NASA CONTRACTOR REPORT

NASA CR-1385

MANIPULATION ERRORS IN FINITE ELEMENT ANALYSIS OF STRUCTURES

by R. J. Melosh and E. L. Palacol

Prepared by
PHILCO-FORD CORPORATION
Palo Alto, Calif.
for Goddard Space Flight Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • AUGUST 1969

NASA CR-1385

MANIPULATION ERRORS IN FINITE ELEMENT ANALYSIS OF STRUCTURES

By R. J. Melosh and E. L. Palacol

Prepared under Contract No. NAS 5-10369 by
PHILCO-FORD CORPORATION
Palo Alto, Calif.

for Goddard Space Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – CFSTI price $3.00

C O N T E N T S

MANIPULATION ERRORS IN FINITE ELEMENT ANALYSIS OF STRUCTURES

By R. J. Melosh* and E. L. Palacol**

SUMMARY

The finite element concept provides the basis for numerical analysis of structures. Implementation of analyses of large practical problems using this concept involve digital computers. The use of computers incurs manipulation errors (round-off and truncation) in the analysis since the computer carries a limited number of significant places in arithmetic. These manipulation errors will increase as the number of calculations are increased. Since the problem sizes are growing, an analysis of the errors induced by the use of digital computer is increasingly important. This report examines these errors for the displacement and force methods of structural analysis for, respectively, systems of finite elements in series and in parallel.

The principal manipulation error involves distortion of the mantissa of the floating decimal representation of the number. This error depends upon the selection of number base, number of places carried, arithmetic mode, and manipulation mode. These characteristics are fixed by the selection of computer hardware and software. Table I identifies these characteristics for several computer systems.

The manipulation error is also influenced by problem scale, structural idealization, the sequencing of joints and elements, selection of coordinate axes, element representations used, choice between the force or displacement method, and the algorithm selected for solving the load-deflection relationships for the structure.

Errors are studied in the input-output, generation and elimination phases of calculation. The input-output phase involves that part of the problem in which the data is introduced into the computer and results are output for analyst interpretation. Input errors are not significant. The most critical input errors arise when decimal fractions are entered. Errors in input can be regarded as changes in the original structure, and consequently their affect can be interpreted by the analyst. Output errors are zero unless all places carried in the digital computer are printed out. If all places are printed,

*Section Manager, Engineering Mechanics, Philco-Ford Corporation, Western Development Laboratory, Palo Alto, Calif.

**Senior Engineer/Scientist, Douglas Missile and Space Division, Culver City, Calif.

the last place may be in error.

Generation errors consist of the errors arising in developing the coefficients in the load-deflection equations. These errors are relatively small for both the force and displacement methods because coefficient generation requires few calculations. Accuracy measures described can be coded to insure the consistency of the coefficients. A larger source of error consists of the introduction of coordinate information which does not permit accurate evaluation of the geometry of the elements of the structure.

The largest manipulation errors are evoked in evaluating the primary unknowns of the structure. For the displacement method these are displacements; for the force method, force redundants.

For the displacement method, this study examines the errors in the analysis of series systems. The worst error arises for these systems in the decomposition (triangularization) of the stiffness array. Error sources in this process include instability of the calculations due to manipulation errors, the accumulation of small errors, and critical arithmetic. Errors in forward and backward substitution to evaluate the displacement unknowns are small in these systems. This study describes criteria for the analyst to minimize manipulation error and equations to bound its magnitudes. Included are descriptions of optimum joints sequencing, element sequencing, selection of idealization, and error estimation formulas. Criteria for software error control include modification of the simultaneous equation solution algorithm and error checks.

The worst error in the force method arises in inverting the force-redundant influence matrix. Errors in triangularizing the geometric matrix are only critical if the structure in kinematically unstable. Algorithm instability and the persistent accumulation of errors are important error sources in resolving the matrix of redundants. Critical arithmetic is less important that for the displacement method. The parallel structural system yields the largest error of any structural system for the solution of the redundants matrix. Analyst criteria include the selection of weighting of the structured redundants to sequence equation treatment, the numbering of elements and the selection of idealization. Programming described to control error includes modifications of the solution algorithm and checks to insure the accuracy of the analysis.

Verification problems consisting of a swept wing and an unswept box are analyzed to validate error bounds for practical structural analysis. Data from this study shows that upper bounds based on the number of calculations are conservative for the displacement method and very conservative for the force method. Study of these problems confirm that parallel systems should be treated by the displacement method and series by the force method to minimize manipulation errors.

Section 1

INTRODUCTION

When the finite element approach to structural analysis was introduced by Levy[1] and Turner[2] fourteen years ago, analysis involved few equations. Currently, analyses involving between 400 and 800 equations are commonplace. Analyses of 1500-2500 equations, which are unusual now, will soon be typical. Thus, though few bad answers have arisen in numerical analyses because of manipulation errors, the probability of answer invalidity is increasing.

In addition to the increase in problem size, the complexity of finite element analyses has increased. In the past, many applications were made to structures for which the analyst could evaluate answer validity by simple calculations, economical tests, or engineering experience. Now, powerful programs like NASTRAN are becoming available. These provide the ability to treat geometric and material orthotropy (sandwich, laminated, wound, and other composites), more complex geometry (shells of arbitrary form, solids), as well as more equations. Checks now available to insure answer validity can be expected to be too gross for these more complex systems.

Even at present problem sizes and with the present problem complexities, deleterious manipulation errors have occured in finite element analyses. These have been treated individually. A comprehensive study of the severity and causes of manipulation errors is required. This study can provide a basis for the analyst to evaluate the accuracy of the analysis with respect to the manipulation error and suggest ways in which he can formulate his problem and/or maximize his answer validity.

Previous studies on manipulation errors pertinent to the structural analysis problem are principally involved with evaluation of the error in the solution of the load-deflection equations. Von Neumann and Goldstine[3] performed an extensive analysis of the errors involved in matrix inversion. Their attack considered the errors involved in the solution to the equations. They defined bounds on the magnitudes of these errors as a function of matrix norms. Turing[4] defined a number of more useable matrix norms ("condition numbers"). He showed that the bounds defined by Von Neumann and Goldstine were much less when expected errors were estimated rather than maxima. Wilkinson[5] adopted a "backward analysis" approach to error evaluation. This approach involves determining bounds of changes in problem formulation rather than bounds on solution error. He found it easier, using this approach, to define a set of bounds for the analysis error. Forsythe[6] has given a good overview of manipulation error problems in linear algebra. He distinguishes dense and sparse matrices as having distinct error problems.

Other authors have treated manipulation errors in structural analyses in particular. Rosanoff and Ginsburg[7] and Rosanoff[8] expounded the basis available for error analysis as related to the structural analysis problem. They cited some examples of analysis error and indicated the complexity of the error analysis problem. Gatewood and Ohanian[9] looked at the manipulation error as a

function of the order of the differential equations being considered. They showed that considering first order equations gave less error than second order differential equations. Moreover, the manipulation errors associated with a pair of second order differential equations was less than that associated with a single fourth order differential equation. Shah[10] evaluated errors in a structural analysis in terms of the eigenvectors of the matrix of the load-deflection equations using the displacement method. His study indicated some general conclusions with respect to error.

This paper is directly concerned with the manipulation errors in conventional force and displacement method analysis of structures. The purposes of this study are as follows:

1. To define the relative importance of all sources of manipulation error in a computer-aided numerical analysis of a structural system using finite elements. The structural systems will be restricted to those with linear response characteristics.

2. To present criteria for the engineer to evaluate the maximum manipulation error that may occur in his analysis and to define ways in which he can formulate his problem to minimize this error.

3. To note error control devices that the programmer may use to reduce manipulation error.

4. To demonstrate the effectiveness of the criteria for error evaluation for practical finite element structural analyses.

The results of the study are presented in five sections. The next section provides the basic definitions for the error analysis and examines the error sources common to both the force and displacement methods of structural analysis. The third section includes the error analysis of the displacement method of structural analysis. The fourth section involves error analysis of the force method. The fifth section examines use of displacement and force criteria in predicting the manipulation error associated with the analysis of a swept wing and an unswept box. The final section of the document contains a summary of report developments.

The valuable assistance of Philip Diether of Philco-Ford and Harvey Puckett of Douglas in formulating test problems is gratefully acknowledged. The assistance of the Ames computing laboratory in performing displacement method calculations was indispensable. The assistance of Stewart Crandall in implementing higher precision analyses was of special help.

Displacement method equation solutions were developed using the SAMIS code available from the University of Georgia. Force method analyses were implemented by Format II, available from Wright-Patterson Air Force Base.

Section 2

GENERAL ASPECTS OF ERROR ANALYSIS

Manipulation errors are caused by using a limited number of places to represent a number. Most engineering data are represented on the computer using floating point numbers. Each representation consists of two parts: a mantissa and an exponent.

In a single operation, manipulation errors can involve exponent exceedance or mantissa distortion. The limited range of the exponent due to the limited number of places alloted, induces "overflow" and "underflow" errors. Overflow occurs when the result of an arithmetic operation is such that the exponent exceeds the value that can be represented. If, for example, exponents bigger than one are inadmissible in a tens base calculation, forming the product of 10^1 and 10^1 would result in overflow. Underflow occurs when the result of the calculations is so small that the exponent is smaller than representable with the number of places provided for the exponent.

Overflow and underflow are not normally critical sources of manipulation error. Usually, the number of places for the exponent is enough so that overflow and underflow do not occur.

If exceedance errors occur, it is conventional to replace an underflowed number with an absolute zero during calculations. Thus, when underflow occurs, the number representing the result is negligible compared with typical numbers. When overflow occurs, however, a number that is infinitely larger than other numbers cannot be substituted for the overflowed number. With some software, the overflowed number is replaced by the maximum number that can be represented on the computer. On others, the number of overflow occurrences is noted to provide a measure of the inadequacy of the calculations to the analyst.

If exponent exceedance errors are important, they can be eliminated by changes to the problem or the software. Scaling or shifting of the problem data base can be used to reduce the required exponent range of calculations. Scaling requires multiplying all numbers by a factor. Shifting involves translating coordinate systems. These same types of changes can be incorporated in software to insure that no exponent exceedance occurs. The optimum problem formulation is that in which the full range of the exponent is used but never exceeded.

Though exponent exceedance errors rarely occur and can be easily sensed and eliminated, the same is not true of mantissa distortion errors. These errors involve the attrition of the mantissa as a consequence of a series of calculations. The attrition may result in exponent exceedance but deleterious errors can exist without over or underflow. The simplest way to detect and eliminate these errors is by performing arithmetic using more places in the mantissa. Determination of the magnitude of these errors is the central concern of this study.

Error Parameters

Manipulation errors depend on the analyst's choice of hardware and software. This choice determines the number base, the arithmetic mode, and the manipulation mode. The number base used in all high-speed computers is two. Both the exponent and the mantissa are represented in a binary mode. Thus, the number 0.5×10^1 in the tens numoer base is represented on the computer as $.101 \times 2^3$. In the representation, only the mantissa (.101) and the exponent (3) are cited. The base of the exponent is implied.

The arithmetic mode of interest is the floating binary mode. Fixed point arithmetic is inherently more accurate using a given number of places to represent a number because more places can be allocated to the mantissa when the exponent is implied. The added programming effort involved in controlling scaling throughout the calculations, however, results in fixed point being an unpopular mode for scientific analyses.

The manipulation mode is defined by the rules for arithmetic. These rules define the precision of the calculations, how the answer is developed, special treatment to the answer and what is done with the remainder in the calculations.

This study will be concerned with single precision manipulation. This means that the number is represented by a single computer word. Most large computers provide for both single and double precision arithmetic in the hardware. Higher than double precision can be attained by software but involves a large time penalty compared with single precision. For example, quadruple precision is over four times more time consuming than double precision though double precision is only 1.4 times more expensive than single precision. The consideration of multiple precision modes does not change the basic error analysis. It means that calculations proceed as if a single number with a larger mantissa has been used. For example, if single precision has a 24 bit (binary place) mantissa, double precision has 48 (IBM 7094).

For this study, it will be assumed that the answers are developed using a single precision accumulator for addition and subtraction. A double precision accumulator will be assumed for multiplication and division. This assumption is true for a number of computers of interest.

After the single or double precision result is obtained, it will be assumed that the result is normalized. Normalization consists of shifting the mantissa to the left as far as necessary so that the lead place contains a non-zero number. It will be assumed that the lead place is a binary place. It is noted that this is not true of the IBM 360 system which uses a hexadecimal (five bit) first place. Binary normalization is common to all other large scale computers, however. Upon completion of normalization, it will be assumed that the remainder is truncated. In truncation the remainder is simply discarded. In rounding, on the other hand, the last place in the mantissa is increased by 1 if the first place of the remainder is non-zero.

Average and maximum errors associated with truncation tend to be greater than those due to rounding. The average truncation error is slightly less than one-half the value of the last place in the result. The average rounding error

is zero. The maximum truncation error is one part in the last place; the maximum rounding error one-half part. Assuming truncation will yield error bounds which would indicate larger errors than assuming rounding.

The manipulation mode assumed is consistent with the mode used in a Fortran IV program operating on the IBM 7094. This mode is more or less the common mode of calculation on all computers.

The characteristics of various computers with respect to manipulation mode are summarized in Table I. This table cites the total work size for number representation, the number of the bits of this word size reserved for the mantissa, and the number of decimal digits that are represented. If the range of the exponent is of interest, the reader can determine the number of digits in the exponent by subtracting the number of bits in the mantissa from floating point word size and considering this as a range on exponent of the number two. For the Burroughs B5500, for example, the number of bits for the exponent is thus eight. (In this particular case one bit is reserved to indicate the existence of data in the storage location.) The largest exponent that can be represented is 2^{128}. It is customary to divide the exponent equally between the positive and negative exponents. Thus, the exponent could range from 2^{-63} to 2^{63}. Converting this into a decimal system, the exponent could vary between 10^{-19} and 10^{+19}.

As indicated by this survey, most of the large scale computers truncate the result of the arithmetic operations. Whether or not rounding or truncation is used depends upon the compiler. In most of the large scale computers, the machine language instruction to perform either rounding or truncation is available. Thus, the entries in the fifth column of the table define treatment of the mantissa in accordance with the Fortran IV compiler for the computer indicated.

Besides his selection of computer hardware and software, the analyst's choice of problem scale, idealization and solution algorithm has an important affect upon the manipulation error in his calculations. As noted, his choice of scales determines whether exponent exceedance will occur and influences the magnitude of mantissa distortion. His idealization of the structure determines the conditioning of the numbers used in the calculation. Of importance are the joint numbering sequence, coordinate axes, the material coefficients, and the element representation selected for the finite elements. The algorithm selected by the analyst defines the sequence in which calculations are performed and the approximateions used in arriving at the solution.

In the sequel errors will sometimes be characterized in their absolute form and sometimes in their relative. Absolute errors are statements of true values of the errors. Relative errors are the ratio of the absolute error divided by some measure of the numbers used in the calculation, e.g., $e = E/A$ where e is the relative error, E is the absolute error and A is the number measure.

Table I

Accuracy of Computers

Computer	Floating Decimal Word Size, Bits	No. of Bits in Mantissa	No. of Digits in Accuracy	Manipulation Mode
Burroughs B5500	48*	39	11.7	Round
CDC 3600	48	36	10.8	Round
CDC 6600	60	48	14.4	Round
GE 265	40	29	8.7	Truncate
Honeywell MH 800	48	40	12.0	Truncate
Honeywell MH 1800	48	40	12.0	Truncate
IBM 7094	36	27	8.1	Truncate
IBM 360	32	24□	7.2	Truncate
Philco 212	48	35	10.5	Round
RCA Spectra 70's	32	24□	8.1	Truncate
Univac 1108	36	27	8.1	Truncate

* One bit indicates data existence.

□ Hexadecimal normalization; thus effectively 23-1/3 bits, on the average.

The objective of this error analysis is to define the magnitude of the errors that arise in calculation. In performing the analysis, each operation will be treated independently. It will be assumed that the input to that operation is precise. The error in the output of that operation will be predicted. Operations include input-out and arithmetic. The error analysis will be performed by analyzing special cases and developing error criteria. These criteria will be evaluated on a set of special problems. Final verification of the criteria will be performed by applying the criteria to practical structural problems and correlating results with computer analyses of these structures.

Input-Output Errors

Input-output errors are the errors involved in communicating problem constants between the analyst or programmer and the computer. Usually input data is developed by the analyst in the form of card-punched information. The programmer supplies input to the calculations in the form of constants defined in his source deck. Output data is transmitted by the computer to the analyst in the form of printouts and punched card data.

Input errors include errors in data truncation and conversion. No matter how accurately the analyst defines his input data, the computer represents this information in a floating point word of limited size. Thus, for example, a ten-digit number is truncated to an 8.3 digit number. Truncation is performed after converting digits. Therefore, in this example, the tenth digit would be disregarded and the ninth represented with some error.

Conversion errors arise because input data is generally introduced in decimal form and must be converted into the binary base system for calculations. No conversion error occurs in transmitting integers. Errors do arise, however, in converting decimal fractions. In fact, a 27 bit representation does not provide for eight digit accuracy due to conversion error[11].

Output errors are caused by transforming the binary representation of the number into the decimal system and reproducing a limited decimal representation for the convenience of the user. These again, are errors of conversion and truncation. In conversion for output, digits beyond those printed or punched are disregarded. In converting the mantissa to a decimal system, errors arise only in digits beyond those represented in the computer, i.e., in the ninth digit in a machine carrying 8.3 decimal places in the mantissa. This is because single precision operations are used to perform the conversion. Since the analyst is normally interested in fewer significant figures than represented on the computer in the mantissa, this error does not appear in the final answer.

Table II provides typical input-output error data. These data were obtained using the IBM 7094. Errors in input transformation for integers is significant only in the last place that can be represented in the computer. This error is a measure of the truncation error alone. Error in converting fractional decimals is seen to be more significant. It is interesting to note

Table II

Input-Output Errors

Number*	Internal Form	Output Form
.1	.099,999,995,6	.099,999,996
.2	.199,999,999,1	.199,999,999
.3	.300,000,000,5	.300,000,001
.4	.399,999,998,2	.399,999,999
.5	.500,000,000,0	.500,000,000
.6	.600,000,001,0	.600,000,001
.7	.700,000,004,3	.700,000,003
.8	.799,999,996,5	.799,999,997
.9	.899,999,998,2	.899,999,999
1.0	1.000,000,000,0	.999,999,994
268435455.	268,435,456.0	268,435,456

*From compiler or source cards.

that data introduced through the compiler by way of input cards may result in a different representation than the same decimal introduced by the programmer. This inconsistency can be a problem if it is required to match input of the analyst with constants supplied by the programmer.

Output errors arise only in the last decimal digit represented on this computer. If less than nine digits are printed for the 27 bit word, the answe is accurate in all digits. Since engineers are seldom interested in more than four significant digits in their answers, this output manipulation error can be disregarded for all computers.

If the analyst is concerned about the magnitude of the error due to input-output error, he can modify his calculations to measure it. This can be achieved by changing input to span the exact numbers to be represented, re-running the analysis, and comparing solutions.

The affect of truncating numbers entered in the source deck is not as easily measured since it would require a recompilation of the program being used. However, these errors are small if the programmer enters constants with maximum precision of the floating words. For example, the constant Π in a 27 bit (8.3 decimal digit) mantissa should be entered as 3.14159265 to minimize the error.

If the analyst insists on eliminating input manipulation errors, this can be achieved for all but irrational numbers by scaling his data by the least common denominator consistent with model scaling laws. He should then choose a computer system which can handle this integer.

It is concluded that input-output errors are not an important source of manipulation errors. The principal errors in input involve conversions. If these errors are important, their effect can be measured by recalculation. The principal errors in output involve truncation. These can be avoided by printing out at least one less significant place than is carried with the precision available. If the analyst requires a complete representation of the number for output, he can always print the number in octal format and perform the conversion himself without errors.

Simple Arithmetic Errors

All errors other than input-output errors arise at the simple arithmetic level. Simple arithmetic consists of single calculations and series of calculations involving a single operation.

Consider first the errors arising in a single addition, subtraction, multiplication and division. The maximum absolute error in such an operation is less than or equal to one part in the last place of the un-normalized answer:

$$|E_m| \leqslant b^{x(A)-p+1} \qquad (2\text{-}1)$$

where b is the number of base of interest, x(A) is the exponent of the answer p is the arithmetic precision (number of significant places in the mantissa).

When two components are added together, the exact sum can contain more places than either addend. Only in this case can the computer result be in error. Thus, errors in the order of the maximum given by equation (1) arise when the exact sum contains more places than either addend. One additional place is sufficient to approach E_m. The following examples show this. The number base assumed is 10. E is the exact value of error.

p=2: $.43x10^1+.29x10^1=.72x10^1$ $E=0$, $E_m=.01x10^1$

p=2: $.12x10^2+.29x10^1=.14x10^2$ $E=.009x10^2$, $E_m=.01x10^2$

p=3: $.123x10^{-1}+.999x10^1=.100x10^2$ $E=.000023x10^2$, $E_m=.0001x10^2$

p=3: $.743.10^0+.999x10^{-3}=.743x10^2$ $E=.000999x10^2$, $E_m=.001x10^2$

When the number base is 2, the error bound given by equation (1) provides a closer estimate than with the number base 10. For example:

p-2: $.11x2^2+.11x2^0=.11x2^2$ $E=.0011x2^2$ $E_m-.01x2^2$

p=3: $.101x^{-3}+.111x2^{-2}=.100x2^{-1}$ $E=.00011x2^{-1}$ $E_m=.001x2^{-1}$

p=4: $.1101x2^3+.1110x2^3-.110x2^4$ $E=.00011x2^4$ $E_m=.001x2^4$

In addition, the relative error is defined as the absolute value of error divided by the sum. Using equation (1) and recognizing that the truncation is always positive, the maximum relative error is given by

$$e_{ma} = b^{1-p} \tag{2-2}$$

Where the second subscript indicates that the operation is addition. The exact sum is bounded by

$$A_E \leqslant A_E + E_m \leqslant A_E + Ae_{ma} \leqslant A \tag{2-3}$$

where A_E is the estimate of the answer. Thus, the solution is always further from zero than the estimate.

In subtraction, the number of places in the answer may be less than the number of places in either the minuend or subtrahend. Then, the subtraction operation involves shifting of the result to yield a normalized mantissa. Modifying equation (1) to include consideration of the shifting of the mantissa after the subtraction, it takes the form

$$E_{ms} = b^{x(A)-p+f+1} \tag{2-4}$$

where f is the number of left shifts (multiplications by b) to normalize the answer. Note that the actual error of subtraction may be positive or negative. The fidelity which this equation defines in upper bound is indicated as follows for decimal and binary based arithmetic. Again, as for addtion, the bound is usually closer when binary arithmetic is used.

b	p	Subtraction		Result	E	E_{ms}
10	2	$.16x10^{1}-.89x10^{0}$	=	$.8x10^{0}$	$.09x10^{0}$	$.1x10^{0}$
10	2	$.16x10^{-2}-.41x10^{1}$	=	$-.40x10^{-1}$	$.006x10^{-1}$	$.01x10^{-1}$
10	3	$.326x10^{2}-.999x10^{1}$	=	$.227x10^{2}$	$.0001x10^{2}$	$.001x10^{2}$
10	3	$.143x10^{-3}-.142x10^{-3}$	=	$.100x10^{-5}$	0.	$.1x10^{-5}$
2	2	$.11x2^{2}-.10x2^{1}=.10x2^{2}$	=	$.10x2^{2}$	0	$.01x2^{2}$
2	3	$.101x2^{-1}-.110x2^{0}$	=	$-.100x2^{0}$	$.0001x2^{0}$	$.001x2^{0}$
2	4	$.1001x2^{6}-.1101x2^{0}$	=	$.1001x2^{6}$	$.1101x2^{0}$	$1000x2^{1}$

Because the results of the subtractions may be zero, it is desirable to define the relative errors as:

$$e_{ms} = \left|E_{ms}\right| \Big/ \sum_{i=1}^{N} \left|n_i\right| \tag{2-5}$$

where $\sum_{i=1}^{N} n_i$ means the sum of the absolute values of the numbers involved in the arithmetic. Using (4) and (5) the exact difference is bounded by:

$$\left|A_E\right| - \left|E_{ms}\right| \leqslant \left|A_E\right| - \sum_{i=1}^{N} n_i \cdot e_{ms} \leqslant \left|A\right| \leqslant \left|A_E\right| \tag{2-6}$$

Thus, the answer is always closer to zero than the estimate.

Subtraction tends to result in smaller errors than addition. Absolute values of errors are the same magnitude, but subtraction involves no error when the exponents of the numbers producing the answer are the same. Therefore, relative errors, as defined by equation (5), will be smaller for subtraction than addition.

This conclusion offends the intuition because the subtraction of nearly equal numbers will result in an answer with few significant figures of accuracy. This deficiency is not subtraction error. It is due embedding the information of interest in the least significant places of the mantissa of the minuend and subtrahend. This operation is one type of "critical arithmetic."

A second type of critical arithmetic occurs in both addition and subtrac-

tion. This occurs when one of the components is small compared with the other, and vital information is lost. In this critical arithmetic, the smaller component is truncated from the left and the compoenent may be lost, e.g.,

$$b=10,\ p=4:\quad .1000\times10^{8}+\ .5000\times10^{4}=.1000\times10^{8}$$

$$b=2,\quad p=3:\quad .101\times2^{3}-.101\times2^{0}=.101\times2^{3}$$

Since the result of multiplication of normalized numbers is always normalized, the limitation of equation (1) to a normalized result can be disregarded for multiplication. It is convenient, however, to define another, less accurate, error measure for multiplication:

$$E_{mm} = AC \tag{2-7}$$

where the subscript m designates multiplication, A is the exact product and C b^{1-p}. The following examples illustrate the effectiveness of equations (1) and (7) in bounding the error for multiplication in the tens and binary bases:

b	p	Product	Result	E	E_m	E_{mm}
10	2	$(.23\times10^{1})(.21\times10^{1})$	$= .48\times10^{1}$	$.03\times10^{-1}$	$.10\times10^{-1}$	$.48\times10^{-1}$
10	2	$(.82\times10^{1})(.13\times10^{1})$	$= .10\times10^{2}$	$.66\times10^{0}$	1.0×10^{0}	1.06×10^{0}
10	3	$(.223\times10^{2})(.684\times10^{2})$	$= .1152\times10^{3}$	$.532\times10^{0}$	1.0×10^{0}	1.52×10^{0}
2	2	$(.11\times2^{0})(.11\times2^{0})$	$= .10\times2^{0}$	$.001\times2^{-1}$	$.1\times2^{-1}$	$.102^{-1}$
2	3	$(.110\times2^{1})(.101\times2^{2})$	$= .111\times2^{2}$	$.010\times2^{1}$	$.1\times2^{1}$	$.111\times2^{1}$
2	4	$(.1101\times2^{-2})(.1001\times2^{3}$	$= .1110\times2^{0}$	$.0101\times2^{-3}$	$.1\times2^{-3}$	$.1110\times2^{-3}$

It can be seen that (7) gives an estimate of the maximum error that is at most twice that given by (1) for binary base arithmetic. It may be ten times greater for decimal arithmetic.

The relative error in multiplication is defined as the error divided by the value of the product. Using equation (7), the relative error takes the form

$$\left| e_{mm} \right| = \left| C \right| \tag{2-8}$$

Bounds for the exact product are

$$\left|A_E\right| \leqslant \left|A\right| \leqslant \left|A\right| + \left|E_{mm}\right| \leqslant \left|A_E\right| + \left|Ae_{mm}\right| \tag{2-9}$$

Thus the answer is always further from zero than the estimate.

Equation (1) defines an error bound for division because, like multiplication, the result is developed in a double precision accumulator. Normalization of the result insures accuracy to as many places as in the original dividend and divisor. Like multiplication, an alternate bound for the quotient error is taken in the form of equation (7). Illustration of the effectiveness of equations (1) and (7) in bounding the error of division is indicated by the following examples:

b	p	Quotient	Result	E	E_m	E_{mm}
10	2	$(.57\times10^{3})(.24\times10^{3})^{-1}$	$= .23\times10^{2}$	$.0073\times10^{-2}$	$.01\times10^{-2}$	$.023\times10^{-2}$
10	2	$(.57\times10^{1})(.68\times10^{-2})^{-1}$	$= -.83\times10^{1}$	$.008\times10^{1}$	$.01\times10^{1}$	$.083\times10^{1}$
10	3	$(.57\times10^{2})(.683\times10^{-1})^{-1}$	$= .84\times10^{3}$	$.001\times10^{3}$	$.01\times10^{3}$	$.084\times10^{3}$
2	2	$(.10\times2^{2})(.11\times2^{1})^{-1}$	$= .10\times2^{1}$	$.0101\times2^{0}$	$.1\times2^{0}$	$.1\times2^{0}$
2	3	$(.110\times2^{-1})(.101\times2^{2})^{-1}$	$= .100\times2^{-2}$	$.011\times2^{-4}$	$.1\times2^{-4}$	$.110\times2^{-4}$
2	4	$(.1101\times2^{3})(.1011\times2^{1})^{-1}$	$= .1001\times2^{3}$	$.0011\times2^{3}$	$.1\times2^{-3}$	$.1001\times2^{-3}$

Again it is seen that the bound given by E_{mm} differs from that of E_m by at most a factor of two when the binary base is involved in the operation.

Relative error for division is defined the same way as for multiplication, i.e., by equation (7) and the quotient is bounded by equation (9). The exact quotient is always farther than zero than the estimate.

It is sometimes desirable to determine error when the exact answer is unknown. Then formulas for the relative error can be used with estimates of the exact answer to predict absolute error magnitudes. These estimates are only as good as the estimates of the exact answers.

Simple Arithmetic Sequences.- A simple algorithm involving arithmetic operations is to perform a series of calculations using either addition, subtraction, multiplication or division. Examination of the truncation error in these sequences provides a basis for bounding errors in more complex sequences.

Assume that it is desired to perform a series addition. The operation consists of performing a number of additions, N, each time adding a component to the previous sum.

Consider first the case where the components are all equal. Then the bounds for the error can be defined by examining the errors when the most and least error prone mantissa are considered. These bounds can then be curve fit to provide simple formulae for evaluating errors in series addition as a function of the number of components and the precision being used.

The mantissa with the maximum error is given by

$$m_m = 2b^{-1} - b^{-p} \tag{2-10}$$

This mantissa results in the maximum absolute and relative error in a series addition. In the tens number base, it gives a component of the form $19 \times 10q^q$ where q is any number within the admissible exponent range and the verniculum indicates that the digits it spans are repeated. The first digit in the mantissa is 1 to maximize the relative error. The remaining digits are nines so that the truncation, which occurs from the right, is a large as possible value.

The mantissa associated with least truncation error is

$$m_L = b^{-1} \tag{2-11}$$

This results in a lower error bound since it involves no error in a series of additions until the number additions is greater than b^p . Then critical arithmetic occurs and the error for each addition is

$$E = b^{-1}\, b^{x(Ac)}$$

Where Ac here is the current sum.

Figure 1 shows the variation of error as a function of N, the number of times the component is added in the series for critical components. The variation of the upper bound with N is typical of errors with any component. The absolute error varies linearly with N in each range in which the sum varies from S to bS where $S = bq^q$ and q is an integer. The range is started by a number of the form $\overline{9}$. Within this range the relative error varies as

$$e_{ma} = \frac{(E_o + nZ\ E_{ma})}{S_o + nZ} \tag{2-12}$$

where E_o is the cumulative error at the beginning of the range

n is the number of components added

Z is the component being added

S_o is the value of the series sum at the beginning of the range. In most ranges, $nE \ll S_o$ and the relative error varies nearly linearly. In the first range, $S_o = E_o = 0$ and the relative error is constant.

Figure 2 shows the relative error as a function of N over a number of ranges. In this case the number base is 10, the precision 8, and the components

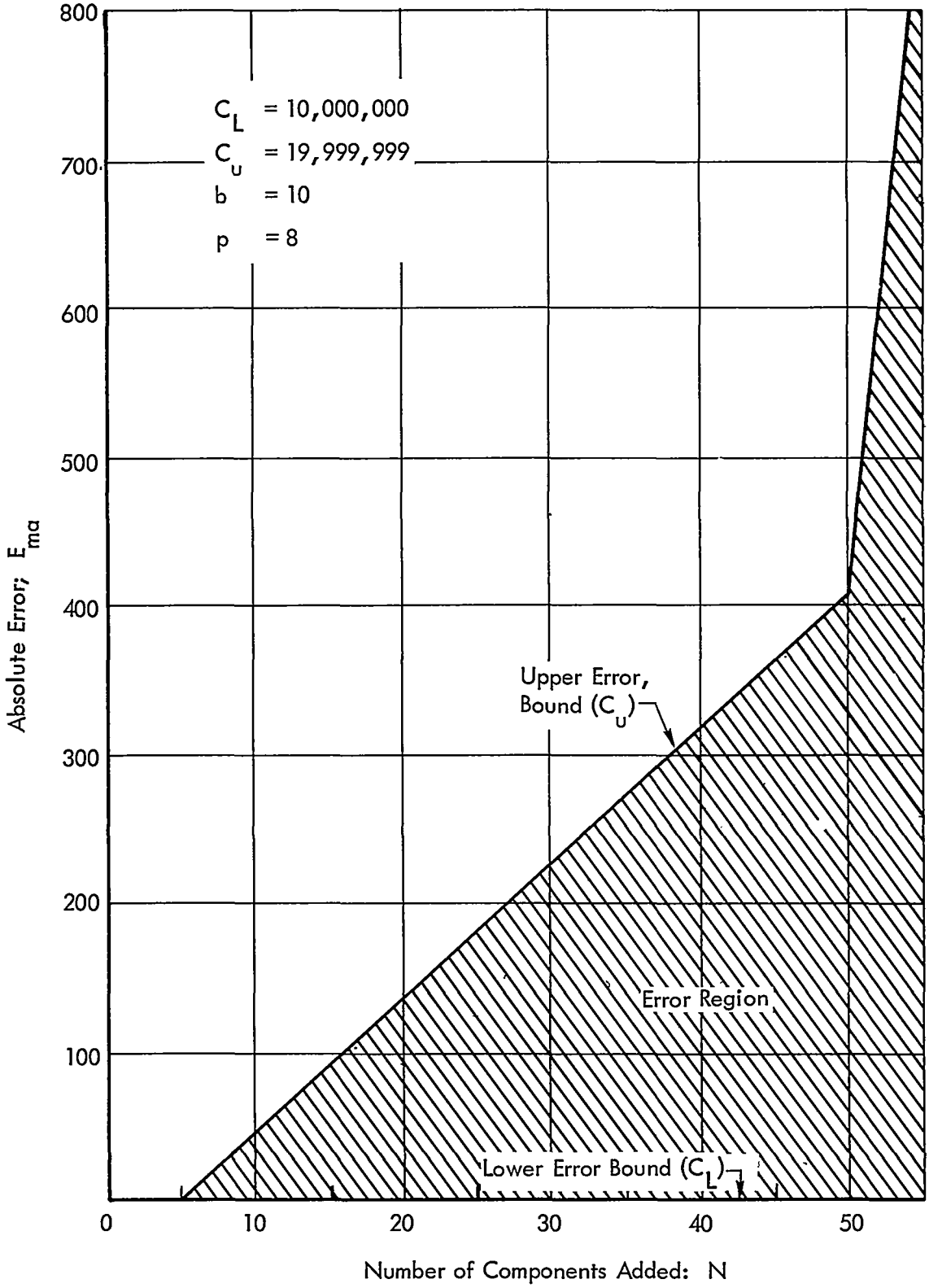


Figure 1. Series Addition Error

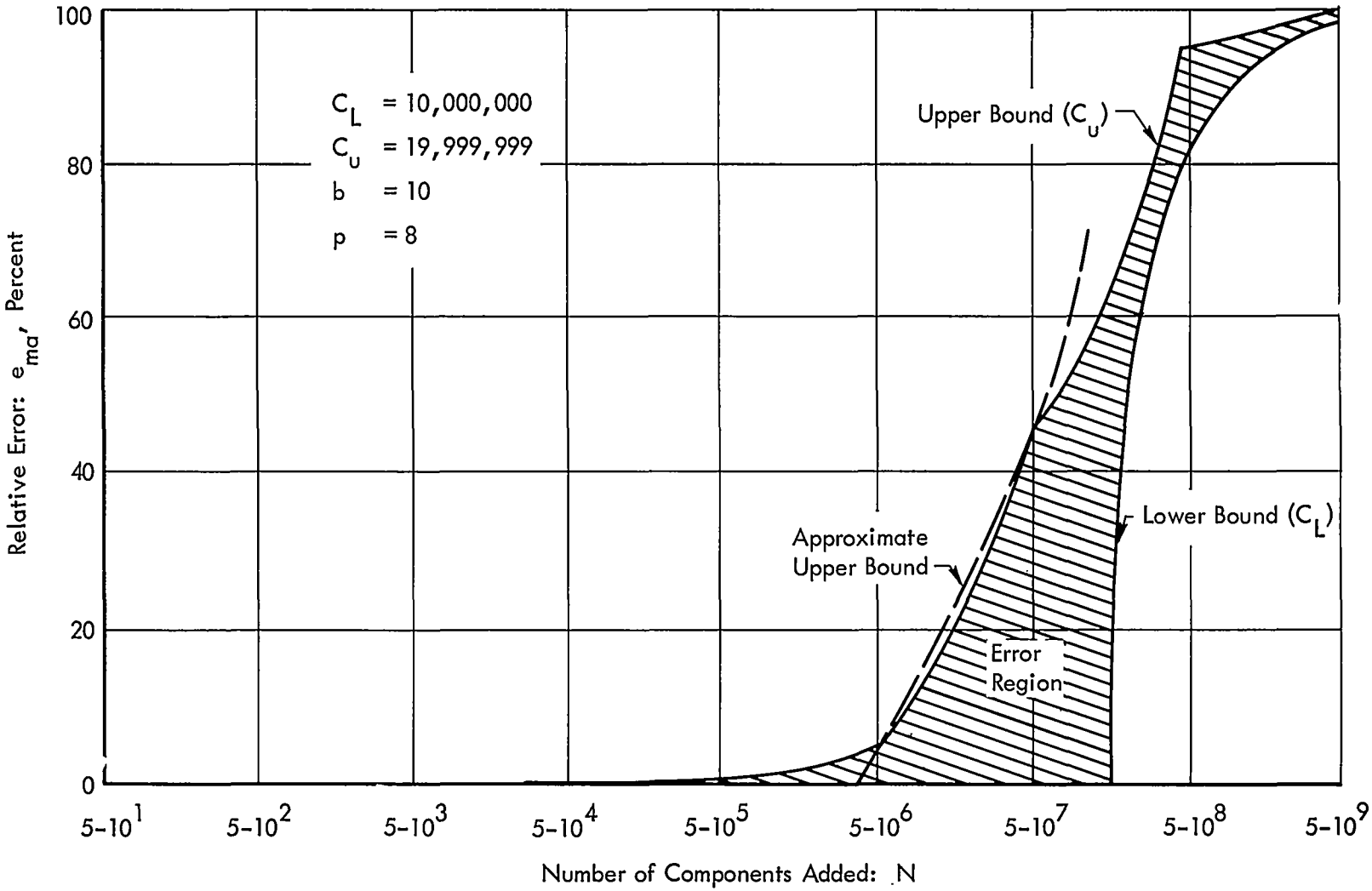


Figure 2. Series Addition Relative Error

considered to define the bounds being added are given by equations (10) and (11). The relative error is negligible when N is less than b^{p-2}. The error is excessive when N is greater than b^p. The critical error range of interest is thus:

$$b^{p-2} \leqslant R_c \leqslant b^p$$

where R_c defines the critical ranges of N.

