188
4-28-82
ms

I-2774

(2)

DR 436

ORNL/CSD/TM-171

UCC-ND

NUCLEAR
DIVISION

A Data-Base Simulator

David J. Pack

DO NOT MICROFILM
COVER

MASTER

OPERATED BY
UNION CARBIDE CORPORATION
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A04; Microfiche A01

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL/CSD/TM--171

DE82 013342

ORNL/CSD/TM-171
Distribution Category UC-32

A DATA-BASE SIMULATOR

David J. Pack

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Date Published: March 1982

COMPUTER SCIENCES
at
Oak Ridge National Laboratory
Post Office Box X
Oak Ridge, Tennessee 37830

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
under Contract No. W-7405-eng-26
for the
Department of Energy

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGMENT	v
ABSTRACT	1
1. INTRODUCTION	1
2. DESCRIPTION OF THE TRUE VALUE GENERATION OPTIONS	4
2.1 Frequency Distribution	4
2.2 Discrete Conditional Frequency Distribution	4
2.3 Analysis of Variance Model	5
2.4 Multiple Regression Model	5
2.5 ARIMA Time Series Model	5
2.6 Membership Within a Defined Constrained Region	8
3. DESCRIPTION OF ERROR VALUE GENERATION OPTIONS	9
3.1 Independent of Error Values for Other Variables	9
3.2 Dependent Upon Error Values for Other Variables in a Specified Group	9
4. INPUT TO TRUE VALUES GENERATION	11
4.1 General	11
4.2 Frequency Distribution Generation	12
4.3 Conditional Frequency Distribution Generation	14
4.4 Analysis of Variance Model Generation	16
4.5 Multiple Regression Model Generation	17
4.6 ARIMA Time Series Model Generation	18
4.7 Constrained Region Generation	21

TABLE OF CONTENTS (Cont'd)

	<u>Page</u>
5. INPUT TO ERROR VALUES GENERATION	23
5.1 No Errors	24
5.2 Independent Errors	24
5.3 Dependent Errors	26
6. INPUT TO DETERMINE DESIRED OUTPUT OBSERVED VALUES	29
APPENDIX A - EXAMPLE INPUT/OUTPUT	31
APPENDIX B - DESCRIPTION OF COMPUTER PROGRAM STRUCTURE	45
APPENDIX C - COMMENTED COMPUTER PROGRAM LISTING	47

ACKNOWLEDGMENT

This work was supported by the Office of Energy Information Validation, Energy Information Administration, U. S. Department of Energy, as part of the Data Validation Research Program at Oak Ridge National Laboratory during the 1980 fiscal year.

ABSTRACT

This document describes the features of and input to a computer program written for the purpose of generating data bases whose data values contain deterministically known errors.

The development of the computer program was motivated by the need to assess automatic data editing procedures for data validation of real data bases. Given that a particular record in a data base violates one or more constraints, these procedures attempt to identify the fields in error and determine "true" values for those fields. Assessment of these procedures depends upon knowledge of the actual errors, knowledge which is rarely available to any meaningful extent with a real data base.

The observed values in the simulated data are the sum of generated "true" values and generated "error" values. For a given variable, "true" data values may be generated by any of the following six methods: frequency distribution, conditional frequency distribution, analysis of variance model, multiple regression model, ARIMA time series model, membership within a defined constrained region. The "error" values for a given variable may be simulated from an independent distribution or from a distribution dependent upon the error values of other specified variables.

The computer program described can be used to satisfy other needs in the area of data simulation beyond the specific need expressed above. Since the addition of errors to the true values is optional, one may readily simulate observed data for variables using one or more of the six previously listed methods. One could, for example, simulate observed data for a variable as the dependent variable in a multiple regression model where the observed data for the independent variable(s) was simulated using selected frequency distributions.

1. INTRODUCTION

This document describes the features of and input to a computer program written for the purpose of generating data bases containing deterministically known errors. The creation of "true values" through one of six fundamental generation options is followed by the creation of "error values" through one of two fundamental generation options and the resultant "observed values" as the sum of true and error.

This computer program was motivated by the need to be able to assess alternative procedures for error localization and true value imputation in the internal consistency checking phase of the validation of an actual data base. Such assessments must depend upon knowledge of the actual errors, knowledge which is rarely available to any meaningful extent with a real data base. The data validation need, however, is only one of many in the area of data simulation that can be satisfied by the program. The addition of errors to the true values is optional. Thus, the computer program described can serve as a general-purpose data simulator.

The fundamental generation options for true values are the following:

- Frequency distribution,
- Discrete conditional frequency distribution,
- Analysis of variance model,
- Multiple regression model,
- ARIMA time series model,
- Membership within a defined constrained region.

All variables' values must be generated from the program (i.e., there is no provision for input of values for selected variables). The discrete conditional frequency distribution, analysis of variance, and multiple regression generations always require the previous generation of one or more variables' values by some means. True value generation options are described verbally in Section 2 and in terms of detailed computer program inputs in Section 4.

The fundamental generation options for error values are:

- Independent of error values for other variables.
- Dependent upon error values for other variables in a specified group.

As noted above, it is not necessary to introduce error into a variable. This might be considered a third fundamental generation option for error values. Error value generation options are described verbally in Section 3 and in terms of detailed computer program inputs in Section 5.

Among the other features of the data base simulator computer program is a capability to produce a record in segments, or subrecords. This feature allows simulation of the data base structure where there is essentially a different data set for each of a sequence of time periods (such as years) where the same variables appear in each time period. The program also allows the user options on whether individual generated variables will be part of the final data base, and, if so, whether they will be treated as real or integer variables. These options on desired output observed values are described in Section 6.

Details of the computer program's overall utilization and structure are included in the appendices to this document. The program is written in standard FORTRAN and has run on an IBM 3033 machine using the following Job Control Language (JCL):

```
/33
//ABCEFGH JOB (12345), 'JOHN DOE BIN D'
//*CLASS CPU91=15S,I0=1.5,REGION=540K
// EXEC FORTHCLG,REGION.FORT=540K,REGION.GO=270K
//SYSPRINT DD DUMMY
//FORT.SYSIN DD *
=PROGRAM
//LKED.SYSLIB DD
//          DD
//          DD
// DD DSN=A.MTH10276.IMSLDP,DISP=SHR
//LKED.SYSIN DD *
/*
//GO.FT05F001 DD *
=DATA
/*
//
```

Note that IMSL Statistical Library routines are required.

The code assumes standard card-printer input-output devices. Appendix A gives an example of the overall input to and output from the program. Broad program structure is outlined in Appendix B. Appendix C contains a commented source code listing for the entire program.

2. DESCRIPTION OF THE TRUE VALUE GENERATION OPTIONS

This section of the document verbally describes the six fundamental generation options for generating the true values of the variables in the simulated data base.

2.1 Frequency Distribution

The user may choose to generate values from frequency distributions of three types:

- User-specified discrete distribution. User inputs number of values, the values themselves, and a probability for each value.
- Uniform distribution. User inputs end points of the interval for possible values.
- Normal distribution. User inputs mean and standard deviation of the values.

2.2 Discrete Conditional Frequency Distribution

This type of generation produces a variable with a finite number of discrete values which are conditioned upon the discrete values taken by another previously generated variable. For example, the newly generated variable may be 7 or 8 with chosen probabilities summing to 1 if the previously generated variable is 0, and the new variable may be 5 or 6 with chosen probabilities summing to 1 if the previous variable is 1.

2.3 Analysis of Variance Model

The analysis of variance model generation segments the variable's values into normally distributed groups with different means but the same standard deviation. Group membership is determined through the values of a previously generated discrete variable. The user inputs the desired number of groups, group means, common standard deviation, variable number and values of the previously generated variable to which group membership is tied.

2.4 Multiple Regression Model

The multiple regression model generation produces values of a variable acting as the dependent variable y in the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e \quad (2.1)$$

$x_1, x_2, \dots, x_k$ are k previously generated variables (the independent variables). The user inputs the variable numbers for the desired x 's, a β coefficient for each x , an intercept value β_0 , and a desired standard deviation for the model error term e .

2.5 ARIMA Time Series Model

The ARIMA time series model generation produces values of a variable in a given time period t as a linear function of values of the same variable in past time periods, error in period t , and error in past time periods. "ARIMA" stands for "autoregressive-integrated-moving average" — a particular class of time series models.

