METHODS & DESIGNS

ALICE: A system for manipulating and analyzing multidimensional data

EDWARD WALKER and JOSEPH A. BAUER, JR.

Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

and

MICHAEL L. GRUBIN

PRIME Computer, Inc., Framingham, Massachusetts 01701

The computer program ALICE solves the two major problems of data manipulation and analysis. First, ALICE allows the user to treat data from an experiment in the form they are generated. Second, mathematical calculations and statistical analyses are included as an intrinsic part of the multidimensional approach to data handling. ALICE accepts raw data in the form and order they were collected; reorganizes, partitions, or selects any subset of them (including a single entry), and arithmetically combines, transforms, or evaluates any formula involving them. Furthermore, learning to use ALICE is simple, even for those who are naive to both computers and data analysis.

Anyone who has performed a simple analysis of a small set of data appreciates the utility of a computational aid, be it a hand calculator, a computer, a pencil and paper, or an assistant. Perhaps this utility is made most obvious by the ubiquity of such aids and the frequency with which they are used. Once the means for computing is at hand, the next most onerous aspect of data manipulation is arranging and numerically converting the data into the form required by the operations to be performed. For example, the wide variety of statistical subroutines available in the BMDP library of programs satisfies almost every requirement for analyzing data, but the subroutines in turn require that the data entries be in a particular form and arrangement. This constraint is common to almost every statistics package or subroutine library.

In creating ALICE, we have removed both of the common problems of data analysis by providing ALICE users not only with the capacity to perform almost any numerical evaluation of the data, but also with the capability for preparing any set of data for evaluation, starting with the data in any form convenient to the user. In principle, ALICE can address any subset of the data and perform any analysis or manipulation which can be expressed as a formula. Furthermore, we have tried to make ALICE readily accessible to new users by insuring that the instruction set and the method of

describing the data are as intuitively obvious as possible.

CHARACTERIZATION OF THE DATA

The data are entered as a continuous stream of numbers which can occur in any order that is convenient to the user. Although some arrangement of the data must be supplied as a starting point, a new description of the data matrix may be specified at any time, and any subset of the data (including a single entry) can be selected or rearranged with respect to other subsets. Therefore, the initial configuration of the data is arbitrary, so far as the program is concerned. In other words, the data may be entered in the form and order they were collected and brought into the form desired for analysis or output by using ALICE instructions.

In the terminology of experimental design, an ALICE dimension is equivalent to a factor, a source of variance, or an independent variable, and the levels of an ALICE dimension are equivalent to the levels of a factor. For example, the data from an experiment in which the heartbeats of six rats were counted after each had received one of three dosages of a drug would be characterized as having two dimensions, dosages and rats. There would be three levels of the dosage dimension and six levels of the rat dimension. The graphic device employed in the ALICE User's Guide would characterize the data from such an experiment as shown in Figure 1. Each datum can be described uniquely by referring to that level of each dimension which contains it. The datum representing the heartbeats of the fifth rat to

The development of ALICE was supported generously by grants held by Professor Richard M. Held [5 R01 MH07642 (1968-1973) and 1 R01 EYO 01101 (1973-1976)]. Recent work on the program has also been supported by a grant held by Professor Merrill Garrett [5 R01 MH07642 (1970-1977)].

Dosage	saline						low						high					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Rat																		
Heart beats	7 0	7 6	6 5	8 8	7 8	9 3	7 9	8 0	7 4	9 0	7 0	9 6	8 3	8 4	6 3	8 7	7 0	9 4

Figure 1. A two-dimensional matrix with heartbeats for six rats receiving three dosages of one drug. Note that the data are arranged in a continuous stream.

receive dosage 2 (70) occupies the second level of the dosage dimension and the fifth level of the rat dimension.

While it is important in performing an analysis to know whether the same or different rats received each dosage, the characterization of the number and arrangement of the data entries shown in Figure 1 applies to all experimental designs, because the same number of data entries are recorded whether or not the design is "nested" or "crossed."

This method of representing the data, unlike many other devices for representing experimental designs, can be expanded readily to include any number of dimensions and any number of levels in each dimension. A particular datum or subset of data can always be referenced easily and conveniently by specifying the level or levels of each dimension which contain that subset. If the experiment described had included three dosages of two drugs, and the pulse had been measured both before and after administration of the drugs, then the description of the expanded matrix would be as shown in Figure 2. In this matrix, the datum 74 (17th from the left) occupies the first level of the drug dimension, the third level of the rat dimension, the second level of the time dimension, and the second level of the dosage dimension. Since any subset of the data can be selected by specifying the dimension and level numbers which refer to it, subsets of the data can be rearranged or operated on simply by specifying what operation

is to be performed and the dimension and level numbers of the subset of data on which the operation is to be done.

