

LEVEL II

(12)

ARO Report 81-2

AD A101442

**PROCEEDINGS OF THE TWENTY-SIXTH
CONFERENCE ON THE DESIGN OF
EXPERIMENTS IN ARMY RESEARCH
DEVELOPMENT AND TESTING**



Approved for public release; distribution unlimited.
The findings in this report are not to be construed
as an official Department of the Army position, un-
less so designated by other authorized documents.

June 1981

DTIC
ELECT
S JUL 16 1981
D
A

Sponsored by

The Army Mathematics Steering Committee
on Behalf of

DTIC FILE COPY

THE CHIEF OF RESEARCH, DEVELOPMENT AND ACQUISITION

81 7 16 099

U. S. Army Research Office

Report No. 81-2

(11) Jun 1981

(12) 6051

(6) PROCEEDINGS OF THE TWENTY-SIXTH CONFERENCE

ON THE DESIGN OF EXPERIMENTS in Armament
Research, Development and Test.

(14) ARS-81-2

Sponsored by the Army Mathematics Steering Committee

HOST

U. S. Army White Sands Missile Range
White Sands Missile Range, N. M.

HELD AT

New Mexico State University, Las Cruces, N. M.

22-24 October 1980.

Approved for public release; distribution unlimited.
The findings in this report are not to be construed
as an official Department of the Army position, un-
less so designated by other authorized documents.

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

The Twenty-Sixth Conference on the Design of Experiments (DOE) in Army Research, Development and Testing was held 22-24 October 1980 and had as its host the U. S. Army White Sands Missile Range (WSMR). Dr. Richard H. Duncan, Technical Director and Chief Scientist, WSMR, made many of the initial plans for this meeting. In particular, he contacted Dr. Harold Law, Associate Academic Vice-President, New Mexico State University and made arrangements with him to hold the conference at his university. The Army Mathematics Steering Committee (AMSC), sponsor of these conferences, would like to thank New Mexico State University for providing such excellent facilities for this meeting.

Dr. Duncan asked Ms. Peggy Hoffer, of the WSMR Plans Office, to serve as chairman on Local Arrangements for this conference; and Mr. Robert Green, of the Instrumentation Directorate, to handle any technical problems associated with the program. These individuals, together with many other members of the WSMR, helped make this, the 26th Conference, a very successful and interesting meeting.

The Subcommittee for Probability and Statistics, chaired by Dr. Douglas B. Tang, is responsible to the AMSC for conducting these Army conferences. Dr. Tang asked Dr. Frank E. Grubbs to be Chairman of the Program Committee for the 1980 conference. One of the first acts of this committee was to select "Data Analysis" as the theme of this meeting. This was a wise choice because of the large amount of analytical and statistical work in testing and modeling performed within the many agencies located on the base of the host installation. At the first meeting of the Program Committee, the following national known scientists were selected as the invited speakers for this year's conference.

<u>Speaker and Affiliation</u>	<u>Title of Address</u>
Professor Francis J. Anscombe Yale University	How Far to go in Looking at Data
Dr. Toby J. Mitchell Union Carbide Nuclear Division	Design of Experiments
Professor W. J. Conover Texas Tech University	The Rank Transformation as a Robust and Powerful Tool for the Analysis of Experimental Data
Professors James R. Thompson, Chih-Chy Fwu, and Richard A. Tapia Rice University	The Nonparametric Estimation of Probability Densities in Ballistic Research
Professor Victor Solo Harvard University	Engineering Time Series Analysis
Professor Richard A. Johnson University of Wisconsin	Stress-Strength Models for Reliability-Overview and Recent Advances

Professor Badrig Kurkjian, Professor of Statistics at the University of Alabama, is at the present time, chairman of the committee to select the recipient of the Samuel S. Wilks Memorial Medal. On 19 June 1980 he advised Dr. Robert Launer, secretary of the Design of Experiments Conference, that Dr. W. Allen Wallis, Chancellor and Professor of Statistics and Economics at the University of Rochester, had been selected as the 1980 Wilks Medalist. This distinguished scientist richly deserves this honor for his contributions to applied statistics.

On 20-21 October 1980, just preceding the start of the DOE conference, a special tutorial on Applied Regression Analysis was held. This tutorial was designed for engineers, scientists and statisticians who are involved in analyzing least squares data, the associated statistical inferences and model building. The instructor for this informative course was Professor Norman Draper, Department of Statistics, University of Wisconsin and the Mathematics Research Center.

The AMSC has asked that these proceedings be distributed Army-wide to enable those who could not attend this conference, as well as those that were present, to profit from some of the scientific ideas presented by the speakers. The members of the AMSC would like to take this occasion to thank all the speakers for their interesting presentations and also the members of the Program Committee for their many contributions to this scientific meeting.