Let Z_t represent the variable's value in period t . If one wishes to generate a non-stationary sequence of Z_t 's, he may employ a model using differences of Z_t , such as $W_t = Z_t - Z_{t-1}$. The word "integrated" refers to the recreation of Z_t from W_t through "integration" — actually summation in the discrete environment. For a stationary sequence of Z_t 's, one may base the model on $W_t = Z_t - \mu$ where μ is the process mean level. Based on the appropriate W_t , the class of time series models utilized is

$$W_t - \phi_1 W_{t-1} - \dots - \phi_p W_{t-p} = a_t - \theta_1 a_{t-1} - \theta_q a_{t-q} \quad (2.2)$$

The left side of (2.2) contains p autoregressive terms with autoregressive parameters ϕ_j . The right side of (2.2) references random error entering the sequence at period $t(a_t)$, $t-1(a_{t-1})$, etc., and contains q moving average terms with moving average parameters θ_j . Defining the mathematical operator B to be such that $B^p W_t = W_{t-p}$ and $B^q a_t = a_{t-q}$, one may write (2.2) as

$$(1 - \phi_1 B - \dots - \phi_p B^p) (1 - B)^d Z_t = (1 - \theta_1 B - \dots - \theta_q B^q) a_t \quad (2.3)$$

where d is the number of differences applied to Z_t (Z_t is replaced by $Z_t - \mu$ if $d = 0$).

The time series models available in this generation program are a generalization of (2.3) taken from another computer program written by the author.¹ The (2.3) class of models potentially contains one

¹Pack, David J., "Computer Programs for the Analysis of Time Series Models Using the Box-Jenkins Philosophy." Distributed by Automatic Forecasting Systems, P.O. Box 563, Hatboro, Pennsylvania 19040. (May 1977).

autoregressive factor (parentheses containing ϕ_j), one difference factor (all d differences are based on subtraction between consecutive values), and one moving average factor (parentheses containing θ_j). If one allows for products of more than one of such factors, and a deterministic trend component θ_0 , the more general class of models is

$$\prod_{i=1}^{\text{NAR}} \phi_i(B) \prod_{j=1}^{\text{NDIFAC}} (1 - B^{s_j})^{d_j} Z_t = \theta_0 + \prod_{k=1}^{\text{NMA}} \theta_k(B) a_t, \quad (2.4)$$

where $\phi_i(B)$ and $\theta_k(B)$ are shortened notation for the parenthesized expressions in (2.3). Z_t values generated from (2.4) may have essentially any pattern of trend and/or seasonality.

User input to ARIMA time series model generation is difficult without past experience using these models. The number of factors of each type (NAR, NDIFAC, NMA in equation above) must be indicated, followed by (d_j, s_j) pairs, if any. Then the number of parameters for each factor, the parameter orders (associated time lags), and parameter initial estimates are supplied. An indicator of trend (θ_0) presence or absence follows, with initial estimate if present. There is clearly a startup problem in (2.4) when $t = 1, 2, 3, \dots$. Thus, initial series levels are generated for a sufficient number of periods from a uniform or normal distribution as specified by the user. Finally, the user must specify the standard deviation of the error term, a_t .

2.6 Membership Within a Defined Constrained Region

Constrained region generation allows the user to generate true values for a variable that lie within a region defined by the user's input of a theoretically unlimited number of constraints. Individual constraints may specify equality (EQ), an exclusive upper bound (LT), an inclusive upper bound (LE), an inclusive lower bound (GE), or an exclusive lower bound (GT). Combined individual constraints may bound the variable on both sides.

The user specifies that generated values are the boundary values plus deviations generated from a particular uniform or normal distribution. When the value is bounded on both sides, the parameters of the chosen distribution are determined from the bounds rather than from user input.

Individual constraints are entered in sequence with the EQ, LT, LE, GE or GT designations indicated above followed by mathematical expressions involving previously generated variables, constants, and the basic operations of addition (+), subtraction (-), multiplication (*), division (/), and exponentiation (**). Previously generated variable number x would be referred to as "Vx." Parentheses may be used as required. Examples of individual input constraints are

$$\begin{array}{ll} \text{GT}(\text{V5} * \text{V4} + \text{V2}/5 - \text{V4}) \\ \text{LT}(\text{V6} ** 1.2 + (\text{V3} - \text{V2})/4.8) \\ \text{LE } \text{V4} + \text{V7} + \text{V8} \end{array} \quad (2.5)$$

The 18 alphabetic characters recognized in the mathematical expressions are the ten digits, a period, the letter "V", opening and closing parentheses, and the symbols for mathematical operations +, -, *, /.

When a constraint is of the inclusive type and is acting alone, the user may also specify the proportion of generated values to be left on the included boundary.

3. DESCRIPTION OF ERROR VALUE GENERATION OPTIONS

There are two fundamental options for generating error values to be added to the true values to produce the observed values in the simulated data base (beyond the option of there being no errors). These options are verbally described in this section.

3.1 Independent of Error Values for Other Variables

Error values for a given variable may be generated independently of those generated for any other variable. The user may choose to generate values from frequency distributions of three types:

- User-specified discrete distribution,
- Uniform distribution,
- Normal distribution.

The chosen frequency distribution is adjusted by a provision of a discrete mass of zero error of the user's desired size. This provision is actually implemented through the user's input of a desired proportion of non-zero error.

3.2 Dependent Upon Error Values for Other Variables in a Specified Group

Error values for a given variable may alternatively be generated dependent upon error values for other variables in a user-specified group of variables. As in independent error generation, the user specifies a frequency distribution and a proportion of non-zero error.

Dependence is introduced in a uniquely defined way through the input of a sequence of conditional probabilities of error in variable number m in the group given the existing pattern of error in the group's first $(m-1)$ variables. Consider a group of three variables. Let E_j represent error in group variable j and $\bar{E}_j$ represent no error in group variable j . The following input is required:

Group Variable 1 - Frequency distribution, $P(E_1)$.

Group Variable 2 - Frequency distribution, $P(E_2)$, $P(E_2/E_1)$.

Note that this input determines $P(E_2/\bar{E}_1)$.

Group Variable 3 - Frequency distribution, $P(E_3)$, $P(E_3/E_1, E_2)$,

$P(E_3/\bar{E}_1, E_2)$, $P(E_3/E_1, \bar{E}_2)$. Note that this

input determines $P(E_3/\bar{E}_1, \bar{E}_2)$.

The general pattern is one of input of $(2^{m-1} - 1)$ conditional probabilities of error in variable m ($m > 1$) given various combinations of error or no error in the first $(m-1)$ variables. The "last" conditional probability, $P(E_m/\bar{E}_1, \dots, \bar{E}_{m-1})$, is fixed by the input of the unconditional $P(E_m)$ and the other conditional probabilities.

There are obvious difficulties associated with the input of the conditional probabilities. First, the input numbers may imply a negative "last" conditional probability, $P(E_m/\bar{E}_1, \dots, \bar{E}_{m-1})$. This will cause the program to stop execution after printing an error message. Second, one must follow a specific order in the entry of the $(2^{m-1} - 1)$ probabilities. This order is most easily specified in terms of 0-1 sequences of length $(m-1)$ (instead of $\bar{E}$ -E sequences). Expansion of these sequences with m

follows a clear pattern illustrated in Table 3.1. For a given m , the first $(m-2)$ columns of the first m rows are the entries under $(m-1)$ with the $(m-2)$ zeroes sequence added at the end and the $(m-1)^{\text{th}}$ column of these rows is always 1. The last $(m-1)$ rows are like the first $(m-1)$ except that the $(m-1)^{\text{th}}$ column of these rows is always 0.

Table 3.1. Conditional Probabilities Order of Entry

$m=2$	$m=3$	$m=4$
1	11	111
	01	011
	10	101
		001
		110
		010
		100

4. INPUT TO TRUE VALUES GENERATION

This section describes in detail the fixed-format card image input to true values generation. An initial section on general input is followed by separate sections on each of the six fundamental generation options. The reader needs to refer only to those sections corresponding to generation options he or she will utilize and the initial section on general input.

4.1 General

The initial general input concerns the number of records to be generated (NREC), the number of subrecords per record (NSREC) (e.g., years a data base might be divided into), and the number of variables per subrecord (NVAR). Titles for the general data base and the subrecords

(if more than 1) are also input. Numeric input should be right-justified within the columns allowed unless otherwise indicated.