Within this range an upper bound fit for the error is given by

$$e_{ma} = 0.9 \times 10^{-p} N \tag{2-13}$$

The fit of this curve for the tens base is indicated on Figure 2. The corresponding formula for the binary base is

$$e_{ma} = 0.5 \cdot 2^{-p} N \tag{2-14}$$

Formulas (13) and (14) indicate relative error in percent if multiplied by 100. Equation (14) indicates that relative error will not exceed five percent until the number of additions exceeds $.1 \times 2^{27} = 26.8 \times 10^6$ when $p = 27$.

Table III shows the small relative error if calculations are resequenced to minimize the error associated with the series addition. The optimum process consists of pairing numbers of equal size in the addition. Thus, for example, if 16 equal components are to be added, 8 pairs are first added, then pairs of these sums are added by four more additions; the results of these are added in two additions and the final result obtained with the final and 15th addition.

Table III shows that the reduction in error due to this optimum sequence of additions is dramatic. This example illustrates the importance of the selection of algorithm in affecting the magnitude of the manipulation error. unfortunately, it is costly to define the sequence of the calculations based upon the component magnitude. Thus, the improvement indicated in this simple example would be expected to be greater than the improvement that would be achieved in practice.

The maximum errors indicated for equal components are also valid estimates for maximum errors when unequal components are involved. The equal component case represents the worst case, since the error accumulates at every step and the accumulation is the maximum that could be attained. Thus, the error involved in the addition of equal components is approached as a limit when unequal components are added.

The maximum addition error defined using the component (10) defines extremes that can be achieved in practice. The error bound given by equation is

Table III

Error Reduction for Optimum Addition

$b = 10$, $p = 8$, $C_u = 19{,}999{,}999$

Relative Error, %

N	Optimum	Bound
10^2	$.113 \times 10^{-4}$	$.400 \times 10^{-4}$
10^3	$.190 \times 10^{-4}$	$.455 \times 10^{-3}$
10^4	$.240 \times 10^{-4}$	$.455 \times 10^{-2}$
10^5	$.300 \times 10^{-4}$	$.455 \times 10^{-1}$
10^6	$.377 \times 10^{-4}$	$.455 \times 10^{0}$
10^7	$.472 \times 10^{-4}$	$.455 \times 10^{1}$
10^8	$.533 \times 10^{-4}$	$.455 \times 10^{2}$
10^9	$.792 \times 10^{-4}$	$.945 \times 10^{2}$
10^{10}	$.843 \times 10^{-4}$	$.994 \times 10^{2}$

an approximation. It gives an upper bound for error except when errors are negligible.

Consider the errors involved in performing a number of subtractions. A series subtraction is defined as a sequence of subtractions such that the result of each subtraction yields an answer which is always opposite in sign to the sign of the next component. This series of subtractions can occur by starting with a large positive number and sequentially subtracting a small component a multiple number of times or it can occur by performing subtractions with alternating signs.

The error bound for serial subtraction can be developed from the bounds for serial addition. The maximum absolute error in subtraction cannot exceed the maximum absolute error of addition since the critical component in addition maximizes error at each step in the calculations. Thus, the maximum absolute error can be defined by considering additions of the mantissa defined by equation (10).

The relative error bound for subtraction may be taken as twice the relative error in addition. This is true because the maximum error does not occur when the minuend and subtrahend have the same exponent. Then the error is zero. Since the denominator for extreme relative error must involve two unequal elements, the worst case is when the smaller part of the denominator is legligible in absolute value compared with the large. Thus, the maximum relative error is bounded by:

$$\begin{aligned} e_{ms} &= 1.8\text{x}10^{-8}N \qquad b = 10,\ p = 8 \\ e_{ms} &= 2^{-p}N \qquad\quad b = 2,\ \ p = p \end{aligned} \tag{2-15}$$

The error for multiplication of a series of factors is the same as for division of a series. The upper bound for the error can be expressed as:

$$E_{mm} = A(1+C)^{N} - A = E_{md} \tag{2-16}$$

then, since C is much less than 1, the error can be satisfactorily approximated by:

$$E_{mm} = ANC = E_{md} \tag{2-17}$$

The corresponding relative error is given by dividing the equation by the answer. Thus, the relative error is given by:

$$e_{mm} = NC = e_{md} \tag{2-18}$$

This can be converted into percent by multiplying by 100. The lower error bound is zero.

Equations (12) and (18) define bounds that cannot be achieved. However, as the precision increases, their accuracy improves. Table IV shows the multipliers involved to maximize relative error in series multiplications in the tens base. Factors were selected by trial and error. Figure 3 provides a plot of exact relative error and estimates obtained by equation (18). These curves show that the bound for estimates is satisfactory when p is four or more. Therefore, it would be expected that equation (18) yields satisfactory upper bounds when b = 2 and p = 27.

In considering the sources of error for a series of operations, the greatest error bounds arise for series multiplications and division. The least error arises in series additions. An extreme bound for the error for any of these sequences of calculations is, therefore, given by the equations:

$$E_m = ANC \quad e_m = NC \tag{2-19}$$

It is noted that when subtractions are involved in the sequence of operations, use of equation (19) to bound the solution error is difficult because the estimate of the answer cannot be used for A. A must be the absolute sum of the numbers subtracted, added, and multiplied. It is also observed that even with the conservative error bound defined by equation (19), many calculations are required before the answer has more than 5% error. When b = 2, and p = 27, 13.4×10^6 calculations would be required.

Though many calculations are required to develop significant relative truncation errors, large analysis errors can still occur. These arise when critical arithmetic is involved. As an example of a sequence involving this error, consider:

$$b=10,\ p=3: \quad .333\times10^{0}+1.0\times10^{6}+.222\times10^{0}\ -1.0\times10^{6}=0.0$$

Here, it is desired to add the first and third numbers. The result for the calculation sequence used is meaningless due to critical arithmetic, though sufficient precision is being used if the sequence of calculations is changed. Note that the relative error of the calculation is small.

Vector Scalar Products.- Wilkinson[12] has developed an error bound for accumulation of inner products. This is given as

$$e_{vm} \approx \frac{\sum_{i=1}^{C} a_i b_i \xi_i}{\sum_{i=1}^{C} a_i b_i} \tag{2-20}$$

$$|\xi_1| = 3C2^{-p}, \qquad |\xi_r| = (C+2-r)2^{-2p}$$

Table IV

Multiplication Error Estimates

b = 10

p = 1

Factors : $(2x9)x(2x9)x(2x9)(2x9) = 1.889568x10^{6}$

Eq. (18): $e_{mm} = 9x1 = 9.0 \quad 900\%$

Exact : $e_{mm} = 1-(\frac{5}{9})^{5} = .947 \quad 94.7\%$

p = 2

Factors : $13x13x(8x14)x(8x14)x(8x14)x(8x14) = 26.62x10^{8}$

Eq. (18): $e_{mm} = 9x\frac{1}{10} = .900 \quad 90.0\%$

Exact : $e_{mm} = 14.62/26.62 = .549 \quad 54.9\%$

p = 3

Factors : $(114x114x110x900x800)x(144x144x110x900x800) = 105.9428x10^{18}$

Eq. (18): $e_{mm} = 9x\frac{1}{100} = .090 \quad 9.0\%$

Exact : $e_{mm} = 5.94/105.9 = .0562 \quad 5.6\%$

p = 4

Factors : $1031x1031x1016x1025x1019x1008x1015x1027x1028 = 1218.4x10^{24}$

Eq. (18): $e_{mm} = 8x\frac{1}{1000} = .008 \quad 0.80\%$

Exact : $e_{mm} = 8.4/1218.4 = .00691 \quad 0.691\%$

p = 8

Factors : $10,007,000x10,007,000x10,497,000x10,100,000x10,157,000x10,020,000$
$x10,156,000 = 10,973,628x10^{42}$

Eq. (18): $e_{mm} = 6x10^{-7} \quad 6x10^{-5}\%$

Exact : $e_{mm} = 6/10,973,628 = 5.47x10^{-7} \quad 5.47x10^{-5}\%$

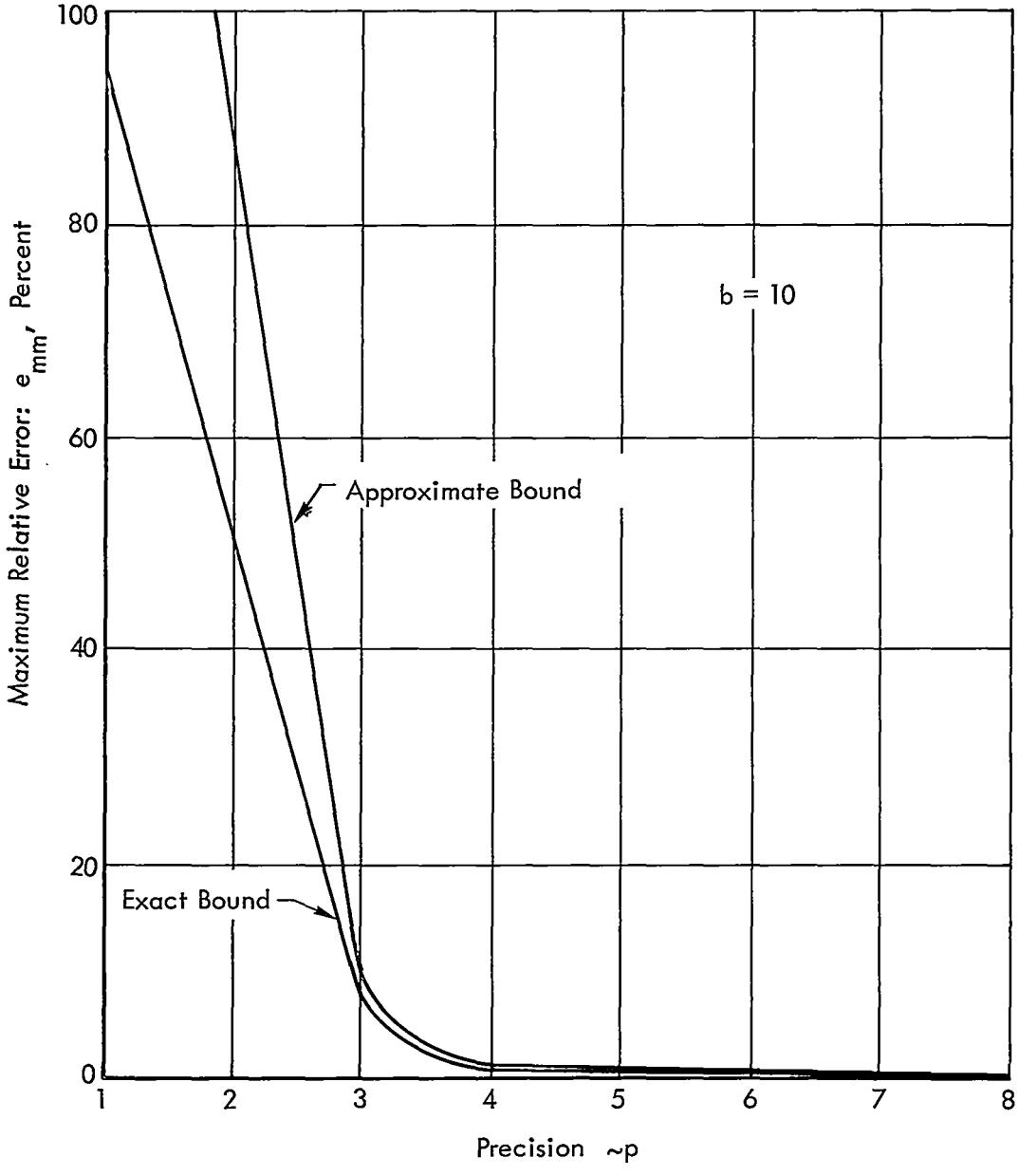


Figure 3. Multiplication Error Estimates

and a_i and b_i are the components of vector A and B. Assume that all the components are equal. Then the relative error is given by

$$e_{vm} = \sum_{i=1}^{C} \xi_i = 3x2^{-2p-1}C(C+3) \tag{2-21}$$

The number of calculations involved is 2C. Therefore, if e = 0.05, the number of calculations must exceed 47.8×10^6, when p = 27. When unequal numbers are involved, Wilkinson estimates that the expected relative error will be

$$e_{vm} \approx C^{1/2}2^{1-p} \tag{2-22}$$

Then, the expected number of calculations for the relative error to be less than five percent would be about 5×10^{14}.

In both cases, the number of calculations to develop significant attrition errors exceeds the bound developed for the simple sequences. Consequently, the probability of significant attrition error will be determined by comparing the number of calculations with 13.4×10^6.

Structural Analysis Errors

Structural analysis involves more complex sequences of calculations. To simplify the analysis of these errors, each of the operations will be considered independently. The operations consist of generations of the coefficients of the structural equations and elimination of the coupling of the equations to define the primary secondary unknowns. The equations of structural analysis are completely defined by the single set of linear algebraic equations.[13]

$$\begin{aligned} D_Q Q + P_Q^T \Delta &= \xi_1 \\ D_X X + P_X^T \Delta &= \xi_2 \end{aligned} \tag{2-23}$$

$$P_Q Q + P_X X = -F \tag{2-24}$$

where D_Q, D_X = substructure flexibility matrices. P_Q, P_X = matrices defining geometric relations between forces P_Q^T, and P_X^T the transpose of P_Q and P_X, Q = unknown internal forces for the determinate substructure and X for the rest of the structure, ξ_1, ξ_2 = interelement distortions F = the vector of loads Δ = the vector of joint displacements. Equations (23) are the internal load deflection relations for the structure. Equations (24) are the equilibrium equations for the system. These constitute the necessary and sufficient equations to define all the internal loads and structural displacements of the system. Disregarding structural idealization, the distinction between the force and displacement method consists only of the manner in which equations (23) and (24)

are solved simultaneously.

In the displacement method equations (23) are solved for Q and X, and the results substituted in (24). Equation (24) neglecting ϵ_1 and ϵ_2, then takes the form:

$$\left[P_Q D_Q^{-1} P_Q^T + P_X D_2^{-1} P_X^T\right]\Delta = F$$

$$\text{or } K\Delta = F \text{ where } K = P_Q D_Q^{-1} P_Q^T + P_X D_X^{-1} P_X^T \qquad (2\text{-}25)$$

and K is called the stiffness matrix. This equation is solved for and the results substituted in equation (23) to evaluate the element forces, Q and X. These equations are used in the form:

$$Q = D_Q^{-1} P_Q^T \Delta \quad X = D_X^{-1} P_X^T \Delta$$

$$\text{or} \begin{Bmatrix} Q \\ X \end{Bmatrix} = S\Delta \text{, where } S = \begin{bmatrix} D_Q^{-1} & 0 \\ 0 & D_X^{-1} \end{bmatrix} \begin{Bmatrix} P_Q^T \\ P_X^T \end{Bmatrix} \qquad (2\text{-}26)$$

The calculations involved in generation consist of those required to develop the coefficients in the stiffness matrices. Generally, these are developed directly rather than by explicitly forming the indicated triple product. Elimination errors are all the errors involved in solving the simultaneous equations (25) to determine and equation (26) to evaluate the unknown forces, Q and X, of the structure.

In the force method Q and X are evaluated first and then displacements are found. Q and X are found by solving equation (24) for Q and substituting the result back in equations (23). Neglecting ϵ_1 and ϵ_2 this gives

$$-D_Q P_Q^{-1} P_X X - D_Q P_Q^{-1} F + P_Q^T \Delta = 0$$

$$D_X X + P_X^T \Delta = 0 \qquad (2\text{-}27)$$

Eliminating Δ from equations (27) gives

$$(P_X^T P_Q^{-1T} D_Q P_Q^{-1} P_X + D_X) X = -P_X^T P_Q^{-1T} P_Q F \qquad (2\text{-}28)$$

This equation is solved for X. Q, found from equation (24), is

$$Q = -P_Q^{-1}\left[P_X X - F\right] \tag{2-29}$$

and Δ is given by,

$$\Delta = -P_Q^{-1^T} D_Q \, Q \tag{2-30}$$

Generation errors are the manipulation errors associated with all operations required to develop the coefficients in equations (24) and (25). Elimination errors are the errors associated with the operations involved in solving the structural equations to define the unknown internal forces X (equation (28)), the forces Q (equation (29)), and the unknown displacements (equation (30)).

These structural analysis manipulation errors will be examined in the next two sections. Section 3 considers displacement method errors and Section 4, force method errors.

Section 3

DISPLACEMENT METHOD ERROR ANALYSIS

This section considers the structural analysis manipulation errors in generating the coefficients in the structural equations and in solving them. It examines the error magnitudes and describes the implications of these errors on problem solutions. It provides the engineering-analyst with guidelines for reducing, estimating and measuring manipulation error. It suggests guidelines for the programmer to reduce and measure errors.

Generation Errors

Generation errors in the displacement method include the manipulation errors incurred in development of the loading, stress, and stiffness matrices for each element of the structure. These errors also include the errors evoked in forming the loading and stiffness matrix for the complete structural system (i. e., the global loading and stiffness matrices).

Generalizing the formulation of Melosh[14], the loading matrices for a finite element are written in the form,

$$
\begin{aligned}
f_{pi} &= I_2 \, C_o \, p_i \\
f_{gi} &= I_3 \, C_o \, M_a \, g_i \\
f_{mi} &= I_3 \, C_o \, M_a \, C_o^{\,T} \\
f_{ti} &= I_3 \, C_o \, D_I \, E_L \, \epsilon_{Ti}
\end{aligned}
\qquad (3\text{-}1)
$$

where

f_{pi} = loading vector for pressure loading of finite element "i".

f_{gi} = loading vectors for field accelerations due to body forces. (D'Alembert forces) in element "i".

f_{mi} = loading matrix due to local accelerations at each joint. (mass matrix) of element "i".

f_{ti} = loading vector for thermal forces treated as body forces in element "i".

and the linear operators given in equation (1) perform the following functions:

C_o transforms displacement coordinates from the global to the local system. The global system is the set of coordinates used in expressing equations (2-23) and (2-24).

I_n performs integration. The subscript n denotes the maximum dimension of the integration space. For example, if n = 2, integration is over a surface.

p_i defines pressure distribution.

g_i defines the acceleration potential over the volume.

M_a defines mass distribution over the volume.

D_I is a differential operator. In this case it transforms displacements into strains.

E_L is a matrix of elastic constants. This transforms strains into stresses.

ϵ_{Ti} defines the distribution of thermal strains over the element before dimensional changes are permitted. Note that the loading resulting from these deformations implies the existence of ϵ_i in equations (2-23)

The stress matrix for a finite element can be formalized as:

$$S_i = I_2 E_L D_I^T C_o^T \tag{3-2}$$

where

S_i is the matrix of stress coefficients used by equation (2-23) to define stresses in the element "i"

The element stiffness matrices are similarly given by:

$$K_i = I_3 C_o D_I E_L D_I^T C_o^T \tag{3-3}$$

It is noted that the coefficients are not usually developed by performing the operations defined by equations (1), (2), and (3) since they can be developed more economically otherwise. Defining the operations this way, however, provides a simple and sufficiently accurate way to estimate the number of calculations involved.

The summation of the element loading and stiffness matrices define the global loading and stiffness matrices:

$$F_q = \sum_{i=1}^{f} U_i^T f_{qi} \qquad q = p,\ q$$

$$F_m = \sum_{i-1}^{f} U_i^T f_{mi} U_i$$

$$K = \sum_{i=1}^{f} U_i^T k_i U_i \qquad (3\text{-}4)$$

where f is the number of finite elements in the structure and the matrix U is a permutation matrix, i.e., a matrix in which each row and column contains only one component equal to 1.0. The permutation matrix has been described by Arqyris[15] as a Boolean matrix. Again, the formal definition of equation (4) is not used in a computer program which develops the structural coefficients because it requires more calculations than necessary. This form is convenient in describing the process.

Attrition errors in developing the loading matrices are negligible. The number of calculations required for the development of the matrices is most when the mass matrix is involved. Here, the number of calculations is of the order $3j^2(2j-1)$ where j is the number of generalized coordinates. The matrices of equation are assumed to be square, fully populated and of order j. For a rectangular prism (an element with a large number of generalized coordinates j = 24, since there are eight joints with three degrees of freedom per joint. Then, the number of calculations would be about 0.081×10^6 including additions, subtractions, multiplications, and divisions. This many calculations are not sufficient to involve significant attrition error since it is much less than the 13.4×10^6, according to the analysis of Section 2.

Development of the coefficients in the loading matrix may, however, involve critical arithmetic. The integrations require calculating the lengths, areas, and volumes of the structural elements using data defining the coordinates of the bounding surfaces of the element. Lengths are obtained by differencing these coordinates; areas and volumes by performing calculations with these differences. Critical arithmetic will be involved if the coordinates describing the boundaries of an element are chosen in a coordinate system so that the difference of the coordinates is incommensurate to the true lengths. For example, if the coordinates of two points on a line are given by (472.1, 0, 0), (472.2, 0, 0), the length of the element must be nearly 0.1 if the error in performing the integration is to be negligible.

Even if the length is satisfactorily described, critical arithmetic can be involved in defining the coefficients in the orientation matrix C_o. The coefficients in this transformation matrix represent cosines of the element surface normals and thus represent ratios of the difference of the coordinates describing the element. If the two lengths defining the direction cosines are a and b, avoidance of critical arithmetic requires that the errors E_a and E_b be such that

$$E_a/a - E_b/b = e_a - e_b$$

is small compared with one.

Critical arithmetic may also be involved in development of the coefficients in the differencing matrix. This matrix, like the integration matrix, requires an accurate representation of the lengths of the element. Again, critical arithmetic will only be involved if the difference in coordinates bounding the element are significantly different from the projections of the element on the axis involved. Utku and Melosh[16] show how this error can destroy measurements of discretization error.

If critical arithmetic is avoided, the small manipulation errors involved in the loading coefficients will be negligible. In linear structure analyses, the change induced in displacements due to a relative error, "e", in loading is of the order of "e." Thus an error in the eighth digit in the loading matrix will only imply an error in the eighth digit in the displacement predictions.

Attrition errors in development of the stress coefficients are also small and can be neglected. The number of calculations is of the order of $36j^2 - 18$. For the solid prism this indicates about 0.21×10^6 calculations. This is small compared with 13.4×10^6, and therefore, the maximum attrition error is negligible.

Critical arithmetic in stress coefficient generation, as for the loading matrix, involves the data describing the location and orientation of the element in three dimensional space. If these errors are avoided, manipulation errors in the stress coefficients will be small.

If errors do arise, their effect may only be local. Errors in the stress coefficients only affect the prediction of stresses for that particular element. Manipulation error in stress coefficients will not affect the accuracy of any deflection predictions per se.

Attrition errors in the development of the element stiffness matrices are small and rarely significant. The number of the calculations required for developing the stiffness matrix is of the order $8j^3 - 4j^2$. For the rectangular prism this indicates 0.108×10^6, calculators, a negligible number compared with 13.4×10^6. Critical arithmetic again involves the basic geometric description of the structure.

Though errors in development of the element stiffness matrices will be small, they will, however, affect the response of the total system. To obtain a measure of the importance of these errors, define a relative error measure, e_R as the energy implied in the rigid body modes divided by the energy in the elastic mode with the smallest energy. If the stiffness coefficients have no error, e_R will be zero. For simplicity, consider that the stiffness matrices are written in the local coordinate system for the element. (This simplification is no restriction since the matrix can always be transferred into this system.) Then the stiffness matrix for a rod element can be taken in form,

$$k_{Ri} = \frac{AE}{a} (1 + E) \begin{bmatrix} 1 + E_{11} & -1 - E_{12} \\ -1 - E_{12} & 1 + E_{22} \end{bmatrix} \tag{3-5}$$

where A is the crossectional area; E, Young's modulus; "a," element length and the E and E_{ij} are the errors contributed by manipulation error. It is noted that the element stiffness matrix for a torque tube and shear panel is the same as for a rod, within a scale factor, so that conclusions for the rod have wider application.

Since the stiffness matrix is symmetric, it is customary to develop only half of this matrix and to reflect it about the main diagonal or to sequence the calculations so that the symmetry of the matrix is insured. Therefore, it is assumed that $E_{ij} = E_{ji}$. The symmetry of the exact matrix about the minor diagonal insures that the $E_{11} = E_{22}$.

Calculating the strain-energy for a rigid translation and for elongation, the error measure can be written as

$$e_{RT} = \frac{(E_{11} - E_{12})}{(2 + E_{11} + E_{12})} \frac{q_T^2}{q_E^2} \approx \frac{(E_{11} - E_{12})}{2} \frac{q_T^2}{q_E^2} \tag{3-6}$$

since E_{11} and $E_{12} << 1$ and

e_{RT} is error ratio for the translation rigid mode

q_T is the amplitude of rigid translation

q_E is the amplitude of elastic deformation

The rigid body and elastic modes are exact regardless of the magnitudes of the errors. The significance of manipulation errors is a function of the amount of rigid body motion involved in the element as compared with the elastic motion as well as the distribution of errors in the matrix. Since the relative errors will be of the order of 2^{-p} compared

with 1, the rigid body motion for the element must be much, much larger than the elastic response if the errors due to the manipulation are to significantly distort the response predictions. It is noted, however, that the relative values of manipulation errors E_{ij} may represent the rigid body mode with negative energy. The implication of negative energy in the computer simulation is significant even when small quantities are involved.

Performing the same type of analysis for the beam as for the rod requires considering an element stiffness matrix of the form

$$K_{Bi} = \frac{6EI}{a^3}(1 + E_i) \begin{bmatrix} 2 + E_{11} & & & \\ 1 + E_{12} & \frac{2}{3} + E_{12} & & \\ -2 - E_{13} & -1 - E_{23} & 2 + E_{11} & \\ 1 + E_{14} & \frac{1}{3} + E_{24} & -1 - E_{12} & \frac{2}{3} + E_{44} \end{bmatrix} \tag{3-7}$$

The beam representation includes two rigid body modes: a translation mode and a rotation. There are thus two ratios to consider. Calculating the eigenvalues and eigenvectors for this stiffness matrix and developing the energies associated with the rigid and elastic modes treating unit length vectors, leads to the ratios,

$$e_{RT} \approx \frac{(E_{11} - E_{13})}{0.142} \frac{q_T^2}{q_E^2}$$

$$e_{RR} \approx \frac{(E_{11} - 4E_{12} + E_{13} - 2E_{14} - 2E_{23} + 2E_{22} + 2E_{24} + 2E_{44})}{0.712} \frac{q_R^2}{q_E^2}$$

$$\approx \frac{(E_{11} - E_{12})}{0.089} \frac{q_R^2}{q_E^2} \tag{3-8}$$

where e_{RR} is the error ratio and q_R the amplitude of the rotational rigid mode. Of the two modes, the error in the rotational energy is much more significant than in translational. Negative energies can also be implied, depending upon the distribution of the relative errors in the element stiffness matrix. Again, the error is important only if the rigid body mode deformation is much, much greater than the elastic, though errors in beam matrices are 24 times more important than for rods, tubes or shear panels.

It is concluded that the effect of stiffness coefficient manipulation error on predictions of response are only important if the element undergoes a large rigid body motion compared with the elastic motion (2^p compared with 1). The errors could be expected to be a problem in the treatment of a cantilever rod or beam in which the load is near the point of fixity so that elements far from the root would undergo only rigid motions.

Errors in forming the global loading matrices are negligible and those in forming the global stiffness matrix are rarely significant. The calculations are the additions and subtractions of a sequence of components. Since few elements are added together to form a particular coefficient in the structural equations (usually between 3 and 10 and, rarely, as many as 15), the truncation error in the accumulation can normally be disregarded.

However, critical arithmetic may be involved in adding incommensurate loading or stiffness coefficients. In the case of the loading coefficients, the relative errors induced by critical arithmetic have an affect of 2^{-p} on predicted deflections. For the stiffness matrix the affect of this critical arithmetic can be deterioration of the process of solving simultaneous equations.

Guidelines for the Analyst. - The significant errors that arise in generation, due to critical arithmetic, can be minimized by proper formulation of the structural problem. This involves locating joints to avoid incommensurate adjacent stiffnesses, sequencing elements to reduce series addition error, and choosing coordinate systems to yield good measures of structural geometry.

If "s" binary place representation of the stiffness of an element is to be retained, the ratio of this stiffness to the total stiffnesses at joints of this element must satisfy

$$2^{s-p} < \frac{k_i}{k_T} < 2^{p-s} \qquad (3\text{-}9)$$

where k_i is the stiffness of element i and k_T is the total stiffness of all elements connected to the joint of element i.

Compliance with this formula insures that the generation error in the stiffness coefficients is satisfactorily small. In accordance with equation (9), if p = 27 and ten bit accuracy (about three significant decimal digits) is required, the stiffness ratio must lie between 2 ± 17 (ie. $7.6 \times 10^{-6} < k_i/k_T < 131,072$). If 20 bit accuracy is required, $0.078 < k_i/k_T < 128$.

If the ratio of stiffnesses are excessive, the analyst can relocate gridpoints so the ratios satisfy equation (9). A simple example of joint location so the stiffness ratio is one is shown for the articulated structure of Figure 4a. Since the stiffness of the rod elements is proportional to AE/a and of beams EI/a^3, the analyst defines the spacing of joints of the structure to attain commensurate stiffnesses at the joint where the rod and beam join by spacing beam joints more closely than rod joints.

Figure 4b shows good location of joints for a membrane or plate. In the case of the membrane, stiffness is explicitly proportional to $tE(12(1-\nu^2)$ and implicitly dependent on the ratio of the sides and the squares and products of the side lengths. In both cases equal stiffnesses are added for a sheet of uniform thickness and isotropic material when joints are located to define square panels.

In securing commensurate stiffness for membranes, the implicit factors can be disregarded. This is shown by the data of Table V. This table lists the x direction, U_{11}, and y direction, V_{11}, diagonal stiffnesses for a membrane as a function of the side ratios. The table is based on Turner's triangular membrane used for the four elements of a rectangular panel. It is assumed that no external loads are applied to the center joint and these coefficients are accordingly eliminated from the stiffness matrix. The second and third columns give the diagonal stiffness for displacements at joint one along the x and y axes. The fourth column of the table measures the error in adding stiffnesses assuming that each panel is independently attached to an adjacent square panel. The relative error in this case is defined as the absolute error divided by the smaller number being summed. These data show that very large panel side ratios result in little relative error. Thus, the network shown in Figure 4b satisfactorily avoids incommensurate stiffnesses for a uniform isotropic panel practically independent of panel side ratios.

Since plate stiffnesses are proportional to length cubed, it is expected that panel ratio will be a more important parameter for plates than for membranes. Using the cubic relation, the ratios indicated in Table V indicate that panel ratios up to two will result in negligible error. Panel ratios up to five have acceptable errors.

If incommensurate stiffnesses are to be added, the analyst can optimize the arithmetic by numbering his elements (which accordingly sequences the additions and minimizes error) so that the smaller stiffnesses are treated first. Figure 4c shows the numbering of panels of a variable thickness plate to achieve this optimization.

In cases where the analyst is concerned with the treatment of structures involving structural elements acting in parallel, the element representation should be chosen so that commensurate stiffnesses are involved. For example, analysis of a wing structure using plate elements for the skin and shear panel elements for the spars and ribs will result in combining stiffnesses proportional to element length squared with stiffnesses proportional to length. If these structures are to be treated, manipulation error will be reduced by representing skins as membranes working in parallel with shear webs (ribs and spars) or by using classical beam spars with the plates.

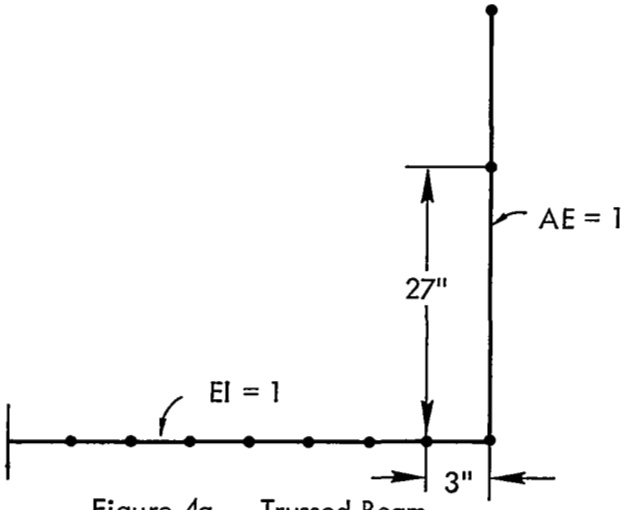


Figure 4a. Trussed Beam

t Constant

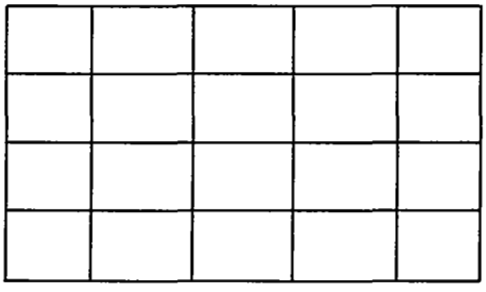


Figure 4b. Panel

Increasing t

Increasing t

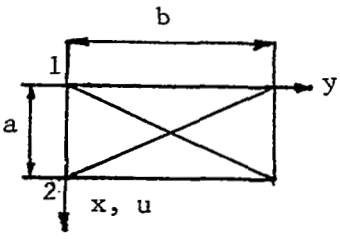
1	2	4	7	11
3	5	8	12	15
6	9	13	16	18
10	14	17	19	20

Figure 4c. Variable Thickness Panel

Figure 4. Joint Location and Element Sequencing

Table V

Relative Membrane Stiffnesses



$$\frac{tE}{(1-\nu^2)} = 1.0$$

$$\nu = 0.3333$$

b/a	u_{11}*	v_{11}*	Rel. Error**
1	.91684372	.91684372	0
2	1.5298992	.82399555	0
4	2.9924205	1.1608030	0
8	5.991504	2.0809400	0
16	11.995061	4.0419431	7×10^{-7}
32	23.997440	8.023968	7×10^{-7}
64	47.998709	16.017984	7×10^{-7}
128	95.999352	32.020992	7×10^{-7}
256	191.9997	64.034496	4×10^{-6}
512	383.9998	128.0652	4×10^{-6}
1024	797.9999	256.1286	4×10^{-6}
2048	1535.999	512.2563	4×10^{-5}

*Diagonal stiffness coefficients.

**When panel is adjacent to a square panel, b = 10, and p = 8.

The analyst should be careful to choose coordinate systems to minimize errors due to critical arithmetic in calculating the lengths and orientation of the elements of the structure. In particular, he should locate the origin of his global coordinate systems near the center of his structure to minimize the span of the coordinate number magnitudes. Coordinate surfaces should be chosen parallel to as many of the elements of the structure as possible to eliminate critical arithmetic in evaluating orientation of these surfaces. (Incidentally, this choice will decrease the number of calculations.) In sensitive cases, the scale of the structure should be selected to maximize the discrimination of the problem geometry description. This is achieved by choosing a scale such that the lengths of the element farthest removed from the origin are satisfactorily represented by the difference of its coordinates. If the option to use local coordinates in describing the geometry is available to the analyst, this capability can be used to eliminate critical arithmetic in the definitions of element geometry and orientation.

If the analyst is introducing element stress and stiffness matrices, these should comply with the programmer's adjustments of generated matrices to eliminate manipulation error. Manipulation error in the loading and stress matrices can usually be neglected.

Guidelines for the Programmer. - To remove analysis inconsistencies, the programmer should adjust final stress and stiffness coefficients. Stress coefficients may be adjusted so that no stress is calculated when rigid body motions are defined.

Stiffnesses should be adjusted to insure that zero energy is involved in rigid body deformations. For rods, for example, the matrix should be forced by making $E_{11} = E_{12}$ in equation (5). In this form, the elastic modes, and rigid body modes, will be exactly represented and exact satisfaction of macroscopic equilibrium is indicated. In the case of a beam stiffness matrix, errors should be adjusted so that the matrix of equation (7) takes the form

$$k_{Bi} = \frac{6EI}{a^3}(1 + E_i)\begin{bmatrix} 2 + 4E_{11} & & & \\ 1 + 2E_{11} & \frac{2}{3} + E_{11} & \text{Sym.} & \\ -2 - 4E_{11} & -1 - 2E_{11} & 2 + 4E_{11} & \\ 1 + 2E_{11} & \frac{1}{3} + E_{11} & -1 - 2E_{11} & \frac{2}{3} + E_{11} \end{bmatrix} \tag{3-10}$$

In this case, both the modes and the energy have been adjusted to insure that the rigid body modes are associated with zero energy. These adjustments guarantee that regardless of what the deformations are, the energy absorbed in rigid body deformation will be precisely zero. Note that by these adjustments, the program forces symmetry of theoretically symmetric matrices.

To protect the analyst from ruinous critical arithmetic in the summing of stiffnesses, a check may be included of the relative size of elements added in calculating the diagonals of the global stiffness matrix. This will eliminate the necessity for the analyst to make this check himself.

Elimination Error

Elimination includes triangularization and forward and back substitution to determine the primary unknowns in the displacement method (displacements) and the calculations to evaluate the secondary unknowns (stresses).

The triangularization may be accomplished by Choleski decomposition. Given a stiffness matrix, the decomposition involves evaluating the matrix L such that

$$L L^T = K \tag{3-11}$$

where L is a lower triangular matrix. This decomposition is always possible if the stiffness matrix, K, is symmetric and positive semi-definite or positive definite.

