flexible so that changes, improvments and addition of more functions can be accommodated without major changes, if any, to the entire package.

iii. Ease of use: the design should be such that any of the three interfaces will be easy to use efficiently. This includes error checking and, in the case of the interactive user interfaces, available online help. Uniform command and function formats in all three modes will be used also.

iv. Efficiency and accuracy: Efficiency of the algorithms used is very important considering that the package to be developed will be used in mini and micro computers with often limited resources and speed of execution problems. Accuracy is also critical so that the user is assured of the quality of the results.

v. Package Transportability: The programs should be written in a way that ensures transportability between varying operating environments. Standard programming policy will be adopted for all modules.

This design will be first implemented on the IBM Personal Computers of the NASA/RECON Project. Parallel development on the DEC VAX-11/780 will also be considered.

III. METHODOLOGY

A highly modular approach will be followed in the design and implementation of this project. This will ensure that many of the design objectives, in particular, flexibility and modifiability, are inherent in the implementation.

For achieving the transportability and modularity goals, the "C" programming language was chosen as the implementation language. It offers good performance characteristics and high modularity which make it most desirable. It is also powerful in character and file manipulation, facts that make it more desirable to use in the second and third phase of the design.

Algorithm selection is critical in the computational parts. Therefore, extensive research will have to be performed in order to determine the most appropriate ones to be used. While a computer approach to manual algorithms can be used, for some cases it is not efficient and better methods should be found.

At this point only the first phase of the statistical package has been totally defined. The interactive and batch interfaces will be designed under the considerations applied to the first phase, with the final goal being to create a common, efficient interface for all three modes. A defined command language for the interactive phase would consist of either a menu-selection procedure, a command language or a combination. Again, the batch interface can be similar to the interactive in terms of command names, arguments, etc, or be a completely different programming language by itself.

For the interactive interface, a spreadsheet configuration like MINITAB is likely to be implemented, with its commands, arguments and options combined to form the batch programming language. Therefore, by making the library interface with a similar structure, the goal of uniformity can be achieved.

As a minimum, the functions shown below are expected to be implemented for the first phase (program library). Then the interactive and batch interfaces can be built on top of the packages. The modularity of the design will allow the addition of new functions and/or the modification of existing ones to be performed efficiently, with no major code changes in the entire program structure.

PROPOSED CONFIGURATION OF PHASE 1

1. Basic Input/Output

   Read from a given file

   Read from terminal

   Write to a given file

   Write to terminal

   Report error/warning messages

2. Basic One-Vector Calculations

   SumX, SumX2, sum2X

   Mean, mode, median

   Variance, standard deviation

   Sort ascending, descending, rank

   Frequency, most/least frequent

   Relative frequencies, signs

   Max-min, local max/min, k-th max/min

3. One-Vector Test Statistics

   Confidence intervals

   z-scores, z-tests

   Proportion tests

   Student's t-test

Small/large sample sizes

4. Basic one-Vector Graphs

    Bar Charts - Histograms

    Frequency graphs

5. Two-Way Statistics

    Hypothesis testing

        Difference of means

        Variance known/unknown

        D-test

        Paired Samples Tests

    Tests for Standard Deviation

    Degrees of Freedom, F-test

    Tests for proportions

6. Two-Vector Graphics

    Plots

    Scatter Grams

    Frequency Plots

    Charts

    X-Y plots

    Distributions

7. Linear Regression and Correlation

    Linear Regression Analysis

    Regression Line calculations

    Correlation Analysis

    Standard Error

8. Multiple Regression

    Exponential regression analysis

    Logarithmic regression analysis

    Parabolic regression analysis

    Multiple Analysis

9. Basic Probability Calculations

    Probability distributions

        Normal distribution

        Binomial distribution

        Poisson distribution

    Probability tests

        p-test Probability confidence intervals

10. Advanced Probability Calculations

    Conditional Probabilities

    Independent Probabilities

    Probability Estimations

    Bayes' Theorem

Basic Combinatoric Calculations

P(m,n) value

C(m,n) value

n! value

Probability distributions of samples

11. Chi Square Analysis

Contingency Tables

Chi Square tests

Chi Square distribution

Lambda Index of association

12. Analysis of Variance

One-way analysis

Two-way analysis

Difference of several means

Total Variance calculations

13. Non-Parametric Tests

Sign test

Mann-Whitney test

Non-parametric ANOVA

IV. SUMMARY

Design, implementation, testing and maintenance of this major software package is expected to generate a support environment for any other activities that require statistical analysis within the NASA/RECON or related DBMS projects. The applicability of statistical analysis methods in information storage and retrieval systems is increasing, ranging from performance measurement and evaluation to natural language text analysis, thus making this project an interesting consideration for further research and development.

The unified environment that this document proposes is expected to further improve the user/system interface and make it more effective. Portability is also provided in order to have a single data analysis environment for more than one hardware configuration.

5.5

1. Report No. 1N-82

2. Government Accession No. 183575 [illegible]

3. Recipient's Catalog No.

4. Title and Subtitle 16 i,

USL/NGT-19-010-900: GENERAL SPECIFICATIONS FOR THE DEVELOPMENT OF A USL NASA PC R&D STATISTICAL ANALYSIS SUPPORT PACKAGE

5. Report Date DATE

August 2, 1984 OVERRIDE

6. Performing Organization Code

7. Author(s)

JINOUS BASSARI AND SPIROS TRIANTAFYLLOPOULOS

8. Performing Organization Report No.

9. Performing Organization Name and Address

University of Southwestern Louisiana
The Center for Advanced Computer Studies
P.O. Box 44330
Lafayette, LA 70504-4330

10. Work Unit No.

11. Contract or Grant No.

NGT-19-010-900

12. Sponsoring Agency Name and Address

13. Type of Report and Period Covered

FINAL; 07/01/85 - 12/31/87

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The USL NASA PC R&D statistical analysis support package is designed to be a three level package to allow statistical analysis for a variety of applications within the USL DBMS NASA Contract work. The design addresses usage of the statistical facilities as a library package, as an interactive statistical analysis system, and as a batch processing package.

This report represents one of the 72 attachment reports to the University of Southwestern Louisiana's Final Report on NASA Grant NGT-19-010-900. Accordingly, appropriate care should be taken in using this report out of the context of the full Final Report.

17. Key Words (Suggested by Author(s))

USL NASA PC R&D Statistical Analysis Package Specifications, PC-Based Research and Development

18. Distribution Statement

19. Security Classif. (of this report)

Unclassified

20. Security Classif. (of this page)

Unclassified

21. No. of Pages

14

22. Price*

*For sale by the National Technical Information Service, Springfield, Virginia 22161