Program Committee

Carl Bates
George E. P. Box

Robert Launer (Secretary)
Douglas Tang (Chairman,
Prob. & Stat. Subcommittee)
Malcolm Taylor
Langhorne Withers

Larry Crow
Walter Foster

Frank E. Grubbs (Program Committee Chairman)

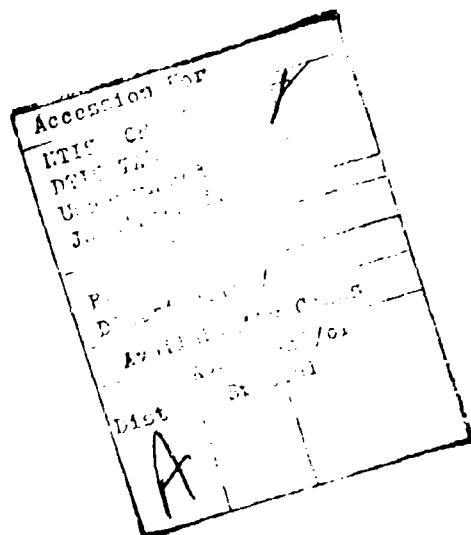


TABLE OF CONTENTS*

TITLE	PAGE
Foreword	iii
Table of Contents	v
Program	ix
HOW FAR TO GO IN LOOKING AT DATA	
F. J. Anscombe	1
THE USE OF RIDGE REGRESSION IN TRAJECTORY ESTIMATION	
William S. Agee and Robert H. Turner	11
AN EIGHT VARIABLE COMPOSIT DESIGN FOR FITTING A SECOND ORDER RESPONSE SURFACE	
Carl B. Bates	25
AN APPLICATION OF ORDER STATISTICS TO TIME-SEQUENCE LOGIC	
William E. Baker and Malcolm S. Taylor	41
RADAR ERROR SIGNAL IMPROVEMENT	
Robert E. Green	55
USE OF THE BILINEAR Z-TRANSFORM IN IMPLEMENTING DIGITAL FILTERS	
Donald W. Rankin	63
INFERENCE PROCEDURES FOR DETERMINING LIFE TIME ESTIMATES OF ADVANCED MATERIALS	
Donald Neal, Edward M. Lenoe and Donal Mason	85
RISKS TO NEIGHBORING FACILITIES	
Paul C. Cox	99
THE 1980 SAMUEL S. WILKS MEMORIAL MEDAL	
Frank E. Grubbs	111
OPTIMAL ESTIMATION TECHNIQUES FOR FORECASTING PROPAGATION PARAMETERS	
K. E. Kunkel and D. L. Walters	121
SOME RESULTS FOR THE UNIVARIATE NORMAL RANDOM LINEAR REGRESSION MODEL PREDICTION THEORY	
D. G. Kabe	125

*This table of contents contains only the papers that are published in this technical manual. For a list of all papers presented at the Twenty-Sixth Conference on the Design of Experiments, see the Program of the meeting.

TITLE	PAGE
ADAPTIVE MEDIAN SMOOTHING William S. Agee and Jose E. Gomez	137
FITTING AN ELLIPSE Donald L. Buttz	165
METHODS FOR APPROXIMATING MATHEMATICAL FUNCTIONS Donald W. Rankin	181
MOS TRAINING COURSE SELECTION CRITERIA: AN APPLICATION OF DISCRIMINANT ANALYSIS Pat Cassady and Lounell Snodgrass	235
THE ARMOR COMBAT FOR MODEL SUPPORT (ARCOMS) FIELD EXPERIMENT Roger F. Willis	245
EXTREME VALUE QUANTILE RESPONSE EXPERIMENTAL DESIGN Jill H. Smith and Jerry Thomas	253
ALTERNATIVE QUANTILE ESTIMATION W. D. Kaigh	287
THE RANK TRANSFORMATION AS A ROBUST AND POWERFUL TOOL FOR THE ANALYSIS OF EXPERIMENTAL DATA W. J. Conover	275
THE NONPARAMETRIC ESTIMATION OF PROBABILITY DENSITIES IN BALLISTICS RESEARCH Chih-chy Fwu, Richard A. Tapia and James R. Thompson	309
TESTABILITY OF LINEAR HYPOTHESES IN NORMAL LINEAR MODELS Gerald S. Rogers	327
THE POTENTIAL UTILITY OF CROSSING A FRACTIONAL FACTORIAL WITH A FULL FACTORIAL IN THE DESIGN OF FIELD TESTS Carl T. Russel	335
SOME REMARKS ON CROSSOVER EXPERIMENTS J. Robert Burge	347
A TIME SERIES ANALYSIS AND MODELING APPROACH OF SENSE AND DESTROY ARMOR (SADARM) RADIOMETRIC (ELECTROMAGNETIC RADIATION) NOISE DATA Richard T. Maruyama	371
THE ROLE OF SPATIAL BANDWIDTH LIMITS IN THE MEASUREMENT AND INTERPRETATION OF SECOND-ORDER STATISTICAL PROPERTIES E. L. Church	387
BOUNDS FOR OPTIMAL CONFIDENCE LIMITS FOR SERIES SYSTEMS Bernard Harris and Andrew P. Soms	411