MATRIX MANIPULATION

ALICE includes a number of "matrix manipulation" instructions which allow the user to bring the data entries into whatever form and arrangement is desired. These instructions are used to interchange dimensions (DINT), to select any or all of the levels of a dimension in the same or a new order (LSEL), to delete any number of levels from a dimension (LDEL), to interchange levels within a dimension (LINT), to eliminate boundaries between adjacent dimensions (FORG), or to redescribe the matrix parameters without rearranging the entries themselves (RNMP). Those who are familiar with the properties of experimental data will recognize that these instructions make it easy to change the hierarchical arrangement of variables in an experiment, to reorder randomly arranged data, to select or delete particular entries or subsets of data, and to change the characterization of the experimental design from one manipulation or analysis to another.

The most far-reaching matrix manipulation instruction is the one which permits the interchange of adjacent dimensions (DINT). The effect of this manipulation is difficult to appreciate without visual aids. Figure 3

Drug Type	D1																							
	R1				R2				R3				R4				R5				R6			
	B		A		B		A		B		A		B		A		B		A		B		A	
	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H
Heart beats	6	6	7	7	7	8	6	6	8	7	8	8	6	6	7	6	7	6	7	6	8	9	8	7
	6	8	1	0	9	3	8	6	0	6	0	4	7	6	2	5	4	3	1	5	5	8	0	7
	9	4	6	8	0	0	3	8	3	3	6	4	9	9	9	9	9	9	9	9	9	9	9	9

Drug Type	D2																							
	R1				R2				R3				R4				R5				R6			
	B		A		B		A		B		A		B		A		B		A		B		A	
	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H
Heart beats	6	6	7	8	8	9	6	6	7	8	8	9	6	7	8	7	7	6	6	6	7	8	7	7
	4	8	0	2	4	4	0	2	4	4	8	4	4	4	4	2	4	4	0	8	6	4	0	8
	4	8	0	2	4	4	0	2	4	4	8	4	4	4	4	2	4	4	0	8	6	4	0	8
	4	8	0	2	4	4	0	2	4	4	8	4	4	4	4	2	4	4	0	8	6	4	0	8

Figure 2. A four-dimensional matrix expanded from the two-dimensional matrix shown in Figure 1 by including Drug Type and Time dimensions. Page limitations prevent arranging the data in a continuous stream, as they would be in ALICE.

Rats	R1								R2								R3												
Drug Type	D1				D2				D1				D2				D1				D2								
Time	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A							
Dosage	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H					
Heart beats	6	6	7	7	7	8	6	6	7	8	8	9	6	6	8	7	8	8	6	6	7	8	8	9	6	6	7	7	
	6	8	1	0	9	3	4	8	0	2	4	4	8	6	0	6	0	4	0	2	4	4	8	4	7	6	2	5	4

Rats	R4								R5								R6											
Drug Type	D1				D2				D1				D2				D1				D2							
Time	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A				
Dosage	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	S	L	H	
Heart beats	7	6	6	8	9	8	6	6	6	7	8	7	7	7	6	7	7	7	7	6	6	9	9	8	6	6	7	9
	1	5	5	8	0	7	0	8	6	4	0	8	9	4	6	8	0	0	4	0	8	8	4	0	3	8	3	3

Figure 3. Figure 2 redrawn to show Dimension 1 (Drug Type) and Dimension 2 (Rats) interchanged by a single DINT instruction. Note the rearrangement of the data entries.

presents the data of Figure 2, with the rat and drug dimensions interchanged. It is easy to imagine that it might be more convenient to collect the data from the hypothetical experiment in the form shown in Figure 2 or in Figure 3, or in some other configuration altogether. In any case, ALICE can accept the data in any form that is convenient for collection and rearrange it to any form that is desired for tabular output or analysis. In practice, some forethought about collecting and recording the data can often simplify the matrix manipulation required, but ALICE can accept even randomly arranged data as a starting point and reorganize them into the form desired for a particular statistical analysis or for conveniently readable tabular output.

ALICE operations are performed on a "current" copy of data matrix. If an intermediate form of the data is to be subjected to more than one further manipulation or analysis, this form of the data matrix can be held in a "buffer" matrix to which the current matrix can be reset at any time. In addition, the data are retained in a "permanent" matrix, which is not altered unless specifically requested by an instruction option. This form of the data can be brought back to the current and buffer matrices at any time. Thus, any number of different manipulations or analyses of a set of data can be performed on the same data in one ALICE run.

