DRAFT

CONF-800820--7
(Draft)

DATA ANALYSIS AND MANAGEMENT FOR THE URANIUM RESOURCE EVALUATION PROJECT

Victor E. Kane

MASTER

Mathematics and Statistics Research Department
Computer Sciences Division
Union Carbide Corporation, Nuclear Division
Oak Ridge, Tennessee 37830

Presented at
The American Statistical Association
Computing Section
Invited Paper Session
Managing Data For Energy and Environmental Research
Houston, Texas - August 11-14, 1980

UNION CARBIDE CORPORATION, NUCLEAR DIVISION
operating the
Oak Ridge Gaseous Diffusion Plant
Oak Ridge National Laboratory
Oak Ridge Y-12 Plant
Paducah Gaseous Diffusion Plant
for the
Department of Energy

ABSTRACT

The Department of Energy has funded a large data collection effort with the purpose of determining the U.S. uranium resources. This Uranium Resource Evaluation (URE) Project required a large data management effort which involved collection, retrieval, processing, display, and analysis of large volumes of data. Many of the characteristics of this data processing system are relevant to other applications, particularly where routine processing involves analyses for input into numerous technical reports. The URE Project computing system has a modular program structure which has enabled a straightforward interface with both special and general graphics and analysis packages such as SAS, BMDP, and SURFACE II. Other topics include cost-effective computing, data quality, report quality computer output, and test versus production program development.

Keywords: Computer processing, data collection, large data sets, data analysis, geographics, graphics.

1.0. INTRODUCTION

Computer processing of large data bases often involves special considerations particularly in a dynamic governmental program where timeliness, flexibility, cost-effectiveness, and data quality are fundamental considerations. The following is a description of the computing system developed to support the Union Carbide Corporation, Nuclear Division (UCC-ND), Uranium Resource Evaluation (URE) Project (see Begovich and Kane, 1980). The resulting URE Project Data Processing System (UREP-DPS) may be of interest to others who anticipate development of an automated system for computer processing. The experience with the UREP-DPS suggests that in developing a new system one should first carefully define the objectives and anticipated characteristics of the computer processing. Then a complete computing system should be designed where the greatest possible flexibility is maintained and the user directs standard program execution requests.

The National Uranium Resource Evaluation (NURE) Program was established by the U. S. Atomic Energy Commission, now the U. S. Department of Energy (DOE), in the spring of 1973. Participation by UCC-ND in the NURE Program began in the spring of 1975 (Arendt, et al, 1979). The objective of the NURE Program has changed from a generally broad systematic reconnaissance of the United States to the current objectives: (1) provide a comprehensive in-depth assessment of the nation's uranium resources for national energy planning, and (2) identify areas favorable for uranium resources. A NURE Program report covering uranium resource

assessment in 116 National Topographic Maps Series (NTMS) 1° x 2° (approximately 7,000 mi^2) quadrangles, which contain 100% of the currently estimated uranium reserves and probable potential resources, is targeted for 1980. The NURE Program is currently scheduled to continue the assessment process well into the 1980's.

2.0. OBJECTIVES AND CHARACTERISTICS OF URE PROJECT COMPUTER PROCESSING

A basic requirement of the URE Project was to provide DOE with technical reports consisting of various tabular and graphical data summaries and analyses for each of the assigned 1° x 2° NTMS quadrangles. A high degree of computerized automation was required since each report consisted of 1,000 to 2,000 samples with each sample having more than 50 measurements. A maximum of two to three reports per month were required to meet the reporting schedule. With these general guidelines, the objective of the computer support effort was formulated as: establish a comprehensive computer system to store, verify, and process data which enables timely and flexible computer processing in a cost-effective manner and ensures the greatest possible data quality while maintaining the security of the data base.

Meeting the objective required knowledge of the anticipated characteristics of the URE Project's data processing requirements. Some of these characteristics included:

1. The types of data collected will be both changing and expanding to meet changing DOE requirements and objectives.

2. A large number (100,000 to 200,000) of samples will be collected with each sample containing more than 50 measurements, some of which are common to all samples. Routine processing will typically involve a subset of 1,000 to 2,000 samples.