TITLE	PAGE
OPTIMAL UPPER CONFIDENCE LIMITS FOR PRODUCTS OF POISSON PARAMETERS WITH APPLICATIONS TO THE INTERVAL ESTIMATION OF THE FAILURE PROBABILITY OF PARALLEL SYSTEMS Bernard Harris and Andrew P. Soms	431
RELIABILITY BASED SAFETY FACTORS FOR CONCRETE STRUCTURES SLIDING STABILITY Paul F. Mlakar	455
A UNIFIED AND UNBIASED ANALYSIS FOR DECISION MAKING IN PATERNITY DISPUTES Paul H. Thrasher	465
AN ALGORITHM FOR TRILATERATION James T. Hall	497
SOCIAL SCIENTIST TECHNIQUE: THE CATALYST FOR OBTAINING OBJECTIVITY FROM SUBJECTIVITY Ronald L. Johnson	515
SOME ASPECTS OF ENGINEERING TIME SERIES ANALYSIS Victor Solo	521
STRESS-STRENGTH MODELS FOR RELIABILITY: OVERVIEW AND RECENT ADVANCES G. K. Bhattacharyya and Richard A. Johnson	531
ON THE INTERPOLATION OF GRAVITY ANOMALIES AND DEFLECTIONS OF THE VERTICAL IN MOUNTAINOUS TERRAIN H. Baussus von Luetzow	549
A SEQUENTIAL k-GROUP RANDOM ALLOCATION METHOD WITH APPLICATIONS TO SIMULATION Andrew P. Soms	565
REGISTERED ATTENDEES	575

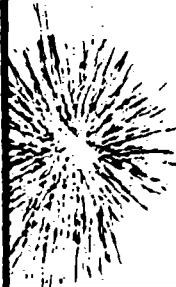
**1980
Twenty-Sixth Conference
on
the Design of Experiments
in
Army Research, Development,
and Testing**



A black and white illustration of a missile launching from a mobile launcher. The missile is shown at an angle, ascending into a wavy sky. To the left of the launcher, there is a stylized drawing of a desert landscape with mesas and a cactus.

**SPONSORED BY
THE ARMY MATHEMATICS STEERING COMMITTEE**

**host
U.S. ARMY WHITE SANDS MISSILE RANGE**



A black and white illustration of a large, multi-pointed explosion or firework display against a dark background.

**HELD AT
NEW MEXICO STATE UNIVERSITY, LAS CRUCES, N.M.**

22-24 OCTOBER 1980

Dear Conference Participant:

It is our pleasure to have you attend the Twenty-Sixth Conference on the Design of Experiments in Army Research, Development, and Testing. We hope this conference will provide new and relevant information which will be useful to you in your future endeavors.

During your stay in Las Cruces we also hope you will take the opportunity to visit the rest of the New Mexico State University campus and White Sands Missile Range.

New Mexico State University, founded in 1888, has grown into a major institution of higher education. In its development, the University has preserved many of the traditions of its land-grant origin while moving toward increased emphasis of the humanities, liberal arts, and natural sciences. The mission of New Mexico State University is to benefit society through education, research, and public service. To carry out its mission, the University operates the Agricultural Experimental Station, the Arts and Sciences Research Center, the Center for Business Research and Service, the New Mexico Energy Institute, the New Mexico Solar Energy Institute, the Engineering Experiment Station, the Mountain Laboratory, the Physical Science Laboratory, and many other educational, research, and service centers.

White Sands Missile Range, established as White Sands Proving Ground on 9 July 1945, represents the largest land mass under control of the United States Army (over two million acres). Together with its remote launch areas in Utah, it allows for missile flights of about 500 miles. Since its establishment, the Range has evolved into one of the most modern test facilities for use by the Defense Advanced Research Projects Agency, Army, Navy, Air Force, National Aeronautics and Space Administration, Department of Energy, and others.

We hope you will have an interesting, enjoyable, and fruitful stay.

Alan A. York
ALAN A. YORK
Major General, USA
Commanding General
US Army White Sands Missile Range

Charles W. Hanna
CHARLES W. HANNA
President
New Mexico State University

X
Best Available Copy

A G E N D A

THE TWENTY-SIXTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN
ARMY RESEARCH, DEVELOPMENT AND TESTING

22-24 October 1980

Host: White Sands Missile Range
Location: Physical Sciences Laboratory,
New Mexico State University

***** Wednesday, 22 October *****

0815-0915	REGISTRATION -- Lobby, Physical Sciences Laboratory
0915-0930	CALLING OF THE CONFERENCE TO ORDER -- Auditorium*
	Dr. Richard H. Duncan, Technical Director and Chief Scientist, White Sands Missile Range
	WELCOMING REMARKS
	MG Alan A. Nord, Commander, White Sands Missile Range
	Dr. Harold A. Dow, Associate Academic Vice-President, New Mexico State University
0930-1200	<u>GENERAL SESSION I</u>
	Chairman - Frank E. Grubbs, Program Committee Chairman, Aberdeen Proving Ground, Maryland
0930-1030	KEYNOTE ADDRESS
	Francis J. Anscombe, Department of Statistics, Yale University
1030-1100	BREAK
1100-1200	Design of Experiments (Title to be announced)
	Toby J. Mitchell, Union Carbide Nuclear Division, Oak Ridge, Tennessee
1200-1330	LUNCH

*All General Sessions, Clinical Sessions and Technical Session will
be held in the Auditorium, Physical Sciences Laboratory.