CONVERSION AND REDUCTION OF THE DATA

Besides the capacity to rearrange and reorganize the data, some ability to transform or arithmetically combine data entries is often required—conversion from rectangular to polar coordinates, reduction of scores for two conditions to differences, calculation of a mean from all or some entries, etc. ALICE provides single

instructions which permit arithmetic, trigonometric, and statistical (mean correction, conversion to variances, etc) transformations of any subset of the data, again specified by dimension and level number. Arithmetic operations can be performed on the data using externally supplied constants, or the data entries themselves may be arithmetically combined or reduced to means, variances, and so forth. A complete list of ALICE instructions appears at the end of this article. These instructions give ALICE the capacity to evaluate any function of the data which can be expressed by a formula.

OUTPUT AND SINGLE INSTRUCTION STATISTICS

The capacity to completely rearrange and convert or reduce data would be useless without the capacity to generate output or perform statistical analyses based on the results of all this manipulation. In both cases, ALICE's capacity for rearrangement has special benefits. In the first place, the data may be written or punched in any form that is convenient for reading or for further analysis.

At any point in an ALICE run, the data may be written in tabular form or punched in their current configuration and form. Similarly, a table listing the data screened for entries which lie outside a specified range (and identified by their matrix parameters), or a table of the data arranged in ascending or descending rank order and in order of occurrence, can be printed.

Common statistical tests are included as single ALICE instructions which cause a summary of the data and the results of the tests to be printed. These instructions include t tests of correlated or uncorrelated samples and for differences from zero. Pearson product-moment

r correlation values are included as part of the output from the t test for correlated samples. An analysis of variance can be performed on any hierarchical, factorial, or mixed design having equal or unequal numbers of observations, using a single simple instruction. Polynomial curve fits of up to ninth order can be performed, and a Pearson r for actual and predicted y values, r^2 goodness of fit statistic, and the coefficients of the appropriate equation printed.

In principle, any test that can be expressed by a formula can be evaluated with a series of ALICE instructions. The User's Guide includes a special section covering the use of ALICE instructions to calculate various nonparametric and parametric statistical tests which are not included as single instructions. The capacity to construct difference scores, arithmetically combine or transform the data, rank data entries, and tally data of particular values makes possible the computation of almost any analysis.

The capacity for matrix manipulation, especially dimension interchange, allows the form of the statistical instructions to be particularly simple and direct. Since the data matrix can be reorganized easily, these instructions can assume a standard matrix configuration. For example, the t tests assume that at the time the test is applied the repeated measure over which the test is done is the last dimension in the matrix, and the samples to be compared are in the next-to-last dimension. Because a standard arrangement of the data is assumed, only the design of the experiment (correlated or uncorrelated samples, or difference from zero in the case of t tests) needs to be specified by the user to perform a particular test.

A standard form of the data matrix can also be assumed in constructing sequences of ALICE instructions. Consequently, if the same analysis or manipulation is to be applied to several sets of data, it can be written once in a general form, and merely duplicated and inserted at the appropriate point in each ALICE run.

DESCRIPTION OF THE PROGRAM AND USER'S GUIDE

ALICE is written in FORTAN IV and consists of approximately 3,500 card images. Versions are currently available for the IBM 360/70, DEC 10, and PRIME family of minicomputers and will be supplied through the authors for distribution costs (approximately \$50).

The ALICE User's Guide presents an introduction to the elementary concepts of experimental design and their application to multidimensional data handling. In addition, the User's Guide contains a complete description of all ALICE instructions, as well as numerous examples in which ALICE is used to analyze data from representative experiments. The User's Guide is also available from the authors.

CONCLUSION

ALICE provides researchers who (should) use computers to manipulate and analyze data with the capacity to enter raw data in the form collected, reorganize the data matrix, mathematically convert the data, arithmetically combine subsets of the data, statistically reduce the data, perform standard statistical analyses, and process data for submission to other programs or for tabular output. In other words, ALICE reduces the problem of analyzing data to its one irreducible aspect: the need to know *what* should be done.

We have found that ALICE is easy to use and that people who are naive to data analysis and the use of computers quickly learn to use ALICE. Several researchers in our department who did not participate in the program's development use ALICE, and we have conducted seminars in which we were able to teach new users to analyze their own data with ALICE after only 6 h of instruction. We are particularly struck by the rapidity with which new ALICE users are able to master the problem of computation and are compelled by using the program to turn to questions of experimental design, the effects of transforming data, and the interpretation of results. In fact, the instructions and method of describing the data matrix have proved to be intuitively understandable and usable to experimenters from a selection of disciplines whose variety astounds us: sensorimotor coordination, psycholinguistics, developmental psychology, electrophysiology, retinoscopy, pharmacology, aphasia, psychoacoustics, etc.

We have compared ALICE's computation of statistical tests with a number of published examples, and we have checked ALICE's performance against our own hand calculations. By both comparisons, ALICE is not only easy to use and indefatigable, but also accurate.