For a large matrix the number of multiplications required to perform the decomposition is equal to nw (w + 1) where n is the matrix order and w the average wavefront. The wavefront is defined as the number of non-zero elements to the right of the diagonal in row r of the matrix when the decomposition has been completed for all rows less than r. About as many additions as multiplications are required in performing the decomposition. In addition, each row of the decomposition requires taking the square root of the diagonal element.

The forward substitution process consists of solving for y in the expression

$$Ly = F \tag{3-12}$$

where F is the global loading matrix loading. For a large array this operation requires n (w + 1) multiplications and division, and an equal number of additions and subtractions.

The back substitution process involves the solution of the equations

$$L^T \Delta = y \tag{3-13}$$

for Δ, the unknown displacements in the structure; i. e., primary unknowns. This operation also requires n (w + 1) multiplications and divisions and an equal number of subtractions and additions.

Calculations of the secondary unknowns involves the multiplication of the element stress matrices by the vector of primary redundants in accordance with equation (2-26). This calculation requires a maximum of 6 (2j - 1) multiplications and additions per element where j is the number of generalized coordinates for the element.

This process of evaluating the primary unknowns can take advantage of the matrix symmetry and sparseness and thus tends to minimize the number of calculations involved. Errors in this elimination include inherited, square root, and attrition error. Attrition errors involving critical arithmetic require special consideration.

Inherited error is the error existing in the coefficients of the matrix due to prior arithmetic; i. e., the generation calculations. As indicated few calculations are involved in developing these coefficients, critical arithmetic can be avoided, and the consistency of coefficients can be insured. The relative inherited errors will usually be less than 2^{-p}. Representation of a physically realizable system is insured.

The errors involved in obtaining square roots are larger than inherited errors. Algorithms for taking the square root are contained in codes of the computer systems software. These codes develop square roots with a maximum absolute error of one part in the last digit of the floating decimal mantissa. The maximum relative errors are less than $+2^{4-p}$. In IBM Fortran IV software, these errors always result in overestimates of the square root. The average relative error is much less than the maximum. For components from 1 to 1029, the square roots have an average relative error of -0.461×10^{-8} and the maximum is -0.148×10^{-7} when p = 27.

Attrition errors cannot be expected to destroy the consistency of results of the elimination process until problems of much larger order than are currently being treated are involved. If attrition errors are to be significant, $2nw^2 + 4nw$ must be greater than or equal to 13.4×10^6. In the analysis of aerospace structures, wavefronts approaching 200 have arisen. This is unusual, however. In analyzing antenna reflectors, the average wavefront is about 50. The maximum wavefront encountered by the authors occurred in analyzing a quadrant of a 90-foot earth station reflector. Here w was 83 for 436 equations. Thus, for these structures a minimum of 950 equations and an expected 2600 equations could be treated without concern for attrition error. It is noted that even if the number of calculations is greater than 13.4×10^6, attrition errors will not necessarily be important since this is an upper bound error to attain a relative error of five percent.

Though consistency of answers may be assured, small attrition errors may incur significant error in response predictions. This can occur due to the propagation of these errors with amplification, the cumulative effect of many errors, or critical arithmetic. The next paragraphs examine the effect of these complications in the numerical analyses of structural systems consisting of a series of finite elements.

Errors in Evaluating Primary Unknowns. - The critical structural systems involved in displacement analyses are pure series systems such as shown in Figure 5a, b, c, d and e. Pure parallel systems are not critical for manipulation errors because they result in small systems of equations whose solution involves few calculations.

The systems in Figure 5 will be examined to determine the magnitudes and characteristics of manipulation errors. Figure 5a shows a simple series rod system. Figure 5b shows a cantilevered beam represented by finite elements. This figure could also represent a series of rectangular prisms. Figure 5c represents a cylindrical shell. The system of Figure 5d consists of a number of rectangular panels representing a membrane. Each panel is subdivided into four triangles and provides resistance only to loads in x-y plane. Figure 5e illustrates a plate system providing resistance only to z direction loads. All systems are clamped on the right edge and free on the left.

In the sequel regular, increasing and decreasing systems are considered. In the regular system, all finite elements have the same stiffness. In the increasing system, the stiffness of each successive element is twice the previous, proceeding from left to right. In the decreasing system, each successive element has half the stiffness of the previous element.

Series rod systems error analysis: The stiffness matrix for the series rod system is given by

$$K = \begin{bmatrix} W_1 (1+\alpha) & & & \\ -W_1 & W_1 + W_2 & \text{Sym.} & \\ 0 & -W_2 & W_2 + W_3 & \\ & & \ddots & \\ & & -W_i & W_i - 1 + W_i \end{bmatrix} \quad (3\text{-}14)$$

where $W_r = A_r E_r / a_r$, $r = 1, 2, 3 \ldots f$, is the axial stiffness of the rod element and f is the number of finite elements (rods). For the rod case f is also the order of the stiffness matrix. α = relative stiffness of a spring located at the free end of the rod.

Figure 5a. Rod

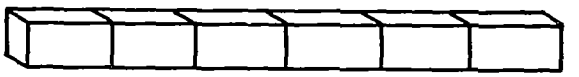


Figure 5b. Beam or Solid

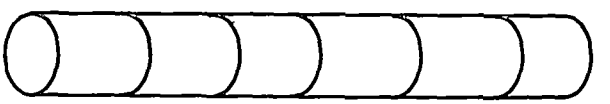


Figure 5c. Cylinder

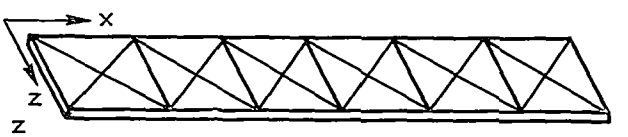


Figure 5d. Membrane

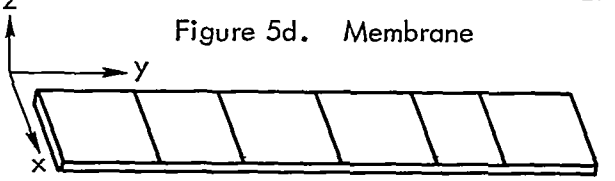


Figure 5e. Plate

Figure 5. Regular Series Structures

Disregarding the inherent error, the first and second diagonals of the decomposition matrix L are given by

$$\ell_{11} = W_1^{1/2} \ (1+\alpha)^{1/2}$$

$$\ell_{22} = W_2^{1/2} \left(1 + \frac{W_1 \alpha}{W_2 (1+\alpha)}\right) \tag{3-15}$$

Relative values of the first and second diagonals of the L matrix can be defined by dividing these terms by the values of the diagonals when the spring stiffness α is zero.

This gives

$$e_{11} = (1+\alpha)^{1/2} - 1$$

$$e_{22} = \left(1 + \frac{W_1 \alpha}{W_2 (1+\alpha)}\right)^{1/2} - 1 \tag{3-16}$$

If α were introduced in the rth equation, equation (16) would also apply for the rth and r + 1 st diagonals. Therefore these equations indicate how an added stiffness in a given diagonal effects the neighboring diagonal.

Assuming that all the W_i are equal (regular system) and interpreting α as an error introduced by manipulation, equations (16) can be studied to show the characteristics of error propagation in the series rod system. This is achieved by solving equations (16) simultaneously to eliminate α and express e_{11} in terms of e_{22}.

A plot of this relationship is shown in Figure 6. This curve shows that when a positive stiffness perturbation ($\alpha > 0$) is introduced in diagonal r successive diagonals retain the perturbation with reduced magnitude. The rate at which the perturbation is damped can be deduced by repeated use of the data in the upper quadrant of Figure 6. The curve shows that positive propagations are bounded by an asymptote at $\sqrt{2} - 1$. Thus, even if a positive relative error of infinite magnitude is introduced at row "r" it will be reduced to a relative value of 0.414 in equation r + 1. The lower left-hand quadrant of Figure 6 shows that a negative perturbation in a diagonal results in an increased negative perturbation in the next diagonal. A negative asymptote exists where the input perturbation has a relative value of -1. Then the next diagonal will have a relative perturbation of minus infinity.

A physical interpretation of the propagation of negative α is that if a load is introduced at joint "r" in the plus direction, then a negative infinite displacement occurs at some joint greater than "r". Thus, in this case, solution of the structural equations is physically meaningless. Mathematically the perturbation has been such that the implied stiffness matrix is indefinite.

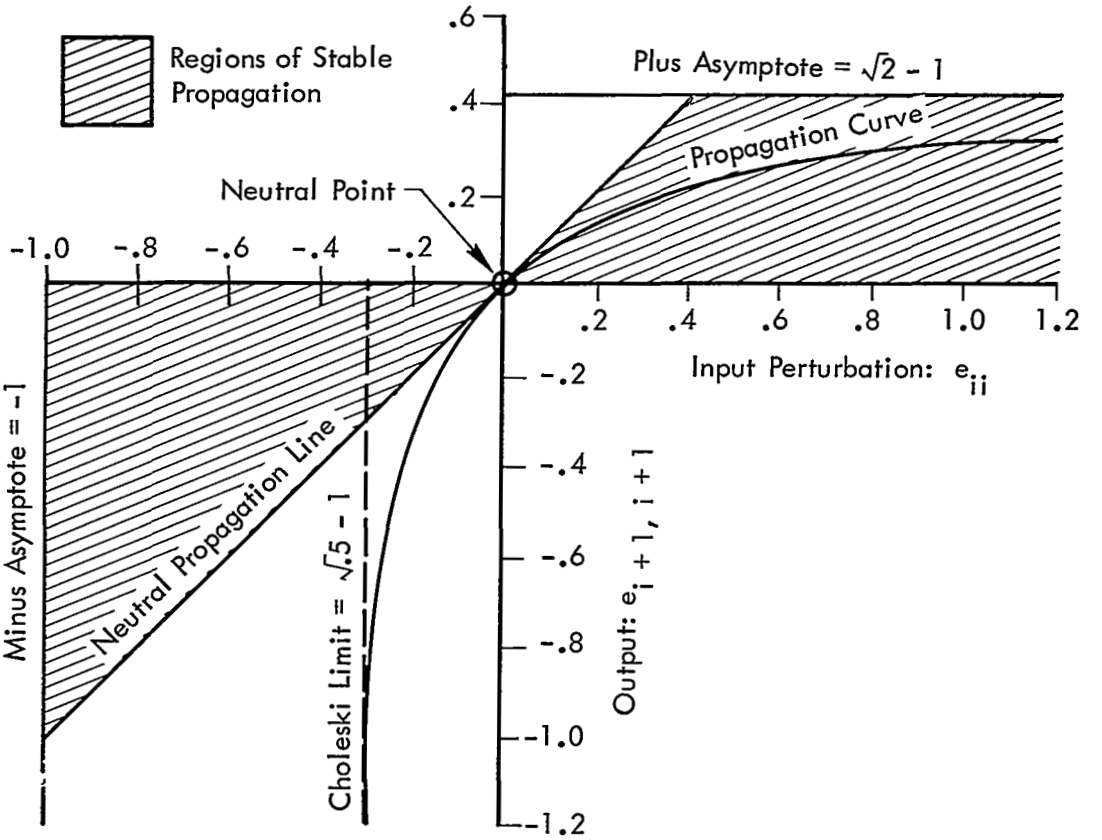


Figure 6. Decomposition Perturbation Propagation Regular Rod

Tighter restrictions can be imposed by considering the singularity of the stiffness matrix, K. One restriction on admissible negative α is that successive equations must not be dependent. This condition is represented by the Choleski limit on Figure 6. It occurs when the relative negative perturbation is -5, since, in this case, $\ell_{22} = 0$.

A still tighter restriction is that the total stiffness matrix must not be singular. The determinant of the regular rod stiffness matrix is $W_1^{2f} (f\alpha + 1)^2$. Therefore, the maximum error, α, that may be introduced in the first equation to avoid singularity must be - 1/f. Conversely, if the error is -2^{-p}, 2^p equations (elements in series) can be treated. The relative error in the determinant is

$$e = 2^{-p} f \tag{3-17}$$

The neutral propagation line shown on Figure 6 defines the regions of stable and unstable propagation. Stable propagation is defined as propagation in which the perturbation is reduced in magnitude in successive diagonals of the decomposition matrix. As shown by the figure, all positive perturbations involve stable propagation whereas all negative errors involved unstable propagation.

Neutral propagation arises due to the fact that the numbers are represented with a finite number of places. Then, a small positive or negative error can persist due to the fact that the nearest finite number is used in the numerical analysis. The greater the number of places involved in the arithmetic, the smaller is the range of neutral propation. In the propagation plot of Figure 7, a neutral stability region exists around the intersection of the axes.

Figure 7 shows plots of the attenuation and amplification of α as a function of the number of successive equations after α is introduced. Amplitudes of stable disturbances are diminished slowly, whereas unstable rapidly increase.

Figure 8 is similar to Figure 6 but is constructed for cases in which the stiffnesses of adjacent elements differ. Study of these curves confirms that positive perturbations, though they increase, never cause singularity. Negative perturbations are deleterious only when in the unstable region. In addition, these data show that if the second element is stiffer than the first, the relative perturbation is decreased in the successive equation and conversely. The figure shows that non-uniform rods can have more than one region in which neutral stability occurs. These are at each intersection of the propagation curve with the neutral line. The plus asymptote and Choleski limits have not been plotted to avoid clutter.

The determinant of the stiffness matrix of the increasing system can be written as a product of a number of factors, only one of which depends upon α. This factor is $(\alpha + 2^{f-1}/(2^f-1))^2$. Therefore, the matrix will be singular for f large, only if $\alpha = -0.5$.

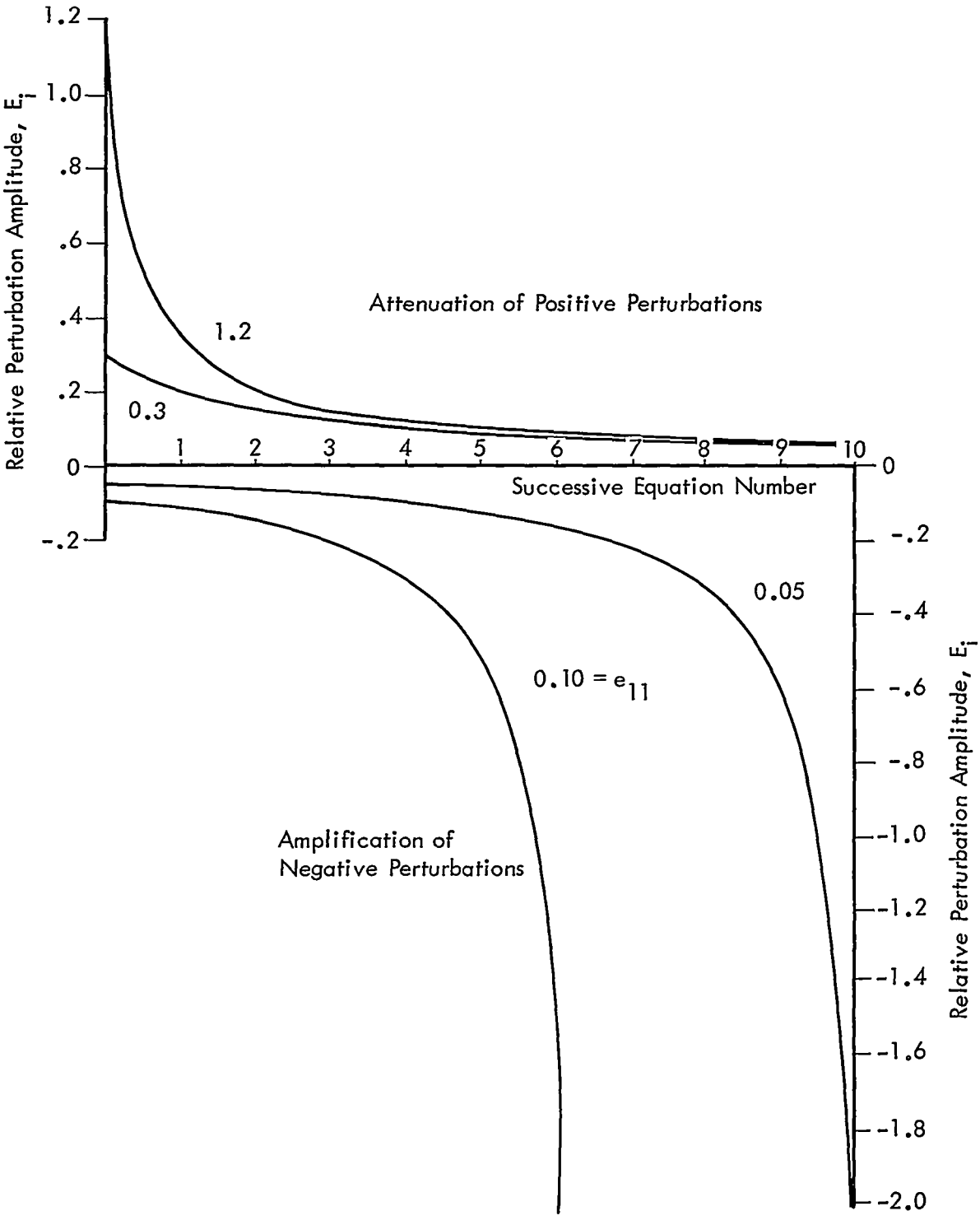


Figure 7. Propagation of Decomposition Errors - Regular Rod System

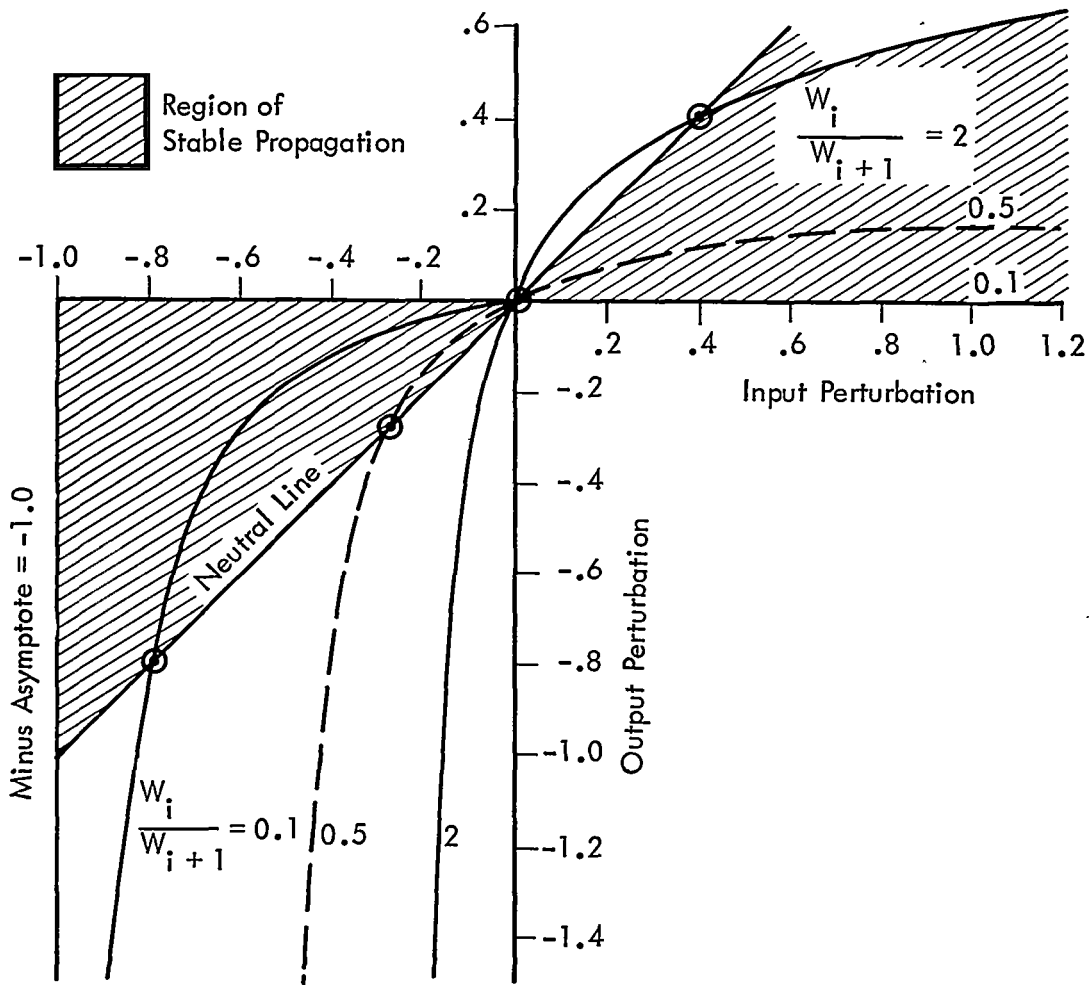


Figure 8. Decomposition Perturbation Propagation - Irregular Rod

The determinant of the stiffness matrix of the decreasing system has an α factor, $(\alpha + (2^{f-1})^{-1})2$. Thus, this matrix will be singular when $\alpha = -2^{-f}$. The decreasing system is therefore more sensitive to error than the increasing system.

Computer experiments verify this analysis. If f = 20 for the decreasing system, the stiffness matrix should be singular if $\alpha = -(2^{20}-1)^{-1}$. Then $e_{11} = 0.-952 \times 10^{-6}$. Computer experiments show that the matrix is positive definite when $e_{11} = -0.482 \times 10^{-6}$ and indefinite when $e_{11} = -0.961 \times 10^{-6}$.

For an irregular structure, growth in the error in the diagonals can be predicted by repeated use of equations (15) or estimated from Figure 8. These ways are simpler than developing the determinant of the stiffness matrix.

The propagation curves of Figures 6 and 8 involve what may be called conservative propagation. It is conservative in the sense that if a succession of equations are treated the sequence of treatment will not affect the relative perturbation propagation through the complete set. The propagation can be made unconservative by introducing new perturbations due to manipulation errors in performing each step of the decomposition.

To estimate the maximum number of rod equations that can be treated without concern for error propagation, it is assumed that a negative error is introduced in each row of the decomposition due to inherent, square root, or attrition error. For a regular rod, propagation is nearly stable so these errors will be summed. Assuming all this error is introduced as a perturbation in the first equation, the number of equations that can be treated for a regular rod without an indication of singularity in the Choleski process is given by $f = 2^{p/2} \approx 6000$ if p = 27. Fewer equations can be treated if the rod elements decrease in stiffness and more if stiffnesses are increasing.

These small perturbations, however, can introduce significant errors in the solution even if they are neutrally stable. To consider the implications of these errors in the decomposition upon the solution for the rod, assume that the typical perturbed equation is given by the difference equation

$$- E_{\alpha} u_{r-1} + 2E_{\beta} u_r - E_{\alpha} u_{r+1} = 0 \tag{3-18}$$

where E_{α} and E_{β} differ from 1 by α and β, and r is the number of the degree of freedom (joint sequence number). Solution of this difference equation is given by

$$u_r = C_1 \left(\frac{E_{\beta}}{E_{\alpha}} + \left[\left(\frac{E_{\beta}}{E_{\alpha}} \right)^2 - 1 \right]^{1/2} \right)^r + C_2 \left(\frac{E_{\beta}}{E_{\alpha}} + \left[\left(\frac{E_{\beta}}{E_{\alpha}} \right)^2 - 1 \right]^{1/2} \right)^r \tag{3-19}$$

Considering the rod pinned at the left and loaded at the right, the boundary conditions are

$$u_o = 0; -E_\alpha u_{f-1} + E_\beta u_f = \frac{P \cdot a}{AE} \tag{3-20}$$

Using these conditions to define the arbitrary constants, C_1 and C_2, letting $E_\alpha = 1 + \alpha$, $E_\beta = 1 + \beta$ with $\alpha \approx \beta << 1$ and expanding terms in a binominal series gives the displacement as

$$U_r \approx \frac{P \cdot a \cdot r}{AE}\left(1 - f\left[\beta - \alpha\right] - \alpha\right) \tag{3-21}$$

When $\alpha = \beta = 0$, equation (21) gives the same displacements at each joint as the solution of the differential equation for the rod, regardless of the number of joints involved. The relative error when $\alpha \neq 0$, $\beta \neq 0$ is given by

$$e \approx -f(\beta - \alpha) \tag{3-22}$$

This error is independent of r; the joint sequence number, and proportional to the total number of equations or finite elements considered.

Suppose displacement boundary conditions are imposed so that $u_o = 0$ and $u_f = 1$. Then the displacements to a first order are r/f and are independent of manipulation error.

The expansion of equation (19) for $\alpha \approx \beta << /$ and $r\alpha << 1$ makes if apparent that the relative error can at worst vary linearly with the joint number. The maximum error will be of the order given in equation (22). Taking $\alpha \approx -\beta \approx 2^{-p}$ the maximum error in the displacements is less than five percent when $p = 27$ if $f < 3.35 \times 10^6$.

Critical arithmetic is involved in the decomposition process in evaluating diagonal elements of the decomposition matrix. These involve calculations of the form

$$\ell_{rr} = \left(K_{rr} - \sum_{j=1}^{r} \frac{\ell_{jr}^2}{\ell_{jj}}\right)^{1/2} \tag{3-23}$$

where the ℓ_{jr} is coefficient in row j column r of the array when row j of the decomposition is being formed. Since all the ℓ_{jj} are positive when K is positive definite, as the decomposition proceeds a particular diagonal is persistently reduced. The stiffness matrix is numerically singular if the total reductions in the diagonal result in a relative zero for the diagonal.

The singularity can be predicted by estimating the decomposition diagonal using its physical interpretation. The rth diagonal is the square root of the stiffness in the rth degree of freedom when all lower freedoms in the set of equations are unrestrained and all higher are fixed. It measures the force to induce a unit displacement in the rth freedom.

Consider use of this criterion for the series rod systems. Assume that equations are sequenced from root to tip. The numerical singularity will be exhibited first at the tip joint. For the regular rod, the square of the last diagonal of the decomposition matrix is, by its interpretation, (AE/fa). Its original value is (AE/a). Then, the set of equations will be singular when

$$\frac{AE}{fa} \leq 2^{-p} \frac{AE}{a} \quad \text{or,} \quad f \geq 2^{p} \tag{3-24}$$

The relative error is given by

$$e = 2^{-pf} \tag{3-25}$$

With joints sequenced from root to tip, the increasing system will result in fewer and the decreasing system greater values of f than given by equation (25). For the increasing system, the deflection of the tip of the f joint rod for unit load when joints inboard are free to move is

$$u_F = \frac{1}{W_1} \sum_{i=1}^{f} 2^{-i} = \frac{2}{W_1} (1 - 2^{-f}) \tag{3-26}$$

The stiffness is the reciprocal of u_f

Since the original value of the f^{th} row stiffness is $W_1 2^{f-1}$ the matrix will be singular when

$$\frac{W_1}{2} (1 - 2^{-f})^{-1} \leq 2^{-p} W_1 2^{f-1} \tag{3-27}$$

or

$$f \approx p.$$

For every reduction in f by 1, one more accurate binary place will occur in the determinant, i. e.

$$e = f/p \tag{3-28}$$

Figure 9 shows the results of computer tests demonstrating the validity of this critical arithmetic analysis for the increasing rod system. Singularity occurs when expected. As f is reduced, the accuracy of the determinant varies nearly as predicted. Variances from the curve can be explained by errors in forming square roots in the Choleski process.

Errors in the forward and backward substitution are relatively small if the inherent error in the coefficients is negligible. In the forward substitution, the worst error is associated with evaluation of the last unknown, y_f. This is formed for the regular rod by calculations of the form

$$y_f = \sum_{j=1}^{f} P_{d_j} \prod_{i=1}^{j} a_i \qquad (3\text{-}29)$$

where $\prod_{i=1}^{j} a_i$ is a series multiplication with j components. The evaluation of y_f requires the summing of terms of a vector scalar product. Each term summed depends upon the series multiplication. As indicated in Section 2, the errors in forming series multiplications are small and errors in taking vector scalar products would be expected to be small. Consequently, errors in forming y_f are negligible unless $2\,fw^2 \geq 13.4 \times 10^6$.

Critical arithmetic may be involved, depending upon the signs of the P_j loadings. When all the loads are in the same direction, y_f involves the series additions of components and, thus, the final sum will have small relative error as long as the number of additions is less than 26.8×10^6. When the signs differ, the relative error of the result will also be small though more difficult to estimate since it is measured by the sum of the absolute values of the series components.

The operations involved in back substitution are comparable to those in forward substitution involving, instead, the summing of the deformations.

It is noted that when the right hand of the rod is assumed to be fixed, a physical interpretation of the substitution processes is possible. The forward substituion defines the loads in each element of the structure. The back substitution involves a summing of the incremental deformations of each of the elements of the structure to obtain the total deflections of the structure.

Since the errors in the forward and back substitution process correspond to the errors in series operations, the following characteristics can be anticipated for each process separately,

1. The absolute error can be expected to change its rate of growth by a factor of 2 when the answers change by a power of 2. Between changes, growth will vary linearly. The calculated answers will always be less than the exact if all loads have the same sign.

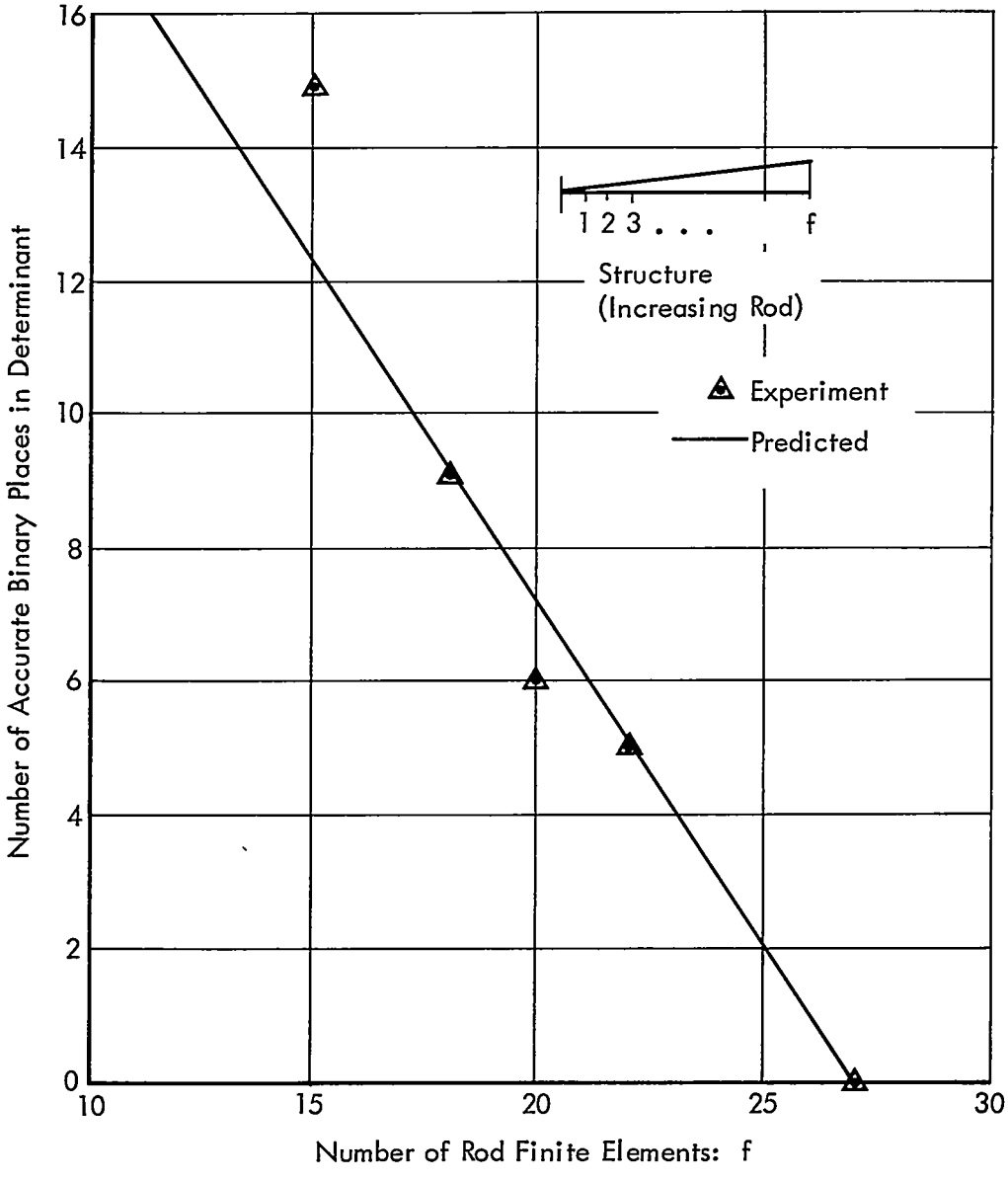


Figure 9. Singularity Criterion Validation

2. Relative errors will increase at worst linearly with the number of components (equations) and will be insignificant for p = 27 unless 13.4×10^6 components are involved, in accordance with equation (2-14).

Confirmation of these characteristics are provided by computer tests for series of rod finite elements. These data are summarized in Figures 10 and 11.

Figure 10 shows the growth of absolute error for back substitution with successive equations in the back substitution process. The decomposition in this case involves no inherent error. These data were developed based on a 400 element rod with a tip load of $(2^{28}-1)$. This loading maximizes the error when p = 27. Predicted deflections are less than the exact at all joints. Changes in slope of the error curve occur when the exponent of the answer changes by a power of two.

The maximum relative error is 1.089×10^{-6} and occurs at the tip because the exact answers vary linearly with the number of the equation being treated. The lower curve of Figure 11 shows the growth of the maximum relative error with the total number of equations treated. These data were obtained by analyzing rods with 100, 200, 400, and 1200 elements. The data exhibits expected characteristics. It is noted that this curve is not bounded by equation (2-14) since the number of components does not lie in the critical range; R_c.

The following characteristics can be anticipated when forward and back substitution are combined so that the errors in forward substitution constitute inherent error to the back substitution:

1. The absolute error will vary linearly with a growth rate equal to the sum of the forward and back substitution growth rates. It will tend to be insensitive to the value of the answer characteristic since the growth rate increases in forward and back substitution and equations are treated in reversed sequence.

2. Relative errors will increase linearly with the number of equations. The relative error is bounded by letting

$$N = \frac{f(f-1)}{2}$$

in equation (2-14) since each of the y_{f-r} are found by adding values involving the sum of (f-r) components where r is the equation sequence number in back substitution. Then, if the error is to be less than five percent, f cannot exceed 7300.

Figure 12 shows the measured growth of absolute error for forward and back substitution combined. Again, the decomposition has no inherent error. These data were developed for the 400 element rod with a load of $(2^{28}-1)$ at every joint. This loading maximizes the error when p = 27. These data exhibit an error which is a linear function of the equation sequence number in the back substitution process.