1330-1500

CLINICAL SESSION A

CHAIRMAN: Louellen Snodgrass, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexico

PANELISTS:

W. J. Conover, Department of Business Administration, Texas Tech University, Lubbock, Texas

W. D. Kaigh, Department of Mathematical Sciences, University of Texas-El Paso

Toby J. Mitchell, Union Carbide Nuclear Division, Oak Ridge, Tennessee

THE USE OF RIDGE REGRESSION IN TRAJECTORY ESTIMATION

William S. Agee and Robert H. Turner, White Sands Missile Range, New Mexico

A SEVEN VARIABLE COMPOSITE DESIGN FOR FITTING A SECOND ORDER RESPONSE SURFACE

Carl B. Bates, US Army Concepts Analysis Agency, Bethesda, Maryland

1330-1500

TECHNICAL SESSION I - "STOCHASTIC MODELING" - ROOM E-1104

CHAIRMAN: Roger Willis, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexico

AN APPLICATION OF ORDER STATISTICS TO TIME-SEQUENCE LOGIC

William E. Baker and Malcolm S. Taylor, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

STOCHASTIC MODELS FOR PAIRS OF WAITING LINES

Mary Anne Maher, Department of Mathematical Sciences, New Mexico State University, Las Cruces

1500-1530

BREAK

Best Available Copy

1530-1700

CLINICAL SESSION B

CHAIRMAN: Jose E. Gomez, White Sands Missile Range, New Mexico

PANELISTS:

Lonnie Ludeman, Department of Electrical and Computer Sciences, New Mexico State University, Las Cruces, New Mexico

Mary Anne Maher, Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico

Richard Tapia, Department of Mathematical Sciences, Rice University, Houston, Texas

RADAR ERROR SIGNAL IMPROVEMENT

Robert E. Green, White Sands Missile Range, New Mexico

USE OF THE BILINEAR Z-TRANSFORM IN IMPLEMENTING DIGITAL FILTERS

Donald W. Rankin, White Sands Missile Range, New Mexico

1530-1630

TECHNICAL SESSION II - "RELIABILITY AND RISK" - ROOM E-1104

CHAIRMAN: James Graves, White Sands Missile Range, New Mexico

INFERENCE PROCEDURES FOR DETERMINING LIFE TIME ESTIMATES OF ADVANCED MATERIALS

Donald Neal, Edward Lenoe and Donald Mason, US Army Materials and Mechanics Research Center, Watertown, Massachusetts

RISKS TO NEIGHBORING FACILITIES

Paul C. Cox, Physical Sciences Laboratory and Experimental Statistics Department, New Mexico State University, Las Cruces

1830-1930

SOCIAL HOUR - Dona Ana Room, Holiday Inn in Las Cruces

1930

BANQUET

Best Available Copy

***** Thursday, 23 October *****

0830-1000

CLINICAL SESSION C

CHAIRMAN: Don Walters, Atmospheric Sciences Laboratory,
White Sands Missile Range, New Mexico

PANELISTS:

Oskar Essenwanger, US Army Missile Command, Redstone Arsenal,
Alabama

Victor Solo, Department of Statistics, Harvard University,
Cambridge, Massachusetts

James Thompson, Department of Mathematical Sciences, Rice
University, Houston, Texas

D. G. Kabe, New Mexico State University and St. Mary's
University, Halifax

A STOCHASTIC MESOSCALE METEOROLOGICAL MODEL

Elton P. Avara, White Sands Missile Range, New Mexico

OPTIMAL ESTIMATION TECHNIQUES FOR FORECASTING PROPAGATION
PARAMETERS

D. L. Walters and K. E. Kunkel, White Sands Missile Range,
New Mexico

0830-1000

TECHNICAL SESSION III - "CURVE FITTING AND SMOOTHING" - ROOM E-1104

CHAIRMAN: Eugene F. Schuster, University of Texas-El Paso

ADAPTIVE MEDIAN SMOOTHING

William S. Agee and Jose E. Gomez, White Sands Missile
Range, New Mexico

FITTING AN ELLIPSE

Donald L. Buttz, White Sands Missile Range, New Mexico

METHODS FOR APPROXIMATING MATHEMATICAL FUNCTIONS

Donald W. Rankin, White Sands Missile Range, New Mexico

1000-1030

BREAK

1030-1200

CLINICAL SESSION D

CHAIRMAN: Jerry Thomas, Ballistic Research Laboratory,
Aberdeen Proving Ground, Maryland

Best Available Copy

PANELISTS:

Francis J. Anscombe, Department of Statistics, Yale University, New Haven, Connecticut

W. J. Conover, Department of Business Administration, Texas Tech University, Lubbock, Texas

Frank Grubbs, US Army Materiel Systems Analysis Activity, Aberdeen Proving Ground, Maryland

MOS TRAINING COURSE SELECTION CRITERIA: AN APPLICATION OF DISCRIMINANT ANALYSIS

Pat Cassidy and Lounell Snodgrass, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexico

ARMOR COMBAT FOR MODEL SUPPORT (ARCOMS)

Roger F. Willis, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexico

1030-1200 TECHNICAL SESSION IV - "QUANTILE STATISTICS" - ROOM E-1104

CHAIRMAN: Robert H. Turner, White Sands Missile Range, New Mexico

EXTREME VALUE QUANTILE RESPONSE EXPERIMENTAL DESIGN

Jill H. Smith and Jerry Thomas, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

A GENERALIZED QUANTILE ESTIMATOR

W. D. Kalbfleisch, Department of Mathematical Sciences, University of Texas-El Paso

1200-1330 LUNCH

1330-1530 GENERAL SESSION II

CHAIRMAN: Robert L. Lauener, US Army Research Office, Research Triangle Park, North Carolina