3. Data for a sample will arrive from multiple sources at different times.

4. A variety of data display and analysis programs will be necessary to prepare the required reports.

5. It will not be practical to use computer scientists for routine processing because of the large volume of data processing.

6. Multiple computer output devices will be necessary to enable:
   (a) report quality computer generated output,
   (b) plotting on different sized paper, mylar, or 35-mm film, and
   (c) microfiche output of data listings.

The computer support objective and the characteristics of URE Project data processing essentially defined many of the required components of the UREP-DPS.

3.0. DATA STORAGE AND ROUTINE PROCESSING

The three types of data collected for each sample are field site data, location data, and laboratory analytical data. Each data type is processed separately by highly automated systems. The data are then merged onto a URE Project master file.

The varying quantity of information associated with each sample (i.e., computer record) required a somewhat complex variable length record structure. While this record structure is complex, it enables efficient storage of data and flexibility in the amount of information associated with a particular sample. This flexibility has been repeatedly utilized during the course of the NURE Program. Prior to release of the data to the public the data are reformated to fixed length records which are inefficient for data storage, but are easily processed.

A subset of the master file, which is located on a computer tape, is placed on an on-line disk following each update of the master file. Note that placing the master file on magnetic tape enables an easy system of rotating tapes to provide a back-up of the master file in the event of computer system anomalies. The disk subset of the master file consists of samples which are not in archival storage. The archival storage consists of samples not being routinely processed and results in a reduction of over 40% in the size of the subset of the master file placed on disk. This archival procedure increases the efficiency in the processing of all URE Project programs. However, it is still possible to directly access the archival samples on magnetic tape, but increased processing time is necessary.

The disk subset of the master file is still large having more than 50,000 samples. Recall that routine processing uses subsets typically less than 2,000 samples. Creation of these subsets requires selection of samples based on numerous subsetting criteria. Thus, a PL/1 search

program was written to select samples based on Boolean logic combinations of various URE Project variables and to write the resulting subset on a mass storage unit. These direct access subfiles are labeled and have only the samples required for a particular area being studied. Additionally, the remarks section from the field form is omitted to shorten the computer record allowing for more compact storage. The above subsetting and subfile creation capability provides flexibility in processing in a cost-effective manner.

4.0. DATA DISPLAY AND ANALYSIS SYSTEM

In order to ensure timely and cost-effective data processing, it was necessary to design a user driven modular system. The basic components of the system are given in Figure 1. The subfile creation process was discussed in the previous section, the remaining components are discussed below.

4.1. USER DRIVEN SYSTEM

Each technical report produced by the URE Project requires approximately 60 to 100 separate executions of various computer programs from the initial data verification stage to the final production of report material. If 2 to 3 reports were to be output per month, a large program "set-up" effort could be expected simply to satisfy production requirements. It was also, of course, necessary to devote a portion of time to program maintenance and development. The only cost-effective option was to make standard requests for the execution of computer programs independent of computer personnel.