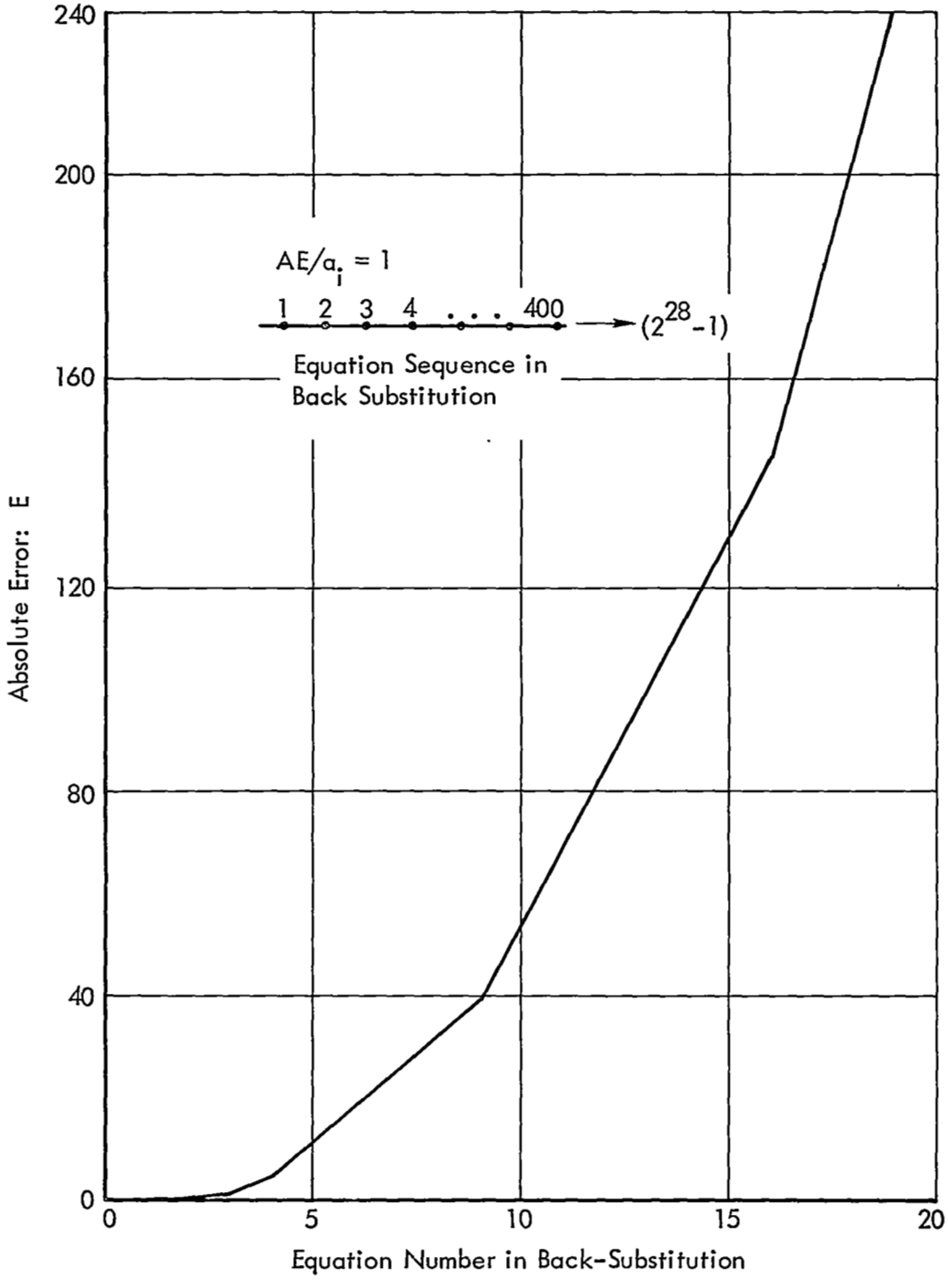


Figure 10. Rod Back Substitution Error

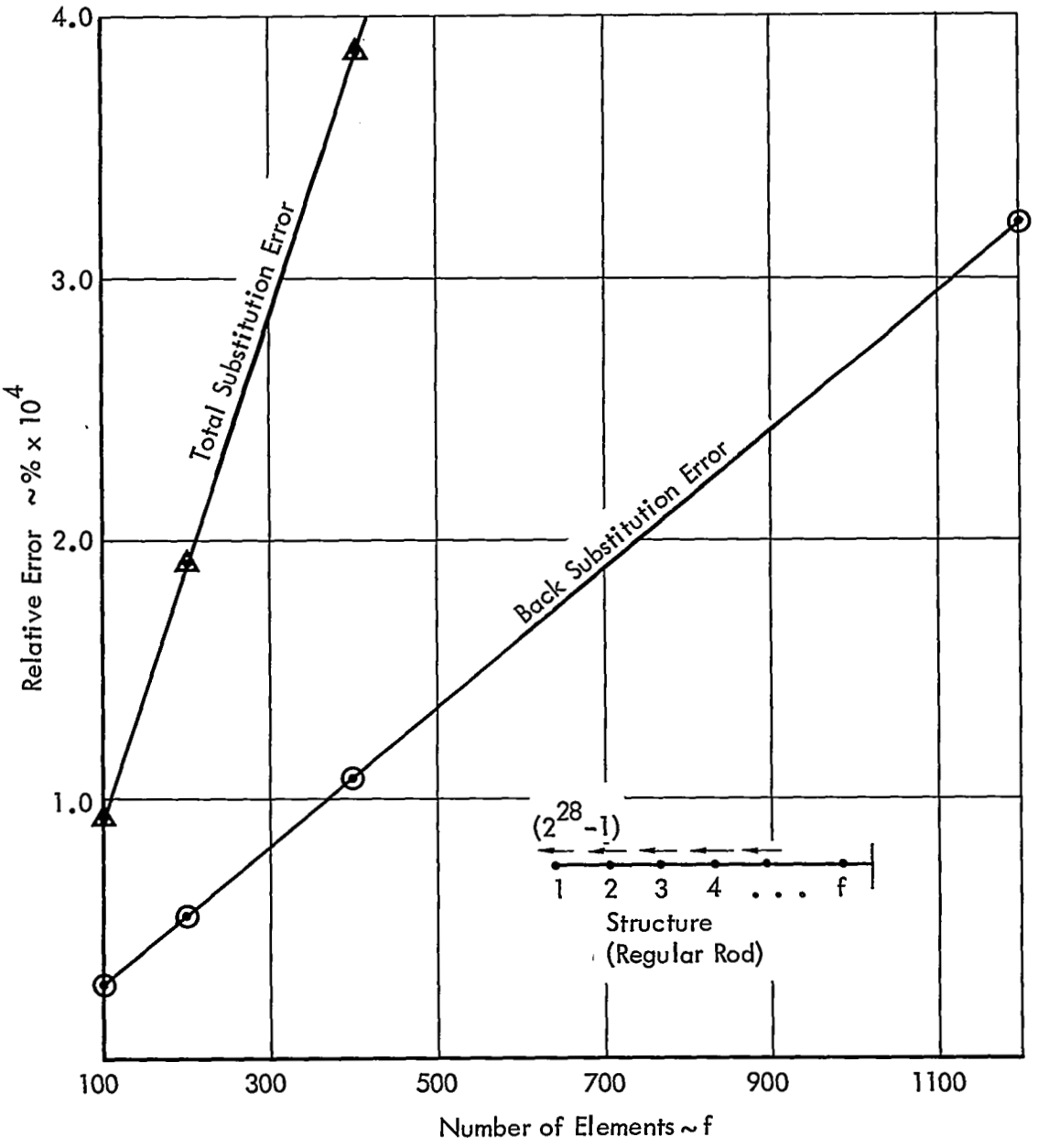


Figure 11. Rod Substitution Errors

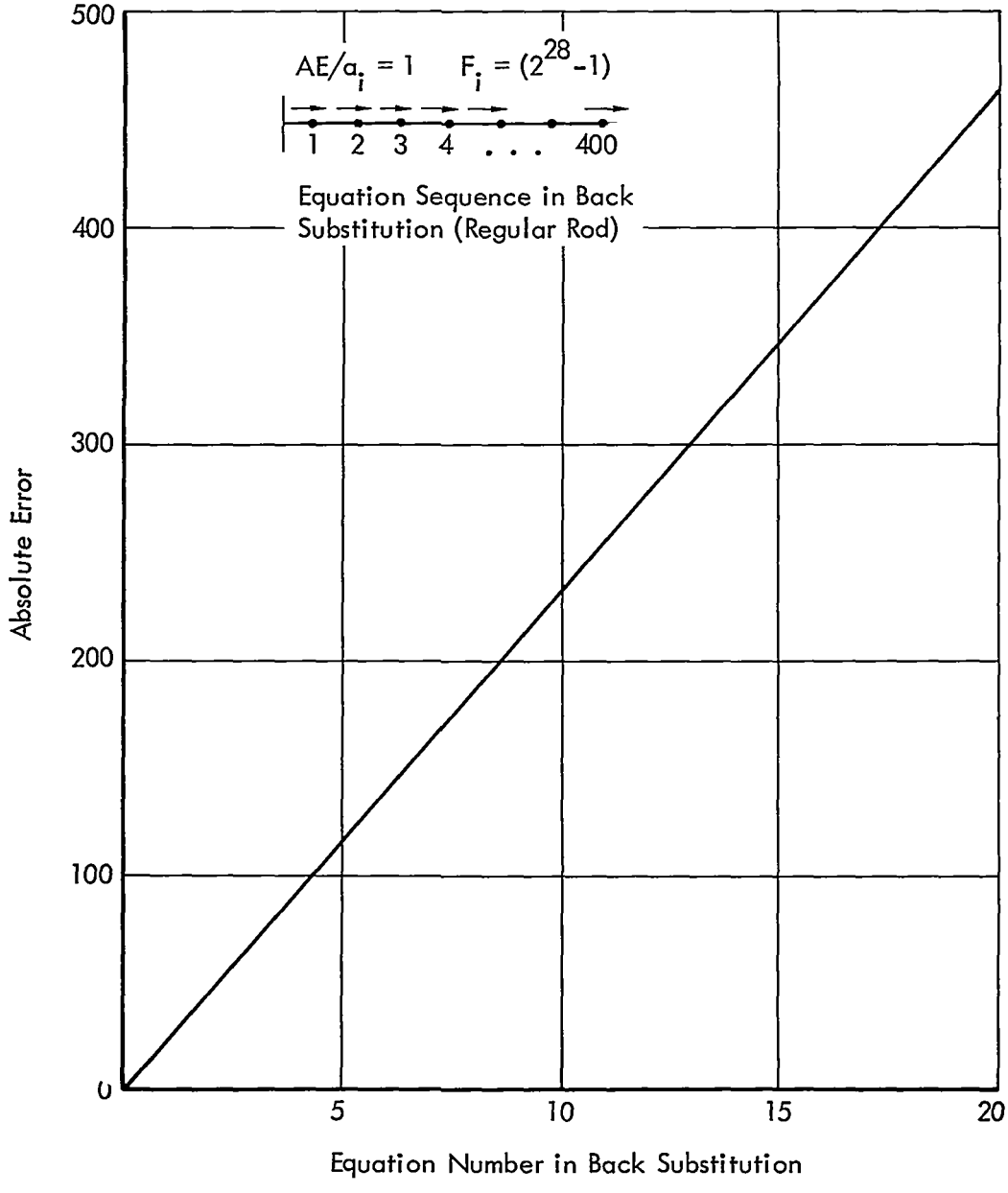


Figure 12. Rod Total Substitution Error

The upper curve of Figure 11 shows the maximum relative error as a function of the number of equations for the combination of forward and back substitution. The errors are small. The relative errors for the combined processes involve 3.57 times the relative error of each process separately. This factor reflects the fact that, in the worst case, the error growth rate and the relative error at a particular growth rate double when the number of components is doubled.

These data demonstrate that the error in substitution in series rod systems is small. The curve of Figure 11 is a straight line. In the worst case, the relative error is given by $e = 0.959 \times 10^{-8} f$. Therefore, the substitution processes in these series rod analysis will yield answers with less than five percent error unless more than 5.2 million elements are involved. This is a better bound than the 7300 obtained by disregarding calculation details.

Series beams systems error analysis: The stiffness matrix for the series beam system is given by

$$K = \begin{bmatrix} \frac{6W_1}{a_1^3}(1+\alpha) & & & \\ \frac{3W_1}{a_1^2} & \frac{2W_1}{a_1}(1+\beta) & \text{Sym.} & \\ -6W_1/a_1^3 & -3W_1/a_1^2 & \frac{6W_1}{a_1^3} + \frac{6W_2}{a_2^3} & \\ 3W_1/a_1^2 & W_1/a_1 & \frac{-3W_1}{a_1^2} + \frac{3W_2}{a_2^2} & \frac{2W_1}{a_1} + \frac{2W_2}{a_2} \ddots \\ 0 & 0 & & \end{bmatrix} \quad (3\text{-}30)$$

where

$W_i = E_i I_i \quad i = 1, 2 \cdots f$

E_i = Young's modulus of segment i

I_i = bending moment of inertia

a_i = segment span

α and β = relative normal displacement and rotational springs at joint 1.

Generalized coordinates are the displacements normal to the beam and the rotation of each joint. Shear deformations are neglected in the representation. Equation (30) is the stiffness matrix for the beam of Figure 5b. Note that there are twice as many equations as there are finite elements. The first of each pair of rows of the stiffness matrix is associated with force equilibrium and is the "force row". The second of each pair is a "moment row"

To examine the propagation of α and β, it will be assumed that when one is non-zero, the other is zero. The relative values of the first and third diagonals of the decomposition matrix assuming $\beta = 0$ are given by

$$e_{11} = (1+\alpha)^{1/2} - 1$$

$$e_{33} = \left(1 + \frac{W_1/a_1^3}{W_2/a_2^3} \quad \frac{\alpha}{(1+\alpha)}\right)^{1/2} - 1 \qquad (3\text{-}31)$$

The relative values of the second and fourth diagonals, assuming $\alpha = 0$, are given by

$$e_{22} = (1+\beta)^{1/2} - 1$$

$$e_{44} = \left(1 + \frac{W_1/a_1}{W_2/a_2} \quad \frac{4\beta}{(1+4\beta)}\right)^{1/2} - 1 \qquad (3\text{-}32)$$

Equations (31) and (32) also define error propagation between any pair of rows r, r + 1 and a second pair r + 2, r + 3 for the series beam when no error has arisen in prior decomposition. Thus, conclusions with respect to these apply at any point of decomposition.

Figure 13 shows a plot of the stability characteristics of the pairs of diagonals. This figure is based on a regular beam, i. e., $W_r = W_{r+1}$. This curve exhibits the same perturbation propagation characteristics as the series rod curves shown in Figure 6. Positive perturbations coverage and negative diverge more rapidly, however. It is noted that the branches shown to the left of the minus asymptote also occur for the rod, though they are not shown on Figures 6 or 8. Figure 13 curves indicate that moment perturbations are more important than force.

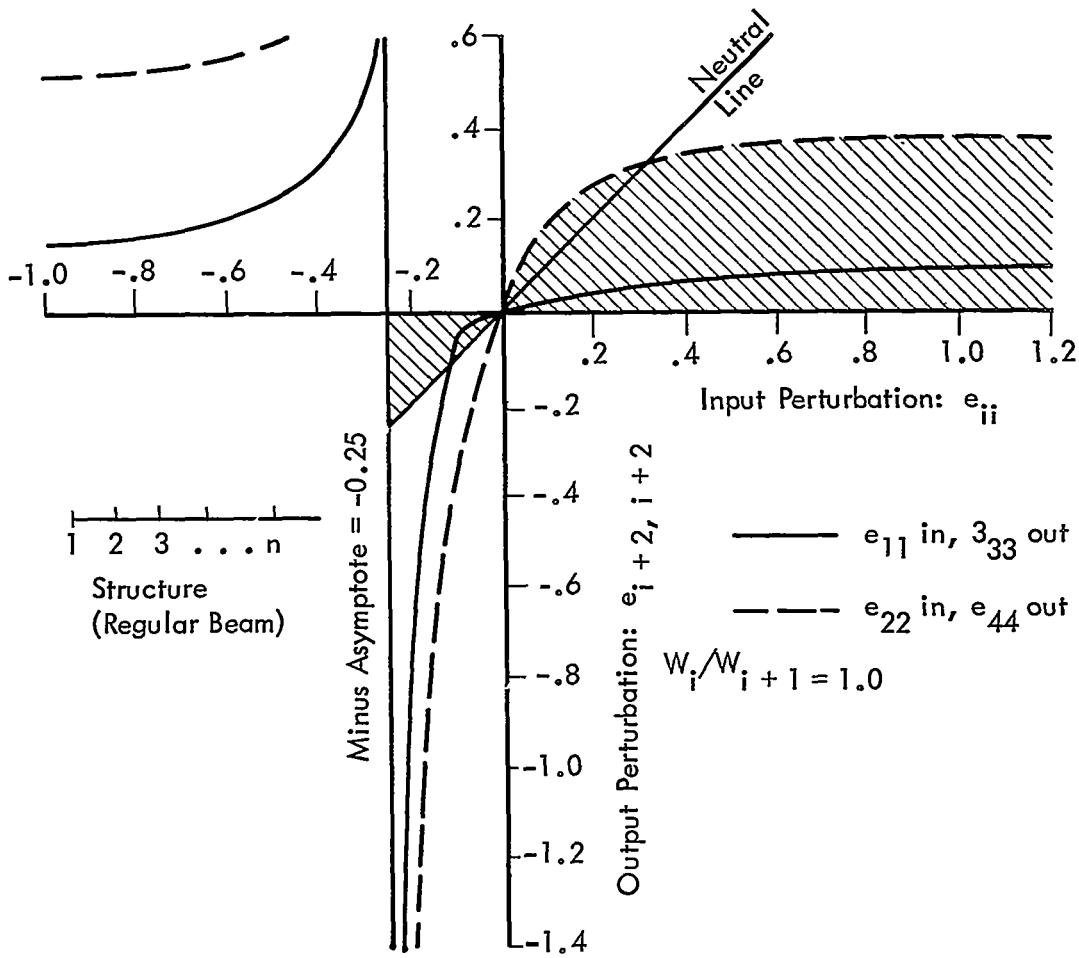


Figure 13. Decomposition Perturbation Propagation

Examining equations (31) and (32), it can be seen that as successive stiffnesses decrease, the process is increasingly sensitive to perturbations. The new propagation curves will be rotated similarly to the way the rod curves of Figure 8 rotate as stiffness ratios decrease. If both α and β are non-zero, it can be shown that curves for e_{33} are not sensitive to small values of α and e_{44} for small β. However, the curves do vary with both of the perturbations.

Because of the coupling between the α and β perturbations, the beam may have more than two neutral stability points. This is reflected by equating the general expression for e_{33} involving both α and β,

$$e_{33} = \left(1 + \frac{W_1/a_1^3}{W_2/a_2^3} \quad \frac{\alpha(1+4\beta)}{4\left(\left[1+\alpha\right]\left[1+\beta\right] - 3/4\right)} \right)^{1/2} - 1 \qquad (3\text{-}33)$$

to e_{11}. Then for the regular beam, regions of neutral stability are seen to arise at $\alpha = 0$ for any β and at $\beta = -1$ for any α.

The number of elements that can be treated without degeneracy if $\alpha \neq 0$ and $\beta = 0$ is twice as many as for the series rods, i.e., if $p = 27$, $f = 134 \times 10^6$. If $\alpha = 0$ and $\beta \neq 0$, the maximum number of elements is 33.5×10^6. If it is assumed that negative errors arise in every equation, 3000 equations (1500 series beam segments) can be treated if $p = 27$.

Consider the implications of persistent small errors in the final structural equations due to decomposition errors. The difference equations are of the form

$$-6w_{r-1} - 3\theta_{r-1}a + 12E_\alpha w_r + E_\alpha \theta_r a - 6w_{r+1} + 3\theta_{r+1}a = 0$$

$$3w_{r-1} + \theta_{r-1}a + E_\alpha w_r + 4E_\beta \theta_r a - 3w_{r+1} + \theta_{r+1}a = 0 \qquad (3\text{-}34)$$

where r and θ are the displacement and rotation of joint "r" $E_\alpha = 1 + \alpha$, $E_\beta = 1 + \beta$, $E_\alpha = 1 + \alpha$ and α, β, and α are small errors of the order of 2^{-p}.

To solve equations (33), let them first be replaced, approximately, by a set of differential equations. Then the solution of the differential equations will be studied. For this purpose let

$$w_r' = \frac{1}{2}(w_{r+1} - w_{r-1}) \approx \frac{\partial w}{\partial r}$$

$$w_r'' = (w_{r-1} - 2w_r + w_{r+1}) \approx \frac{\partial^2 w}{\partial r^2}$$

$$\theta_r' = \frac{1}{2}(\theta_{r+1} - \theta_{r-1}) \approx \frac{\partial \theta}{\partial r}$$

$$\theta_r'' = (\theta_{r-1} - 2\theta_r + \theta_{r+1}) \approx \frac{\partial^2 \theta}{\partial r^2}$$

then equations (27) can be written as

$$-2w_r'' + 4\alpha w_r + 2\theta_r a + \frac{\alpha}{3}\theta_r a = 0$$

$$-6w_r' + \alpha w_r + \theta_r'' a + (6 + 4\beta)\theta_r a = 0 \qquad (2\text{-}35)$$

Since each of the errors will be small, it will be assumed that the effect of α, β, and γ can be examined separately and the effects of simultaneous occurrence evaluated by superposition.

Table VI summarizes the four cases of interest. The second column of this table summarizes the characteristic equations of the set of differential equations for each case. The third column of the table cites the general expression for the solution of the equation. This expression is obtained by expanding the solution of the equations in powers of $r\lambda$, dropping terms of higher order than $r^5\lambda^5$ and using the characteristic equation to eliminate terms involving λ^4.

Examination of the expressions in Table VI provides some characterization of the effects of manipulation errors in final response predictions. It is noted first that to a first order, all the constants in column three with the same subscript are equal, i.e.,

$$A_o = B_o = C_o = D_o \quad \text{when } \alpha,\ \beta,\ \gamma \ll 1$$

Further, it is noted that the first two constants in each equations are determined, to a first order, by the displacement boundary conditions. The other two constants are defined by loading or displacement boundary conditions of redundant supports.

It can be seen that the solution of Case 1, where all the errors are zero, is the exact solution of the differential equation for a beam. It is easy to show that this solution is also the exact solution of the difference equations for the finite element. Thus, the replacement of the pair of difference equations by differential equations is exact, under the assumption of no manipulation errors, regardless of the number of equations involved in the analysis.

Case 2 involves errors occurring in every moment equation. The error is assumed to exist on the diagonal of the stiffness matrix. In this case, since both the constant β_2 and β_3 cannot be zero, the error is proportional to βr^2 regardless of what displacement boundary conditions are involved. If all the $B_k \geq 0$, i.e., for a cantilever, the error reduces deflections at all joints, if $\beta \geq 0$. Conversely, if manipulation error is less than zero, all deflections are increased.

In Case 3, it is assumed that persistent manipulation errors are introduced in the force equations on the diagonal. The examination of the expanded expression for displacement indicates that depending on boundary conditions the error may vary as αr^4 or αr^2 when $r\alpha$ is small. If all the arbitrary constants are greater than or equal to zero, the error reduces deflections at all joints if the persistent error is greater than zero, and conversely. The error contributions to the elastic modes (subscripts 2 and 3) are less than those due to β, but for some displacement boundary conditions the rigid modes are distorted.

Table VI

Solution of Beam Error Equations

$$-2w_r''+4\alpha w_r+2a\theta_r'+\frac{\gamma}{3}a\theta_r = 0$$

$$-6w_r'+\gamma w_r+a\theta_r''+(6+4\beta)a\theta_r = 0$$

Case	Characteristic Equation	General Expression for w	Particular Solution
1	$\lambda_i=0;i=1,2,3,4$	$A_0+A_1r+A_2r^2+A_3r^3$	$\frac{Pa^2r^3}{3EI}\left(\frac{3f}{2r}-\frac{1}{2}\right)$
2 $\beta\neq0$	$\lambda_1=\lambda_2=0;\lambda_3^2=\lambda_4^2=-4\beta$	$B_0+B_1r+B_2r^2(1-\frac{r^2\beta}{3})+B_3r^3(1-\frac{r^2\beta}{5})$	$\frac{Pa^3r^3}{3EI}\left(\frac{3f}{2r}-\frac{1}{2}-\frac{2f^3\beta}{r}\right)$
3 $\alpha\neq0$	$\lambda_i^4=-2\alpha\lambda_i^2-12\alpha$	$C_0(1-\frac{r^4\alpha}{2})+C_1r(1-\frac{r^4\alpha}{10})+C_2r^2(1-\frac{r^2\alpha}{6})+C_3r^3(1-\frac{r^2\alpha}{10})$	$\frac{Pa^3r^3}{3EI}\left(\frac{3f}{2r}-\frac{1}{2}-\frac{f^2\alpha}{2}\left[\frac{7f-3}{r}\right]\right)$
4 $\gamma\neq0$	$\lambda_i^4=4\gamma\lambda_i+\frac{1}{2}\gamma^2$	$D_0(1+\frac{r^4\gamma^2}{48})+D_1r(1+\frac{r^3\gamma}{6}+\frac{r^5\gamma^2}{240})+D_2r^2(1+\frac{r^3\gamma}{15})+D_3r^3(1+\gamma r)$	———

Case 4 involves the persistence of γ at a constant value. This error arises in the coupling terms between the displacement and rotation coordinates in both the force and moment equations. It can be seen that the error involves terms in $r\gamma$ and $r^3\gamma$ primarily. Thus, whereas Case 1, 2 and 3 involve even error functions, this case involves an odd function. Errors may decrease or increase response predictions, depending upon boundary conditions. Generally, these errors will be more significant than the β errors and less significant than α errors.

Considering all the cases together, the error may vary up to the fourth order in the number of equations involved. It may be either positive or negative at any given station, depending upon problem boundary conditions. In the worst case, the errors accumulate and the maximum elastic error is of the order $f^4\alpha/10$. Therefore, with $\alpha \approx 2^{-27}$ it can be anticipated that in the worst case about 190 beam elements in series can be treated without answer invalidity due to the accumulative effect of small errors.

The fourth column of Table VI lists the particular solutions for a tip-loaded cantilevered beam for the first three cases. These equations are solutions when $f\lambda << 1$. Case 1 shows that the exact solution is obtained when no manipulation error exists. Case 2 and Case 3 show that the error in deflection predictions vary as the third power of the total number of beam elements, inversely with the joint sequence number, and linearly with error magnitude. Near the tip of a typical beam (where $r \rightarrow f$, $f >> 1$), the error reduces deflections by a magnitude proportional to $f^2\alpha$ of $f^2\beta$.

Critical arithmetic for decomposition series beams of matrices takes the same form as for series rods. The most significant errors are made in evaluating the diagonal elements of the decomposition matrix.

Consider use of the singularity criteria for the series beam. For the regular beam, with equations sequenced from root to tip, the square of last diagonal is by its interpretation $3EI\,(fa)^3$. Its original value is $12EI/a^3$ Then the set of equations will be singular when

$$\frac{3EI}{f^3 a^3} \leq 2^{-p}\,\frac{12EI}{a^3} \quad \text{or} \quad f >> 2\,\frac{p-2}{3} \tag{3-36}$$

Thus, when $p = 27$, 370 finite elements will yield meaningless predictions of deflections.

As for rods, the increasing system results in fewer and the decreasing in greater values of f than given by equation (36). The force row will involve critical arithmetic before moment since the force diagonal depends on the number of elements cubed while rotation depends on the number to the first power. Singularity will occur theoretically when the ratio of original to final stiffness is greater than 2^p. This is given by

$$K_f\, w_f \geq 2^p = 268 \times 10^6 \text{ when } p = 27 \tag{3-37}$$

where K_f is the stiffness of the last beam segment in displacement and w_f is the tip displacement.

The first three columns of Table VII summarize the evaluation of $K_f w_f$ for the increasing series beam when f = 12, 14, 15, 16, and 17. w_f is evaluated by integrating the beam differential equation for the increasing beam stiffness distribution. These data show that only 16 segments can be treated before singularity occurs and answers become meaningless.

A maximum relative error, due to decomposition, can be defined by

$$e = w_f 2^{(x (K_f) + 1 - p)} \tag{3-38}$$

Note the similarity of this equation (2-1). Equation (38) is an error bound established by assuming that an error of one part occurs in the last binary place of the original stiffness coefficient, K_f and dividing this error by the stiffness of the final decomposition diagonal. This formula is only meaningful before singularity occurs.

The last four columns of Table VII summarize the calculations of the relative error predicted by equation (38) and values obtained from computer measurements. Computer measurements used Gauss triangular decomposition. At each joint, the force equilibrium equation was written before moment equilibrium. Computer results are much better than expected. Results are obtained beyond predicted signularity. Actually, singularity was not indicated until 19 elements were involved in the beam. Nevertheless, the critical arithmetic bound is regarded as excellent. It provides a close bound for error when the errors are small and it shows the trend of increasing error. It's imprecision is attributed to the sensitivity of this calculation, on the computer, to manipulation error.

The calculation for singularity can be simplified at the loss of some accuracy. This is achieved by assuming the beam is uniform with the smallest stiffness to evaluate w_f. This approach indicates singularity when f = 14 and the relative error is given by $e = f (p-3)^{-1}$.

Relative errors in substitution for beams have characteristics similar to rod substitution errors. The worst errors arise when each joint is loaded with a force and a reinforcing moment with a load valued at the critical component (2^{28}-1 for p = 27).

Figure 14 is a plot of the maximum relative error in substitution for a beam as a function of the number of beam elements involved. The decomposition was exact. The relative error varies linearily with the number of series beam segments. The relative error is given by

$$e = 0.134 \times 10^{-7}$$

If this error is to be less than five percent, less than 3,700,000 beam segments can be treated.

Table VII

Increasing Beam Critical Arithmetic Error

f	$w_{f,\ Exact}$	K_f	$K_f w_f$	$2^{p-1-X(K_f)}$	e	$w_{f,\ Computed}$	e,Actual
12	449.329102	$6x2^{11}$	$5.53x10^6$	8192	5.50%	425.54143	5.29%
14	633.332107	$6x2^{13}$	$31.1x10^6$	2048	30.9%	470.11828	25.8%
15	737.332720	$6x2^{14}$	$72.5x10^6$	1024	71.9%	324.53154	56.0%
16	849.333027	$6x2^{15}$	$167.x10^6$	512	166%	220.24451	74.1%
17	969.331800	$6x2^{16}$	$380.x10^6$			133.74655	86.2%

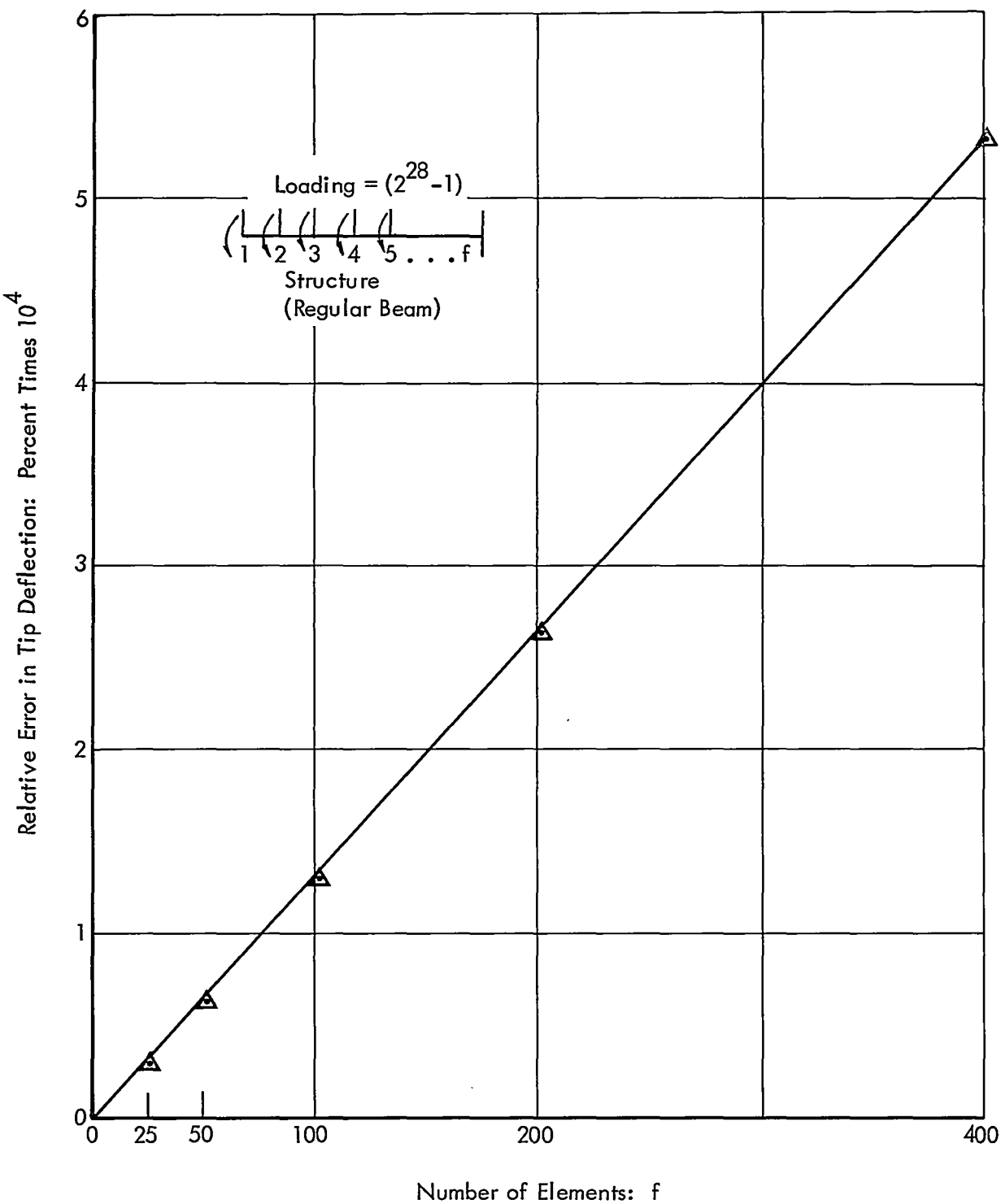


Figure 14. Beam Total Substitution Errors

Other series systems: The effects of manipulation errors for series rods and beams can be extrapolated to other series systems. It is argued that the rod and beam systems represent the basic behaviors of any structural system. This is in fact the basis for the lattice and framework analogies proposed by Hrennikoff [17]. These have been used with some success for analyzing structural systems. Another argument suggesting extrapolation is valid, is that errors in rod and beam analyses differ in intensity rather than characteristics.

Series systems of torque tubes and shear panels will exhibit the same errors as series rods. Series systems for the finite element for the zeroth harmonic of the cylinder given by Percy. et al[18], membrane representations such as those of Turner, et al[2] Argyris[19], and Hrennikoff[17], and representations of elastic solids such as those of Melosh[20] will exhibit errors with intensities similar to those of the rod systems. Errors will be more significant, however, due to the increase in attritition error caused by greater stiffness matrix wavefronts, as the element goes from a one to a three dimensional representation.

Table VIII cites the results of experiments of some of these series systems. These results verify expectations. Results for the series rod are included for comparison. Since the stiffnesses for the cylinder were imprecise (because computer generated rather than input), these results reflect inherent errors which do not exist in results for the membrane and cube. In the Turner membrane representation a square panel model was developed first and the central joint equations eliminated before stiffnesses were summed. Indicated errors are only those due to manipulation. Idealization error is eliminated by developing the exact solution for system behavior based upon its intrinsic exact representations. For the membrane, for example, the case of uniaxial tension is treated since the elements represent this exactly if determinante boundary conditions and uniform loading are treated. The same situation exists for the rod, cylinder and rectangular prism.

The fourth and fifth columns of the table permit comparing the exact solutions with those calculated. These data confirm the trend of increasing error with the increase in the average wavefront of the stiffness matrix. Examining the membrane results, it is also apparent that with the same density, the manipulation error varies with the choice of idealization.

The last column of Table VIII indicates the number of elements that were treated on the computer before singularity was sensed. Data in this column also shows that analyses breakdown due to manipulation error is aggravated be increasing the average wavefront of the matrix.

Series systems of plates and the first cylinder harmonic can be expected to behave like beam systems. Analysis of these systems will be sensitive to errors in the decompsition process and accuracy may be controlled by critical arithmetic. Attrition errors will be worse than those for beams because of the increased stiffness matrix density.

Table VIII

Series Systems Manipulation Errors

Element Type	f[a]	n[b]	w[c]	Exact Tip δ	Calculated	Calculated	s[d]
Rod	100	100	1.99	.152587	.152588	.152584	>4000
Cylinder, Zeroth	100	402	6.46	.212209	.212206	.212199	>100
Membrane, Square							
Turner	100	401	6.42	.143051	.143045	.143000	500>s>400
Argyris	100	401	6.42	.889079	.888131	.888643	150>s>100
Hrennikoff	100	401	6.42	.286102	.285927	.285882	>500
Prism, 1x2x3	25	305	18.0	.260417	.233177	.233177	49
Beam	100	200	3.48	.666667	.666666	.655364	800>s >400
Cylinder, First	25	102	6.32	.803014	.625960	.695000	17
Plate, Square	20	120	9.22	.733321[e]	.756491	.757842	>200

[a] f = Number of finite elements

[b] n = Number of equations

[c] w = Average wave front

[d] S = Number of elements causing singularity (p=27)

e Calculated solution, p=36

Table VIII shows solution results for beam, cylinder first harmonic and square plate series systems. The plate finite element is that of Bogner, Fox, Schmit[21]. Errors indicated are due only to manipulations in elimination. The data shows that as the average matrix wavefront increases, error increases. Early failure of the cylinder analysis as a attributed to its large inherent error. Only eight significant decimal digit accuracy was used to define the stiffness matrix.

Series systems of curved shell elements are regarded as mixed systems since they involve parallel subsystems of rod-like (membrane) and beam-like (plate) behavior. Mixed systems are beyond the scope of this study.

Errors in Evaluating Secondary Unknowns. - These are due to the fact that the operation intrinsically involves critical arithmetic. Determination of stresses from displacements is essentially a differentiation operation. The characteristic form of the calculation is

$$s_i = t \frac{\Delta_r - \Delta_{b+1}}{x_r - x_r + 1}$$

where t is a factor involving material constants and x_r is a coordinate of an element joint. As the network interval approaches zero, successive displacements (Δ_r and Δ_{r+1}) approach each other. The stress predictions become meaningless as the evaluation of the stresses involves the subtraction of two components which are nearly equal. Since the useful information is contained in the lower bits of the Δ_r, critical arithmetic is intrinsically involved.

Anderson and Christiansen[22] point out that in the case of an incompressible material (Poisson's ratio nearly 0.5) critical arithmetic is involved even when the joint spacing is large. Stresses calculated for nozzles under nearly pure dilatational deformation indicate an oscillating sign stress pattern throughout the network. This type of response aggravates the critical arithmetic error.

Guidelines for the Analyst. - The analyst's formulation of his structural problem has an important influence on the amount and distribution of manipulation error. His selection of joint numbering, joint spacing, and structural idealization can insure his analysis accuracy almost independent of the computer and program he selects. Idealization defines the mathematical model of each finite element, thus fixing matrix sparseness and critical arithmetic involved.

In displacement method computer codes the sequence in which joints are numbered corresponds with the sequence in which equations are treated in the elimination process. This enables the analyst to eliminate critical arithmetic in the decomposition for series systems. If non-optimum sequencing is used, the singularity criteria previously discussed can be used to estimate manipulation error.

Optimum joint sequencing is defined as that which minimizes the ratio of the stiffness diagonal over its final value during decomposition; for the diagonal with the largest ratio. This definition seeks to avoid critical arithmetic. It is consistent with the predictions and experimental results reported here for rods and beams.

Based on this definition, optimum joint sequencing for regular cantilevered series rod and beam systems can be proved to involve numbering from the free edge toward the support.

If all the edges of the structure are restrained, the joint numbering sequence has less affect on critical arithmetic than for a structure with a free edge. This is illustrated for the series rod by the data in Table IX. In addition, these data show that numbering from the midstation toward both supports or numbering from one end to the other results in the same diagonals for the decomposition matrix for a regular rod.

The data in Table IX also show that for the regular rod with both ends clamped, optimum joint numbering involves sequential numbering of adjacent joints. Since this is also the case for optimum numbering with a free edge, and since this type of numbering tends to minimize matrix wavefront, it is hypothesized that optimum joint numbering from the point of view of manipulation error is that which minimizes the stiffness matrix wavefront and proceeds from free edges, when they exist.

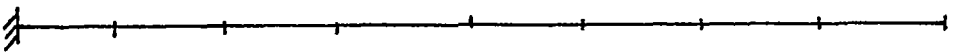
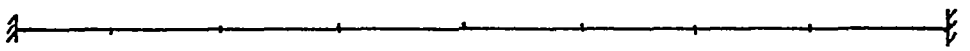
The effect of varying stiffness is to increase the importance of proper joint numbering as demonstrated by the increasing rod and beam systems. With improper numbering, numerical singularity can occur with only two elements if the elements differ in stiffness drastically.

The guideline of numbering from the free edge also is valid for other systems. This conclusion is based on comparing analysis errors using the data in the fourth, fifth and sixth columns of Table VIII. In column five, joints are sequenced from the free edge; in column six, from the fixity. The comparison shows that generally numbering from the free edge gives smaller error. The cylinder exception is discounted because of inherent error in input. The Argyris membrane exception can be attributed to the capriciousness of manipulation error. The maximum relative error for this membrane, in accordance with equation (38), is 12.3%.

The analyst should space his joints so that as the joint numbers increase, the element stiffnesses increase. Table X shows the results of rod analyses for irregular structures. The relative errors are indicated to be smaller when stiffness increases as joint numbers increase. The problems considered are rods with elements whose stiffness doubles or is multiplied by one-half for each sequential element going from tip to root. The structures are loaded with a force of $(2^{28}-1)$ at the tip and errors measured for tip displacements. The reduction in error due to preceding in the direction of increasing stiffness is less important than the proper joint numbering for these small problems. However, as the order of the equations becomes large, propagation instability is more probable in the system where element stiffnesses are decreasing and proper joint spacing becomes more important in avoiding stiffness decreases.