THE RANK TRANSFORMATION AS A ROBUST AND POWERFUL TOOL FOR THE ANALYSIS OF EXPERIMENTAL DATA

W. J. Conover, Department of Business Administration, Texas Tech University, Lubbock, Texas

Best Available Copy

THE NONPARAMETRIC ESTIMATION OF PROBABILITY DENSITIES IN
BALLISTICS RESEARCH

James R. Thompson, Chih-chy Fu, and Richard A. Tapia,
Department of Mathematical Sciences, Rice University,
Houston, Texas

1530-1600

BREAK

1600-1700

TECHNICAL SESSION V - "DESIGN OF EXPERIMENTS AND LINEAR
MODELS"

CHAIRMAN: Robert Green, White Sands Missile Range, New
Mexico

TESTABILITY OF LINEAR HYPOTHESES IN NORMAL LINEAR MODELS

Gerald S. Rogers, Department of Mathematical Sciences,
New Mexico State University, Las Cruces, New Mexico

THE POTENTIAL UTILITY OF CROSSING A FRACTIONAL FACTORIAL
WITH A FULL FACTORIAL IN THE DESIGN OF FIELD TESTS

Carl T. Russell, US Army Cold Regions Test Center, Seattle,
Washington

SOME REMARKS ON ANALYSIS OF CROSSOVER EXPERIMENTS

J. Robert Burge, Department of Biostatistics, Walter Reed
Army Institute of Research, Washington, DC

1600-1700

TECHNICAL SESSION VI - "SPECIAL APPLICATIONS OF SPECTRAL
ANALYSIS" - ROOM C-1104

CHAIRMAN: Paul C. Cox, Physical Sciences Laboratory and
Experimental Statistics Department, New Mexico
State University, Las Cruces, New Mexico

A TIME SERIES ANALYSIS AND MODELING APPROACH TO SENSE AND
DESTROY ARMOR (SADARM) RADIONLITRIC (EMR) NOISE DATA

Richard T. Maruyama, Ballistic Research Laboratory, Aberdeen
Proving Ground, Maryland

THE ROLE OF SPECTRAL LIMITS IN THE MEASUREMENT AND INTER-
PRETATION OF SECOND-ORDER STATISTICAL PROPERTIES

E. L. Church, US Army Armament Research and Development
Command, Dover, New Jersey

Best Available Copy

1600-1730

TECHNICAL SESSION VII - "RELIABILITY" - ROOM EB-015

CHAIRMAN: Paul Thrasher, White Sands Missile Range, New Mexico

DEVELOPMENT OF AN IDEALIZED GROWTH CURVE MODEL

Larry H. Crow, US Army Materiel Systems Analysis Activity, Aberdeen Proving Ground, Maryland

OPTIMAL CONFIDENCE BOUNDS ON SYSTEM RELIABILITY

Bernard Harris, University of Wisconsin-Madison and Andrew P. Sons, University of Wisconsin-Milwaukee

*****Friday, 24 October *****

0800-0930

CLINICAL SESSION E

CHAIRMAN: Carl Bates, US Army Concepts Analysis Agency, Bethesda, Maryland

PANELISTS:

Bernard Harris, Mathematics Research Center, University of Wisconsin-Madison

Richard A. Johnson, Department of Statistics, University of Wisconsin-Madison

RELIABILITY-BASED SAFETY FACTOR FOR CONCRETE STRUCTURES
SLIDING STABILITY

Paul F. Mlaker, US Army Waterways Experiment Station, Vicksburg, Mississippi

FULL UTILIZATION OF BLOOD TESTS IN PATERNITY DISPUTES

Paul Thrasher, White Sands Missile Range, New Mexico

0800-0930

TECHNICAL SESSION VIII - "STOCHASTIC MODELING II" - ROOM E-1104

CHAIRMAN: Peggy Hoffer, White Sands Missile Range, New Mexico

AN ALGORITHM FOR TRILATERATION

James T. Hall, White Sands Missile Range, New Mexico

SOCIAL SCIENTIST TECHNIQUE - THE CATALYST FOR OBTAINING
OBJECTIVITY FROM SUBJECTIVITY

Ronald L. Johnson, US Army Mobility Equipment Research and Development Command, Ft. Belvoir, Virginia

Best Available Copy

0930-1230

GENERAL SESSION III

CHAIRMAN: - Douglas B. Tang, Department of Biostatistics,
Walter Reed Army Institute of Research, Washington,
DC

OPEN MEETING OF THE AMSC SUBCOMMITTEE ON PROBABILITY AND
STATISTICS

1000-1030

BREAK

1030-1230

GENERAL SESSION III (Continued)

ENGINEERING TIME SERIES ANALYSIS

Victor Solo, Department of Statistics, Harvard University,
Cambridge, Massachusetts

STRESS-STRENGTH MODELS FOR RELIABILITY-OVERVIEW AND RECENT
ADVANCES

Richard A. Johnson, Department of Statistics, University of
Wisconsin, Madison

1230

ADJOURN

* * * * *
PROGRAM COMMITTEE

Carl Bates
George E. P. Box

Larry Crow
Walter Foster

Frank E. Grubbs (Program
Committee Chairman)

Robert Launer (Secretary)
Douglas Tang (Chairman,
Prob. & Stat. Subcommittee)
Malcolm Taylor
Langhorne Withers

HOW FAR TO GO IN LOOKING AT DATA*

F. J. Anscombe
Yale University
New Haven, Connecticut

All analysis of statistical data involves a balancing feat. On the one hand, we do not examine the data with an empty mind, there are specific questions we want to obtain answers to (that's why we assembled the data in the first place), we have preconceptions about the data, in examining the data we should not forget what we were looking for. But on the other hand, we should have an open mind, we should let the data speak for themselves, we should not just assume that some theoretical model is appropriate without checking. How far should we go in responding to unexpected features of the data, how far should we let the data control the kinds of things we do; or how far should we trust our preconceptions?