1. $\overline{1} \overline{2} \dots \overline{29} \overline{30}$
 cc 1-10 Number of records (NREC).
 cc 11-20 Number of subrecords. Enter '1' if records not segmented (NSREC).
 cc 21-30 Number of variables to be generated per subrecord (NVAR).
2. One card with title describing general data base. Title may be typed anywhere within the 80 columns on the card.
3. (Required only if cc 11-20 of 1. above is > 1.)
Repeat 3. for each subrecord (NSREC times).
One card with title describing subrecord. Title may be typed anywhere within the 80 columns on the card.

4.2 Frequency Distribution Generation

The first variable(s) in a data base will often be generated from a selected frequency distribution since this generation option does not require the preceding generation of any other variables.

1. $\overline{1} \overline{2} \overline{3} \overline{4} \overline{5} \overline{6} \overline{7} \dots \overline{38} \overline{39} \overline{40}$
 cc 8-10 Variable number (NV).
 cc 13-20 Variable name with maximum of eight characters (VNAME).
 cc 30 Indicator for desired frequency distribution (INT). Use discrete distribution (INT = 1), uniform distribution (INT = 2), or normal distribution (INT = 3).

cc 39-40 Code as 'EQ' if want variable values equal on all sub-records. Code as 'GE' if want variable values generated separately for each subrecord. Ignore if number of sub-records NSREC = 1 (SRGS).

2. (Required only if cc 30 of 1. above is 1.)

1 2 9 10

cc 1-10 Number of possible values for variable. Maximum is 100 (NVA).

3. (Required only if cc 30 of 1. above is 1.)

1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the NVA possible values for variable. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (V).

4. (Required only if cc 30 of 1. above is 1.)

1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the NVA probabilities of the possible values for variable. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (P).

5. (Required only if cc 30 of 1. above is 2 or 3.)

1 2 19 20

cc 1-10 For uniform distribution (INT=2), the lower end point of value interval. For normal distribution (INT=3), the mean (P(1)).

cc 11-20 For uniform distribution (INT=3), the upper end point of value interval. For normal distribution (INT=3), the standard deviation (P(2)).

For both entries, no decimals are assumed if no decimal point punched.

4.3 Conditional Frequency Distribution Generation

Conditional frequency distribution generation must follow a preceding generation of values for one or more discrete variables to be used as the conditioning variable(s). The structure of the following input logically requires that the input stream be repeated two or more times in order that the generated variable has a value on every record.

1. $\overline{1} \overline{2} \overline{3} \overline{\frac{C}{4}} \overline{\frac{F}{5}} \overline{6} \overline{7} \dots \overline{38} \overline{39} \overline{40}$

cc 8-10 Variable number (NV).

cc 13-20 Variable name with maximum of eight characters (VNAME).

cc 29-30 Number of conditioning variables. Maximum is 20 (INT).

cc 39-40 Code as 'EQ' if want variable values equal on all sub-records. Code as 'GE' if want variable values generated separately for each subrecord. Ignore if number of sub-records NSREC=1 (SRGS).

2. $\overline{1} \overline{2} \dots \overline{9} \overline{10}$

cc 1-10 Number of discrete values allowed the generated variable during this input stream. Maximum sum of this number and numbers of values of conditioning variables is 100 (NVAL(1)).

3. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the NVAL(1) possible values for generated variable during this input stream. Use additional cards of more than eight values. No decimals are assumed if no decimal point punched (V).

4. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the NVAL(1) probabilities of the possible values for generated variable during this input stream. Use additional cards of more than eight values. No decimals are assumed if no decimal point punched (P).

5. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the INT variable numbers for the variables which serve as conditioning variables. Use additional cards if more than eight values (IV).

6. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the INT numbers of values involved in the conditioning for each conditioning variable. Use additional cards if more than eight values (NVAL(2), NVAL(3), etc.).

7. Repeat 7. for each conditioning variable (INT times).

$\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., for the j^{th} conditioning variable, enter the NVAL($j+1$) individual values to be conditioned upon. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (V).

4.4 Analysis of Variance Model Generation

Analysis of variance model generation is a special kind of conditional frequency distribution generation. Values are generated from normal distributions with the same standard deviation but different means that are conditional upon the value of one previously generated discrete variable. There is thus always one conditioning variable. The following input stream will be used only once.

1. $\overline{1} \overline{2} \overline{3} \overline{4}^A \overline{5}^N \overline{6} \overline{7} \cdot \cdot \cdot \cdot \overline{38} \overline{39} \overline{40}$
 cc 8-10 Variable number (NV).
 cc 13-20 Variable name with maximum of eight characters (VNAME).
 cc 29-30 Number of groups in the generation (i.e., number of different means). Maximum is 50 (INT).
 cc 39-40 Code as 'EQ' if want variable values equal on all subrecords. Code as 'GE' if want variable values generated separately for each subrecord. Ignore if number of subrecords NSREC=1 (SRGS).

2. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$
 In cc 1-10, 11-20, 21-30, etc., enter the INT desired group means. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (V).

3. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{9} \overline{10}$
 cc 1-10 The common group standard deviation for the normal distributions. No decimals are assumed if no decimal point punched (P(2)).

4. 1 2 9 10

cc 1-10 The variable number for the conditioning variable which determines group means (IV(1)).

5. 1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the INT values of the conditioning variable to be conditioned upon. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (V).

4.5 Multiple Regression Model Generation

Multiple regression model generation must always be preceded by the generation of values for the variables that serve as independent variables in the multiple regression model equation.

1. 1 2 3 R E 4 5 6 7 38 39 40

cc 8-10 Variable number (NV).

cc 13-20 Variable name with maximum of eight characters (VNAME).

cc 29-30 No. of independent variables. Maximum is 20 (INT).

cc 39-40 Code as 'EQ' if want variable values equal on all sub-records. Code as 'GE' if want variable values generated separately for each subrecord. Ignore if no. of sub-records NSREC=1 (SRGS).

2. 1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the INT variable numbers for the variables which serve as independent variables. Use additional cards if more than eight values (IV).

3. 1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the INT desired coefficients for the independent variables in the regression model. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (BETA).

4. 1 2 19 20

cc 1-10 The intercept parameter in the multiple regression equation (P(1)).

cc 11-20 The standard deviation of the multiple regression equation error term (P(2)).

For both entries, no decimals are assumed if no decimal point is punched.

4.6 ARIMA Time Series Model Generation

No other form of generation need precede the ARIMA time series model generation. The reader is referred to the long discussion of ARIMA time series models in Section 2 as an aid in fully understanding the following input.

1. 1 2 3 T S 6 7 38 G E
4 5

cc 8-10 Variable number (NV).

cc 13-20 Variable name with maximum of eight characters (VNAME).

cc 28-30 Number of time series observations per subrecord (if only one subrecord, the number of time series observations in entire series). Maximum back ordering (i.e., time lagging) in expanded model plus product of this number and number of subrecords must be 200 or less (INT).

2. $\overline{1} \overline{2} \dots \overline{29} \overline{30}$

cc 1-10 Number of autoregressive factors in model. This is 'NAR' of equation (2.4) (MFAC(1)).

cc 11-20 Number of differencing factors in model. This is 'NDIFAC' of equation (2.4) (MFAC(2)).

cc 21-30 Number of moving average factors in model. This is 'NMA' of equation (2.4) (MFAC(3)).

3. (Required only if cc 11-20 of 2. above is non-zero.)

$\overline{1} \overline{2} \dots \overline{??}$

cc 1-10 Number of differences of first type in model. This is d_1 of equation (2.4) (ND(1)).

cc 11-20 Order of differences of first type in model. This is s_1 of equation (2.4) (IOD(1)).

Repeat in cc 21-40 for second differencing factor if there is one (d_2, s_2) and in cc 41-60 for third differencing factor if there is one (d_3, s_3). (Maximum number of differencing factors is 3.)

4. (Required if sum of cc 1-10 and cc 21-30 of 2. above is non-zero.)

$\overline{1} \overline{2} \dots \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the (MFAC(1) + MFAC(3)) numbers indicating the number of individual parameters in each autoregressive and moving average factor (considering the factors in order from left to right). Maximum number of entries, i.e., maximum (MFAC(1) + MFAC(3)), is 6 (INC).