Table IX

Effect of Joint Sequencing for Regular Rods

STRUCTURE								
Joint No.	1	2	3	4	5	6	7	8
l_{ii}^{2*}	2	3/2	4/3	5/4	6/5	7/6	8/7	1/8
k_{ii}/l_{ii}^{2**}	1	1.33	1.50	1.60	1.67	1.71	1.75	16.0
Joint No.	8	7	6	5	4	3	2	1
l_{ii}^{2*}	1	1	1	1	1	1	1	1
k_{ii}/l_{ii}^{2**}	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
STRUCTURE								
Joint No.	1	2	3	4	5	6	7	
l_{ii}^{2*}	2	3/2	4/3	5/4	6/5	7.6	8/7	
k_{ii}/l_{ii}^{2**}	1.00	1.33	1.50	1.60	1.67	1.71	1.75	
Joint No.	6	4	2	1	3	5	7	
l_{ii}^{2*}	7/6	5/4	3/2	2	4/3	6/5	8/7	
k_{ii}/l_{ii}^{2**}	1/71	1.60	1.33	1	1.50	1.67	1.75	
Joint No.	1	5	2	7	3	6	4	
l_{ii}^{2*}	2	1	2	1/2	2	1	2	
k_{ii}/l_{ii}^{2**}	1	2.00	1	4.00	1	2.00	1	

*Square of the Decomposition Diagonal

**Stiffness Diagonal/Decomposition Diagonal Squared

Table X

Error Magnification for Variable Stiffness

Series Rods

No. of Segments	Increasing Systems Tip Deflection Exact	Calculated	% Error	Decreasing Systems Tip Deflection Exact	Calculated	% Error
15	$(2^{15}-1)(2^{28}-1)$	$(2^{15}-1)\ 2^{28}-2^{15}$	0	$(2^{15}-1)(2^{28}-1)$	$(2^{15}-1)\ 2^{28}-2^{19}$	5.22×10^{-6}
20	$(2^{20}-1)(2^{28}-1)$	$(2^{20}-1)\ 2^{28}-2^{20}$	0	$(2^{20}-1)(2^{28}-1)$	$(2^{20}-1)\ 2^{28}-11.2^{20}$	3.72×10^{-6}
40	$(2^{40}-1)(2^{28}-1)$	$(2^{40}-1)\ 2^{28}-2^{41}$	$.372\times10^{-6}$	$(2^{40}-1)(2^{28}-1)$	$2^{68}-17.2^{40}$	5.96×10^{-6}
100	$(2^{100}-1)(2^{28}-1)$	$(2^{100}-1)2^{28}-2^{101}$	$.372\times10^{-6}$	$(2^{100}-1)(2^{28}-1)$	$2^{128}-17.2^{100}$	5.96×10^{-6}

The selection of the mathematical representations for a continuum elements also has an important effect on manipulation error. Table XI compares analyses of a membrane using Turner and Argyris representations for the series problem. The Turner representation results in less relative manipulation error for all ratios of the panel sides.

The manipulation errors tend to reduce in membrane and plate systems as the panels become elongated in the direction of the series. This is evident from the data in the last two rows of Table XI. This suggests that errors in mixed systems may be larger than those in pure series systems. Moreover, it implies that one of the most effective ways the analyst can use to reduce manipulation errors is to introduce displacement constraints. In the case of the uniformly loaded series membrane using the Turner, Argyris or Hrennikoff model, this reduces manipulation errors to those of the rod system.

Manipulation errors can be reduced by idealizing the structure with lower order difference representations according to Gatewood and Ohanian[9]. For the beam, they demonstrated by numerical experimentation that manipulation error was reduced when the single fourth order difference representation was replaced by a pair of second order equations. Further error reduction occurred when the pair of second order equations were replaced by four first order equations. This study confirmed this trend theoretically. The absolute error for the rod equations (a first order set) at worst is proportional to the equation number. The relative error is proportional to the total number of equations. Thus, the analyst should choose idealizations involving many low order difference equations to minimize his maximum error.

Having formulated this problem, the analyst must determine what precision is required in computer calculations to insure the desired accuracy. Based on the needed precision, the analyst can select an adequate computer configuration: hardware and software.

Formulas developed for series rod and beam systems can be used as guidelines for evaluating required precision. Table XII summarizes these formulas for relative error for extreme cases for the regular structures. These relative errors can be assumed to be additive.

Multiple source propagation error is based on an error of 2^{-p} in every diagonal of the decomposition. In evaluating the expected error it is assumed that this occurs only only half the diagonals. The accumulation of persistent small errors in the decomposition is based on errors in every row of the decomposition of magnitude 2^{-p}. The expected error formula assumes that these errors assume an average value of one-half the maximum. Substitution error is based on the critical component for the load. The expected value is taken as the average error of the critical component load and an ideal load.

Assume that critical arithmetic is avoided. Then, for the rod systems of more than 265 elements, the propagation of unstable errors in decomposition is the most important error source. When many elements are involved, the required precision to insure less than five percent error can be estimated by

Table XI

Effect of Panel Side Ratios

Formulation Idealization	p	Calculated Tip Deflection[a] 8/1 Panel	1/1 Panel	1/8 Panel
Turner[b]	Exact	.725002	.143051	.113281
	27	.724982	.143045	.113281
	e	.0152%	.0042%	0
Argyris[b]	Exact	.177816	.889079	.277837
	27	.177763	.888131	.277842
	e	.0299%	.106%	.0018%
Bogner[c]	Exact	.952381	.761905	.186012
	36	.820100	.761928	.190452
	27	.765259	.756491	.190713
	e	19.5%	.710%	2.15%

[a]Due to uniform load at tip. Structure fixed at high numbered joints, Gauss solution.

[b]Membrane, 100 series elements, 401 equations.

[c]Plate, 20 series elements,120 equations.

Table XII

Regular Series Systems Errors

	Rod Systems		Beam Systems	
Error Source	Error Bound	Expected	Error Bound	Expected
Multiple Source Propagation	$f^2 2^{-p}$	$\frac{1}{4}f^2 2^{-p}$	$16f^2 2^{-p}$	$4f^2 2^{-p}$
Persistent Accumulation	$f2^{-p}$	$\frac{1}{2}f2^{-p}$	$f^4 2^{-p}$	$\frac{1}{2}f^4 2^{-p}$
Substitution	$265\ f2^{-p}$	$133\ f2^{-p}$	$360\ f2^{-p}$	$180f2^{-p}$
Critical Arithmetic	$f2^{-p}$	$f2^{-p}$	$f2^{-\frac{(p-2)}{3}}$	$f2^{-\frac{(p-2)}{3}}$

$$p > \ln_2 (20\, f^2) \qquad \text{Bound}$$

$$p > \ln_2 (5\, f^2) \qquad \text{Expected} \qquad (3\text{-}39)$$

The equations show that single precision arithmetic for all the computers of Table I is adequate for treating large rod systems.

For beam systems with many elements, persistent small manipulation errors in the decomposition process is the most important error source. The required precision to insure less than five percent error is given by

$$p > \ln_2 (20\, f^4) \qquad \text{Bound}$$

$$p > \ln_2 (10\, f^4) \qquad \text{Expected} \qquad (3\text{-}40)$$

Application of equations (39) or (40) to practical structural problems requires judgment. Each equation is based on a worst case series system (with no critical arithmetic), whereas the practical structure involves mixed series and parallel subsystems. The equations are also written in terms of the number of finite elements whereas the manipulation error must depend on the number of non-zero coefficients in the stiffness matrix. Assuming that series systems are critical, and rewriting equation (40) in terms of the number of equations for the beam given

$$p > \ln_2 (0.625\, f^4) \qquad \text{Expected} \qquad (3\text{-}41)$$

this equation can be used to estimate the required precision in a practical structural analysis.

It is observed that all errors in Table XII are proportional to 2^{-p} except those due to critical arithmetic. This suggests that analyses with different levels of precision can be performed to estimate manipulation errors.

Figure 15 shows measurements demonstrating the validity of this approach for a regular series rod and several regular series beams. The scale for precision has been chosen so that if all errors are proportional to 2^{-p}, the errors for a particular structure will plot as a straight line passing through the intersection of the given axes.

This figure shows that the manipulation error varies with 2^{-p} over most of the tested range. Failure of some of the lines to pass through the intersection of the axes suggests some nonlinearity when errors are very small, due to use of a discrete number system.

These results indicate that analyses with several levels of precision can be used to estimate manipulation errors. If three levels are used, the assumption that error is proportional to 2^{-p} can be checked. If only two levels are used, the proportionality must be assumed to predict error magnitudes.

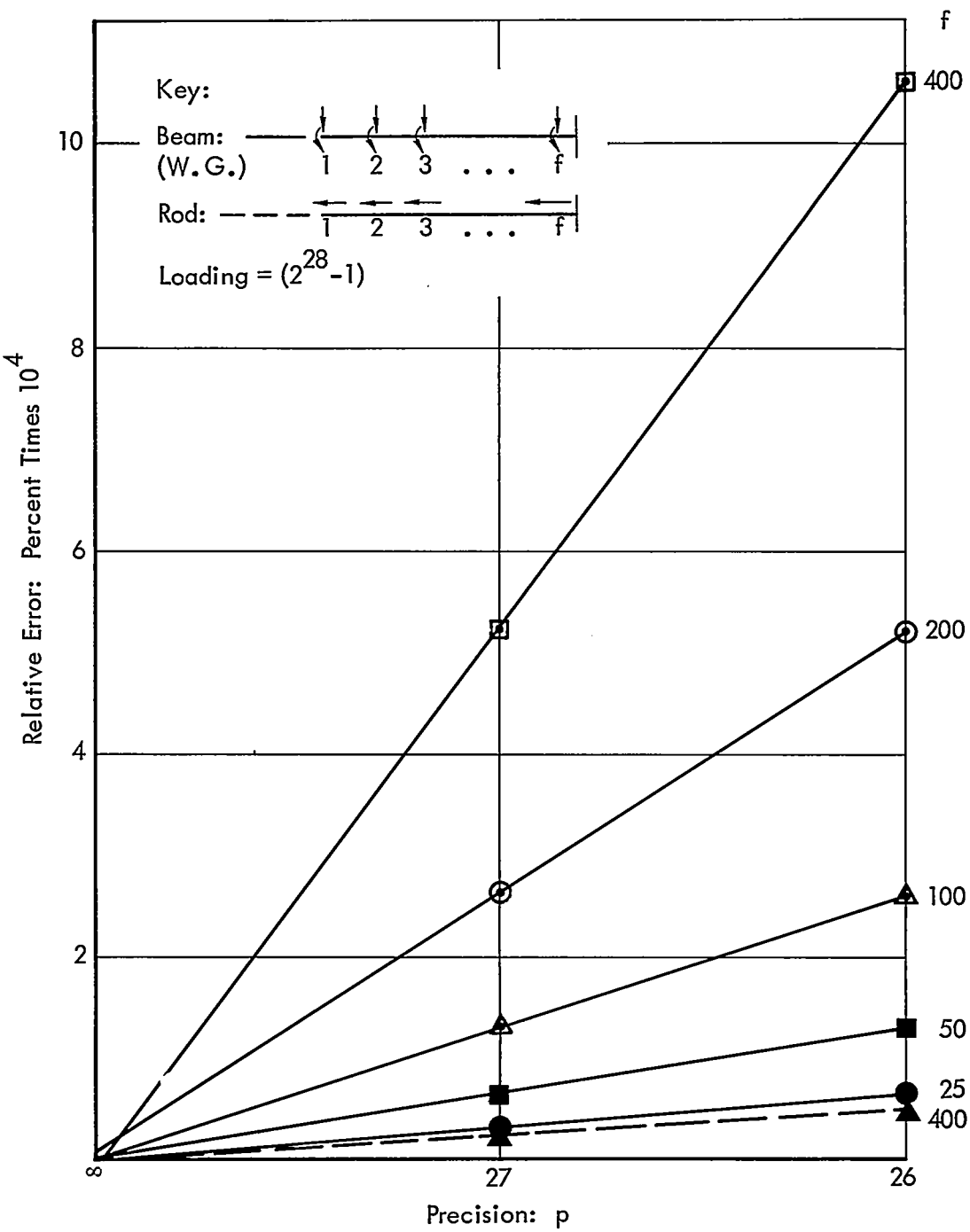


Figure 15. Effect of Precision on Error

Guidelines for the Programmer. - Though the analyst fixes the sequence in which the equations of a joint are introduced into the elimination process, the details of the solution algorithm are selected by the computer programmer. It is these details which he can modify to improve the accuracy of the structural analysis. In addition, there are a number of checks that the computer programmer can include in his code to indicate the accuracy of the numerical results.

One of the most significant algorithm changes is to increase the precision of the arithmetic involved. One device for making this improvement is simply to make available multiple precision arithmetic. This study has shown that single precision arithmetic on a 27 bit mantissa machine is satisfactory for problems up to about 3000th order if the analyst is careful in program formulation. The use of multiple precision arithmetic results in a computer code that can handle smaller problem sizes or problems of the same size at much more machine time because of the added core space required to retain multiple precision numbers.

An examination of the decomposition process shows that all the elements of the decomposition matrix are less than one. Since this is true, the bits reserved for the exponent can be implied and fixed decimal arithmetic used throughout with the decimal point positioned at the left of the number representation. This will effectively permit an increase of at least 20 percent in the mantissa for the machines listed in Table I. This algorithm change requires no additional storage space over the standard algorithm.

Since the largest single source of error arises in forming square roots, it might be thought that use of double precision to develop square roots would result in improved accuracy. Figure 16 shows that this is not the case. The double precision square roots result in twice as much error as expected for the increasing series rod systems. The explanation is that truncating the double precision root to single precision insures that the errors, when they are non-zero, are negative. Since the process is unstable for negative errors, the small errors are magnified resulting in larger errors in the result than for the single precision case where only positive errors can occur.

The squre root error can be entirely avoided if a modified Gauss triangularization process is selected instead of the Choleski process. Let the decomposition be written as

$$K = L\,D\,L^{T} \tag{3-42}$$

where L = a lower triangular matrix with ones on the diagonal

D = a diagonal matrix

The elements of D can be stored along the diagonal of L and the ones implied so that no additional space is required over Choleski decomposition. No square rooting is required. The accuracy of this modification for series rods is indicated in Figure 16. The predicted accuracy assuming no manipulation error correlates well with that measured with this modified Gauss decomposition. Table XIII provides a comparison of the error in a series beam analysis for Choleski and Gauss for a regular series beam with 100 segments with unit lateral loads and for the increasing series beam. The error in the Choleski is not always worse than the Gauss error. However, as the matrix approaches numerical singularity, the advantage of Gauss becomes increasingly evident.

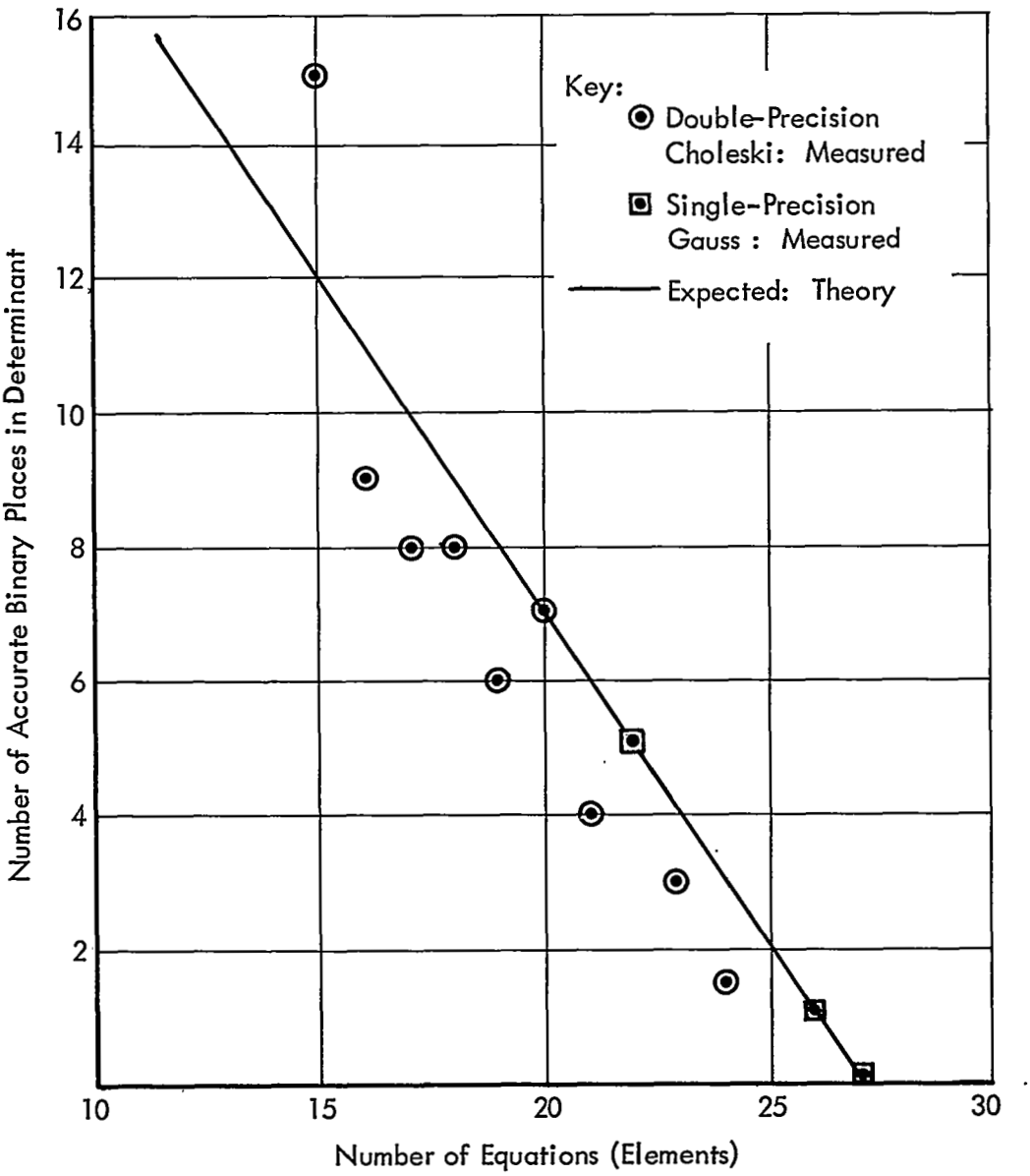
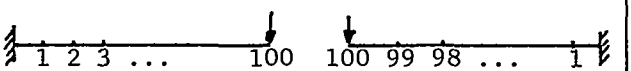


Figure 16. Effect of Algorithm on Rod Singularity Tests

Table XIII

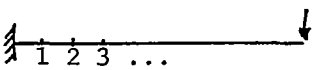
Comparison Between Choleski and Gauss Error

Regular Beam



Loading Joint	Displ. Joint	Exact Displ.	Gauss Displ.	Choleski Displ.	Gauss Displ.	Choleski Displ.
1	1	.666667	.665973	.665382	.666667	.666764
	100	99.6667	96.1824	89.6148	99.6667	100.051
100	1	99.6667	96.1826	89.6147	99.6667	100.051
	100	666,667	645,046	595,514	666,665	669,453

Increasing Beam



f	^{w}f, Exact	^{w}f, Gauss	^{w}f, Choleski	E Chol./E Gauss
12	449.329102	425.54143	461.42364	.524
14	633.332107	470.11828	718.19141	.520
15	737.332720	324.53154	1344.5670	1.475
16	849.333026	220.24451	Not positive definite	-----

Though the analyst specifies the equation sequence by his joint numbering, the sequence in which the degrees of freedom are treated at a joint affects the accuracy of solution. The importance of treating force equilibrium first for the beam can be seen in Figure 17. This figure shows the relative errors at several stations on the 100 segment beam as a function of precision for each equation sequence. This figure shows that though errors may sometimes be greater when forces are treated first, the errors behave better.

The intrinsic advantage of treating force equilibrium equations before moment equations is evident by looking at the forward or back substitution equations. When force equilibrium is stated first, the first of the equation pair can be solved independently of the second. This equation is a first order equation of the form,

$$\theta_r - \theta_{r-1} = M_r \tag{3-43}$$

Since all the M_r are less than the exact moments because of truncation, the θ_r are all underestimated (for this argument all loads are considered reinforcing).

When moment equilibrium is stated first, each pair of equations is coupled. Eliminating displacements in the pair of back substitution equations gives a second order difference equation of the form,

$$\theta_r - 2\theta_{r-1} + \theta_{r-2} = M_r - M_{r-1} \tag{3-44}$$

Thus, solution of these equations is sensitive to the rate of error growth as well as magnitude and hence more sensitive to the moment errors. Since the error rate is negative, rotations (and displacements) may be overestimated or underestimated even though loads are reinforcing, as indicated in Figure 17.

Another way in which the errors in the elimination process can be reduced is by making a change of variables in the equilibrium equations. If the displacement variables are replaced by change of displacement, critical arithmetic will be largely removed in the decomposition process. This can be accomplished explicitly by a change of variables. It is accomplished implicitly by optimum equation sequencing.

Another possibility for minimizing manipulation error is for the computer code to provide the capability to resort the equations so that they are handled in an optimum sequence. This solution is not an acceptable solution. For series systems for example, there are only two reasonable sorting sequences to preserve optimum bandedness. If joints are numbered toward the fixity, equations are treated in an optimum sequence. This suggests that minimum bandwidths imply an optimum sort, both from the point of view of efficient data handling and minimum manipulation error. Since resorting of large matrices is inefficient on the computer and a near optimum sequencing can easily be specified by the analyst cognizant of the topology of his structure, automatic equation resequencing is unattractive.

Several attempts have been made to minimize the affects of critical arithmetic in the determination of stresses in the displacement method. These have involved some technique for smoothing the stress estimates so that big jumps in stresses do not occur between elements. This smoothing is achieved by taking cognizance of the stresses in neighboring elements to condition the estimates of stress in a given element.

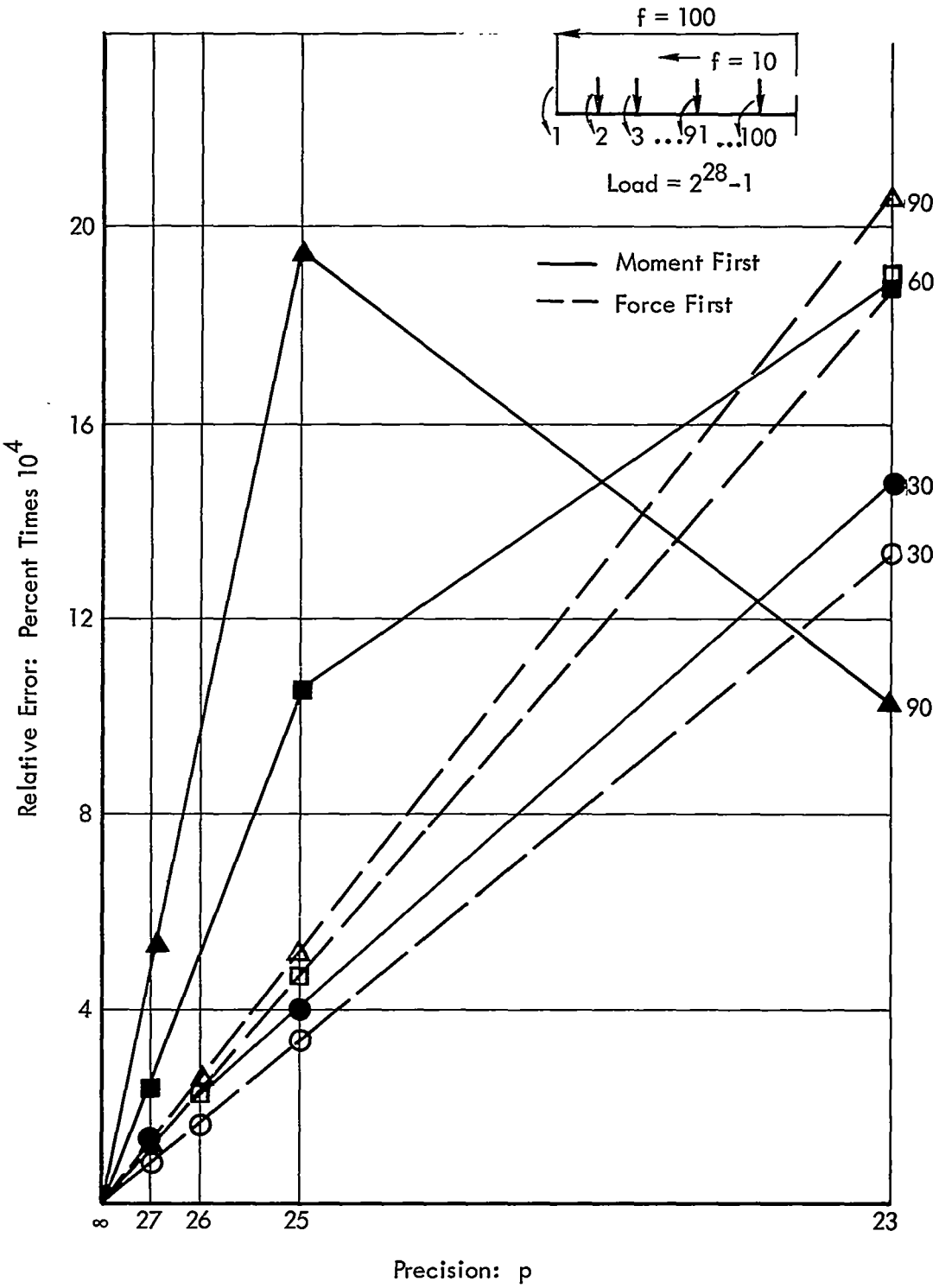


Figure 17. Effect of Equation Internal Order on Error

Turner, Martin, and Weikel[23] have shown by numerical experimentation that better estimates of the stresses are obtained for membranes by averaging the stresses for the elements meeting at a joint. Utku[24] has found that a least squares fit to the stress values is helpful in improving the estimates.

There are four checks that should be included in a displacement method computer code to validate solutions. The first consists of a singularity test. The studies described demonstrate that the relative value of the diagonal after the decomposition process compared with its value before is the best single measure of the accuracy of the decomposition process. In addition, the last diagonal of the matrix provides the best measure of the accuracy of the determinant of the matrix. Therefore, a simple comparison between the diagonal before and after the decomposition process can lead to measures of the analysis accuracy.

A second computer based test can be developed considering the error propagation characteristics described earlier in this section. A measure of the propagation stability can not only determine the accuracy of predicted structural response but also can anticipate the singularity or non-positive definiteness of the stiffness array.

A third, and already popular test involves equilibrium checks. These checks measure the error in the solution of the simultaneous equations by resubstituting the solution in the original equations (2-25).

A fourth check is the evaluation of stress calculations with respect to manipulation error. This check is easily incorporated and will define, for the analyst, the relative accuracy of the calculated stresses.

One check which has been used in the past is based upon Maxwell reciprocity. This check consists of comparison of the off diagonals of the flexibility matrix to determine analysis accuracy. Test problems in this study have convincingly demonstrated that this test is not a sufficient test of accuracy. For example, a regular 400 segment cantilevered beam fixed in the last equations gives no significant figures accuracy for deflection under a tip normal load or a load adjacent to the fixity. Maxwell reciprocity, however, is satisfied for seven significant digits.

A number of investigators have suggested the use of eigenvalue ratios as a measure of the manipulation error of the matrix. Test problems have shown that these criteria are not directly related to the principal manipulation error inducing phenomena in the elimination process: critical arithmetic. This is proven by the fact that when the equation of series rod systems are sequenced from tip to root the solution has negligible error compared to their sequencing from root to tip. Though resequencing does not change the matrix eigenvalue ratios, errors vary dramatically. Moreover, as the number of segments of the rod increase, the eigenvalue ratio (the conditioning number) will increase monotonically. Nevertheless, the accuracy for solving problems of order of 50 for the rod systems is comparable to the accuracy for solving those of the order of 1200 when the structure is numbered from the free edge to the root.

Section 4

FORCE METHOD ERROR ANALYSIS

This section considers the structural analysis manipulation errors in generating the coefficients in the structural equations and in eliminating coupling to evaluate the primary and secondary unknowns. In the force method the primary unknowns are the element forces, X and Q. The secondary unknowns are the structural deflections, Δ in equations (2-23) and (2-24). This section provides guidelines for the engineering analyst and programmer for reducing, estimating, and measuring manipulation error.

Although the sets of variables in the displacement and force methods of structural analysis are dual to each other, there are significant differences in computational detail in the solution process. The principal differences arise in the transformations of variables and in solving the resulting set of linear simultaneous equations. On this basis, the error analysis of the displacement method cannot be completely carried over for the force method.

Generation Error

Generation errors in the force method include the manipulation errors incurred in the development of the following coefficient matrices:

(1) The loading coefficient matrix F

(2) The geometric assembly matrix P. This matrix includes the partitions P_Q and P_X of equation (2-24).

(3) The flexibility matrices d_i for each element in local coordinates. When arranged in block diagonal form along the main diagonal, these matrices form matrices D_Q and D_X of equation (2-23).

(4) The element deformations in local coordinates.

The coefficients of matrix F are direction cosines of load vectors. If they are direct-input, input conversion and truncation error is involved. Often, however, the coefficients depend upon the coordinates of two points defined by the analyst to describe the direction of a load vector. Then the calculations in generating the elements of the F matrix can involve critical arithmetic.

Similarly, the calculations for the elements of the P matrix involve calculation of lengths and ratios of lengths. Lengths are obtained by differencing coordinates of points. Critical arithmetic will be involved if the coordinates are defined such that their difference is not an accurate measure of element length. Furthermore, since the cosines represent ratios of two lengths, to avoid critical arithmetic, as noted in Section 3, the difference of the relative error in the calculation of the lengths must be small compared to one.

The generalized coordinates used in the development of the element flexibilities are based only on elastic deformations and do not include rigid body motion. The errors in the matrix of elastic constants used for calculating element flexibilities are due primarily to input errors. This statement also applies to the generation errors involved in the development of the matrix of element deformations, e_T.

The maximum number of calculations to generate an element of the P matrix is in the order of 130 calculations. This many calculations correspond to that required to calculate the kick forces or out-of-plane forces for a warped shear panel in terms of the element generalized force that is entered into the P matrix. The detailed expressions for the calculations are given in reference 26.

On comparing the various flexibility matrices of finite elements, e.g. those derived in reference 27, the maximum number of calculations to generate an element of the matrices D_Q and D_X correspond to that of a tetrahedron finite element and is in the order of 125 calculations.

It can be seen that if critical arithmetic is avoided generation errors are small. The number of calculations is much less than 13.4×10^6 for each coefficient. In accordance with the analysis of Section 2, this many calculations would be required before error would exceed five percent with a 27 binary place mantissa.

Elimination Error

The general structural equations are given in Section 2. This outline of the solution by the force method is appropriate when the redundant forces are preselected since the ordering of the determinate and redundant forces establishes the corresponding partitions in the coefficient matrix in equation (2-23) and equation (2-24). However when the redundant forces are automatically selected by the computer, the solution process is modified.

In the force method, the structural equations are regarded as two sets of simultaneous equations. The first set of equations are the equilibrium equations (2-24). This set of equations requires partitioning the sparsely populated and unsymmetric geometry assembly matrix P into P_Q and P_X by automatic selection of the redundants. The P_Q matrix involves forces in a statically determinate substructure. The standard way of reducing the structure to a determinate structure by suitable releases or "cuts" is equivalent to obtaining a "particular solution" to the equilibrium equations. This solution satisfies the conditions of equilibrium but not the boundary conditions of the problem, i.e., the continuity of contiguous elements. Thus, the solution of the equilibrium equations contains arbitrary parameters which are determined from continuity considerations.

The partitioning process is accomplished by the Jordan diagonalization method. This is similar to the familiar Gauss elimination method. The only difference is that in the former the coefficient matrix is reduced to a diagonal (or to a unit matrix) whereas in the latter it is reduced to a triangular array. In either case the process is equivalent to performing elementary row operations or premultiplying by a sequence of matrices.

The partitioning approach is to find and diagonalize a set of linearly independent columns of the dimension of the row space of P (the number of equilibrium equations). If a null linearly dependent column or a zero pivot (diagonal) element is encountered in the elimination process, column interchanges are performed. If insufficient, non-null columns are found, at least one of the equilibrium equations is dependent. If enough columns are found and the P_X partition is not null, the columns of P_Q comprise the selected stable determinate substructure and the columns of P_X comprise the selected redundants. The dimension of the column space of P_X is the degree of indeterminacy of the structure. If P_X is null, the structure is statically determinate. In this case the equilibrium equations (2-24) are the only relevant set of equations; the continuity equations (2-23) are irrelevant.

The redundant forces are evaluated from the second set of equations. These equations are compatibility conditions. They result in equations (2-28) which involve the coefficient array

$$P_X^T P_Q \, {}^{-1}T_{D_Q} P_Q \, {}^{-1}P_X + D_X = \delta_{XX} \tag{4-1}$$

The matrix, δ_{XX}, is positive definite and symmetric since it is obtained by a congruent transformation of the positive definite matrices, D_Q and D_X. If the set of redundants are not orthogonal, the matrix δ_{XX} is densely populated. If the redundants form an orthogonal set, the matrix, δ_{XX} is a diagonal matrix.

In solving (2-28) for the redundant forces, X, Jordan diagonalization is also used. Since the matrix δ_{XX} is positive definite, pivoting is not required. It is sometimes incorporated with the intent of minimizing manipulation error.

The next step of the elimination process involves the use of equation (2-29) to evaluate the internal forces, Q. Significant attrition errors can result from this calculation if the determinate base structure and corresponding redundant stress system are not properly chosen. This occurs when the effect on the determinate base structure of the redundants is comparable in magnitude to that of the external loads. The calculations involve subtraction of numbers of approximately the same magnitude.

Finally the calculations described by equation (2-30) define displacements. No critical arithmetic is involved in the calculation. The product of the first pair of matrices on the right-hand side of equation (2-30) is preserved from the elimination and multiplied by the internal forces to assess the displacements.

Thus, error analysis for elimination error in the force method will be concerned with errors in partitioning the geometric assembly matrix, P, and solving for redundant forces by triangularizing the matrix δ_{XX}. Errors for these arrays will be examined separately.

Error Analysis for the Geometric Assembly Matrix

Errors in eliminating the redundant interaction to evaluate the unknowns include inherent and attrition error. Inherent error is the error existing in the coefficients of the matrices due to prior arithmetic. In the case of the P matrix (the geometric assembly matrix) the prior arithmetic are generation calculations. As described above, few calculations are involved in developing these coefficients and critical arithmetic can be avoided. The relative inherent error will be less than 2^{-p}.

An important source of error in selecting redundants is the numerical singularity of the geometric assembly matrix. The matrix P_Q is a square array of coefficients of the unknown "determinate" forces in the equations of equilibrium (2-24). Singularity of this matrix can arise from a linear dependency or inconsistency of the equations of equilibrium, i.e., of the rows of the P matrix. A particular case is when at least one of the element coordinate forces is not a generalized force. In the force method the forces in a disassembled element fall into two groups: element reactions and element forces. Only element forces may occupy a column in the P matrix. If element reactions were inadvertantly included in the P matrix a dependency would exist. A physical interpretation of the singularity of the geometric assembly matrix is that the structure is unstable. This implies the formation of a mechanism.

Assuming that the idealized structure is physically stable, a condition of local instability can result from the following cases:

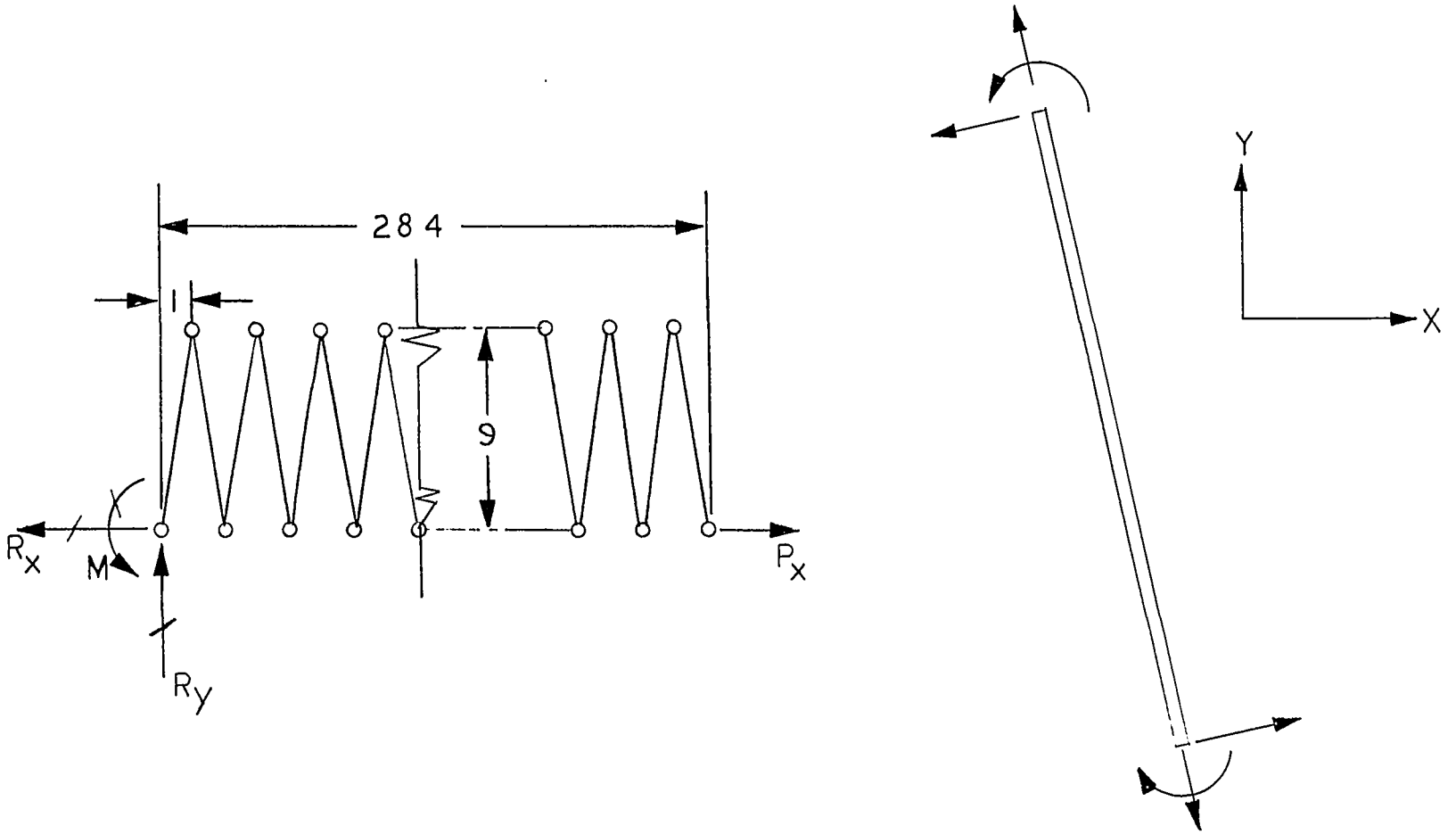
(1) Generation errors involving critical arithmetic resulting from the calculation of lengths and orientation of elements. For example, two axially loaded pin-connected rods which are supposedly collinear are unstable if the ends are offset.