Now let me digress from that theme for just a moment. I have not only been invited to speak at this conference, I have been asked to give the keynote address. I think that a keynote address also involves a balancing feat. On the one hand, a keynote address should be inspirational, or if not quite inspirational at least interesting, or if not quite interesting at least fairly intelligible. On the other hand, it is not one of the regular invited papers or other real business of the conference. It should not try to steal their thunder. It should not be too weighty or indigestible--it should be hors d'oeuvre rather than a main course. To perform this balancing feat, I shall raise some questions that do not seem to be talked about much, but which all of us are aware of, and which therefore should be somewhat interesting to consider. In the interest of digestibility, I shall mostly refrain

*Prepared in connection with research supported by the Office of Naval Research (contract N00014-75-C-0563).

from answering the questions--because in most cases I have no idea how to answer them. But at the end of the talk I will give a little information about one of these questions that is possibly not well known, and to that extent the talk will have a little content.

Forecasting time series. Earlier this year at a meeting I heard a talk on short-term forecasting of business time series. The talk, which was very well delivered, threw out a challenge rather vividly, and I have thought about it quite a bit since. I'd like to indicate the gist of the talk. (I have not seen any write-up of the talk, what I'm saying may not adequately represent the speaker's views, and so I will refrain from mentioning his name.) The situation considered was this. He had several time series relating to a business company, production, sales, various things, and also public economic series; I think they were all quarterly series, and they went back a good many years. The object was to forecast some of the series for the next 1 or 2 or 3 quarters after the last observed value. A standard method would be to use Box-Jenkins technology--fit a parametric class of models to the data by maximum likelihood and use the fitted model to make the forecast. The parametric class is quite wide, and if in fact it is wide enough to represent reality adequately this procedure will be just about the best possible. The procedure is fully describable, or programmable, and therefore can be implemented completely by computer. The speaker didn't want to do that. He had imbibed the spirit of John Tukey's Exploratory Data Analysis. (Tukey does not discuss forecasting in his book, and I do not know whether what the speaker did was similar to what Tukey might have done.) He plotted his time series and took a good look at them. He noticed that round

about 1973-74 the series seemed to change character. Something happened to the economy (there was the oil crisis), and he judged that subsequent behavior of the series was unlike the preceding behavior. So he felt that only data after the change should be used in forecasting, although there were not many readings, and he proceeded to make (I think) some simple extrapolations by eye. He expressed the opinion that it was better to do a rough-and-ready job with just a few relevant readings than a fancy job with a lot of readings that were mostly irrelevant. If indeed there was a big enough change in the functioning of the economy in 1973-74 to make preceding data uninformative about later behavior, then he certainly had a case. How can you tell? It's a question about whether reality is better described in terms of Box-Jenkins parametric models, or better understood by someone who makes judgments based on plots and general background information, judgments that cannot (I think) be computer-programmed. The speaker seemed to think that obviously the second was the case. I don't think it's obvious either way. Some things could be done to investigate the matter, but nothing very easy or very quick.

How much should we look at the data? Everyone agrees that we must sometimes, to some extent, look at the data. Suppose we entertain some probabilistic model, or more modestly some way of thinking about a phenomenon and possible observations. If this model or way of thinking is not vacuous, there are some logically possible observations that conflict with it--otherwise it tells us nothing. Therefore it behooves us to see whether the observations are consistent. Particular instances of this are very well known.

So we must look at the observations a bit. The trouble is, it is easy

to be puzzled and misled if we look at them very much. A sample (from a population or probability distribution or stochastic process) has many individual features that do not reflect its source and would not persist if the sample size were much increased. Given adequate computing power, the question of how much to look, how far to go, outstrips available significance tests or other critical apparatus that might aid our judgment. To refrain from examining the data because we do not know how to evaluate what we see, that surely is foolish. To assume without evaluation that everything seen is important, is foolish too.

The matter is brought home to me whenever I take a small random sample from some distribution or process. Suppose I generate 50 observations from $N(0,1)$. One of the readings is bigger than 3 and the estimated variance is large. I take another sample and this time there are no outliers, the largest reading is only about 1.5 and the sample looks as if it came from a uniform distribution. I wonder if my program is wrong, so I go over the program again and I take a much larger sample and make a goodness-of-fit test, and then things seem O.K.

I'd like to show what happened the last time I tried a simulation. I was wanting to illustrate the difference between a stationary random sequence that was jointly normally distributed (jointly Gaussian) and a stationary random sequence with the same autocorrelations (same moments of first and second orders) and marginal normal distribution for any one member that was not jointly normal and would have a less temporally homogeneous behavior. The simplest examples would be Markov sequences, and I began with a lag-1 serial correlation coefficient equal to $\frac{2}{3}$. Figure 1 shows the first 60

members of a jointly normal sequence; and Figure 2 shows the first 60 members of a type of "jump" process which behaves in stretches like a jointly normal random sequence with a bit higher lag-1 serial correlation (actually $\frac{3}{4}$), but every now and then there is a break and the next reading is independent of its predecessors---sometimes at the break there is a big jump. So the two sequences should look similar except for occasional big jumps in the second process. Figure 2 looks as it is expected to, but not Figure 1---the first 20 readings move around a lot (with one very big jump between the tenth and eleventh reading), then the later readings are much less mobile. I can't help thinking that the speaker I heard on time-series forecasting would have identified a change in the economy round about reading no. 20.