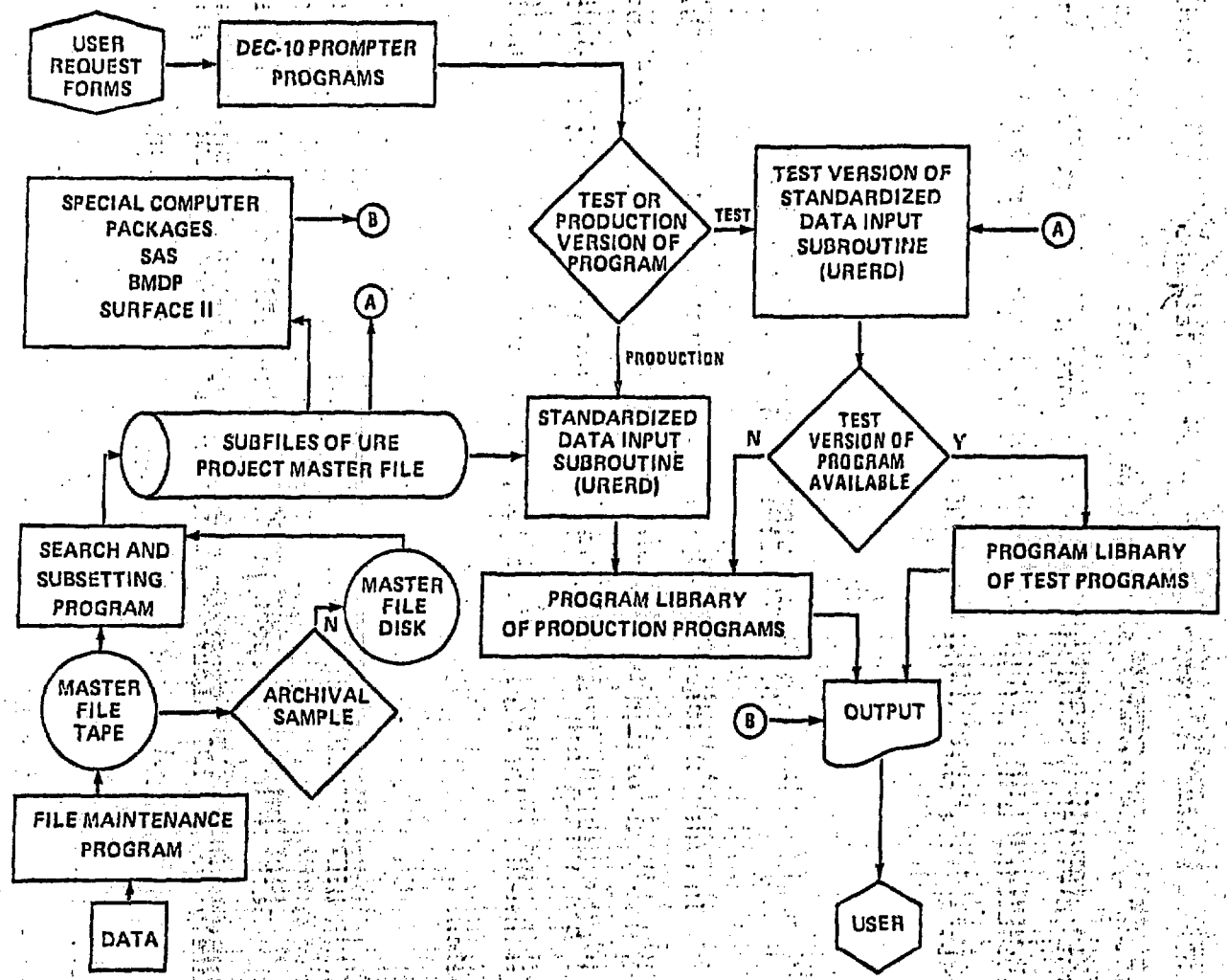

Figure 1

DATA PROCESSING SYSTEM FOR DATA DISPLAY AND ANALYSIS

The UREP-DPS was designed so that the user, generally a geologist, would complete a request form having questions concerning the desired output. Basically, these questions identify the particular program options that are desired. The completed request form is then given to nontechnical personnel who execute a DEC-10 (Digital Equipment Corp. DEC-system-10) prompter program using terminals located at the URE Project site. These programs ask for standard preliminary information such as subfile identifier, type of output, etc. and then prompt for answers to the questions on the request form. After all questions have been completed, the prompter program forms a file of 80 column card images with the appropriate options indicated. Also, the required job control language is formed so that the type of output devices are identified and the correct files read. The DEC-10 file is then submitted by two terminal commands from the DEC-10 computer to the IBM 360/195 computer. The output arrives at the URE Project site the following day.