(2) Structures which are nearly unstable. An example of these is a truss with a panel whose length to depth ratio is such that the diagonal becomes ineffective or lost in the numerical calculations resulting in the formation of a mechanism.

Instability can also occur when the redundant forces are not properly selected, i.e., the statically determinate structure remaining after removal of redundants is unstable. In this case the coefficient matrix for the statically determinate forces P_Q is singular.

The singularity error for the coefficient matrix of the determinate structure is not critical for a parallel system. The matrix is a scalar for rod systems and at most a well-conditioned 2 x 2 matrix for beam systems. In a series system, there are no redundants, i.e., P_Q is identical to P. Then, the relative flexibilities of the structural elements do not contribute to the singularity error in the coefficient matrix. This matrix expresses the combined effects of the geometrical position of the member with respect to the coordinate system and of the incidence of the members at the joints. It thus represents a linear transformation of the base vectors of the element-forces space into the basis of the vector space of the joint-applied loads. The ordering of the elements or joints affects the disposition of elements of the transformation matrix but not their values.

For a series system the coefficient matrix is also well behaved. This is confirmed by the solution for the structure shown in Figure 18 in which the angle between contiguous elements is made small. Each bar carries axial and bending restraints at each end. The significant results are the pivot values of the Jordan diagonalization of the P matrix. These maintained a value of 1.0 throughout.

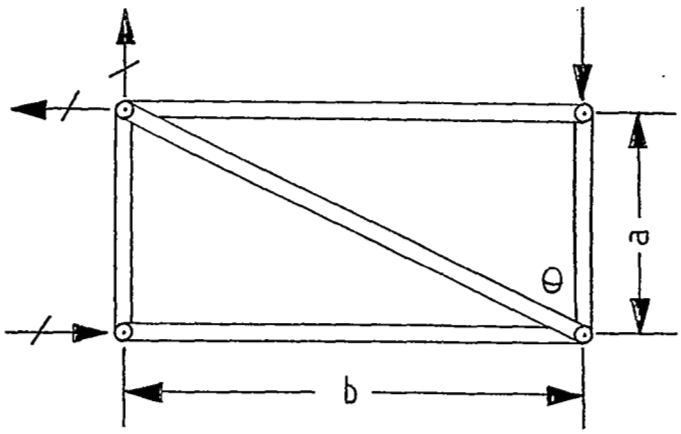


(a) Idealized Structure

(b) Internal Loads
Typical Bar

Figure 18 - Beam Elements in Series

Consideration of a mixed system involving parallel and series elements may perhaps yield a more meaningful criterion for the numerical singularity of the P matrix. To discover this criterion, a typical rectangular panel made up of four edge bars and one diagonal bar is analyzed with the ratio a/b reduced by a factor of 2 in successive solutions, where a is the width and b is the length of the panel. See Figure 19.



Vectors with strokes are reactions

Figure 19 - Rectangular Panel

In the program used for the solutions the format for the input of coordinate values is prescribed in fixed point manner with six places provided for the fractional part of the value. One can see that the length 'a' would be zero (and the structure a mechanism) when values of coordinates fall below 10^{-6}. In the case at hand since the whole part had two additional significant figures, cos θ becomes unity when $a/b \leqq 10^{-8}$. If the input data are expressed in floating point manner the ratio of a/b for creation of a mechanism is smaller.

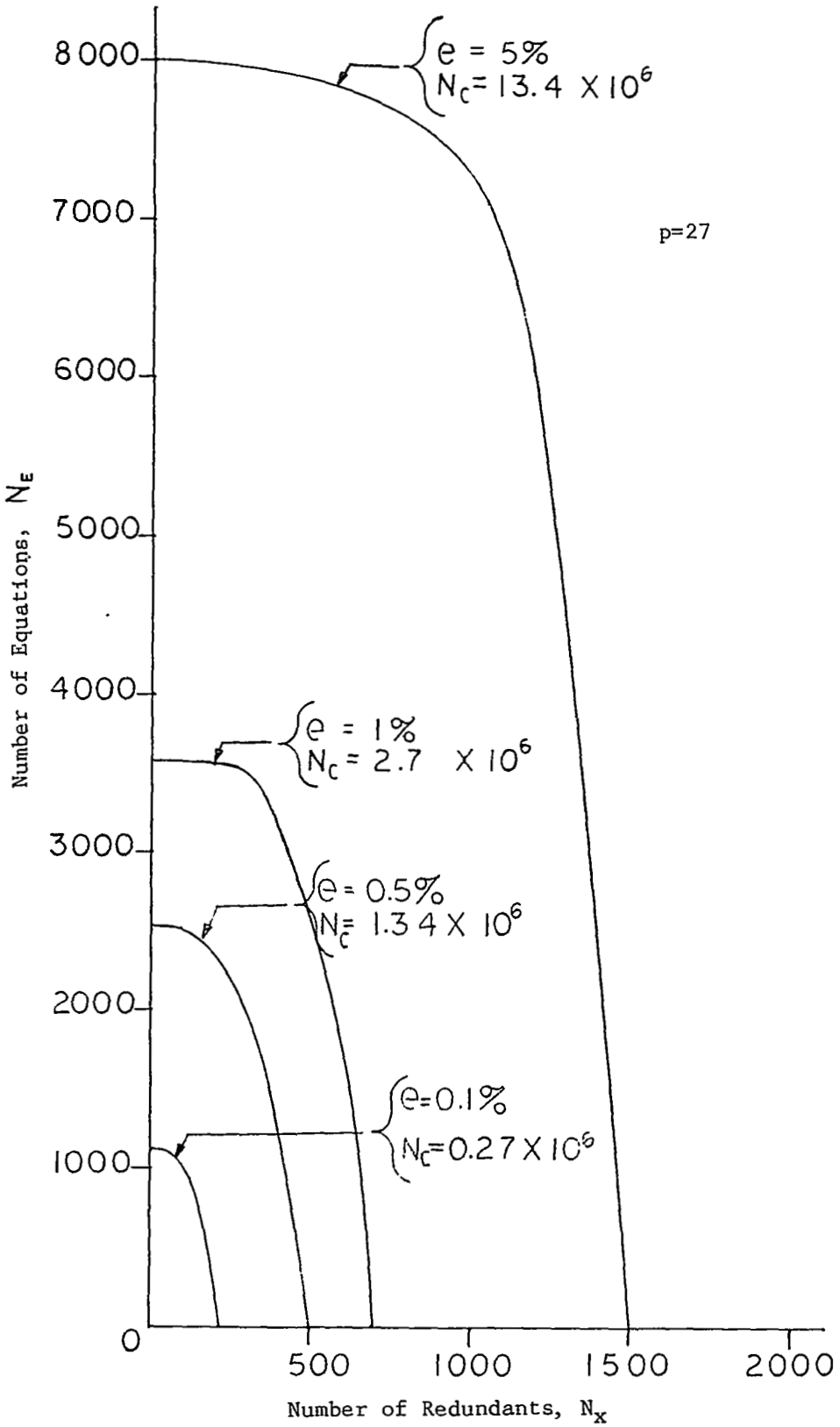
Therefore, it is seen that the numerical singularity of the geometric assembly matrix will not arise until the coordinates define degenerate structural geometry. The error in computations evaluating length, angles, areas, and volumes will be of the order of 2^{-p}. Only when the actual quantities are of this order, will the error destroy solution validity.

Error Analysis for the Redundants Matrix

The errors incurred in partitioning the geometric assembly matrix are inherent errors for the redundants matrix, δ_{xx}. These errors may be significant when many force unknowns are treated since they can induce unstable error propagation. The redundants matrix, like the stiffness matrix, must be positive definite if it represents a realizable structure. Error propagation will have characteristics like those described in Section 3 for the displacement method. However, since the δ_{xx} matrix is full, the differential equation approach described in Section 3 to evaluate the implications of the error propagation cannot be used.

Like the stiffness matrix, the redundants matrix could exhibit numerical singularity. The worst case system for singularity of the redundants matrix is for the parallel system. The parallel structure consists of "N" collinear members joined to common points at their ends. Numerical singularity for parallel rods and beam systems with equal and unequal flexibilities will be examined.

Figure 20 - Envelope of Relative Error and Number of Calculations



Parallel Rod Systems.- For the parallel system of N rods the structure is statically indeterminate to the N-1 degree. Assume member 1 is the determinate substructure. The redundants matrix equation (4-1) is then given by

$$\delta_{XX} = \begin{bmatrix} D_Q + D_{X1} & & & \text{SYM} \\ D_Q & D_Q + D_{X2} & & \\ D_Q & D_Q & D_Q + D_{X3} & \\ - - - & - - - & - - - & - - - \\ D_Q & D_Q & D_Q - - D_Q & + D_{XN} \end{bmatrix} \qquad (4\text{-}3)$$

where

$$D_Q = \frac{L}{AE} \text{ flexibility of determinate structure}$$

$$D_{Xi} = \frac{L}{AE} \text{ flexibility of } i^{th} \text{ redundant}$$

Assuming equal flexibilities of bars we can write

$$D_Q = D_{Xi} = \frac{L}{AE} \quad i=1, 2, \ldots N-1 \qquad (4\text{-}4)$$

Assuming decreasing flexibilities of bars and taking the most flexible bar as the determinate structure we can write

$$D_Q = \frac{L}{AE}$$

$$D_{Xi} = \frac{1}{2^{i-1}} D_Q \quad i = 1, 2, \ldots N-1 \qquad (4\text{-}5)$$

Considering increasing flexibilities of bars and taking the most rigid bar as the determinate structure we can write

$$D_Q = \frac{L}{AE}$$

$$D_{Xi} = 2^{i-1} D_Q \quad i = 1, 2, \ldots N-1 \qquad (4\text{-}6)$$

Consider now the triangularization of δ_{XX}. The general k^{th} diagonal element after k-1 Gauss-Jordan reduction represents the deflection at the cut in the direction of redundant k with all elements j, j=1 to k-1 closed and all elements m, m=k to N open.

Using the above physical interpretation, the following expression for the general k^{th} diagonal element after k-1 reductions can be derived

$$\delta_{kk} = D_{i+1} + \frac{\prod_{i=1}^{k} D_i}{\sum_{i=1}^{k} \prod_{i=1}^{k-1} D_i} \tag{4-7}$$

where

D_i = flexibility matrix of i^{th} rod element

$\prod_{i=1}^{k} D_i$ = continuous product of the flexibility matrices of the k redundant elements

$\sum_{i=1}^{k} \prod_{i=1}^{k-1} D_i$ = sum of the k continuous products of the flexibility matrices of the k redundant elements taken k-1 at a time

Using equation (7) for N rods in parallel with equal flexibilities D, we obtain

$$\delta_{NN} = D + \frac{D^N}{N(D^{N-1})} = D(1 + \frac{1}{N}) \tag{4-8}$$

The number of elements N at which numerical singularity occurs is when the second term in equation 7 or $\frac{1}{N}$ in equation 8 is smaller than the numerical value of the last recorded bit in the computer number representation. Assuming floating point arithmetic and letting p = 27,

$$\frac{1}{N} = 2^{-p-1}$$

$$\text{or } N = 2^{28} = 268 \times 10^6 \text{ elements} \tag{4-9}$$

Parallel Beam System.- Consider the same parallel system except the elements connecting the two end points are now beams instead of rods. The flexibility matrix to represent the elastic behavior of each beam may be characterized by two generalized forces. Each force may refer to a single independent force at a coordinate degree of freedom.

For the parallel system of beams there are two independent equations of equilibrium and 2N unknown element forces. Thus, the structure is indeterminate to 2(N-1) degrees. Assuming beam 1 is the determinate structure, the redundants matrix, equation (4-1) is also given by equation (4-3). However, D_Q = 2 x 2 flexibility matrix of determinate structure and D_{Xi} = 2 x 2 matrix of i^{th} redundant.

A linear transformation always exists from one set of generalized coordinates to another. The set of generalized coordinates shown in Figure 21 is used for the beam for this array.

Considering equal flexibilities of beams the elements of δ_{XX} in equation (3) can be written as

$$D_Q = D_{Xi} = \frac{L^3}{3EI}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \tag{4-9}$$

Considering decreasing flexibilities such that the determinate structure is most flexible, the elements of δ_{XX} in equation (3) are

$$D_Q = \frac{L^3}{3EI}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

$$D_{Xi} = \frac{L^3}{3EI}\,\frac{1}{2^{i-1}}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \tag{4-10}$$

$$i = 2, 3 \ldots N-1$$

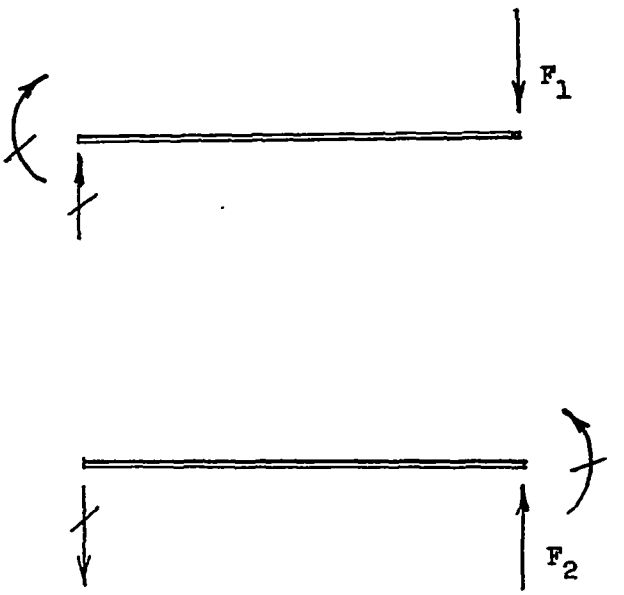
Considering increasing flexibilities such that the determinate structure is most rigid, the elements of δ_{XX} in equation (3) are

$$D_Q = \frac{L^3}{3EI}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

$$D_{Xi} = \frac{L^3}{3EI}\,2^{i-1}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \tag{4-11}$$

$$i = 2, 3 \ldots N-1$$

The prediction of numerical singularity for parallel beam systems follows the logic for rods and leads again to equation (7). The order of the matrix at which it is singular is the same for beams and rods. For beams, however, the element flexibility matrix D is a 2 x 2 matrix whereas for rods, D was a scalar. Thus, the number of parallel beam elements at which δ_{XX} is singular is only one half the corresponding number of parallel rod elements, i.e., 134×10^6.



Vectors With Stroke are Element Reactions

F_1, F_2 = Generalized Coord.

Figure 21- Generalized Coordinates For Beams

Since it is impractical to verify the singularity of the δ_{XX} matrix for 134×10^6 elements, a test problem of 32 parallel beams of equal flexibilities was used to verify the criteria. By inspection of the pivot values in the Jordan-diagonalization of δ_{XX} tabulated in Table XIV.

$$(\delta_{XX})_K = D(1 + \frac{1}{k+1}) \tag{4-12}$$

which confirms the previous relationship, equation (8).

For the case of unequal flexibilities, the previous relationship may also be applied. However, it is quite apparent from the input data what order the rows of the matrix are almost identical. The δ_{XX} matrix is a full matrix consisting of groups of 2 x 2 corresponding to the element flexibility matrix of the determinate beam. In addition, each 2 x 2 block along the main diagonal are incremented by the 2 x 2 element flexibility matrices of the redundants. For the case where the determinate beam is the most flexible, it is easy to see that when the difference between the elements of the flexibility matrix of the determinate beam and elements of the flexibility matrix of the N^{th} redundant beam is approximately 2^{-27}, the rows of the δ_{XX} matrix are practically identical which renders it singular. The matrix order when this occurs is 54. This is confirmed by an actual computer run. (Table XV). Since in the case of unequal flexibilities in which the determinate structure is the most rigid the numerical singularity will incur more elements than for the case of equal flexibilities, no computer runs were made for the case of increasing flexibilities.

Assuming a precision p = 27 bits, the main conclusions from the singularity error analysis of the matrix, δ_{XX}, for a parallel system of rods and beams are:

(1) Considering beam elements of decreasing flexibilities and assuming the determinate structure to be the most flexible element, very few elements (54) are required to cause singularity, i.e., meaningless answers due to critical arithmetic.

(2) Considering beam elements of equal flexibilities, a substantial number of elements (134×10^6) are required before numerical singularity occurs.

(3) Considering beam elements of increasing flexibilities and assuming the determinate structure is the least flexible (most rigid), the number of elements to cause singularity are even much greater than in (2) above.

(4) The number of rod elements to cause singularity is twice the number of beam elements for each of the three corresponding cases since the order of the matrix at which it is singular is the same for beams and rods. For beams, however, the element flexibility matrix D is a 2 x 2 matrix whereas for rods, D was a scalar.

Table XIV Pivot Values In Triangularization of δ_{xx}

For Parallel Beams, Equal Flexibilities

PIVOTS IN ORDER OF SELECTION

COL	PIVOT	COL	PIVOT	COL	PIVOT	COL	PIVOT
1	.10000+01	2	.75000-00	3	.75000-00	4	.56250-00
5	.66667-00	6	.50000-00	7	.62500-00	8	.46875-00
9	.80000-00	10	.45000-00	11	.58333-00	12	.43750-00
13	.57143-00	14	.42857-00	15	.56250-00	16	.42188-00
17	.55556-00	18	.41667-00	19	.55000-00	20	.41250-00
21	.54545-00	22	.40909-00	23	.54167-00	24	.40625-00
25	.53846-00	26	.40385-00	27	.53571-00	28	.40179-00
29	.53333-00	30	.40000-00	31	.53125-00	32	.39844-00

Table XV Pivot Values In Triangularization Of δ_{xx}

For Parallel Beams, Diminishing Flexibilities

PIVOTS IN ORDER OF SELECTION

COL	PIVOT	COL	PIVOT	COL	PIVOT	COL	PIVOT
1	.10000+01	2	.75000-00	3	.46667-00	4	.35000-00
5	.23810-00	6	.17857-00	7	.12157+00	8	.91176-01
9	.61584-01	10	.46188-01	11	.31013-01	12	.23260-01
13	.15565-01	14	.11674-01	15	.77974-02	16	.58480-02
17	.39025-02	18	.29268-02	19	.19522-02	20	.14641-02
21	.97633-03	22	.73225-03	23	.48823-03	24	.36617-03
25	.24412-03	26	.18310-03	27	.12206-03	28	.91548-04
29	.61034-04	30	.45776-04	31	.30510-04	32	.22883-04
33	.15257-04	34	.11443-04	35	.76215-05	36	.57161-05
37	.38127-05	38	.28521-05	39	.18994-05	4	.14227-05
41	.95168-06	42	.70585-06	43	.48379-06	44	.36085-07
45	.24012-06	46	.17958-06	47	.12708-06	48	.95185-07
49	.61476-07	50	.46071-07	51	.37707-07	52	.20821-07

MATRIX DELXX IS SINGULAR

In summary, the ideal situation is to select the most rigid element as the determinate structure. To apply this criteria derived for a worst case parallel system to a practical structure, the criteria may be stated as follows: The optimum determinate base structure is a stable substructure which is as rigid (large structural stiffness) as possible. This also corresponds to an optimum selection of redundants. With the use of accurate and stable solution algorithms, these optimum conditions would minimize manipulation errors.

The proper choice of the base structure and corresponding redundant stress system is the most important strategy in the force method. The improper selection of the determinate structure can result in meaningless answers due to the following conditions:

(1) Singularity of the P_Q matrix caused by an implied kinematic instability of the determinate structure.

(2) Singularity of the δ_{XX} matrix when the determinate structure is very flexible.

(3) Significant attrition errors in the calculation of internal element forces Q in equation (2-29) when the effect of the loads and the redundants on the base structure are comparable in magnitude.

Guidelines for the Analyst

The guidelines that may be provided for the analyst pertain to the basic input data which affects generation errors. These are related to the proper idealization of the structure, proper choice of reference systems, and accurate description of the geometry of the structure. As previously noted in section 3, these aspects are within the control of the analyst in properly formulating the problem in order to avoid critical arithmetic that causes numerical singularity and subsequent invalidation of answers.

The analyst should locate the global coordinate system near the center of his structure to minimize the span between coordinate points which define the boundaries of elements. If necessary the scale of the structure should be selected such that the element farthest removed from the origin are satisfactorily represented by the difference of its coordinates.

In the idealization of the structure the numbering of the elements defines the ordering of the element forces which correspond to the columns of the P matrix as well as the ordering of the flexibilities of elements in the D matrix. The numbering of elements defines the sequence in which flexibilities are added. To minimize the errors and avoid critical arithmetic, the joints should be located such that adjacently ordered elements should have commensurate flexibilities. If incommensurate flexibilities are to be added, the analysts can optimize the arithmetic by numbering his elements, adding the smaller flexibilities first.

In the idealization of the structure it is also desirable for the analyst to have an idea of the possible load paths for the external loads to the points of supports. With this idea in mind the joints can be located and the elements can be numbered such that those elements which comprise the more direct load paths are included in establishing the determinate base structure. The significance of this criterion cannot be overemphasized. It is a complementary measure for an adequate structure-cutter whose essential features will be discussed under "Guidelines to the Programmer." In general this base structure is comprised of elements with the smallest flexibilities and in addition account for the connectivity of all the structural elements. In this case since the effect of the redundants will be minimum, the attrition errors in the calculation of internal element forces Q in equation (2-29) will also be minimum.

In addition, attrition errors in triangularizing the P matrix will be small, because the number of calculations is small. For a general practical problem the total number of calculations in the elimination process and evaluation of internal element forces depends on two parameters. These are the number of independent equations of equilibrium, N_E or the row dimension of P and the number of unknown element forces N_F or the column dimension of P. Since the number of redundants, N_X corresponds to the difference between the column and row dimensions of P an alternative set of parameters would be the number of independent equations of equilibrium and the number of redundant forces.

Consider first a pure series system. Since this is statically determinate the number of independent equations of equilibrium equals the number of unknown element forces. The total number of calculations depends solely on the order of the matrix P. For the number of calculations to be of the order of 13.4×10^6 the number of unknowns will be approximately 8000. This number is considered to be an upper bound for the number of unknown element forces. This is based on the assumption that the P matrix is five percent dense and that the maximum number of calculations for an element force is twice the average calculation for all elements of the solution vector.

Consider next a pure parallel system of beams. Most of the calculations will pertain to the inversion of the redundant matrix δ_{XX}. For the number of calculations to be of the order of 13.4×10^6 the number of redundants will be approximately 1500. This number is considered to be an upper bound for the number of redundants. The result is based on the assumption that the δ_{XX} matrix is full and that the maximum number of calculations for a redundant is twice the average calculation for all the redundant forces.

For a more practical problem the determination of the total number of calculations is more involved. In carrying through the solution process of the general structural equations presented in Section 2, the total number of calculations for one element force under one loading condition has the following general form: (see figure 20)

$$N_c = \frac{2}{N_E+N_X}\left\{\frac{1}{10}\left[N_E^3 + N_E^2 N_X\right] + N_E N_X^2 + 3N_X^3\right\} \qquad (4\text{-}2)$$

where

N_C = total number of calculations

N_E = independent equations of equilibrium

N_X = total number of redundants

Equation (4-2) is based on the following assumptions:

(1) the matrix P is five percent dense while the matrix δ_{XX} is full.

(2) The maximum number of calculations for one element force is twice the average calculation for all the element forces.

Additional guidelines to the analyst pertain to making good choices of redundancies where they are not automatically established by the computer. The only general rule that can be given in the good choice of redundancies is that their effect should be as localized as possible, i.e., their effect should not propagate throughout the entire structure. The objectives in making a good choice of redundants are as follows:

(a) To have a stable base structure.

(b) To have a well conditioned δ_{XX} matrix using as a criteria $\delta_{ii} \gg \delta_{ik}$ for all k.

(c) To minimize the amount of calculations in the inversion of δ_{XX} by reducing the number of non-zero elements to minimum.

(d) To minimize the effect of the redundants on the determinate structure in order to reduce attrition errors in the calculations of internal element forces to minimum.

This will be illustrated with simple examples.

Consider the continuous beam shown in figure 22. We discuss three alternative choices of redundancies.

(a) In Figure 22 a redundants X_1 and X_2 are taken as the reactions at the supports A and B. The δ_{XX} matrix for this system is

$$\delta_{XX} = \begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix} = \frac{L^3}{6EI} \begin{bmatrix} 24 & 9 \\ 9 & 4 \end{bmatrix} \tag{4-13}$$

The above choice of redundants is a remarkably bad choice since $\delta_{12} \gg \delta_{22}$.

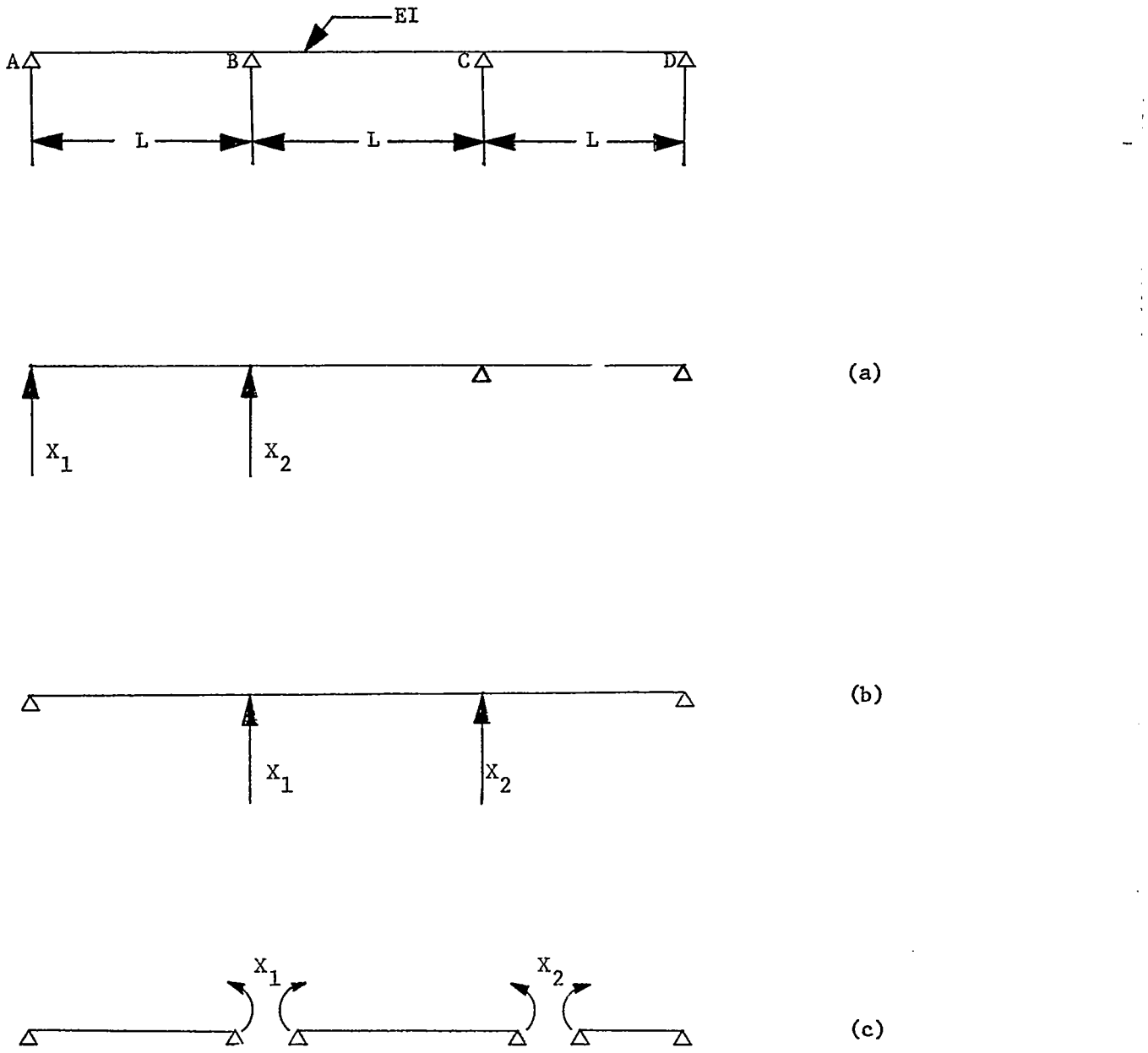


Figure 22 - Alternative choices of redundants for continuous beam

(b) In Figure 22b redundants X_1 and X_2 are taken as the reactions at the intermediate support B and C. For this system

$$\delta_{XX} = \frac{L^3}{18EI}\begin{bmatrix} 8 & 7 \\ 7 & 8 \end{bmatrix} \tag{4-14}$$

This is still a bad choice since all the δ's are of the same order of magnitude.

(c) In Figure 22c X_1 and X_2 are taken as the bending moments at supports B and C. Then

$$\delta_{XX} = \frac{L}{12EI}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \tag{4-15}$$

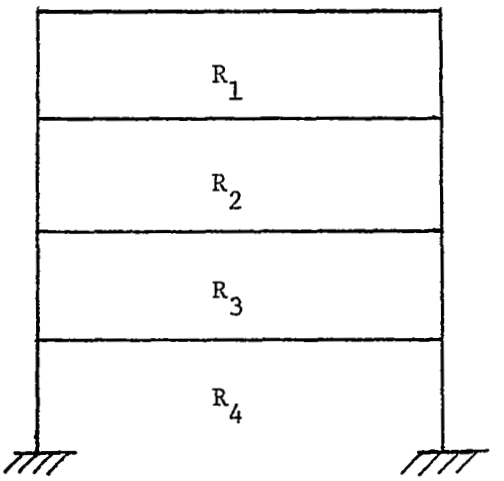
The choice (c) is clearly the most suitable choice of redundants. The differences in the above systems become even more pronounced when the number of spans is increased. For example for six spans the corresponding matrices for release system (b) and (c) are as follows:

$$(\delta_{XX})_b = \frac{L^3}{150EI}\begin{bmatrix} 160 & 325 & 200 & 115 \\ 225 & 360 & 340 & 200 \\ 200 & 340 & 360 & 225 \\ 115 & 200 & 225 & 160 \end{bmatrix} \tag{4-16}$$

$$(\delta_{XX})_c = \frac{L}{6EI}\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix} \tag{4-17}$$

The coefficients in δ_{XX} for choices (a) and (b) tend, for a large number of spans, to become linearly dependent.

As another illustration of a good and bad choice of redundancies consider the plane frame shown in Figure 23. We discuss two alterative choices of redundancies for making the frame statically determinate. The structure is a plane frame with four rings so that it is 12 times statically indeterminate.



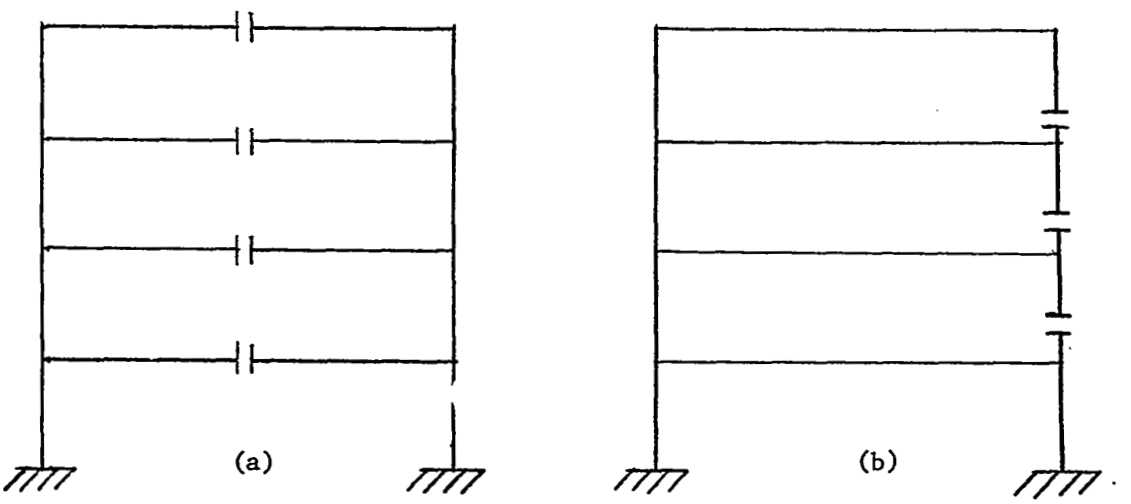


Figure 23 - Alternative choices of redundants for plane frame

(a) In Figure 23a a cut is made of the center of each of the beams. It is seen that each of the arbitrary constants at the release of the top beam will produce a bending moment diagram throughout the length of the columns so that each ring R_1 to R_4 is affected. Thus, all influence coefficients δ_{11}, δ_{12}, δ_{13}, ... $\delta_{1,\ 12}$ will exist, there being three arbitrary constants for each ring. In the case of the second beam from the top the arbitrary constants of this cut will affect the beam itself and the whole length of the column from this beam to the base so that each ring R_1 to R_4 is affected. Thus, all influence coefficients δ_{41}, δ_{42}, δ_{43} ..... $\delta_{4,\ 12}$ will exist. Each succeeding beam will affect adjacent rings and all other rings below the beam. We therefore find that each component of the matrix δ_{XX} exists, i.e.,

$$\delta_{XX} = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \cdots & \delta_{1,\ 12} \\ \delta_{21} & \delta_{22} & \delta_{23} & \cdots & \delta_{2,\ 12} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \delta_{12,\ 1} & \delta_{12,\ 2} & \delta_{12,\ 2} & \cdots & \delta_{12,\ 12} \end{bmatrix} \tag{4-18}$$

(b) Consider now the other release system where one of the columns in each ring is cut as shown in Figure 23b. In this system the arbitrary constants of ring R_1 will affect only the members of ring R_1. Therefore the following influence coefficients of ring R_1 will exist:

$$R_{11} = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} \\ \delta_{31} & \delta_{32} & \delta_{33} \end{bmatrix} \tag{4-19}$$

Similarly for the arbitrary constants of ring R_2 the influence coefficients exist as follows:

$$R_{22} = \begin{bmatrix} \delta_{44} & \delta_{45} & \delta_{46} \\ \delta_{54} & \delta_{55} & \delta_{56} \\ \delta_{64} & \delta_{65} & \delta_{66} \end{bmatrix} \tag{4-20}$$

and so on for the four rings.

However, since the second beam from the top is a shared member, there is an interaction between the ring R_1 and R_2 so that the following linking terms arise:

$$R_{12} = \begin{bmatrix} \delta_{14} & \delta_{15} & \delta_{16} \\ \delta_{24} & \delta_{25} & \delta_{26} \\ \delta_{34} & \delta_{35} & \delta_{36} \end{bmatrix} \qquad (4\text{-}21)$$

In the case of a ring sharing members with rings on either side, two such sets of linking terms arise. The final pattern of the δ_{XX} matrix can be written as follows:

$$\delta_{XX} = \begin{bmatrix} R_{11} & R_{12} & 0 & 0 \\ R_{21} & R_{22} & R_{23} & 0 \\ 0 & R_{32} & R_{33} & R_{34} \\ 0 & 0 & R_{43} & R_{44} \end{bmatrix} \qquad (4\text{-}22)$$

This form of the matrix is usually called a continuant matrix. For all sets of simply connected rings a choice of redundancies can be made such that the above pattern applies. It is noted that the above pattern also results from the continuous beam using moment releases at the supports.

As a third illustration in the analysis selection of redundancies we consider the case where it is possible to separate the redundants into groups such that there is coupling of redundants only within each group but there is no coupling of redundants between groups.

For example, consider the generalized loading of plane frames such as the one shown in Figure 24. In this case the arbitrary constants separate into two sets.

(a) The first set consists of the moment X_1 and associated shear X_2 for bending in the plane of the portal and the direct thrust X_3.

(b) The second set consists of the moment X_4 and associated shear X_6 for bending normal to the plane of the portal and the torsional moment X_5.

Note that the arbitrary constants in one group are not coupled to those of the other. Thus, the matrix δ_{XX} is as follows:

$$\delta_{XX} = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & 0 & 0 & 0 \\ \delta_{21} & \delta_{22} & \delta_{23} & 0 & 0 & 0 \\ \delta_{31} & \delta_{32} & \delta_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_{44} & \delta_{45} & \delta_{46} \\ 0 & 0 & 0 & \delta_{54} & \delta_{55} & \delta_{56} \\ 0 & 0 & 0 & \delta_{64} & \delta_{65} & \delta_{66} \end{bmatrix} \qquad (4\text{-}23)$$

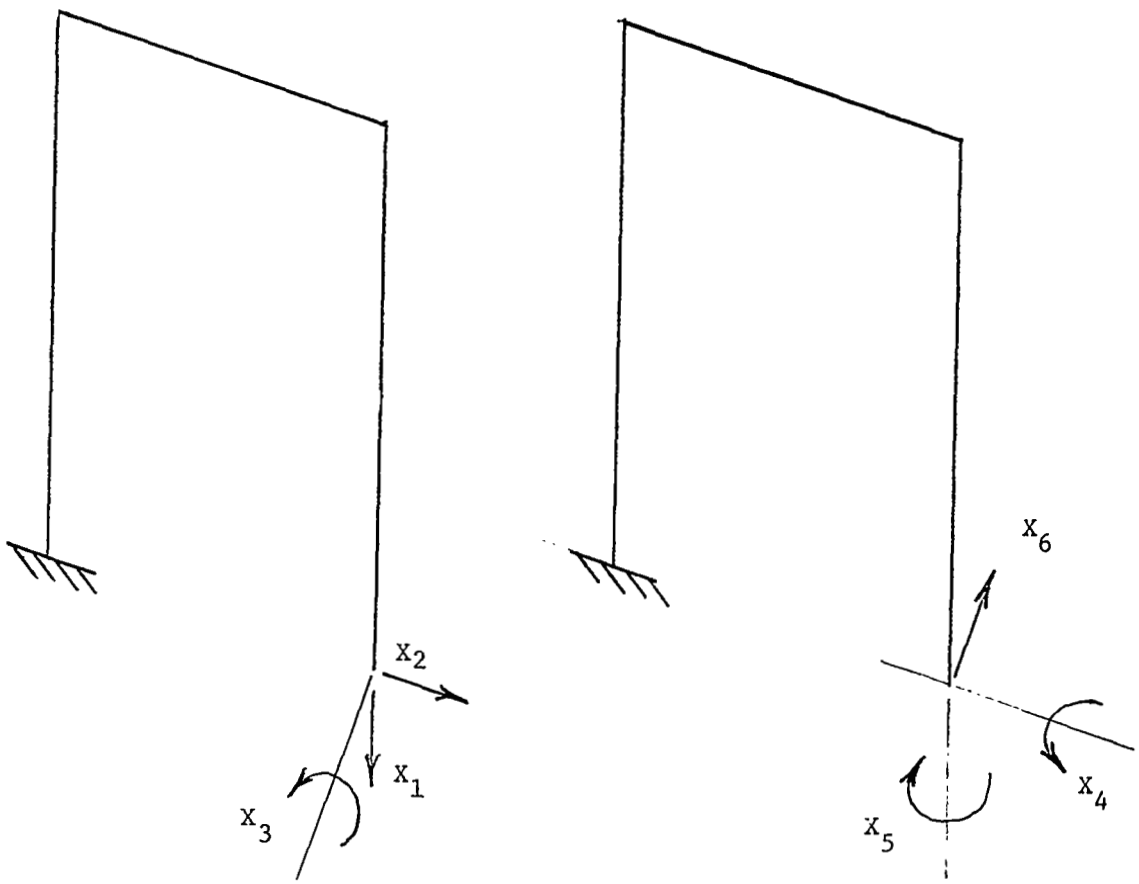


Figure 24 - Separable groups of redundants

$$\delta_{XX} = \begin{bmatrix} \delta'_{XX} & 0 \\ 0 & \delta''_{XX} \end{bmatrix} \tag{4-23}$$

Thus, using equation (4-23) and by partitioning and expanding the equation for calculating the arbitrary constants, equation (2-28) can be broken into two independent sets of simultaneous equations each of which will involve a smaller number of unknowns than the original set.