APPENDIX A: THE SET OF ALICE INSTRUCTIONS

Conversion

These instructions perform arithmetic conversions or algebraic transformations on the subset of the data specified by the Dimension (D) and Level (L) numbers.

ABSV	D, L, L, L:	Absolute value
SINE	D, L, L, L:	Sine of angle
COSN	D, L, L, L:	Cosine of angle
TANG	D, L, L, L:	Tangent of angle
ARCS	D, L, L, L:	Arccosine
ARCC	D, L, L, L:	Arccosine
ARCT	D, L, L, L:	Arctangent
LNAT	D, L, L, L:	Natural Log (base e)
LTEN	D, L, L, L:	Log (base 10)
ALBE	D, L, L, L:	Natural Anti-log (base e)
ALBT	D, L, L, L:	Anti-log (base 10)
ADDC	D, M, L, L, L:	Add Constant(s)
Constants		(SF16.6)
SUBC	D, M, L, L, L:	Subtract Constant(s)
Constants		(SF16.6)
MULC	D, M, L, L, L:	Multiply by Constant(s)
Constants		(SF16.6)
DIVC	D, M, L, L, L:	Divide by Constant(s)
Constants		(SF16.6)
STOR	M:	Store in Temporary Area
POWR	D, L, L, L:	Raise to Integer Power
Power		(I4)
ROOT	D, L, L, L:	Extract Integerth Root
Root		(I4)
SBST	D, L, L, L:	Substitute Value 2 for Value 1 with Range
Value 1	Value 2	Range
		(3F8.3)

MISS	D:	Missing Data Mean Substitution
Value	Range	(2F8.3)
MCOR	D:	Arithmetic Mean Correction
GCOR	D:	Geometric Mean Correction
ZCOR	D:	Z-Score Conversion

Reordering

These instructions reorder the data according to the Dimension and Level numbers specified.

DINT	D&D	Interchange Adjacent Dimensions
LINT	D,L&L:	Interchange Levels
LSEL	D,L,L,L:	Level Selection
LDEL	D,L,L,L:	Level Deletion

Matrix Description Operations

These instructions change the parameters of the data matrix.

FORG	D:	Forget Dimension Boundary
RNMP	:	Read New Matrix Parameters
	Name=nl,name=nl	

Intermatrix Operations

These instructions copy data from one matrix into another.

HOLD	:	C-Matrix→B-Matrix
BACK	:	P-Matrix→B-Matrix→C-Matrix

Derivation/Reduction

These instructions yield a new matrix derived arithmetically or statistically from the matrix preceding the instruction.

LADD	D,L+L,L+L,L+L:	Level Addition
LSUB	D,L-L,L-L,L-L:	Level Subtraction
LMUL	D,L*L,L*L,L*L:	Level Multiplication
LDIV	D,L/L,L/L,L/L:	Level Division
PAMN	D,L,L,L:	Partial Arithmetic Mean
PGMN	D,L,L,L:	Partial Geometric Mean
PPRD	D,L,L,L:	Partial Product
PSUM	D,L,L,L:	Partial Sum
PTAL	D,L,L,L:	Partial Tally of Entries Not Equal to Value
Value	Range	(2F8.3)

PVAR	D,L,L,L:	Partial Variance
PSTD	D,L,L,L:	Partial Standard Deviation
TOTL	D:	Total
MEAN	D:	Mean
GEOM	D:	Geometric Mean
MEDN	D:	Median
VARN	D:	Variance
STAN	D:	Standard Deviation

Output

These instructions write out the matrix or the results of a statistical procedure performed on the matrix.

RNTC	:	Read New Title
Title		
LIST	N:	Primary Output (Print)
PNCH	N,SU,BL:	Alternate Output (Punch)
(Format)		
SCRN	D,L,L,L:	Screen Data (Range Check)
Min	Max	(2F8.3)
POLY	N M:	Polynomial Curve Fit
TCOR	N:	T test for Correlated Means
TUNC	N:	T test for Uncorrelated Means
TZER	N:	T test for Mean Diff. from Zero
RKLS	N:	Rank in Descending Order
RKSL	N:	Rank in Ascending Order
ANOV	N,ND,PR,FR,RE:	Analysis of Variance
UNEN	N,ND,PR,FR:	Unequal N Analysis of Variance
Name=nl,Name=nl, . . .		Matrix Parameters
s1 s2 s3 . . .		# of Subjects/Group (2014)

Program Control

These instructions tell the program to proceed to the next ALICE run or stop execution.

NEXT	:	Begin Next ALICE Run
STOP	:	Terminate Execution

(Received for publication May 11, 1976;
accepted May 11, 1976.)