5. (Required if sum of cc 1-10 and cc 21-30 of 2. above is non-zero.)

1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter for each individual autoregressive and moving average parameter (i.e., INC(1) + INC(2) + ... separate entries) the order of the parameter (i.e., its associated time lag). Use additional card if more than eight entries. Maximum number of entries is 10 (IOPA).

6. (Required if sum of cc 1-10 and cc 21-30 of 2. above is non-zero.)

1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter a value for each individual autoregressive and moving average parameter (i.e., INC(1) + INC(2) + ... separate entries). Use additional card if more than eight entries. Maximum number of entries is 10. No decimals are assumed if no decimal point punched (UPA).

7. 1 2 9 10

cc 1-10 Enter a 0 to indicate no trend parameter in the time series model. Non-zero value indicates there is a trend parameter. The trend parameter is θ_0 of equation (2.4) (ITREND).

8. (Required if cc 1-10 of 7. above is non-zero.)

1 2 9 10

cc 1-10 Value for the trend parameter. No decimals are assumed if no decimal point punched (TREPA).

9. 1 2 9 10

cc 10 Indicator for desired frequency distribution of generated series initial levels (NDIST). Use uniform distribution (NDIST=2), or normal distribution (NDIST=3).

10. 1 2 19 20

cc 1-10 For uniform distribution (NDIST=2), the lower end point of value interval. For normal distribution (NDIST=3), the mean (P(3)).

cc 11-20 For uniform distribution (NDIST=2), upper end point of value interval. For normal distribution (NDIST=3), the standard deviation (P(4)).

For both entries, no decimals are assumed if no decimal point punched.

11. 1 2 9 10

cc 1-10 The standard deviation of the time series model error term. No decimals are assumed if no decimal point punched (P(2)).

4.7 Constrained Region Generation

Constrained region generation must follow other forms of generation if the constraints employed involve references to other variables, which will be the typical case. It is possible, however, that the constraints may contain only numerical constants, and in this case constrained region generation may precede any form of generation.

There is one aspect of constrained region generation that differs from the other fundamental generation options and should be carefully

noted. It concerns cases where the generated data base is to contain more than one subrecord. Values for the second and subsequent subrecords can only be generated by provision of an appropriate stream of input cards as described below. This is in contrast to the pattern of Sections 4.2-4.6 where an 'EQ' or 'GE' code in cc 39-40 of the first input card caused values to be generated for second and subsequent subrecords by either setting them equal to first subrecord values or generating them in the same manner as first subrecord values were generated.

1. $\overline{1} \overline{2} \overline{3} \overline{4} \overline{5} \overline{6} \overline{7} \cdot \cdot \cdot \cdot \cdot \overline{28} \overline{29} \overline{30}$

cc 8-10 Variable number (NV).

cc 13-20 Variable name with maximum of eight characters (VNAME).

cc 21-30 Number of constraints to be input to form constrained region for this particular generated variable. No maximum (INT).

2. Repeat 2. for each constraint (INT times).

$\overline{1} \overline{2} \cdot \cdot \cdot \cdot \cdot \overline{??}$

cc 1-10 In the case of a single inclusive type constraint (i.e., LE or GE), the proportion of generated values allowed on the constraint boundary. Leave blank in other cases. No decimals are assumed if no decimal point punched (PROP).

cc 19-20 Code as 'EQ' if equality constraint, as 'LT' if exclusive upper bound constraint, as 'LE' if inclusive upper bound constraint, as 'GE' if inclusive lower bound constraint, and as 'GT' if exclusive lower bound constraint (TG).

cc 21-?? The mathematical expression of the constraint utilizing 18 recognized alphabetic characters — the 10 digits, period, letter "V," opening and closing parentheses, and the symbols for mathematical operations +, -, *, /. See equation (2.5) for examples (C).

3. $\overline{1\ 2} \cdot \cdot \cdot \overline{9\ 10}$

cc 10 Indicator of desired frequency distribution of generated values from constrained region boundaries (NDIST). Use uniform distribution (NDIST=2), or normal distribution (NDIST=3).

4. (Required when constraint(s) bound(s) the generated variable below or above but not both.)

$\overline{1\ 2} \cdot \cdot \cdot \overline{19\ 20}$

cc 1-10 For uniform distribution (NDIST=2), the lower end point of value interval. For normal distribution (NDIST=3), the mean (P(1)).

cc 11-20 For uniform distribution (NDIST=2), the upper end point of value interval. For normal distribution (NDIST=3), the standard deviation (P(2)).

For both entries, no decimals are assumed if no decimal point punched.

5. INPUT TO ERROR VALUES GENERATION

This section describes in detail the fixed-format card image input to error values generation. An initial section on input when no errors are to be generated, i.e., the true values are also to be the observed

values in the data base, is followed by separate sections on each of the two fundamental generation options for error. The reader needs to refer only to those sections corresponding to generation options he or she will utilize. Input to error values generation immediately follows input to true values generation.

5.1 No Errors

If the program user wishes only to use one or more of the six fundamental generation options to generate true values without error, he or she still needs to provide some simple input to indicate that no errors are to be generated.

$$1. \quad \underline{1} \quad \underline{2} \quad \cdot \cdot \cdot \cdot \cdot \quad \underline{9} \quad \underline{0}$$

$$2. \quad \underline{1} \quad \underline{2} \quad \cdot \cdot \cdot \cdot \cdot \quad \underline{9} \quad \underline{0}$$

5.2 Independent Errors

A frequent assumption in the examination of errors in several variables is that the errors in one variable occur independently of the errors in other variables. The first input to independent error generation (1. below) is the number of variables for which independent errors are to be generated. The rest of the input (2. through 6. below) must be repeated for each of these variables, conditional on the type of frequency distribution used for error generation.

1. $\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{9} \overline{10}$

cc Number of variables for which independent errors are to
be generated (NIEV).

2. Repeat 2.-6. for each variable with independent error (NIEV times).

$\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{29} \overline{30}$

cc 1-10 Variable number (NV).

cc 20 Indicator for desired frequency distribution of generated
error (NDIST). Use discrete distribution (NDIST=1), uni-
form distribution (NDIST=2), or normal distribution
(NDIST=3).

cc 21-30 Proportion of non-zero error for this variable. No deci-
mals are assumed if no decimal point punched (PROP).

3. (Required only if cc 20 of 2. above is 1.)

$\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{9} \overline{10}$

cc 1-10 Number of possible error values. Maximum is 100 (NVA).

4. (Required only if cc 20 of 2. above is 1.)

$\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the NVA possible values for
error. Use additional cards if more than eight values. No decimals
are assumed if no decimal point punched (V).

5. (Required only if cc 20 of 2. above is 1.)

$\overline{1} \overline{2} \cdot \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., enter the NVA probabilities of the
possible values for error. Use additional cards if more than eight
values. No decimals are assumed if no decimal point punched (P).

6. (Required only if cc 20 of 2. above is 2 or 3.)

1 2 19 20

cc 1-10 For uniform distribution (NDIST=2), the lower end point of value interval. For normal distribution (NDIST=3), the mean (P(1)).

cc 11-20 For uniform distribution (NDIST=2), the upper end point of value interval. For normal distribution (NDIST=3), the standard deviation (P(2)).

For both entries, no decimals are assumed if no decimal point punched.

5.3 Dependent Errors

Dependence among errors in variables in a given group is introduced through the input of a sequence of conditional probabilities of error in variable m in the group given the existing pattern of error in the group's first $(m-1)$ variables. For the first variable in a group only the unconditional probability of error is entered. For the m^{th} variable, $m > 1$, $(2^{m-1} - 1)$ conditional probabilities must be entered. The pattern of input of these conditional probabilities was explained in detail in Section 3.

The first input to dependent error generation (1. below) is the number of groups of variables to be considered. The rest of the input (2. through 8. below) must be repeated for each of these groups. The second input (2. below) is the number of variables in a given group. The rest of the input (3. through 8. below) must be repeated for each of these variables.

1. $\overline{1} \overline{2} \cdot \cdot \cdot \overline{9} \overline{10}$

cc 1-10 Number of separate groups of variables to be considered (NDEG).

2. Repeat 2. - 8. for each separate group (NDEG times).

$\overline{1} \overline{2} \cdot \cdot \cdot \overline{9} \overline{10}$

cc 1-10 Number of variables in given group (NVIG).