I was so disgusted with the untypical behavior of the first plot that I scrapped them both and tried again, this time with a higher lag-1 serial correlation ($\frac{7}{8}$) which I thought would cramp the style of the jointly normal process and make it behave better. Both plots (Figures 3 and 4) looked reasonably "typical" of what I expected. (But note the apparent change in direction of the jointly normal plot around reading no. 33.)

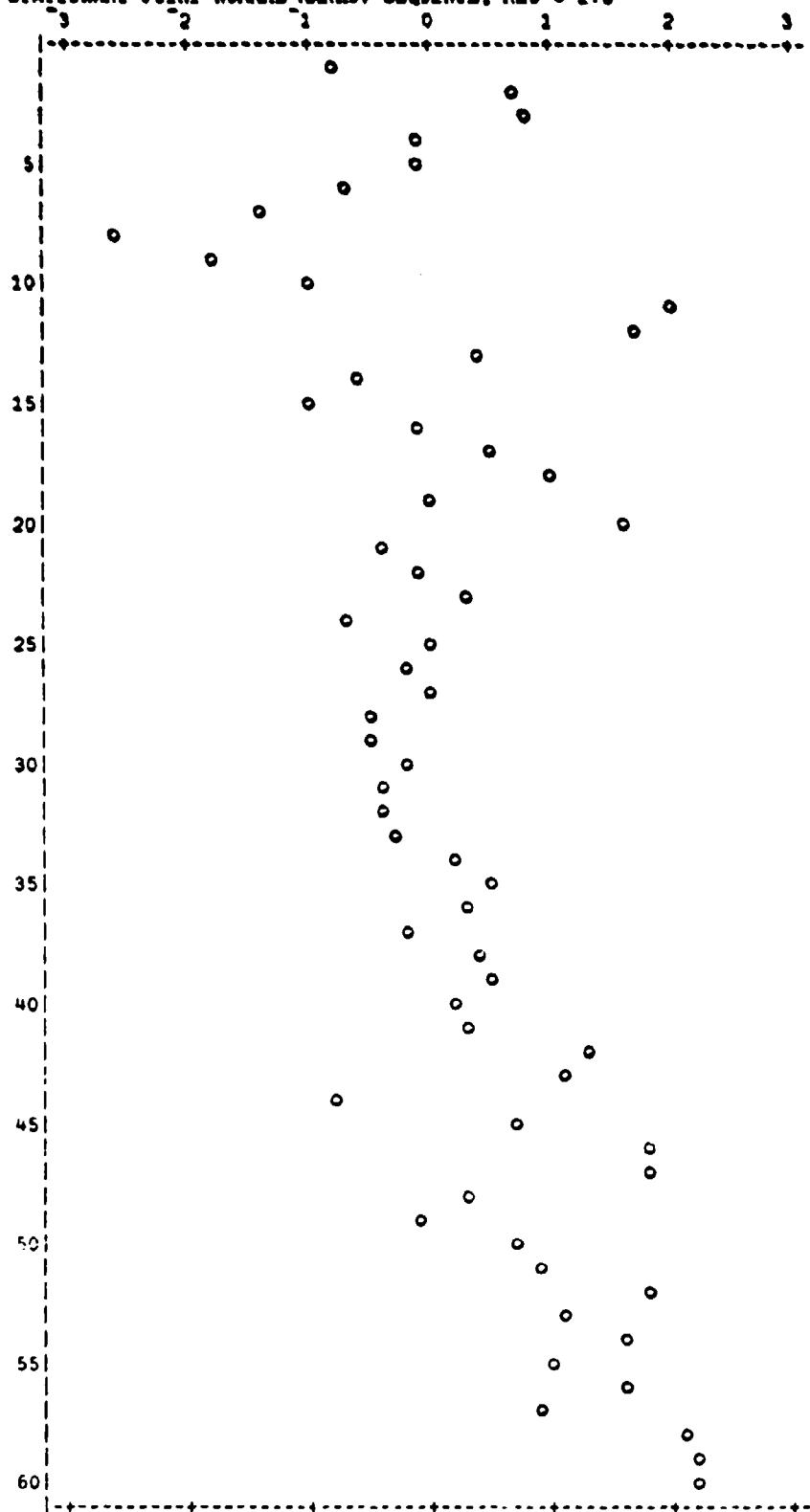
How many explanatory variables in regression? The question arises in different connections---sometimes very troublesome, sometimes easy to answer. The easy case is a planned factorial experiment of the classic Fisherian kind designed to permit estimation of the main effects of various factors and all sorts of interactions. Often what happens is that a few main effects and perhaps a few interactions are large and interesting and need to be duly reported and understood, while the rest are small and for most purposes can

be ignored---it wasn't known in advance which effects would be large and interesting, which small and ignorable. Usually the effects are orthogonal; their meaning does not depend on what other effects are estimated. Provided there are a few degrees of freedom for estimation of error, and provided we don't challenge the appropriateness of the model (structure) being fitted, there is an easy answer to the question, how many effects to estimate: estimate them all and then ignore any that aren't interesting.