The above system has great flexibility to use various computer output capabilities. It is cost-effective since it only involves nontechnical personnel and hence allows effective use of computer scientists' time for maintenance and development tasks. Also, reexecution of programs is minimized since program input is checked on entry and no alignment is necessary into traditional card columns.

4.2. STANDARDIZED DATA INPUT SUBROUTINE (URERD)

The URE Project master file structure is reasonably complex, having numerous computer variables and variable length records. Since there

were a number of computer programmers involved in the URE Project support effort, it was desirable to have a single subroutine to read the input data. This subroutine served to standardize the computer variable names (for over 200 variables) across programs. Also, the labels associated with the computer variables were standardized [e.g., "URANIUM(PPM)"] so that all computer printing and plotting labeled the variables in the same manner. Additionally, it was possible to create various mathematical combinations of quantitative variables to use in statistical analyses such as trend surface analysis, factor score plots, residual plots, etc. Finally, the treatment of censored data (concentrations below the laboratory detection limit) and missing data was standardized.

The resulting URERD subroutine enabled cost-effective use of available programming talent with minimal lost time frequently necessary for new personnel to become familar with a complex data structure. More importantly, new variables and other data related changes could be made to the system by simply changing the URERD subroutine rather than altering numerous individual programs. This capability has been repeatedly utilized by the URE Project and provides the necessary flexibility in handling the data. A final attribute of the modular data input design is that the system can be easily modified to read data from other sources in different data formats. All that is necessary is another data input subroutine, say URERD*, be used. Various identifiers would also be necessary for the prompter to call for linkage of URERD*, rather than URERD, during execution of the standard programs.

4.3. MODULAR PROGRAM DESIGN

Programs in the UREP-DPS are constructed in modules which are primarily FORTRAN subroutines. Essentially, these components are the building blocks of individual programs. There are three advantages to using the component approach. The first is that there are often certain common procedures that are necessary in two different programs. The best example is the URERD subroutine which reads the data and is common to all programs. Thus, once a subroutine module is developed and documented, it is possible for other programming efforts to use the routine. With a number of computer scientists involved in the UREP-DPS, the modular design has reduced program development time and hence reduced costs.

The second advantage of the subroutine structure is that it is possible to minimize the computer storage required for the execution of every program through use of the linkage editor overlay capability. Minimizing the storage space decreases the queueing time required for a program in the computer system. Thus, the user obtains his results faster. Overlaying only the URERD subroutine, after the data have been read, saves over 90K of core.

The third advantage of the modular design is that it enables development and modification of computer programs while the URE Project simultaneously runs production work using old versions of programs. Both test and production programs are part of the same system and accessible through a simple command from the DEC-10 prompter. Two factors necessitate the formal development of test and production versions of programs. First, the URE Project reporting schedule required use of programs in the UREP-DPS until a very short time before reports go to

reproduction. Delay of even a few days is often unacceptible. Second, it is extremely difficult to test programs on all possible combinations of data structures and program options. After thorough testing by a programmer, it is highly desirable for users to test programs for a period fo time. The dual testing has proven to be a very useful feature of the UREP-DPS.

Implementation of the test and production system is straightforward and simply requires two separate program libraries. If a test version of a program is desired, the DEC-10 prompter indicates the test library in the job control language. However, if no test version of a program is available, the default production library is automatically accessed. The DEC-10 prompter simply indicates the production library for the production version of a program. The changing data structures require a large number of changes to the URERD subroutine. The changes to this module can easily be tested by placing the test URERD module on the test library. Thus, flexibility and responsiveness to URE Program changes is attained while not impacting report schedules.

5.0. DATA QUALITY

The concern for the quality of the data reported to the public is central to the field collection, laboratory analysis, data processing, and data analysis efforts of the URE Project. Each sample collected in the NURE Program represents a relatively large geographic area and it is imperative that a sample properly reflect the environment from which it

is taken. It should be emphasized that even with the extensive effort and concern for the quality of the data, some errors are to be expected.