3. Repeat 3. - 8. for each variable in given group (NVIG times).

$\overline{1} \overline{2} \cdot \cdot \cdot \overline{29} \overline{30}$

cc 1-10 Variable number (NV).

cc 20 Indicator for desired frequency distribution of generated error (NDIST). Use discrete distribution (NDIST=1), uniform distribution (NDIST=2), or normal distribution (NDIST=3).

cc 21-30 Proportion of non-zero error for this variable (i.e., unconditional probability of error). No decimals are assumed if no decimal point punched (PR(JJ)).

4. (Not required for the first variable in given group.)

$\overline{1} \overline{2} \cdot \cdot \cdot \overline{??}$

In cc 1-10, 11-20, 21-30, etc., for the m^{th} variable in given group ($m > 1$), enter the $(2^{m-1} - 1)$ conditional probabilities of error in variable m given various possible patterns of error in first $(m-1)$ variables. (See Section 3 for discussion of specific order of entry.) Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (PR).

5. (Required only if cc 20 of 3. above is 1.)

1 2 9 10

cc 1-10 Number of possible error values. Maximum is 100 (NVA).

6. (Required only if cc 20 of 3. above is 1.)

1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the NVA possible values for error. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (V).

7. (Required only if cc 20 of 3. above is 1.)

1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the NVA probabilities of the possible values for error. Use additional cards if more than eight values. No decimals are assumed if no decimal point punched (P).

8. (Required only if cc 20 of 3. above is 2 or 3.)

1 2 19 20

cc 1-10 For uniform distribution (NDIST=2), the lower end point of value interval. For normal distribution (NDIST=3), the mean (p(1)).

cc 11-20 For uniform distribution (NDIST=2), the upper end point of value interval. For normal distribution (NDIST=3), the standard deviation (P(2)).

For both entries, no decimals are assumed if no decimal point punched.

6. INPUT TO DETERMINE DESIRED OUTPUT OBSERVED VALUES

The concluding program input gives the user the flexibility to say that any generated variable may be ignored in outputting observed values, or that a generated variable may be output as either a real variable (continuous) or an integer variable (discrete). This final input immediately follows input to error values generation.

1. 1 2 ??

In cc 1-10, 11-20, 21-30, etc., enter the NVAR (the last number on the very first input in Section 4.1) indicators of how individual variables are to be treated in output observed data base. Indicator assumes values of 0, 1, 2 with following pattern:

0 - ignore this variable,

1 - real variable,

2 - integer variable.

Use additional cards if more than eight values (IE).

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

APPENDIX A

EXAMPLE INPUT/OUTPUT

The latter pages of this Appendix provide an example of the complete input to and partial output from utilizing the data base simulator program to simulate one data base. This example will be discussed in terms of separate segments of input, as described in Sections 4, 5, and 6 of this document.

Input to True Values Generation

1. One hundred records are to be generated. Each record will be divided into three subrecords. Each subrecord will contain 18 variables. Thus, there are 54 variables to be dealt with in total.
2. Data base title.
3. Title for each subrecord.
4. Variable 1, named SAMPLEID, is generated from a uniform distribution on the interval [1,1000]. Because of the 'EQ' code, Variable 19 and Variable 37 will have the same values as Variable 1.
5. Variable 2, named FUEL, is generated from a discrete distribution with five possible integer values 1, 2, 3, 4, 5 with probabilities .1, .2, .4, .25, .05, respectively. Variables 20 and 38 will have the same values as Variable 2.
6. Variable 3, named CAPACITY, is generated from a normal distribution with mean 75,000 and standard deviation 15,000. Variables 21 and 39 will have the same values as Variable 3.
7. Variable 4, named TYPEPLT, is partially generated via a conditional frequency distribution. In this partial generation it will be given

the single value 1 with probability 1 if Variable 2 takes any of the three values 1, 2, 3. Variables 22 and 40 will have the same values as Variable 4.

8. Variable 4 generation is completed by assigning the single value 2 with probability 1 if Variable 2 takes either of the two values 4, 5. Variables 22 and 40 will have the same values as Variable 4.
9. Variable 5, named ANOVAV, is generated via an analysis of variance type model. Values will be chosen from a normal distribution with mean 100 or mean 200, standard deviation 15, where the mean utilized depends on whether Variable 4 takes the value 1 or 2. Because of the 'GE' code, Variable 23 and Variable 41 will have values generated in the same manner as Variable 5's (but the values usually will not equal those of Variable 5).
10. Variable 6, named CONSUM, is generated via a multiple regression model. Values will be .6 times the value of the one independent variable, Variable 3, plus an intercept of 0, and an error from a normal distribution (automatic) with standard deviation 2000 (mean zero automatic). Variables 24 and 42 will have values generated in the same manner.
11. Variables 7-12 are generated as six observations of a time series in the first subrecord. They will appear with the names STOCK1, STOCK2, ..., STOCK6. The ARIMA time series model used in the generation contains one autoregressive factor, one difference factor, one moving average factor. The single difference has order 6. The model contains one autoregressive parameter and one moving average parameter, each of them having order 1, the first taking a

value .7000 and the second taking a value .4000. There is no trend parameter. Thus, the model is

$$(1 - .7B)(1 - B^6)Z_t = (1 - .4B)a_t$$

in the notation of Section 2. Initial levels are generated from a uniform distribution on the interval [100,400]. The error term, a_t , has standard deviation 5 (automatically normal with mean zero). Variables 25-30 and 43-48 are continuations of the time series on the second and third subrecords. (This continuation is automatic, i.e., use of 'EQ' in place of 'GE' will not produce the expected result.)

12. Variables 13, 31, 49, named TOTSTOCK, are produced in a constrained region in an equivalent manner. They must be specified separately, however, since the 'EQ' and 'GE' codes are not operative in constrained region generation. In this case, each variable is the sum of six previously generated variables plus a normally distributed deviation with mean 100 and standard deviation 10 that will result in all of the generated values being different from the original sum of six variables.
13. Variables 14, 32, 50, named GEN, are each constrained to be less than a single specified previously generated variable. Deviations from the constraint boundary will be generated from a uniform distribution on the interval [5000,10000]. No generated values will be left on the boundary.
14. Variables 15, 33, 51, named CON15, are each constrained to be between upper and lower bounds which are each functions of previously

generated variables. Placement of generated values between the bounds will be accomplished with a normal distribution (user input "3") whose parameters will depend on the actual bounds in a particular case.

15. Variables 16, 34, 52, named CON16, are each constrained to be less than the minimum of two functions of previously generated variables. Deviations from the constraint boundary will be generated from a uniform distribution on the interval [100,1000]. No generated values will be left on the boundary.
16. Variables 17, 35, 53, named CON17, are each constrained to be greater than a function of three previously generated variables. Deviations from the constraint boundary will be generated from a normal distribution with mean 25 and standard deviation 5. No generated values will be left on the boundary.
17. Variables 18, 36, 54, named CON18, are each constrained to be greater than or equal to zero. Deviations from the constraint boundary will be generated from a normal distribution with mean 50 and standard deviation 10. A proportion .400 of the generated values will be left on the boundary, i.e., will be equal to zero.

Output of True Values Generation

1. Generated true values for all variables generated, by record number, by subrecord. The illustrated output is a partial output for this example showing 18 variables for 26 of the 100 record numbers for the first of three subrecords (year 1977).

Input to Error Values Generation

1. Three variables are to have error values generated with a presumption of independence of error between the variables.
2. Variable 1 is to have non-zero errors in 20% of the records. The errors will be generated from a normal distribution with mean 50 and standard deviation 20.
3. Variable 3 is to have non-zero errors in 10% of the records. The errors will be generated from a normal distribution with mean 500 and standard deviation 100.
4. Variable 2 is to have non-zero errors in 30% of the records. The errors will be generated from a discrete distribution with two possible values, 1 and 2, with probabilities .700 and .300, respectively.
5. One group of variables is to have error values generated with a presumption of dependence of error between the variables in the group.
6. The group of variables contains three variables.
7. The first variable in the group is Variable 6. This variable is to have non-zero errors in 10% of the records. The errors will be generated from a normal distribution with mean 2000 and standard deviation 200.
8. The second variable in the group is Variable 7. This variable is to have non-zero errors in 20% of the records. The conditional probability of Variable 7 error given Variable 6 error is .500. The errors will be generated from a normal distribution with mean 100 and standard deviation 20.