The separation of δ_{XX} into two parts depends upon the following:

(a) The arbitrary constants shall be labelled in such a way that the copuled elements are consecutively ordered together in one group and no coupling occurs between elements of different groups.

(b) The axes of both the arbitrary constants X and the member element forces are related to an orthogonal set of axes oriented to conicide with the principal axes of the structure and also that all the shear centers of the members of the structure lie in one plane.

The above concepts of selecting redundancies illustrated by some simple examples can be readily applied to structures which are comprised of relatively simple finite elements and which are relatively regular in geometry. For more general types of structures with more complex geometry the choice of redundants may not be quite obvious. In many cases it is best when the equations are ill conditioned, to select a different basic system and corresponding redundancies X which are statically equivalent with any previous choice of redundancies X.

In cases where the above guidelines fails to yield a well conditioned δ_{XX} matrix, a transformation from a previous choice of redundants to an orthogonal set of redundants [28, 29, 30] can be performed. Although this approach yields a diagonal δ_{XX} matrix and is therefore very well conditioned there are a considerably greater number of calculations involved. The procedure is as follows:

Let f_X be the matrix of element forces (including both statically determinate forces and redundants) resulting from unit values of the redundants. Then

$$f_X = \begin{bmatrix} -P_Q^{-1} P_X \\ I \end{bmatrix} \tag{4-24}$$

The columns of f_X comprise the base vectors for the previously chosen redundants.

The next step is to orthogonalize these base vectors by the Gram-Schmidt orthogonalization procedure. The desired set of orthogonal redundants is obtained as follows: The first redundant of the desired set is chosen equal to the first column vector of f_X, i.e.,

$$\bar{f}_1 = f_1$$

The succeeding set of desired redundants are related to the given set as follows:

$$\begin{aligned} \bar{f}_2 &= f_2 + \bar{f}_1 C_{12} \\ \bar{f}_3 &= f_3 + \bar{f}_1 C_{13} + \bar{f}_2 C_{23} \\ \bar{f}_r &= f_r + \bar{f}_1 C_{1r} + \bar{f}_2 C_{2r} + \ldots \bar{f}_{r-1} C_{r-1} \end{aligned} \tag{4-25}$$

It remains to determine the coefficients C_{ij} such that δ_{XX} is diagonalized. i.e., in the congruent transformation

$$\delta_{XX} = \bar{f}_x^T D \bar{f}_x \tag{4-26}$$

the condition to be satisfied is

$$(\delta_{XX})_{ij} = (\delta_{XX})_{ji} = 0 \quad i \neq j \tag{4-27}$$

Using this condition the coefficient C_{ij} in the transformation is

$$C_{ij} = - \frac{\bar{f}_i^T D f_j}{\bar{f}_i^T D \bar{f}_i} \quad j > i \tag{4-28}$$

Beginning with the known vector $\bar{f}_1 = f_1$, the constant C_{12} is computed then the vector $\bar{f}_2$. Having obtained $\bar{f}_2$ the constant C_{13} and C_{23} are calculated and so on.

Note that the coefficients C_{ij} depend only on the nature of the structure.

In the continuity equation the unknown redundants X are readily obtained because the diagonal matrix δ_{XX} is easily inverted. Finally, the element forces are obtained from

$$F = \bar{f}_X \bar{X} + f_o \tag{4-30}$$

where

$$f_o = \begin{bmatrix} -P_Q^{-1} F \\ 0 \end{bmatrix} \tag{4-31}$$

One interesting property of the transformation to an orthogonal set of redundants is the fact that the matrix f_o is invariant to the class of transformation

$$\bar{f}_X = Cf_X$$

where C is a non-singular matrix.

The main significance of the above approach is that in spite of the ill-conditioning of the equations the transformation to an orthogonal set of redundants avoids critical arithmetic due to any numerical singularity of the δ_{XX} matrix. This is done however at the expense of a considerably larger number of calculations which incurs manipulation errors.

Guidelines to the Programmer

Guidelines for the programmer includes related **strategies** to properly sequence calculations to minimize errors. These revolve around the central theme of arriving at the most suitable stable base structure so the best set of redundant forces is used.

The following generally related areas in which guidelines to the programmer will help avoid critical arithmetic and minimize manipulation errors in the force method are:

(1) Scaling of matrices prior to the solution process.

(2) Programming for elimination of extraneous equations in the equation set.

(3) Programming to detect, minimize and control errors in the solution process by the Jordan diagonalization method.

(4) Programming to mathematically cut the structure to obtain a stable and rigid determinate base structure and the set of redundant forces that optimizes the conditioning of the associated coefficient matrix.

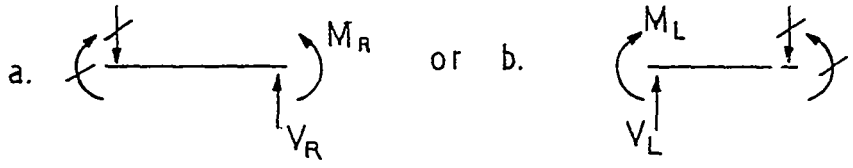
Depending upon the algorithm selected, items (4) and (3) can either be separate or combined.

In the elimination process there are related strategies that are essential to minimize, detect, and control manipulation errors and to insure success in the Gauss elimination method or Jordanian diagonalization method when used with unsymmetric matrices. These strategies are (a) scaling and (b) use of pivoting in the elimination process. Numerical studies by Wilkinson [31-32] show that by the use of scaling and pivoting, the Gauss elimination process can be stabilized to prevent break down.

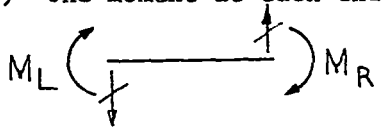
Consider first scaling. As an illustrative example, consider different possible choices for the generalized coordinates to represent the flexibility matrix of a beam. The elastic behavior of the beam may be

characterized by two generalized forces which may be a single independent force at a coordinate degree of freedom or a linear combination of such forces. The following sets of generalized forces are possible: (The arrows with the strokes are the element reactions).

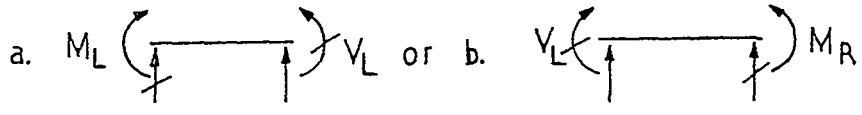
(1) One shear force and one moment at one end:



(2) One moment at each end

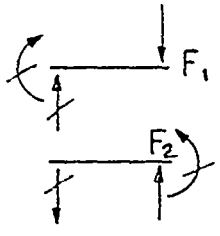


(3) One shear force at one end and one moment at the other end



An alternative set of generalized forces based on a linear combination of the forces in set (1) above is as follows:

(4)



It can be readily seen that the generalized forces F_1 and F_2 in the above set is a linear combination of the generalized forces in 1a and 1b.

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} V \\ M \end{bmatrix} \qquad (4\text{-}32)$$

The respective flexibility matrices for the above sets of generalized forces are as follows:

1a or 1b: One shear force and one moment at an end

$$D = \frac{L}{EI}\begin{bmatrix} \theta & u \\ 1 & \frac{L}{2} \\ \frac{L}{2} & \frac{L^2}{3} \end{bmatrix} \qquad (4\text{-}33)$$

Scaling the displacements u by $\frac{1}{L}$ and the forces by L we obtain

$$D = \frac{L}{ET}\begin{bmatrix} \theta & \frac{u}{L} \\ 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix} \qquad (4\text{-}34)$$

(2) One moment each end

$$D = \frac{L}{3EI}\begin{bmatrix} \theta_1 & \theta_2 \\ 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \qquad (4\text{-}35)$$

(3) One shear force at one end and one moment at the other end

$$D = \frac{L}{EI}\begin{bmatrix} \theta & \frac{u}{L} \\ 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix} \qquad (4\text{-}36)$$

(4) Linear combination of forces in set 1

$$D = \frac{L^3}{3EI}\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \qquad (4\text{-}37)$$

Using the ratio of the maximum and minimum eigenvalues as a measure of matrix conditioning it is evident that the flexibility matrices given by (35) and (37) are better conditioned than those given by (34) and (36) since the eigenvalue ratios in the former are much smaller and closer to one than the latter. Since a linear transformation always exist from one set of generalized forces to another, by proper scaling one can seek to obtain the most suitable form of flexibility matrix to use. The above results for a beam show that using either two shear forces or two moments will yield a better conditioned matrix than using a combination of a shear force and moment.

Consider next the problem of scaling the matrix A in the solution of $Ax = b$. Alternate terms for scaling are preconditioning and equilibration. This is normally done by multiplying the rows and columns of the matrix by factors such that the elements of the matrix have approximately the same magnitude. More rigorously, if the condition of the matrix A is defined by some measure, the non-singular diagonal matrices C_1 and C_2 can scale A by the transformation

$$A \rightarrow C_1^{-1} A\, C_2$$

so as to reduce the condition of $(C_1^{-1} A\, C_2)$ to as low a value as is reasonably possible. Powers of the floating point base are usually used for scale factors, to avoid the introduction of rounding errors in the scaling.

Although some studies[33] have been made on the problem of finding C_1 and C_2 to minimize the condition of $(C_1^{-1} A\, C_2)$, it turns out that the solution involves computation of A^{-1} to find a reasonable scaling. Thus, from a practical standpoint, the above method is open to question. The only advice that one is sometimes given is to pick C_1 and C_2 so that the resulting matrix $C_1^{-1} A\, C_2$ has elements of approximately the same magnitude or has its maximum element in each row and column (in absolute value) in the interval (.1,1) in whatever base one is using.

Another device for minimizing manipulation error is modified pivoting. Pivoting consists of interchanging rows, columns, or rows and colums so that the diagonal element in the i^{th} row, after i reductions is non-zero. This is a necessary part of Jordan reduction for an arbitrary non-singular matrix. In modified pivoting, an attempt is made to find a large diagonal. The pivot element is the diagonal element found.

There are three pure pivoting stratagies:

(a) Complete pivoting in which at each stage one selects as a pivot some element a_{ij} of maximum value among all the remaining elements of the matrix.

(b) Partial pivoting with row interchange in which at each stage one selects as a pivot some element a_{ij} of maximum absolute value among the first column of the remaining elements of the matrix.

(c) Partial pivoting with column interchange in which one selects as a pivot some element a_{ij} of maximum absolute value among the first row of the remaining elements of the matrix.

Positive definite matrices, such as the matrix δ_{XX} in the continuity equations, do not require pivoting in the Jordan process though modified pivoting may be advantageous.

Although complete pivoting has invariably provided success and small error bounds, the penalty paid is much longer computer times than partial pivoting. Studies by Wilkinson[31] and previous experience with practical types of problems indicate that the use of partial pivoting should be satisfactory. However, since in floating point computation it is generally not easy to determine if some number is effectively zero or not, it is desirable to use as a safeguard a certain "pivot tolerance." If in a certain row (if partial pivoting by row interchange is used) or if in a certain column (if partial pivoting by column interchange is used) no pivot greater than the "pivot tolerance" can be found, then that row or column is considered to be a linear combination of the other rows or columns in which pivots have already been chosen. A pivot tolerance of 10^{-5} is adequate for a 27 bit mantissa.

Scaling and pivoting are actually related. From a theorem by Bauer[33], if the ordered set of pivotal elements is selected in advance, scaling of matrix A by powers of the floating point base does not change a single digit of significance of any intermediate or final number in the solution of $Ax = b$ by Gaussian elimination. Thus, the only possible effect of the scaling of A on the rounding errors must occur through changing the order of pivots.

Probably the most important of all guidelines for the programmer for minimization of manipulation errors is in programming to obtain the best set of redundant forces. Recall that this operation is related to the partitioning of the matrix P. In the selection of this set, it implies that the associated base structure is stable and flexibility matrix for the redundant set of forces is best conditioned. The term conditioning of a matrix is based on some kind of measure. A well conditioned matrix generally implies small errors in the elimination process.

It is well known that the choice of the redundant forces in a statically indeterminate structure is not unique. The only restriction is that for any choice of redundant stress system, the associated determinate structure must be stable. This implies that the coefficient matrix for the determinate forces must be non-singular.

Starting with the matrix P in the cutting of the structure mathematically, it is desirable to weigh the elements of the P matrix by the relative flexibilities of the elements. The idea is to bias the choice of redundant forces in such a way that the more rigid members will comprise the base structure and the more flexible members will comprise the redundants. It can be deduced from the results of the test problems and the error analysis for the redundants matrix, δ_{XX}, that ideally the resulting base structure should be the most rigid of all possible alternative choices. In this case the load path is most direct. The effect of the redundants is minimum since the stresses in the base structure will be approximately the same as the stresses in the final structure. If the base structure is relatively flexible the redundant forces have a significant effect on the stresses of the base structure and critical arithmetic may be involved.

It is noted that most rigid elements do not necessarily build up the most rigid, stable base structure. The connectivity of the elements also enters into consideration. The weighting factors suggested aims to develop a base structure incurring low, not least, manipulation error.

Elimination of extraneous equations in the equations of equilibrium should be automated to eliminate inadvertent numerical singularities. This may be accomplished in the generation of the elements of the coefficient matrix P to yield linearly independent rows or in the Jordanian elimination process.

In the most general case six equations per joint or free body are required. In many instances, however, fewer than six are independent. For example, a plane pin jointed truss requires only two equations per joint. Writing more than two equations results in a linearly dependent set of equations.

The basic problems are twofold:

(1) To determine linear dependency of the equations. This is manifested if in the triangularization of the P matrix one or more dependent equations of the form 0 = 0 will result. Due to round-off errors, however, the numbers may not be exactly zero.

(2) To determine if the equations are consistent by investigating if sufficient internal restraint is provided for each applied load. In the triangularization of P, with simultaneous reduction of $F = F_o$, one or more equations of the form $0 = f_i$, where f_i is a non-zero component of the applied load may result. In this case the equilibrium equations are inconsistent or incompatible and no solution exists. The model of the physical system is mathematically unstable. If the equilibrium equations defined by the coefficient matrix P is to be consistent and have linearly independent rows to avoid singularity, the extraneous equations of the form $0 = f_i$ must be eliminated or modified by introducing additional structural restraint.

The elimination of extraneous equations may be achieved in two ways. In the first the programming provides a means of inspecting the individual joints to determine whether the internal element forces are null, colinear, or coplanar and write only the appropriate number of equations. Since the above conditions may not be exactly satisfied due to input-output error, a cut-off criteria should be used in the vector evaluation to determine if the degree of divergence from any one of the above conditions (i.e., null, colinear, or coplanar) is sufficient to warrent inclusion of the related equilibrium equations. An upper limit of the cut-off criteria define vector magnitudes which are sufficient to warrant inclusion of equations. A lower limit of the cut-off criteria defines vector magnitudes which are considered insignificant and permit elimination of equations. Note that the lower limit of the cut-off criteria is intended to compensate for the errors in the generation of the coefficients and do not introduce additional errors. The upper limit of the cut-off criteria is quite arbitrary. It is intended to detect errors due to poorly idealized joints or in defining coordinates. From experience, the equations representing divergence from the condition of null, colinear, or coplanar ranging from .001 to .003 are often due to the above mistakes indicating erroneous idealization and formulation. Thus, an upper limit cut-off criteria of .003 is considered satisfactory.

The second way is to sense extraneous equations directly in the Jordanian elimination process. In this method, an equivalent criteria is used by specifying a certain "pivot tolerance" to determine if some number is effectively zero or not. This method of elimination of extraneous equations, however, requires row interchanges rather than column interchange in the partial pivoting strategy and is therefore usually uneconomical. If in a certain row no pivot greater than a certain pivot tolerance can be found then that row is considered to be a linear combination of the other rows and pivots that have already been chosen.

As final guidelines to the programmer, there are checks that should be included in the force method computer program to validate solutions. The first consists of a singularity test. The pivot values in the triangularization of the P and δ_{XX} matrix should be checked. If they are less than the prescribed tolerance level of 10^{-5} the matrix is considered singular and error notes should be printed out.

The following computer based tests which can be included are similar to those already discussed in Section 3 for the displacement method.

(1) Test for error propagation characteristics in triangularization process.

(2) Equilibrium check.

In addition to the above tests the following checks are peculiar only to the force method.

(1) Back-substitution check in continuity equations. This is one measure of the accuracy in the calculation of redundants.

(2) Evaluation of displacement calculations with respect to manipulation error.

Some investigators[7, 34] have suggested the use of eigenvalue ratios as a measure of manipulation error. The present analysis and test problems have shown that these criteria is not always an adequate approach since it does not reflect any numerical singularity in the two coefficient matrices of concern. The matrix eigenvalue ratios for a structure does not change regardless of the sequence of solving the equations. It has been shown in this analysis however that the errors vary dramatically depending on how the equilibrium equations are partitioned, i.e., the physical cutting of the structure.

Section 5

VERIFICATION ANALYSES

This section describes the displacement and force analyses of a practical wing. The purpose of these analyses is to determine the magnitude of manipulation error in a typical analysis and to apply the guidelines of Sections 3 and 4 to a practical structural analysis. In addition, these analyses will provide a comparison between displacement and force method manipulation errors.

Description of Problems

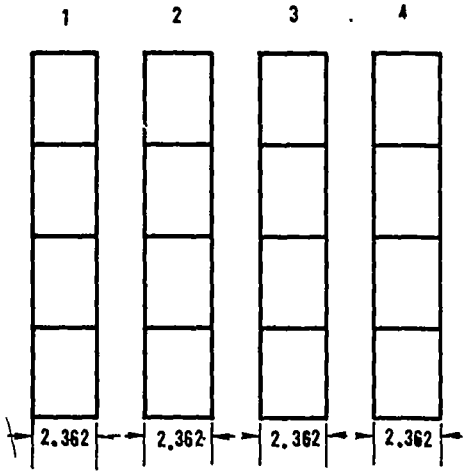
Figure 25 shows the geometry of the structure. The structure consists of five main spars and four ribs supporting variable thickness skins. Table XVI defines the geometry and material properties of the elements of the structure.

A model of this structure has been fabricated and analyzed. These data indicate that the basic network (identified by solid and dashed lines in Figure 25) leads to satisfactory predictions of behavior.

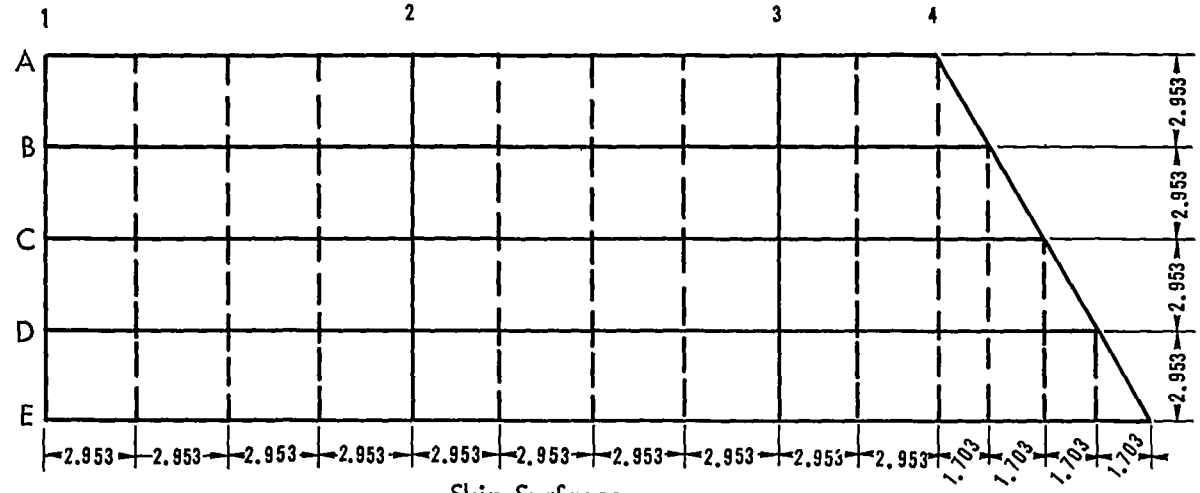
The boundary conditions for this study are summarized in Table XVII. The upper table defines the loading conditions when the box represents a swept wing. The wing has a 30 degree sweep and represents the structural box of a low aspect ratio surface. The swept end is fully fixed. All of the six loads act normal to the wing.

The second set of boundary conditions consist of six loadings under the assumption of full fixity along the long edge of the box. The structure is then an unswept box. This alternate fixity condition changes the ratio between series and parallel elements in the structure. In the first boundary condition there are between ten and fourteen series elements and nine parallel elements. With the fixity along the long edge, there are between ten and fourteen parallel elements and nine series elements.

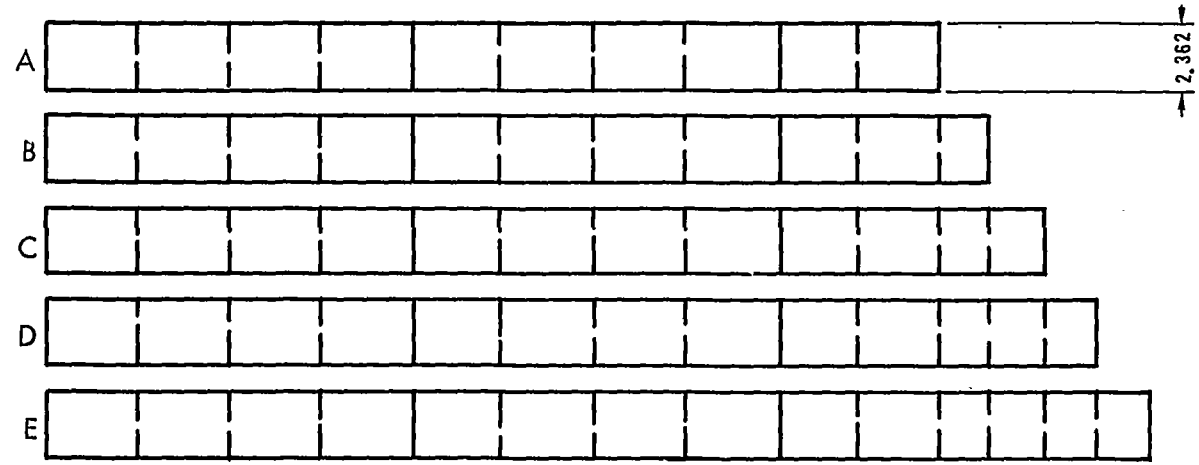
Displacement Method Analysis. - The structure shown in Figure 25 is idealized as shear panels, membranes, and rods for the displacement method analysis. The webs of the spars and ribs are treated as rectangular shear panels. The rectangles are divided into two triangles, and each is represented by a triangular membrane. The model of Turner, et al[2], is degenerated to a shear panel by choosing only a non-zero shear modulus for the elastic coefficients. The Turner membrane representation is also selected for the skin elements. In this case, an isotropic material is defined. Stiffners fastening the spar and rib webs to the skins and wing verticals are represented by rod elements.



Ribs



Skin Surfaces



Spars

Figure 25. Swept Wing Geometry

Table XVI

Properties of Structural Elements

Element	Bar Area in^2	Panel Thickness in	Modulus, E $psi \cdot 10^{-6}$	Modulus $psi \cdot 10^{-6}$	Poisson's Ratio
Spars A,E Caps	0.0652	---	10.525	---	---
Spars B,C,D Caps	0.0466	---	10.525	---	---
Ribs 1.4 Caps	0.0652	---	10.525	---	---
Ribs 2.3	0.0466	---	10.525	---	---
Spar & Rib Panels	---	0.059	9.814	3.774	0.3
Skin Panels	---	0.118	10.525	4.048	0.3

Table XVII

Loading Conditions - Swept End Fixed

Loading Condition	Joint	Direction	Magnitude
1	A-1	-Z	1103.75
	E-1	-Z	1103.75
2	A-1	-Z	2207.50
3	A-1	≠Z	1.00
4	E-1	≠Z	1.00
5	A-3	≠Z	1.00
6	E-3	≠Z	1.00

Loading Conditions - Long Edge Fixed

Loading Condition	Joint	Direction	Magnitude
1	A-4	-Z	1103.75
	A-1	-Z	1103.75
2	A-4	-Z	2207.50
3	A-4	≠Z	1.00
4	A-1	≠Z	1.00
5	D-4	≠Z	1.00
6	D-1	≠Z	1.00

It is noted that in a case of the displacement method analysis the idealization involves no "lumping" of material. Panel thicknesses in the skin and webs are taken directly from the geometry of the basic structure. Cross sections of the rods are selected to match the geometry of the model structure.

Figure 26 depicts the displacement method network and idealization selected for the analysis. The idealization consists of 130 joints. At each joint displacement in three orthogonal directions are admitted, thus providing for a maximum of 390 degrees of freedom. The idealized wing is comprised of 516 elements: 182 shear panels, 192 skin panels and 172 rods. The whole wing structure is considered in the analysis.

Not only was the box analyzed under two different displacement boundary conditions, but two different analyses were run for each boundary condition. The first analysis consisted of solution of the problem, using a 27 bit mantissa to represent numbers. Calculations were performed on the IBM 7094. The second displacement analysis for each boundary condition was run with a mantissa of 23 bits. This size mantissa was selected to simulate the IBM 360 Fortran IV arithmetic. Only 23 bits were used to adjust for the hexi-decimal normalization of the IBM 360.

Figures 27 and 28 show relative error contours for the two boundary conditions. These data are developed by dividing the difference between the 23 and 27 bit solutions by the 27 bit answer. This provides a measure of the relative manipulation error. Assuming the manipulation error is proportional to 2^{-p}, the measure is 15 times the error in the 27 bit solution.

For both boundary conditions, all normal deflection predictions using the 23 bit mantissa were greater than for the 27 bit solution. Relative errors were minimum where deflections were greatest and maximum where they were least. The ratio of maximum to minimum relative error was about ten in each condition.

Maximum error for the 23 bit solution is .73% for the swept wing and .062% for the unswept box. This result is found by scaling the errors of Figures 27 and 28 by 16/15. Using equation (3-40) it is determined that this error must be less than five percent if 68 bits are used in the mantissa. The formula is satisfied, but very pessimistic. Equation (2-15) provides a better bound on error. The table below shows the calculation data and the predicted error bound and actual maximum relative errors.

Structure	No. Eqs.	$w_{AVG.}$	No. Calculations	e_{max} (2-15)	e_{max} (Actual)
Swept Wing	360	34.0	453,800	1.06%	0.73%
Unswept Box	300	28.1	261,100	0.61%	0.062%

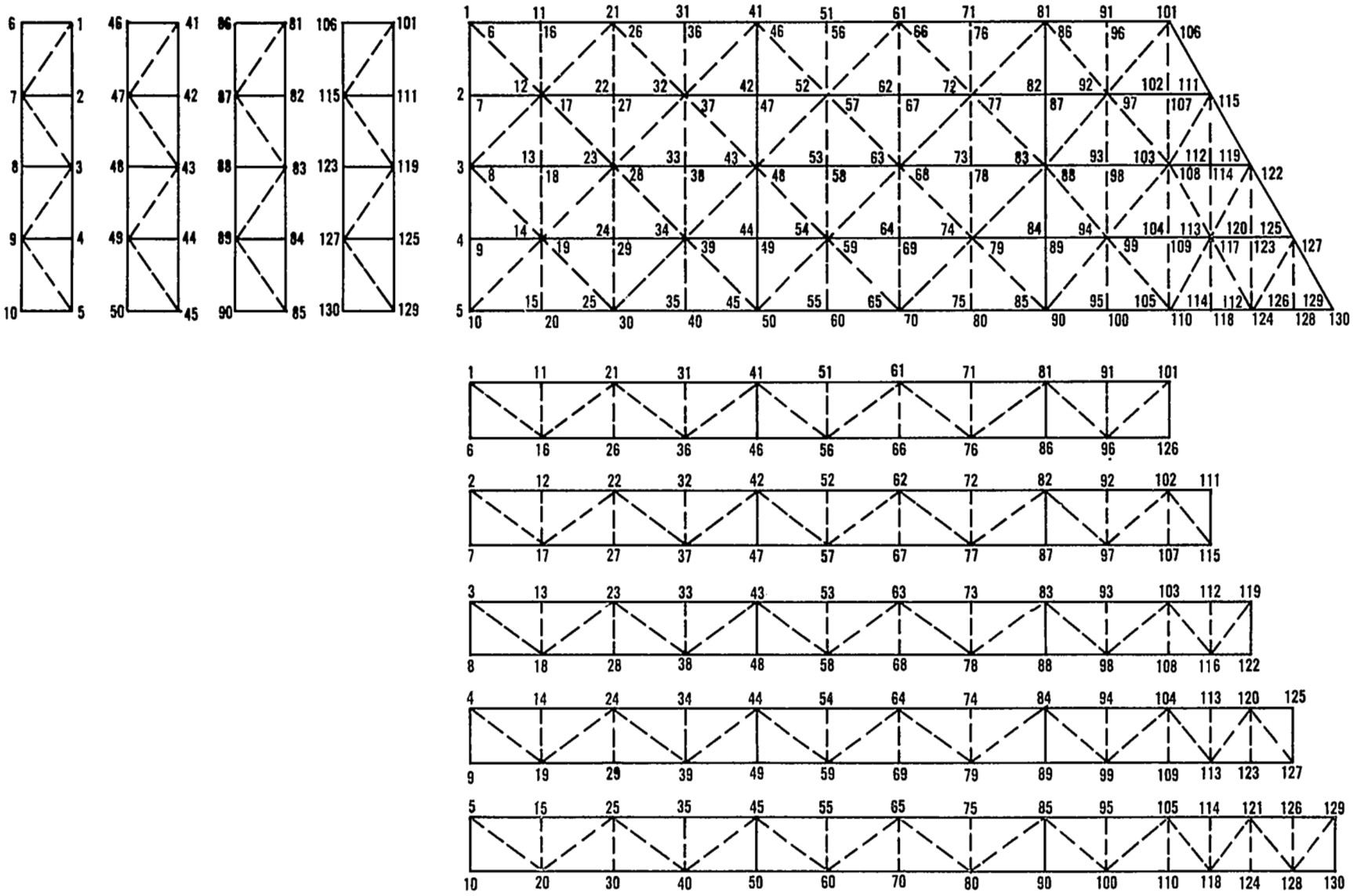


Figure 26. Displacement Method Idealization

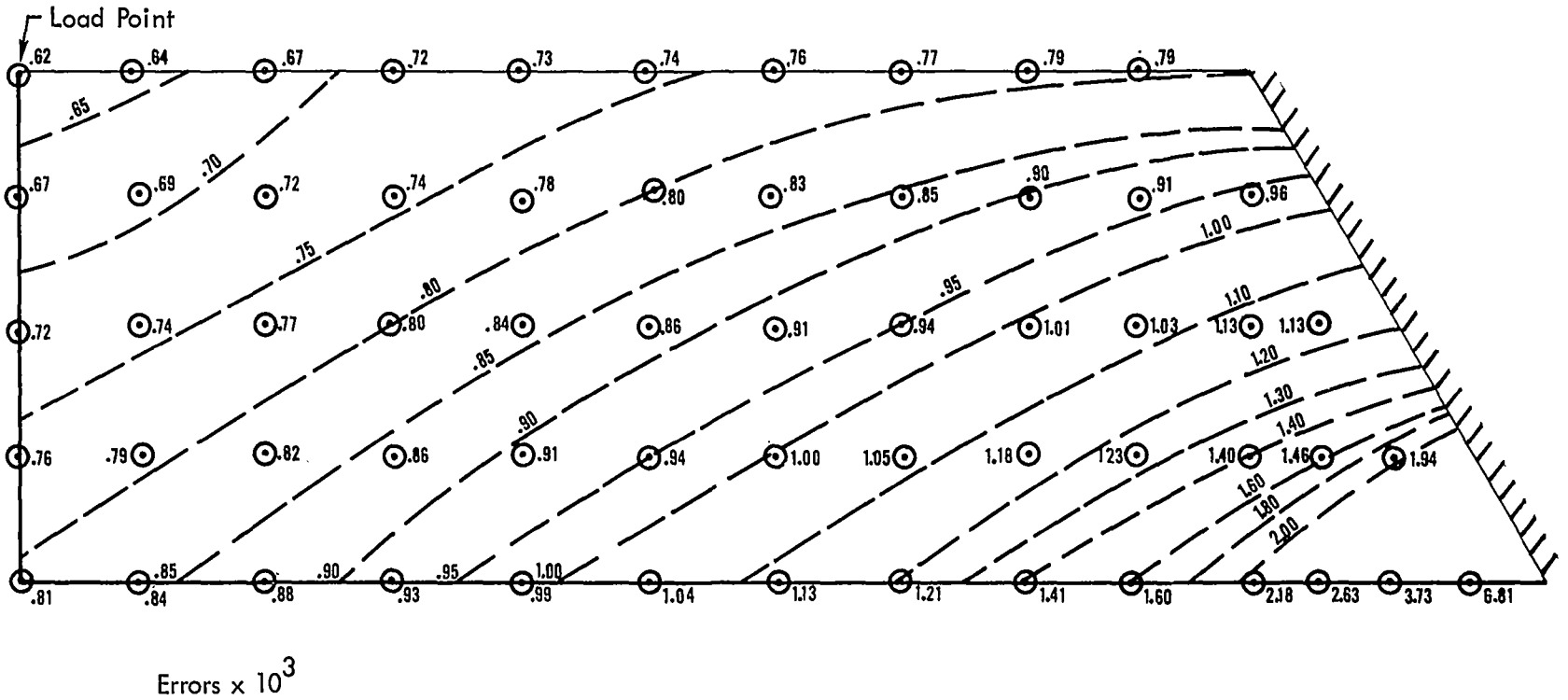


Figure 27. Relative Error Contours, Swept Wing

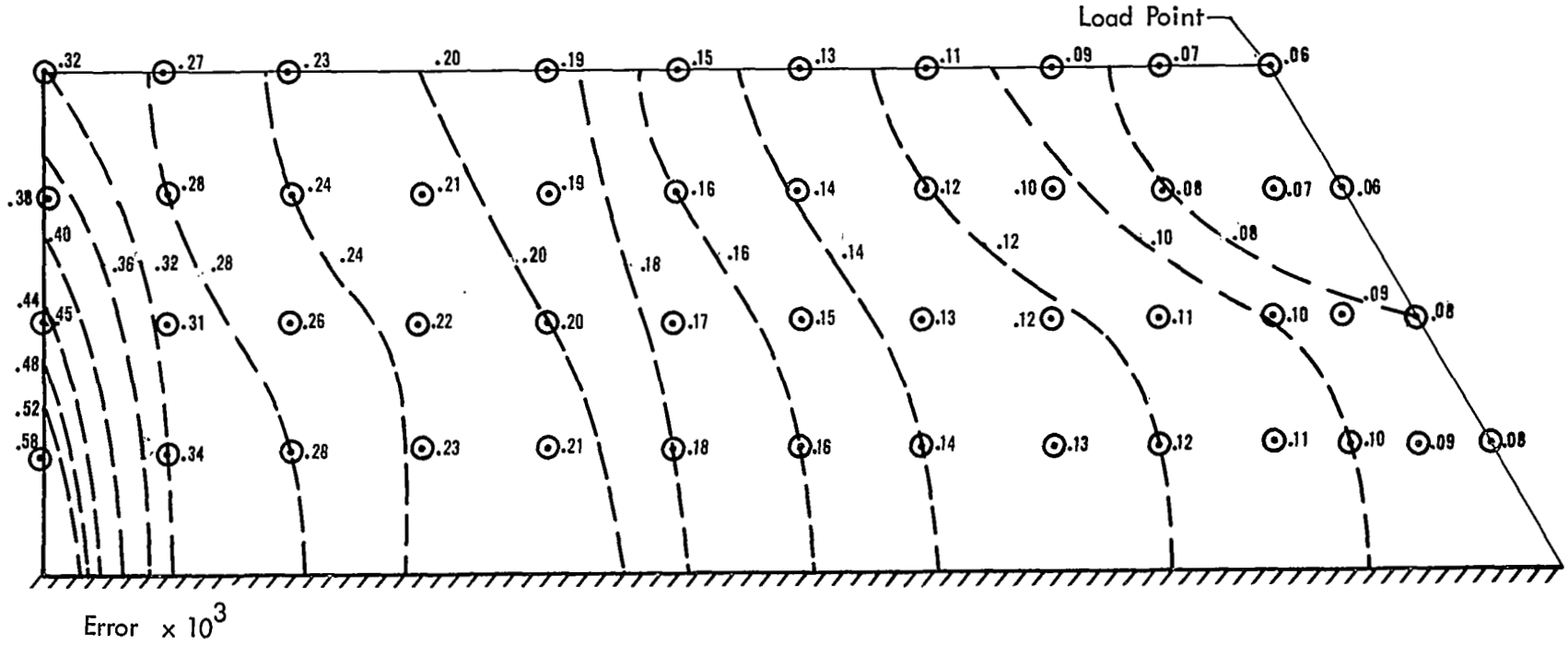


Figure 28. Relative Error Contours, Unswept Box

Table XVIII lists the influence coefficients for unit loadings described by loads 3, 4, 5, and 6 in Table XXI for each boundary condition. These data are based on a 23 bit mantissa and involve only loads and displacement normal to the box. The underlined elements in this table have the biggest error in satisfying Maxwell's reciprocity theorem. In both the swept wing and unswept box, these terms relate the tip influence to a joint adjacent to the root of the cantilever.