The field site data are collected on a form which has a "check-off" system of assigning categorical information. This format has a number of advantages of both encoding and decoding information. Preprinted computer assigned sample number labels are provided to the field samplers to minimize transcription problems. For all data entering the master file, standard checks of coded information against master lists are performed along with range checks of quantitative data. Additionally, a separate program performs checks of data fields just prior to public release of the data. The latitude/longitude coordinates of the sampling locations are determined from an automated digitization system that employs selective redigitization in a quality control effort.

Standard laboratory procedures are used for calibrating and monitoring instrumentation. Blind submission of samples from quality control standard batches is performed on a continuing basis. The UREP-DPS provides standard quality control charts and summaries of quality control data. After a data set is complete, multivariate outliers are identified using principal component procedures developed by Hawkins (1974), and Jackson and Mudholkor (1979). Approximately 1% of the total data set, selected from the list of outliers, are reanalyzed in the laboratory. Results from the reanalyses are examined to determine whether additional samples should be reanalyzed and if corrections to the original sample measurements are warranted.

6.0. DATA DISPLAY AND ANALYSIS PROGRAMS

Over 20 programs are routinely accessed by the UREP-DPS to perform: data verification, data quality determinations, tabular and graphical data displays, data summaries, measures of association between both chemical parameters and sampling stations, and data evaluations. Report quality computer listings and 35 mm film are often used as output media for report preparation.

Data analyses requiring specialized statistical knowledge arise occasionally and are performed by statisticians separate from the UREP-DPS. Procedures such as factor, discriminant, and regression analysis are processed by the Statistical Analysis System (SAS 1979) which accesses the URE Project subfiles. Occasionally, the Biomedical Computer Programs (Dixon and Brown 1979) and SURFACE II (Sampson 1975) are also used to perform special analyses. These two computer packages provide a broad spectrum of statistical analyses and enhance the data evaluation capabilities of the URE Project. A cluster analysis package, DENDRØ (Larson and Begovich, 1980), is used to perform various types of cluster analyses.

7.0. SUMMARY

The time and cost involved in developing a large computer analysis system necessitates careful initial planning. Topics such as timeliness, flexibility, cost-effectiveness, and data quality are fundamental concerns. However, experience with the development of the UREP-DPS suggests other important concerns are: user processing, subset creation, archival

storage, test versus production program execution, and integration of specialized computer packages. In any case, initial planning the design of the overall system is of paramount importance.

REFERENCES

1. Arendt, J. W., Butz, T. R., Cagle, G. W., Kane, V. E., and Nichols, C. E., *Hydrogeochemical and Stream Sediment Reconnaissance Procedures of the Uranium Resource Evaluation Project*, Union Caribde Corporation, Nuclear Division, Oak Ridge Gaseous Diffusion Plant, Oak Ridge, Tennessee, K/UR-100 (December 1979).

2. Begovich, C. L. and Kane, V. E., *Data Display and Analysis Programs in the URE Project Data Processing System*, Union Carbide Corporation, Nuclear Divison, Oak Ridge Gaseous Diffusion Plant, Oak Ridge, Tennessee, K/UR-45 (August 1980).

3. Dixon, W. J. and Brown, M. B., *BMDP-79 Biomedical Computer Programs P-Series*, University of California Press, Berkeley (1979).

4. Hawkins, D. M., "The Detection of Errors in Multivariate Data Using Principal Components," *Journal of the American Statistical Association, Vol. 69*, pp. 340-344 (1974).

5. Jackson, J. E. and Mudholkor, G. S., "Control Procedures for Residuals Associated with Principal Components," *Technometrics, Vol. 21*, pp. 341-349 (1979).

6. Larson, N. M. and Begovich, C. L., *A Revised User's Manual for the Hierarchical Cluster Analysis Code DENDRØ*, Union Carbide Corporation, Nuclear Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee (To Be Published).

7. Sampson, R. J., *Surface II Graphics System*, Kansas Geological Survey, Lawrence (1974).

8. *SAS User's Guide*, Statistical Analysis System Institute, Inc., Raleigh (1979).