9. The third variable in the group is Variable 13. This variable is to have non-zero errors in 30% of the records. The conditional probabilities of Variable 13 error are the following: given error in both Variables 6 and 7, 1.000, given error in Variable 6 only (first in group), .500, given error in Variable 7 only (second in group), 1.000. The errors will be generated from a normal distribution with mean 300 and standard deviation 50.

Output of Error Values Generation

1. Generated error values for all variables generated (zeroes assumed for variables not mentioned in error value generation input), by record number, by subrecord. The illustrated output is a partial output for this example showing 18 variables for 26 of the 100 record numbers for the first of three subrecords (year 1977).

Input to Determine Desired Output Observed Values

1. For the 18 variables in each subrecord, it is indicated that Variables 5, 11, 12, and 17 are to be dropped out in the final output (input 0), Variables 1, 2, and 4 are to be output as integer values (input 2), and all other variables are to be output as real values (input 1).

Output of Observed Values

1. Observed values for the 14 variables whose output was requested, by record number, by subrecord. The illustrated output is a partial output for this example showing 14 variables for 26 of the 100 record numbers for the first of three subrecords (year 1977). Note variables SAMPLEID, FUEL, and TYPEPLT are output as integer values.

 REPRODUCTION OF INPUT LEADING TO GENERATED TRUE VALUES

100	3	19	GENERATED DATA BASE NUMBER ONE					1
								2
								3
								4
								5
								6
								7
								8
								9
								10
FD 1	SAMPLEID	2	EQ					
1.000	1000.000							
FD 2	FUEL	1	EQ					
5								
1.000	2.000	3.000	4.000	5.000				
0.100	0.200	0.400	0.250	0.050				
FD 3	CAPACITY	3	EQ					
75000.000	15000.000							
CF 4	TYPEPLT	1	EQ					
1								
1.000								
1.000								
2								
3								
1.000	2.000	3.000						
CF 4	TYPEPLT	1	EQ					
1								
2.000								
1.000								
2								
2								
4.000	5.000							
AN 5	ANOVAV	2	GE					
100.000	200.000							
15.000								
4								
1.000	2.000							
RE 6	CONSUM	1	GE					
3								
0.600								
0.0	2000.000							

TS	7	STOCK	6	GE
	1		1	
	1		6	
	1		1	
	1		1	
	0.700	0.400		
	0			
	2			
	100.000	400.000		
	5.000			
VC	13	TOTSTOCK	1	
	0.0	EQV7+V8+V9+V10+V11+V12		
	3			
	100.000	10.000		
VC	31	TOTSTOCK	1	
	0.0	EQV25+V26+V27+V28+V29+V30		
	3			
	100.000	10.000		
VC	49	TCTSTOCK	1	
	0.0	EQV43+V44+V45+V46+V47+V48		
	3			
	100.000	10.000		
VC	14	GEN	1	
	0.0	LTV3		
	2			
	5000.000	10000.000		
VC	32	GEN	1	
	0.0	LTV21		
	2			
	5000.000	10000.000		
VC	50	GEN	1	
	0.0	LTV39		
	2			
	5000.000	10000.000		
VC	15	CON15	2	
	0.0	GT(V14-2000)		
	0.0	LTV3+4000		
	3			
VC	33	CON15	2	
	0.0	GT(V32-2000)		
	0.0	LTV21+4000		
	3			
VC	51	CON15	2	
	0.0	GT(V50-2000)		
	0.0	LTV39+4000		
	3			

11

12

13

14

VC	16	CON16	2
0.0		LT(V6**1.2+V6)	
0.0		LTV3	
100.000	2		
VC	34	CON16	2
0.0		LT(V24**1.2+V24)	
0.0		LTV21	
100.000	2		
VC	52	CON16	2
0.0		LT(V42**1.2+V42)	
0.0		LTV39	
100.000	2		
VC	17	CON17	1
0.0		GT(V5*V4+V2/5-V4)	
25.000	3		
VC	35	CON17	1
0.0		GT(V23*V22+V20/5-V22)	
25.000	3		
VC	53	CON17	1
0.0		GT(V41*V40+V38/5-V40)	
25.000	3		
VC	18	CON18	1
0.400		GE0.0	
50.000	3		
VC	36	CON18	1
0.400		GE0.0	
50.000	3		
VC	54	CON18	1
0.400		GE0.0	
50.000	3		
50.000	10.000		

15

16

17

39

TRUE VALUE PRINTOUT
GENERATED DATA BASE NUMBER ONE
YEAR 1977

	SAMPLEID	FUEL	CAPACITY	TYPEPLT	ANOVA	CONSUM	STOCK1	STOCK2	STOCK3	STOCK4
1	0.10815E 03	0.20000E 01	0.87329E 05	0.10000E 01	0.71634E 02	0.56599E 05	0.65163E 02	0.95610E 02	0.93859E 02	0.35850E 02
2	0.59425E 03	0.30000E 01	0.48725E 05	0.10000E 01	0.10252E 03	0.28880E 05	0.24533E 03	0.17036E 03	0.15443E 03	0.14716E 03
3	0.81552E 03	0.50000E 01	0.90036E 05	0.20000E 01	0.20683E 03	0.54873E 05	0.28955E 03	0.32695E 03	0.34567E 03	0.30864E 03
4	0.34207E 03	0.40000E 01	0.67213E 05	0.20000E 01	0.20031E 03	0.41515E 05	0.27720E 03	0.24184E 03	0.26640E 03	0.26875E 03
5	0.43179E 02	0.30000E 01	0.87599E 05	0.10000E 01	0.11696E 03	0.52795E 05	0.34145E 03	0.34796E 03	0.31902E 03	0.32162E 03
6	0.61707E 03	0.50000E 01	0.87057E 05	0.20000E 01	0.21696E 03	0.51498E 05	0.14545E 03	0.12623E 03	0.80117E 02	0.76349E 02
7	0.69972E 03	0.20000E 01	0.92803E 05	0.10000E 01	0.87607E 02	0.54080E 05	0.26158E 03	0.23910E 03	0.26923E 03	0.31026E 03
8	0.16292E 03	0.50000E 01	0.88316E 05	0.20000E 01	0.19296E 03	0.51407E 05	0.39053E 03	0.36233E 03	0.33953E 03	0.41916E 03
9	0.78868E 02	0.20000E 01	0.73439E 05	0.10000E 01	0.11488E 03	0.45642E 05	0.30592E 03	0.28468E 03	0.32439E 03	0.27863E 03
10	0.32415E 02	0.20000E 01	0.90312E 05	0.10000E 01	0.12610E 03	0.60270E 05	0.16080E 03	0.20488E 03	0.16511E 03	0.18630E 03
11	0.51296E 03	0.20000E 01	0.68187E 05	0.10000E 01	0.89650E 02	0.39241E 05	0.36390E 03	0.34053E 03	0.40445E 03	0.39240E 03
12	0.82881E 02	0.30000E 01	0.73365E 05	0.10000E 01	0.89861E 02	0.43696E 05	0.30643E 03	0.29454E 03	0.32807E 03	0.25953E 03
13	0.54955E 03	0.10000E 01	0.66690E 05	0.10000E 01	0.87855E 02	0.38870E 05	0.29951E 03	0.26019E 03	0.21132E 03	0.23105E 03
14	0.77643E 03	0.30000E 01	0.73577E 05	0.10000E 01	0.11314E 03	0.42891E 05	0.41648E 03	0.37201E 03	0.34951E 03	0.40162E 03
15	0.63861E 03	0.40000E 01	0.70821E 05	0.20000E 01	0.21254E 03	0.40068E 05	0.29660E 03	0.30899E 03	0.26920E 03	0.30555E 03
16	0.12438E 02	0.30000E 01	0.60240E 05	0.10000E 01	0.12027E 03	0.36516E 05	0.27896E 03	0.27402E 03	0.29538E 03	0.27564E 03
17	0.43173E 03	0.30000E 01	0.53014E 05	0.10000E 01	0.12463E 03	0.33484E 05	0.18594E 03	0.20446E 03	0.24778E 03	0.21964E 03
18	0.49645E 03	0.30000E 01	0.69709E 05	0.10000E 01	0.10109E 03	0.38905E 05	0.22669E 03	0.18868E 03	0.15248E 03	0.18712E 03
19	0.32875E 03	0.30000E 01	0.74047E 05	0.10000E 01	0.85067E 02	0.40770E 05	0.14217E 03	0.19386E 03	0.20225E 03	0.18848E 03
20	0.70169E 01	0.30000E 01	0.80469E 05	0.10000E 01	0.11419E 03	0.48775E 05	0.16388E 03	0.16565E 03	0.14103E 03	0.16176E 03
21	0.22857E 03	0.40000E 01	0.60713E 05	0.20000E 01	0.18333E 03	0.35903E 05	0.42603E 03	0.42723E 03	0.37725E 03	0.39640E 03
22	0.53372E 03	0.50000E 01	0.72255E 05	0.20000E 01	0.21120E 03	0.44052E 05	0.17498E 03	0.14067E 03	0.18499E 03	0.19860E 03
23	0.33658E 03	0.20000E 01	0.70211E 05	0.10000E 01	0.10886E 03	0.45163E 05	0.13765E 03	0.14259E 03	0.18333E 03	0.11527E 03
24	0.70202E 03	0.30000E 01	0.71850E 05	0.10000E 01	0.10547E 03	0.42322E 05	0.20766E 03	0.21377E 03	0.24805E 03	0.27565E 03
25	0.87120E 03	0.30000E 01	0.84626E 05	0.10000E 01	0.14541E 03	0.52083E 05	0.35566E 03	0.40325E 03	0.38161E 03	0.38982E 03
26	0.32936E 02	0.30000E 01	0.74137E 05	0.10000E 01	0.91148E 02	0.46008E 05	0.17685E 03	0.11758E 03	0.15478E 03	0.14160E 03