The table below shows the reciprocity relative error and the manipulation errors for these coefficients. The reciprocity relative error was calculated by dividing the error by the 27 bit mantissa answer. This answer satisfied reciprocity to six parts and the seventh digit. The relative manipulation error was obtained by dividing the difference in the influence coefficient for the 23 and 27 bit solutions by the 27 bit result. This table shows that with higher manipulation error, the error in satisfying Maxwell's theorem is greater. However, they also show that the symmetry error is a poor measure of manipulation error. It is noted that the reciprocity check for the unswept box would be expected to have larger error than that of the swept wing because the coefficient is of small magnitude.

Reciprocity and Manipulation Error

(23 bit error/27 bit result)

	Swept Wing	Unswept Box
Reciprocity Relative Error	$.191 \times 10^{-5}$	2.37×10^{-5}
Relative Manipulation Error	84.4×10^{-5}	$383. \times 10^{-5}$

Table XVIII

Influence Coefficients (p=23)*

Swept Wing

Load At:

Deflected Joint	A1	E1	A3	E3
A1	$.25830959^{-3}$	$.19804210^{-3}$	$.24087872^{-4}$	$.17200505^{-4}$
E1	$.19804193^{-3}$	$.26603381^{-3}$	$.13673694^{-4}$	$.34577300^{-4}$
A3	$.24087919^{-4}$	$.13673729^{-4}$	$.14555457^{-4}$	$.24086030^{-5}$
E3	$.17200531^{-4}$	$.34577366^{-4}$	$.24086052^{-5}$	$.24834058^{-4}$

Unswept Box

Load At:

Deflected Joint	A4	A1	D4	D1
A4	$.30936913^{-4}$	$.22232770^{-5}$	$.66991787^{-5}$	$.32952533^{-6}$
A1	$.22232679^{-5}$	$.41814419^{-4}$	$.40422528^{-7}$	$.92233677^{-5}$
D4	$.66991779^{-5}$	$.40421568^{-7}$	$.92155460^{-5}$	$.68137396^{-7}$
D1	$.32952078^{-6}$	$.92233622^{-5}$	$-.68138106^{-7}$	$.94475617^{-5}$

* Exponents in table imply a multiplier of 10 with that exponent; 3.g. $.32^{-4}$ means $.32 \times 10^{-4}$.

Force Method Analysis. - The same four analyses were run for the force method as for the displacement method. The "lumped parameter" approach was taken in the idealization. This resulted in bars under axial restraint whose areas were a summation of true frame members and stringers plus portions of adjacent skin or web material. The skin and webs were represented by panels which (in the idealization) could carry shear only. If follows then, that the bar areas used in the force method are greater than those of the displacement method and that more bars exist.

Figures 29, 30, and 31 show the basic components used for the idealization. These are 130 joints (nodes), 293 bars, and 168 panels.

It becomes evident that the computation of percent relative error $(100(F_{27} - F_{23})/F_{27})$ is not relevant for answers close to zero since relatively small answers are unreliable. Since answers are reported as nodes only, the possibility of answers near zero increases as the structural idealization is made finer. For the structural examples used in this report the grid is coarse enough so that no extremes in percent relative error encountered but there is a definite trend toward larger percent relative error for the smaller values of forces.

For the stiffness method of solution, the primary unknowns are displacements and these were used in showing the error characteristics. For the flexibility method, the primary unknowns are forces and therefore errors related to force answers are reported. Those chosen are the bar axial stresses (in the upper surface) which are effectively spar caps. For the swept wing these run in the long direction of the structure and for the unswept box, in the short direction.

The behavior of the error characteristics can best be shown by plotting the difference in answers $(F_{27} - F_{23})$ from the 27 bit solution against the 27 bit answers as shown in Figure 32 and 33. Note that the abcissa are differences in answers and not percent relative error. For the swept wing (Figure 32) it can be seen that the answers for spars A, B, and C show relatively small differences (absolute values are used) regardless of the force magnitudes, whereas for spars D and E the differences are relatively large and increase roughly proportional to the force magnitudes. In attempting to account for the existence of this phenomena it was discovered that spars D and E were retained as part of the statically determinate structures whereas spars A, B, and C were cut, i.e., they were redundant.

For the unswept wing the relative error associated with the largest force is .004% and that associated with the smallest force is .164% for the 23 bit solution. These data are obtained by scaling difference errors by 16/15. Averaging the five values at rib 2 (approximately mid-length) yields .0951% relative error.

For the unswept box (Figure 33) the variation of differences between 27 and 23 bit answers taken from root to free end vary roughly proportional to the force magnitudes. The slope, which is a rough indication of relative

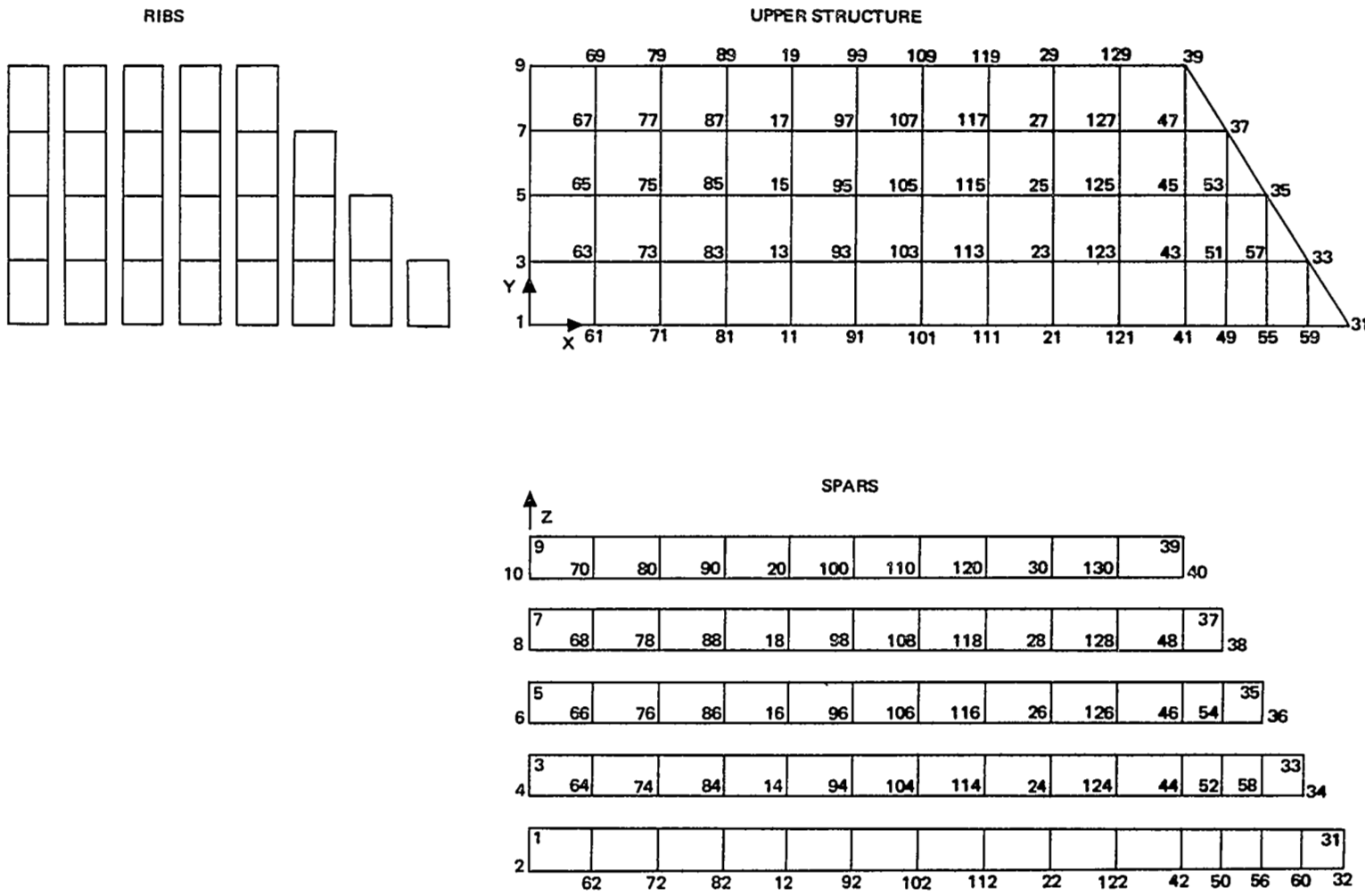


Figure 29. Joint Identification Numbers (Idealized Structure)

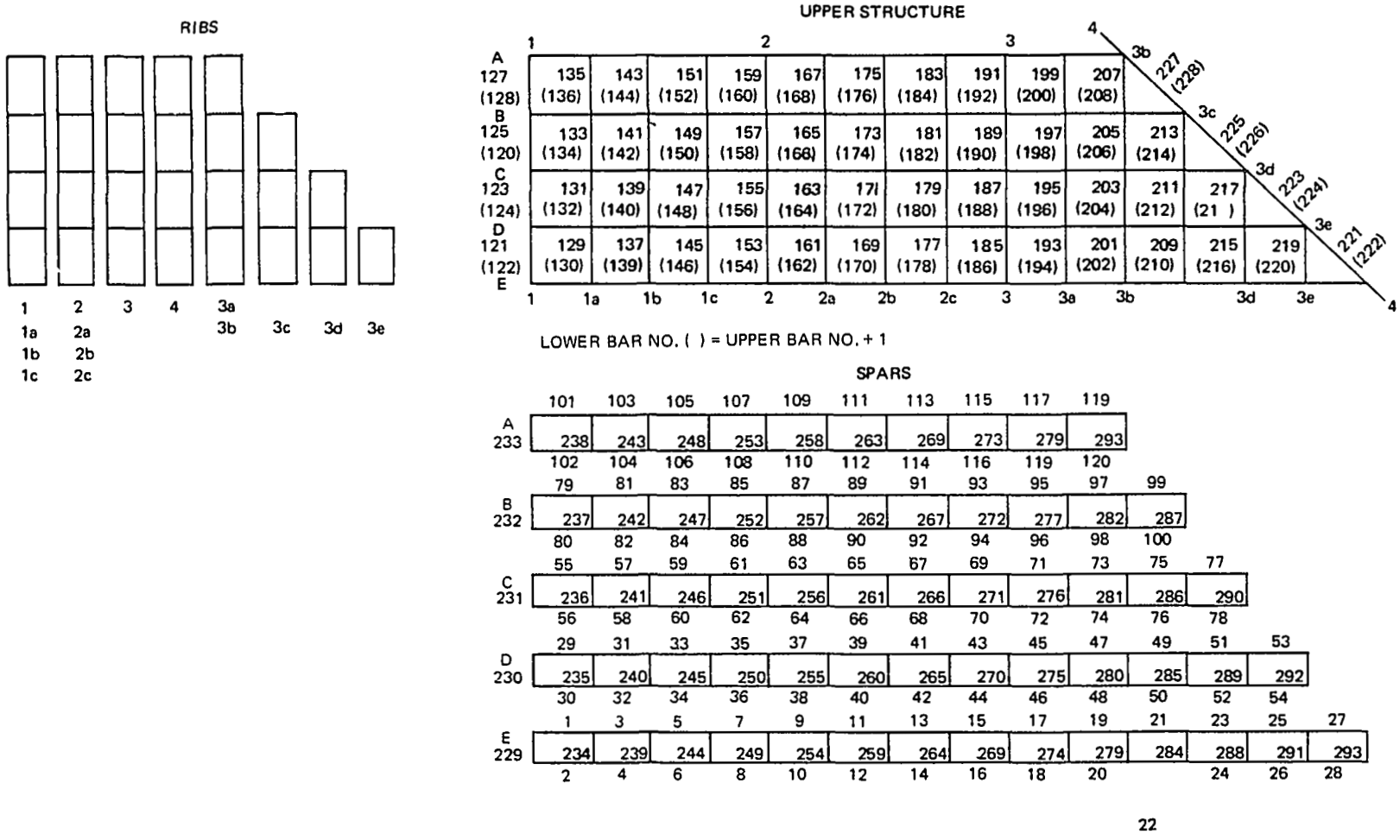


Figure 30. Bar Identification Numbers (Idealized Structure)

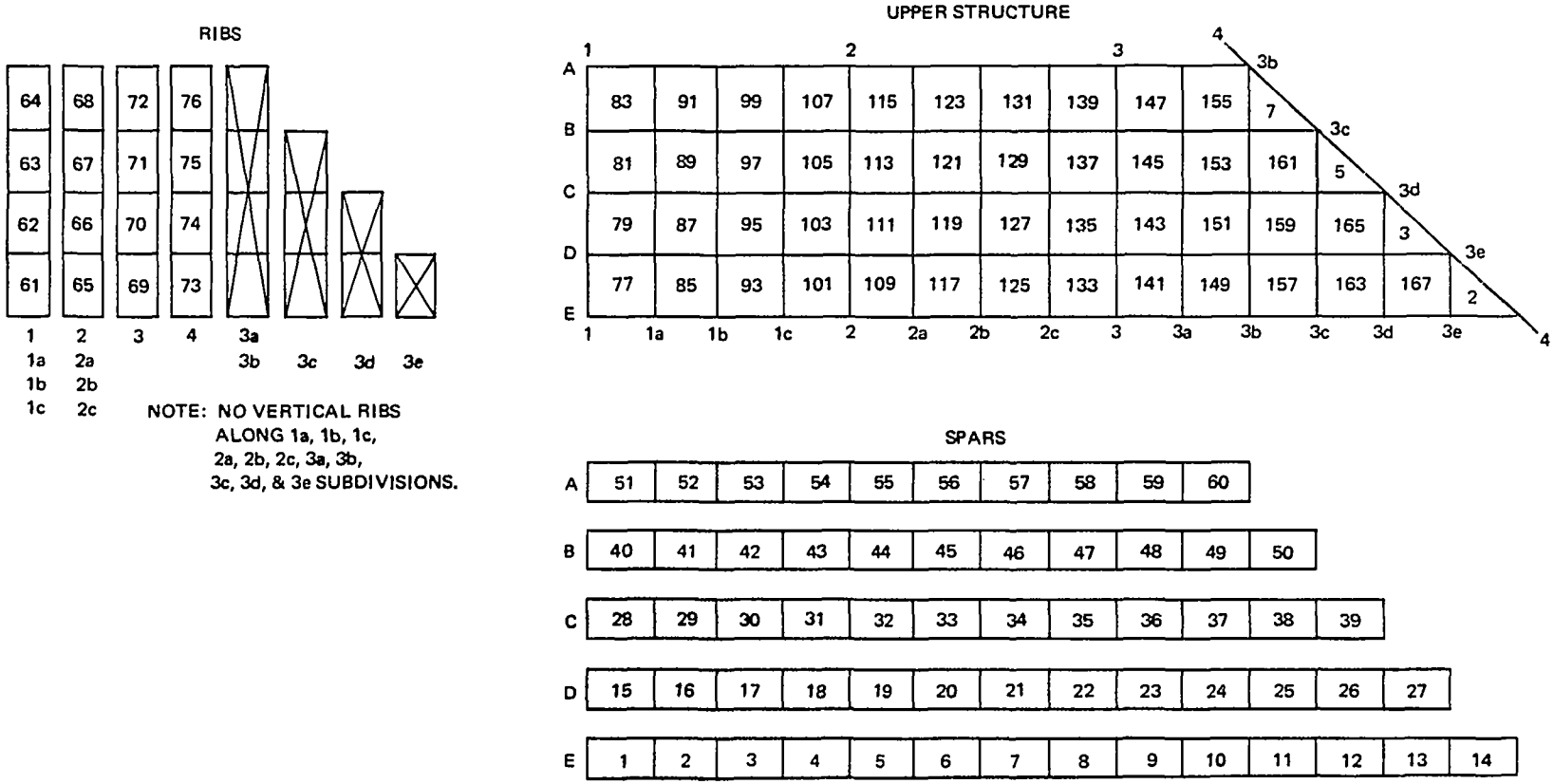


Figure 31. Panel Identification Numbers (Idealized Structure)

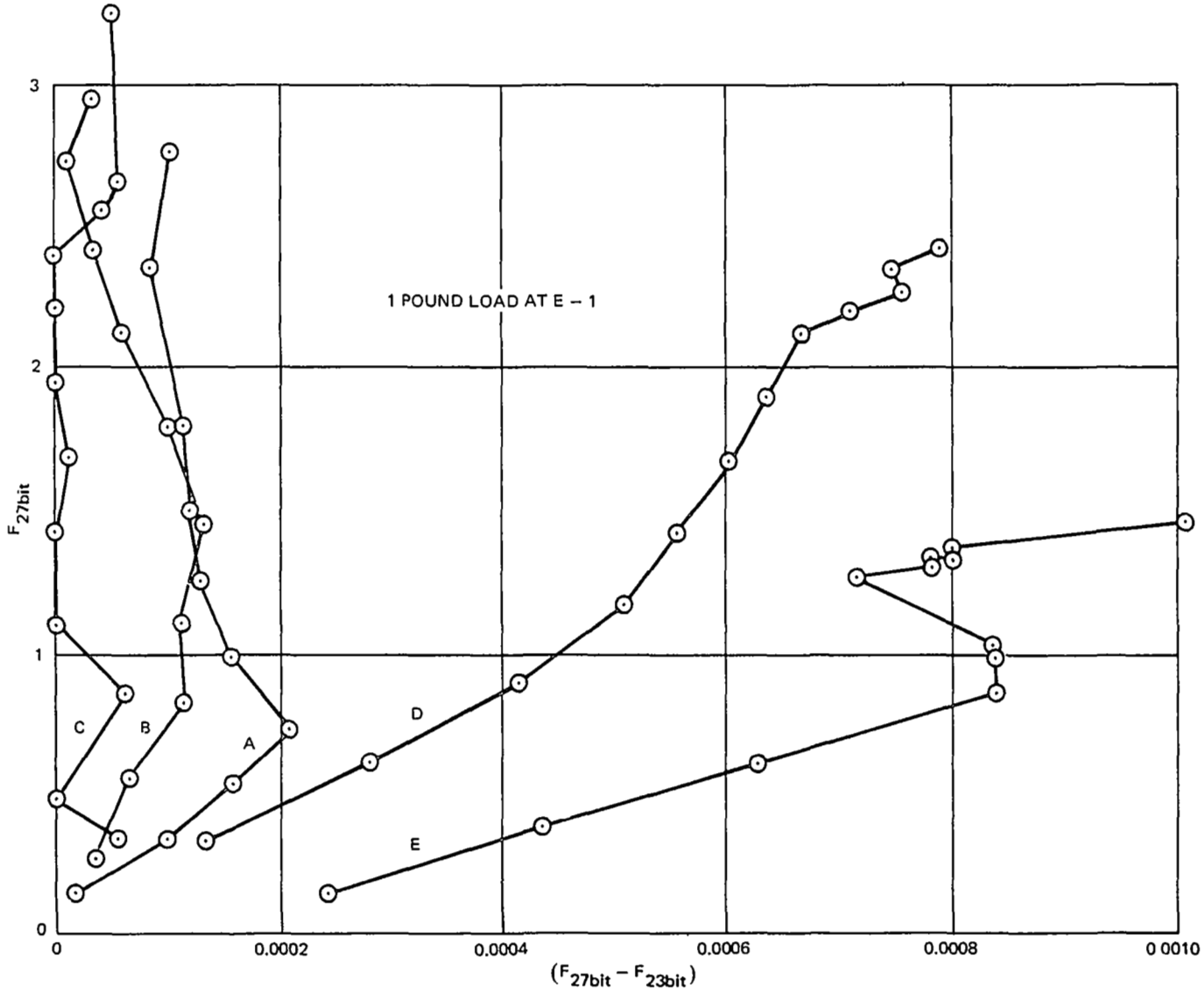


Figure 32. Variation of Deviation with Bar Forces (Swept Wing)

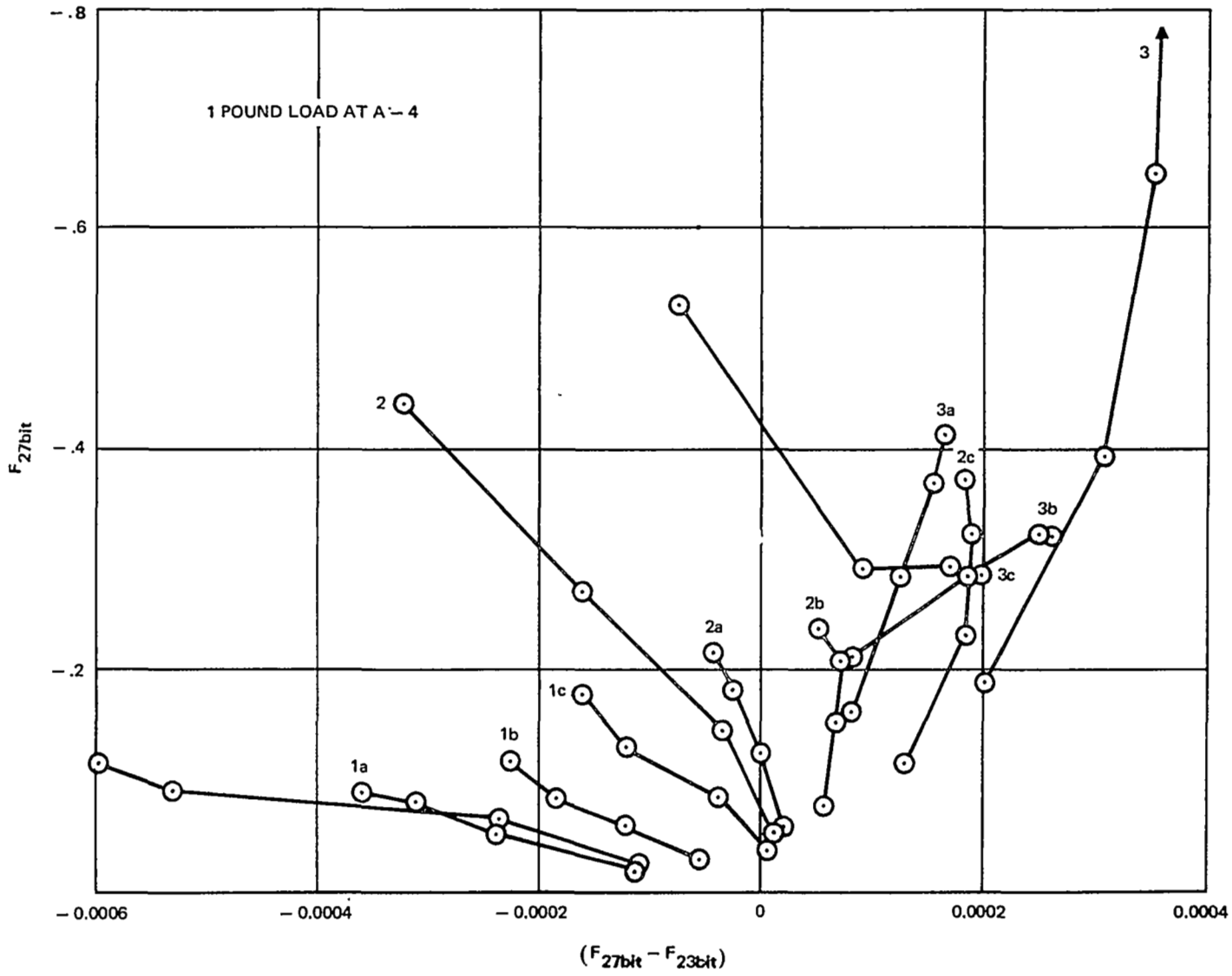


Figure 33. Variation of Deviation with Bar Forces (Unswept Box)

error, is different for each line. For the 23 bit solution, the relative error associated with the largest force is .068%, with the smallest force, and the mean for rib C is .137%.

The table below summarizes the problem characteristics and predictions of upper bound error based on equation (2-15). These data indicate that error bounds are very conservative for the force method.

Structure	Eqs.	Unknowns	Redund.	Calculations	e_{max}(2-15)	e_{max}(actual)
Swept Wing	390	491	101	1.71×10^6	4.09%	.164%
Unswept Box	390	551	161	3.58×10^6	8.38%	.606%

Table XIX lists the influence coefficients (23 bit solution using the flexibility method) for unit loads 3, 4, 5, and 6 in Table XVII for both structures. The underlined elements show the largest deviation from satisfying maxwell's reciprocity theorem.

The table below compares the maximum reciprocity relative error and the manipulation relative error for the same items. Fair correlation exists. Comparing these values with those of the stiffness method shows larger errors for the flexibility method. This may be attributed to the fact that deflections are secondary answers in the flexibility method. This is corroborated by the increase indicated for manipulation error over the values cited above for force manipulation error.

Reciprocity and Manipulation Error

(23 bit error/27 bit solution

Relative Reciprocity Error	175.3×10^{-5}	$12720. \times 10^{-5}$
Relative Manipulation Error	125.4×10^{-5}	$5407x9 \times 10^{-5}$

Table XIX

Influence Coefficients

Flexibility Method (23 Bit Solution)

Swept Wing

Load At

Deflection Joint	A1	E1	A3	E3
A1	$.20341133^{-3}$	$.16268703^{-3}$	$.16619749^{-4}$	$.15458976^{-4}$
E1	$.16270924^{-3}$	$.22122845^{-3}$	$.97645824^{-4}$	$.30316180^{-4}$
A3	$.16617989^{-4}$	$.97806405^{-5}$	$.918121294^{-5}$	$.12552637^{-5}$
E3	$.15465255^{-4}$	$.30263036^{-4}$	$.12534983^{-5}$	$.18139210^{-4}$

Unswept Box

Load At

Deflection Joint	A4	A1	D4	D1
A4	$.25654212^{-4}$	$.13474491^{-5}$	$.46184286^{-5}$	$.10113581^{-7}$
A1	$.13626413^{-5}$	$.31488911^{-4}$	$-.24631481^{-6}$	$.54631855^{-5}$
D4	$.46403147^{-5}$	$-.24606016^{-6}$	$.74720010^{-5}$	$-.11576549^{-6}$
D1	$.11399606^{-7}$	$.54632010^{-5}$	$-.11489374^{-6}$	$.59473158^{-5}$

Comparison of Displacement and Force Method Errors

The table below furnishes a comparison of measured and predicted bounds for manipulation errors in evaluating the primary unknowns. Errors are for the 23 bit analysis. These data show that the displacement method incurs smaller errors for the unswept box and the force method smaller for the swept wing. Since errors are less than bound predictions, critical arithmetic was not an important error source in either analysis. Suppose the selection of analysis method is to be such that manipulation error is minimized. Then, these data suggest that the displacement method has smaller errors for structures whose subsystems act predominately in parallel (box) and force method with series subsystems (wing). Arithmetic in the force method is indicated to be better optimized since measured errors are a smaller fraction of the upper bound errors.

Analysis Relative Error

(Percent)

	Stiffness Method (Deflections)		Flexibility Method (Forces)	
	Wing	Box	Wing	Box
With largest answer	.066	.0664	.0014	.0682
With smallest answer	.73	.064	.164	.606
Mean for mid-rib (Rib 2 or C)	.091	.0195	.0951	.137
Error bound	1.06	.61	4.09	8.38

Section 6

CONCLUSIONS

Table XX summarizes and characterizes the errors examined for the displacement and force methods. Input, generation and output errors are similar for the two methods. Elimination errors examined are the same but are of differing importance in the two approaches. Note that the solution process chosen for displacement method permits separating attrition error into decomposition cumulative and substitution attrition error.

Table XXI summarizes some of the guidelines the analyst can use for minimizing error. This table is concerned with modification of the problem to descriptive data since this is at the analyst's disposal. Guidelines for input, output, and generation errors can normally be disregarded.

Particular criteria to fix required arithmetic precision in a structural analysis have been estimated for series systems in the displacement method and for parallel systems for the first method. These criteria appear in Sections 3 and 4. On the basis of the evidence presented in Section 5, these criteria must be regarded as very conservative.

Based on the study described in the previous sections, the following conclusions are drawn:

(1) Elimination is the most important error source. Input errors, except for decimal fractions, can be interpreted by the analyst in terms of a modified structural model. These errors usually are negligible. Generations are small since relatively few calculations are required per coefficient in the structural equations. Lack of discrimination in the coordinant data is the largest single source of generation error can cause significant errors. Output errors are negligible unless as many digits are printed out as are contained in the computer representation of the number.

(2) Considering series systems as critical for the displacement method and parallel systems as critical for the force method, the following characteristics were observed for elimination errors:

(a) In both methods, the solution can be invalidated due to numerical singularity, unstable propagation of manipulation errors, cumulative triangularization (decomposition) attrition errors, and attrition errors in the substitution processes.

(b) While singularity errors are important in the displacement method, they are relatively unimportant for the force method.

(c) Attrition errors are important for the displacement method and the force method. Cumulative attrition errors are important for

the displacement method, particularly for systems of equations of higher than first order. Substitution errors are not. Because of the large number of calculations involved in the force method manipulations, attrition errors are important for the force method diagonalization.

(3) As presently practiced, the force method intrinsically has lower manipulation error than the displacement method. In the force method, redundants are often selected automatically and equations sequenced in an attempt to minimize manipulation errors. The displacement method has no equivalent operation. The sequencing of equations is entirely at the disposal of the engineering analyst. The force method uses the Gauss-Jordan reduction process which is more accurate than the Choleski process used in some displacement analyses. The force method uses few and simple element representations which involve low manipulation error while the displacement method may use a broad class of representations, some of which may incur large manipulation errors.

(4) With optimum error control, single precision arithmetic is regarded as adequate for the analysis of problems up to 5000 order. Lacking optimum control, higher precision arithmetic can be used. The worst case structure in the displacement method for single precision on the IBM 360 permits treatment of 60 series beam elements of 1300 series rod elements. Few practical structures being analyzed today have more than 60 elements in series, although very small sets of equations can be involved with hundreds of elements in series. Since the error is reduced by a factor of two for every added bit in the mantissa, the 48 bit word machines (Philco 212, Honeywell MH 800, CDC 3600, and Burroughs B5500) will involve negligible manipulation error except for pathological problems.

(5) Manipulation error bounds based on the number of calculations can provide a measure of error when critical arithmetic is avoided. Equation (2-15) gives fair error estimates for both the displacement and force analysis. It does not indicate the proper selection of analysis method, if manipulation error is to be minimized, because arithmetic is better optimized in the force method. These conclusions are based on the verification analyses. These analyses also confirm that structures composed predominately of parallel subsystems should be analyzed by the displacement method. Structures with most subsystems in series should be analyzed by the force method to minimize manipulation errors.

Table XX

Summary of Importance of Manipulation Errors

Errors	Page References	Displacement Method	Force Method
Input - Output	9	Negligible except for unusual cases	
Truncation	9, 11	Input, controlled by analyst; output, by coder	
Conversion	9, 11	Input, only significant for decimal fractions; output, only for last digit of number representation	
Generation	28, 84	Small, controllable, and easily measured	
Coordinate discrimination	30,84, 99	Important in comparative studies	
Transformation	27, 87	Negligible since few operations/coefficient	
Series Addition	15,34, 99	Negligible if small components added first	
Elimination		Large errors possible	
Unstable propagation	41,57, 83	Can be sensed and controlled so need not be an output	
Cumulative attrition	48, 60	Large and possibly unavoidable for high order and large sets of difference equations	
Numerical singularity	49,63, 69 87,90, 99	Small with optimum joint numbering	Usually very small
Substitution attrition	51,64, 81	Small	Small with optimum redundant selection
Secondary unknown	69,132	Small, but may involve critical arithmetic	Small

Table XXI

Analyst's Guidelines

Error	Displacement Method Guides	Force Method Guides
Input	Scale to minimize truncation error and decimal fraction input.	
Generation	Use local coordinates for structural elements.	
	Local origin of global system at centroid of structure.	
	Choose coordinate surfaces parallel to structural surfaces.	
	Number most flexible elements first.	
	Avoid adjacent incommensurate stiffnesses.	Avoid adjacent incommensurate flexibilities.
Elimination	Number joints from free edge.	Number stiff determinate substructure first starting at fixity.
	Number toward stiffer structure.	
	Avoid Choleski algorithm.	Avoid Choleski algorithm.
	Average stresses.	
Output	Disregard last converted digit if computer prints entire representation.	

REFERENCES

1. Levy, S., "Structural Analysis and Influence Coefficients for Delta Wings," Journal of the Aeronautical Sciences, Vol. 20, No. 7, pp. 449-454, July, 1953.

2. Turner, M. J., Clough, R. W., Martin, H. C., and Topp, L. J., "Stiffness and Deflection Analysis of Complex Structures,"Journal of the Aeronautical Sciences," Vol. 23, No. 9, pp. 805-823, September, 1956.

3. Von Neumann, J., and Goldstine,H. H., "Numerical Inverting of Matrices of High Order," Bull. Amgr. Math., Soc. Vol. 53, 1947, pp. 1021-1099.

4. Turing, A. M., "Rounding-Off Errors in Matrix Processes," Quart. J. Mech. and Physics, Vol. 1, September, 1948, pp. 287-308.

5. Wilkinson, J. H., "Error Analysis or Direct Methods of Matrix Inversion," ACM Journal, Vol. 8, No. 3, July, 1961, pp. 281-330.

6. Forsythe, G. E., "Today's Computational Methods of Linear Algebra," SIAM Review, Vol. 9, No. 3, July, 1967, pp. 489-515.

7. Rosanoff, R. A. and Ginsburg, T. A., "Matrix Error Analysis for Engineers," Proceedings of Conference on Matrix Methods in Structural Mechanics, Dayton, Ohio, December, 1965, pp. 887-910.

8. Rosanoff, R. A., and Radkowski, P. P., "Research Directions in Matrix Error Analysis," AIAA 5th Aerospace Sciences Meeting, New York, January, 1967, paper 67-142, 20 p.

9. Gatewood, B. E., and Ohanian, N., "Examples of Solution Accuracy in Certain Large Simultaneous Equation Systems," Conference on Matrix Methods in Structural Mechanics.

10. Shah, J. M., "Ill-Conditioned Stiffness Matrices," Journal Struct. Div., ASCE, Vol. 92, No. ST6, December 1966, pp. 443-457.

11. Goldberg, I. B., "27 Bits are not Enough for 8 Digit Accuracy," Communications of the ACM, Vol. 10, No. 2, February, 1967, pp. 105, 106.

12. Wilkinson, J. H., "Rounding Errors in Algebraic Processes," Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

13. Kosko, E., "The Equivalence of Force and Displacement Methods in Matrix Analysis of Elastic Structures," Proc. of Conf. on Matrix Methods in Structural Mechanics, AFFDL-TR-66-80, WPAFB, December, 1965.

14. Melosh, R. J., "Basis for Derivation of Matrices for the Direct Stiffness Matrix," AIAA Journal, Vol. 1, No. 7, July, 1963, pp. 1631-1637.

15. Argyris, J. J., "Continua and Discontinua," Proceedings of Conference on Matrix Methods in Structural Mechanics, Dayton, Ohio, December, 1965, pp. 11-190.

16. Utku, S. and Melosh, R. J., "Behavior of Triangular Shell Element Stiffness Matrices Associated with Polyhedral Deflection Distributions," paper No. 67-114, AIAA 5th Aerospace Sciences Meeting, New York, January 23-26, 1967, 18 p.

17. Hrennikoff, A., "Solution of Problems in Elasticity by the Framework Method," J.A.M., December 1941, pp A-169 - A-175.

18. Percy, J. H., Pian, T. H., Navaratna, D. R., and Klein, S., "Application of the Matrix Displacement Method to the Linear Elastic Analysis of Shells of Revolution," ASRL TR 121-7 MIT, Cambridge, Mass., Jan. 1965, 73 p.

19. Argyris, J. H., "Energy Theorems and Structural Analysis," Part 1, General Theory, Aircraft Engineering, Vol. XXVII, April. 1955, pp 125-133.

20. Melosh, R. J., "Structural Analysis of Solids," Jour. Struct. Div., ASCE, Vol. 89, No. ST4, Aug. 1963, 1p 205-224.

21. Bogner, F. K., Fox, R. L. and Schmit, L. A., "The Generation of Inter-element - Compatible Stiffness and Mass Matrices by the Use of Interpolation Formulas," presented at 2nd Conference on Matrix Methods in Structural Mechanics, WPAFB, Oct. 26-28, 1965, 61 p.

22. Anderson, J. M. and Christiansen, H. N., "Behavior of the Finite Element Stiffness Method for Nearly Incompressible Materials," presented at meeting of Interagency Committee on Solid Rocket Propulsion, JPL, Pasadena, Calif., Dec. 5-6, 1967.

23. Turner, M. J., Martin, H. C., and Weikel, R. C., "Further Development and Applications of the Stiffness Method", presented at AGARD Structures and Materials Panel, Paris, France, July 1962, 120 p.

24. Utku, S., "Computation of Stresses in Triangular Finite Elements," JPL, TR No. 32-948, Pasadena, Calif., June 1966, 29 p.

25. Dwyer, P. S., Linear Computations, John Wiley & Sons, Inc., 1951

26. Morris, R. C., "Format II - Second Version of Fortran Matrix Abstraction Technique, Technical Report AFFDL-TR-66-207, Volume III," December 1966.

27. Przemieniecki, J. S., Theory of Matrix Structural Analysis, McGraw-Hill, 1968.

28. Filho, F. V., "Orthogonalization of Internal Force and Strain Systems," Conference on Matrix Methods in Structural Mechanics," Dayton, Ohio, December 1965, pp. 385-391.

29. Argyris, J. H., and Kelsey, S., Modern Fuselage Analysis and the Elastic Aircraft, Butterworths, London, p. 101, 1963

30. Gillis, P., and Gerstle, K. H., "Analysis of Structures by Combining Redundants," ASCE Structural Division Journal, ASCE, Volume 87, No. STI, pp. 41-56, January 1961.

31. Wilkinson, J. H., The Algebraic Eigenvalue Problem, Clarendon Press, 1965

32. Wilkinson, J. H., The Solution of Ill-Conditioned Linear Equations, Mathematical Methods for Digital Computers, pp. 65-93, Wiley, 1961.

33. Bauer, F. L., "Optimally Scaled Matrices," Numer. Math. 5, 1963, pp. 73-87

34. Chan, H. M., "Matrix Errors Analysis and an Estimate of Solution Accuracy of the Force Method in Finite Element Structural Analysis," Douglas Report No. DAC 58763, December 15, 1967.