	STOCK5	STOCK6	TOTSTOCK	GEN	CON15	CON16	CON17	CON18
1	0.16233E 03	0.83051E 02	0.53586E 03	0.80673E 05	0.83675E 05	0.86598E 05	0.96951E 02	0.49318E 02
2	0.20384E 03	0.18274E 03	0.11038E 04	0.42164E 05	0.44112E 05	0.48393E 05	0.11979E 03	0.0
3	0.25815E 03	0.25567E 03	0.17846E 04	0.80108E 05	0.84546E 05	0.89059E 05	0.44429E 03	0.57604E 02
4	0.25862E 03	0.27755E 03	0.15904E 04	0.59865E 05	0.61347E 05	0.66976E 05	0.41877E 03	0.50156E 02
5	0.38273E 03	0.33080E 03	0.20436E 04	0.80318E 05	0.83289E 05	0.87017E 05	0.15120E 03	0.0
6	0.10544E 03	0.13490E 03	0.66848E 03	0.78216E 05	0.80492E 05	0.86545E 05	0.45917E 03	0.63993E 02
7	0.25326E 03	0.24261E 03	0.15761E 04	0.85547E 05	0.91886E 05	0.92461E 05	0.11395E 03	0.32875E 02
8	0.35760E 03	0.38863E 03	0.22578E 04	0.82899E 05	0.88495E 05	0.87602E 05	0.41502E 03	0.47947E 02
9	0.28269E 03	0.34684E 03	0.18232E 04	0.64397E 05	0.71323E 05	0.72982E 05	0.13352E 03	0.0
10	0.19131E 03	0.21166E 03	0.11201E 04	0.81398E 05	0.86117E 05	0.89983E 05	0.15312E 03	0.0
11	0.39716E 03	0.36251E 03	0.22610E 04	0.61881E 05	0.64837E 05	0.67923E 05	0.11220E 03	0.0
12	0.33030E 03	0.31084E 03	0.18301E 04	0.64071E 05	0.70288E 05	0.72912E 05	0.11054E 03	0.71238E 02
13	0.27978E 03	0.24748E 03	0.15293E 04	0.57100E 05	0.62216E 05	0.66270E 05	0.10986E 03	0.0
14	0.38534E 03	0.37051E 03	0.22958E 04	0.68323E 05	0.71239E 05	0.73462E 05	0.13909E 03	0.53299E 02
15	0.28970E 03	0.27617E 03	0.17466E 04	0.62583E 05	0.69169E 05	0.70055E 05	0.44843E 03	0.65360E 02
16	0.26511E 03	0.33231E 03	0.17214E 04	0.52467E 05	0.57168E 05	0.59508E 05	0.14381E 03	0.47680E 02
17	0.18523E 03	0.16411E 03	0.12075E 04	0.47062E 05	0.51187E 05	0.52136E 05	0.15178E 03	0.29446E 02
18	0.22862E 03	0.21085E 03	0.11944E 04	0.62510E 05	0.67407E 05	0.68807E 05	0.12812E 03	0.57666E 02
19	0.15925E 03	0.15873E 03	0.10447E 04	0.66660E 05	0.69122E 05	0.73444E 05	0.11302E 03	0.50435E 02
20	0.13393E 03	0.17210E 03	0.93834E 03	0.75379E 05	0.78930E 05	0.80125E 05	0.13612E 03	0.0
21	0.43241E 03	0.41929E 03	0.24786E 04	0.53686E 05	0.57213E 05	0.60287E 05	0.39423E 03	0.47543E 02
22	0.14981E 03	0.16729E 03	0.10164E 04	0.65875E 05	0.70753E 05	0.71865E 05	0.44043E 03	0.54041E 02
23	0.14180E 03	0.16467E 03	0.88530E 03	0.64437E 05	0.70448E 05	0.69954E 05	0.13160E 03	0.34288E 02
24	0.21498E 03	0.28189E 03	0.15060E 04	0.66018E 05	0.68796E 05	0.71313E 05	0.12748E 03	0.61374E 02
25	0.33913E 03	0.37608E 03	0.22456E 04	0.76034E 05	0.81988E 05	0.83745E 05	0.16809E 03	0.0
26	0.13075E 03	0.13846E 03	0.86002E 03	0.67785E 05	0.72550E 05	0.73447E 05	0.11948E 03	0.0

 REPRCOUCTION OF INPUT LEADING TO GENERATED ERRORS

				1
3				2
1	3	0.200		3
50.000	20.000			4
3	3	0.100		5
500.000	100.000			6
2	1	0.300		7
2				8
1.000	2.000			9
0.700	0.300			
1				
3				
6	3	0.100		
2000.000	200.000			
7	3	0.200		
0.500				
100.000	20.000			
13	3	0.300		
1.000	0.500	1.000		
300.000	50.000			

ERROR VALUE PRINTOUT
GENERATED DATA BASE NUMBER ONE
YEAR 1977

	SAMPLEID	FUEL	CAPACITY	TYPEPLT	ANOVA	CONSUM	STOCK1	STOCK2	STOCK3	STOCK4
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.15408E 04	0.73335E 02	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.86560E 02	0.0	0.0	0.0
4	0.26152E 02	0.10000E 01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.83833E 02	0.0	0.0	0.0
6	0.0	0.10000E 01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.30535E 02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.20000E 01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.20000E 01	0.0	0.0	0.0	0.0	0.10768E 03	0.0	0.0	0.0
14	0.0	0.0	0.43820E 03	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.0	0.10000E 01	0.0	0.0	0.0	0.0	0.10926E 03	0.0	0.0	0.0
18	0.0	0.10000E 01	0.41614E 03	0.0	0.0	0.17872E 04	0.0	0.0	0.0	0.0
19	0.67474E 02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	0.54356E 02	0.10000E 01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.10000E 01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.71646E 02	0.20000E 01	0.47207E 03	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0	0.20014E 04	0.11728E 03	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.0	0.10000E 01	0.0	0.0	0.0	0.0	0.74964E 02	0.0	0.0	0.0
	STOCK5	STOCK6	TOTSTOCK	GEN	CON15	CON16	CON17	CON18		
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
2	0.0	0.0	0.30376E 03	0.0	0.0	0.0	0.0	0.0		
3	0.0	0.0	0.26031E 03	0.0	0.0	0.0	0.0	0.0		
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
5	0.0	0.0	0.30443E 03	0.0	0.0	0.0	0.0	0.0		
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
13	0.0	0.0	0.32384E 03	0.0	0.0	0.0	0.0	0.0		
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
16	0.0	0.0	0.34272E 03	0.0	0.0	0.0	0.0	0.0		
17	0.0	0.0	0.33191E 03	0.0	0.0	0.0	0.0	0.0		
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
24	0.0	0.0	0.30840E 03	0.0	0.0	0.0	0.0	0.0		
25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
26	0.0	0.0	0.34254E 03	0.0	0.0	0.0	0.0	0.0		

```

*****
REPRODUCTION OF INPUT LEADING TO OUTPUT OBSERVED VALUES
*****
      2      2      1      2      0      1      1      1      1
      1      1      0      0      1      1      1      1      1
      0      1      0      0      1      1      1      1      1
*****

```

OBSERVED VALUE PRINTOUT
GENERATED DATA BASE NUMBER ONE
YEAR 1977

SAMPLEID	FUEL	CAPACITY	TYPELT	CONSUM	STOCK1	STOCK2	STOCK3	STOCK4	TOTSTOCK
1	108	2 0.87329E 05		1 0.56599E 05	0.65163E 02	0.95610E 02	0.93859E 02	0.35850E 02	0.53586E 03
2	594	3 0.48725E 05		1 0.30421E 05	0.31867E 03	0.17036E 03	0.15443E 03	0.14716E 03	0.14076E 04
3	815	5 0.90036E 05		2 0.54873E 05	0.37611E 03	0.32695E 03	0.34567E 03	0.30864E 03	0.20449E 04
4	368	5 0.67213E 05		2 0.41515E 05	0.27720E 03	0.24184E 03	0.26640E 03	0.26875E 03	0.15904E 04
5	43	3 0.87599E 05		1 0.52795E 05	0.42529E 03	0.34796E 03	0.31902E 03	0.32162E 03	0.23420E 04
6	617	6 0.87057E 05		2 0.51498E 05	0.14545E 03	0.12623E 03	0.80117E 02	0.76349E 02	0.66848E 03
7	699	2 0.92803E 05		1 0.54080E 05	0.26158E 03	0.23910E 03	0.26923E 03	0.31026E 03	0.15761E 04
8	162	5 0.88316E 05		2 0.51407E 05	0.39053E 03	0.36233E 03	0.33953E 03	0.41916E 03	0.22578E 04
9	78	2 0.73439E 05		1 0.45642E 05	0.30592E 03	0.28468E 03	0.32439E 03	0.27863E 03	0.18232E 04
10	62	2 0.90312E 05		1 0.60270E 05	0.16080E 03	0.20488E 03	0.16511E 03	0.18630E 03	0.11201E 04
11	512	4 0.68187E 05		1 0.39241E 05	0.36390E 03	0.34053E 03	0.40445E 03	0.39240E 03	0.22610E 04
12	82	3 0.73365E 05		1 0.43696E 05	0.30643E 03	0.29454E 03	0.32807E 03	0.25993E 03	0.18301E 04
13	549	3 0.66690E 05		1 0.38870E 05	0.40719E 03	0.26019E 03	0.21132E 03	0.23105E 03	0.18532E 04
14	776	3 0.74015E 05		1 0.42891E 05	0.41648E 03	0.37201E 03	0.34951E 03	0.40192E 03	0.22958E 04
15	638	4 0.70821E 05		2 0.40068E 05	0.29660E 03	0.30899E 03	0.26920E 03	0.30595E 03	0.17466E 04
16	12	3 0.60240E 05		1 0.36516E 05	0.27896E 03	0.27402E 03	0.29538E 03	0.27564E 03	0.20641E 04
17	431	4 0.53014E 05		1 0.33484E 05	0.29520E 03	0.20446E 03	0.24778E 03	0.21994E 03	0.15354E 04
18	496	4 0.70125E 05		1 0.40692E 05	0.22669E 03	0.18868E 03	0.15248E 03	0.18712E 03	0.11944E 04
19	396	3 0.74047E 05		1 0.40770E 05	0.14217E 03	0.19386E 03	0.20225E 03	0.18848E 03	0.10447E 04
20	7	3 0.80469E 05		1 0.48775E 05	0.16388E 03	0.16565E 03	0.14103E 03	0.16176E 03	0.93834E 03
21	282	5 0.60713E 05		2 0.35903E 05	0.42603E 03	0.42723E 03	0.37725E 03	0.39640E 03	0.24786E 04
22	533	6 0.72255E 05		2 0.44052E 05	0.17498E 03	0.14067E 03	0.18499E 03	0.19860E 03	0.10164E 04
23	408	4 0.70683E 05		1 0.45163E 05	0.13765E 03	0.14259E 03	0.18333E 03	0.11527E 03	0.88530E 03
24	702	3 0.71850E 05		1 0.44324E 05	0.38804E 03	0.21377E 03	0.24805E 03	0.27656E 03	0.18144E 04
25	871	3 0.84626E 05		1 0.52083E 05	0.35566E 03	0.40325E 03	0.38161E 03	0.38982E 03	0.22456E 04
26	32	4 0.74137E 05		1 0.46008E 05	0.25181E 03	0.11758E 03	0.15478E 03	0.14160E 03	0.12026E 04

GEN CON15 CON16 CON18

1	0.80673E 05	0.83675E 05	0.86598E 05	0.49318E 02
2	0.42164E 05	0.44112E 05	0.48393E 05	0.0
3	0.80108E 05	0.84546E 05	0.89059E 05	0.57604E 02
4	0.59865E 05	0.61347E 05	0.66976E 05	0.50156E 02
5	0.80318E 05	0.83289E 05	0.87017E 05	0.0
6	0.78216E 05	0.80492E 05	0.86545E 05	0.63993E 02
7	0.85547E 05	0.91886E 05	0.92461E 05	0.32875E 02
8	0.82859E 05	0.88495E 05	0.87602E 05	0.47947E 02
9	0.64357E 05	0.71323E 05	0.72982E 05	0.0
10	0.81358E 05	0.86117E 05	0.89983E 05	0.0
11	0.61881E 05	0.64837E 05	0.67923E 05	0.0
12	0.64071E 05	0.70288E 05	0.72912E 05	0.71238E 02
13	0.57100E 05	0.62216E 05	0.66270E 05	0.0
14	0.68323E 05	0.71239E 05	0.73462E 05	0.53299E 02
15	0.62583E 05	0.69169E 05	0.70055E 05	0.65360E 02
16	0.52467E 05	0.57168E 05	0.59508E 05	0.47680E 02
17	0.47062E 05	0.51187E 05	0.52136E 05	0.29446E 02
18	0.62510E 05	0.67407E 05	0.68807E 05	0.57666E 02
19	0.66660E 05	0.69122E 05	0.73444E 05	0.50435E 02
20	0.75379E 05	0.78930E 05	0.80125E 05	0.0
21	0.53686E 05	0.57213E 05	0.60287E 05	0.47543E 02
22	0.65875E 05	0.70753E 05	0.71865E 05	0.54041E 02
23	0.64437E 05	0.70448E 05	0.69954E 05	0.34288E 02
24	0.66018E 05	0.68796E 05	0.71313E 05	0.61374E 02
25	0.76034E 05	0.81988E 05	0.83745E 05	0.0
26	0.67785E 05	0.72550E 05	0.73447E 05	0.0

APPENDIX B

DESCRIPTION OF COMPUTER PROGRAM STRUCTURE

The subprogram structure of the data base simulator computer program is diagrammed below. The six fundamental generation options for true values are centered in the subprograms FDIST, CFD (both discrete conditional frequency distribution and analysis of variance model), REGRES, TSERIE, and RAIN. Subprograms GGUBFS and GGNQF are part of the IMSL Statistical Library, generating uniform random deviates and normal random deviates, respectively. All other subprograms are single-purpose utility routines called by only one subprogram. Individual subprogram purposes are explained in depth in the commented listing of Appendix C.

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

APPENDIX C
COMMENTED COMPUTER PROGRAM LISTING

The microfiche attached on page 51 of this report contains the commented computer program listing for the data base simulator computer program.

Subprograms are listed in the following order:

MAIN
CFD
EQUATE
FDIST
RAINT
RNUM
VALUE
REGRES
TSERIE
EQCOEF
MULTBP
NCHOSE

Subprograms MAIN, RRAINT, and TSERIE are commented throughout. Other subprograms are commented only at the beginning — to define purpose and arguments.

ORNL/CSD/TM-171

A DATA BASE SIMULATOR
David J. Pack
MICROFICHE ATTACHED