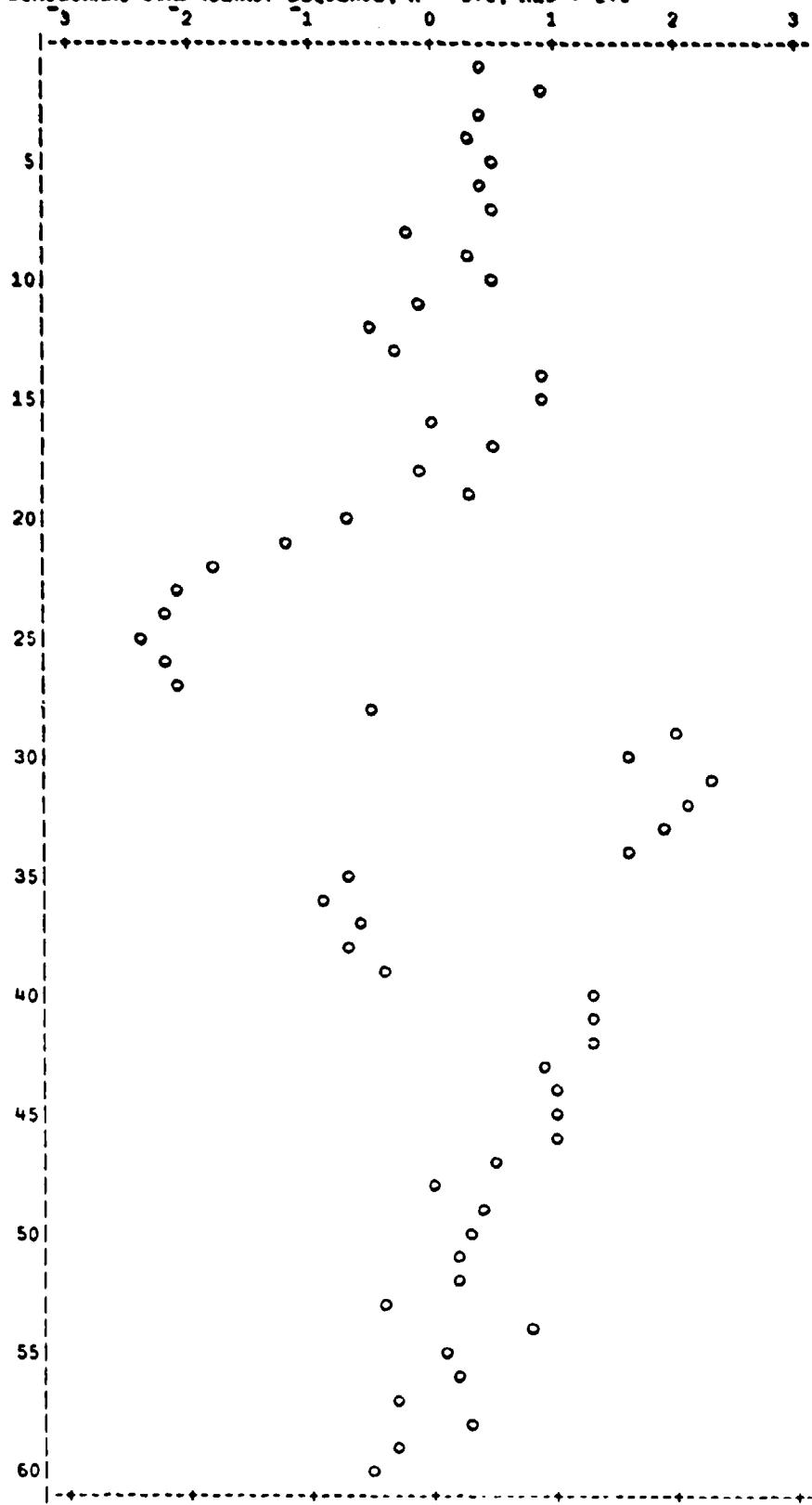
The proviso that the appropriateness of the model being fitted isn't challenged is important. I've already said that it behooves us to verify that models being fitted are consistent with the data. Suppose we want to check whether the (hypothetical) unexplained "error" term in the model seems to be something like i.i.d. normal. The obvious thing to do is calculate residuals from the estimated effects or regression relation. How closely the residuals reflect the hypothetical errors in the model depends on how many effects or regression coefficients have been estimated. If many have been estimated, there is a central limit effect---each residual is an average of many errors, and does not principally reflect one. To have informative residuals, small effects should not be estimated but left in the residuals.

[The talk concluded by presenting some material from Appendix 2 of the author's forthcoming book, Computing in Statistical Science through APL (Springer). A rule for deciding how many effects to estimate, due to J. W. Tukey, for the purpose of obtaining informative residuals, was described. Then two further rules were described, one of them based on C. L. Mallows's C_p statistic, designed to permit good prediction of unobserved values of the dependent variable. It was pointed out that these two purposes, informative residuals and good prediction, though at first glance quite different, are really closely related, and the three rules often lead to the same result.]

STATIONARY JOINT NORMAL MARKOV SEQUENCE. RHO = 2/3



STATIONARY JUMP MARKOV SEQUENCE, $A = 0.9$, $RHO = 243$



STATIONARY JOINTLY NORMAL MARKOV SEQUENCE, RHO = 7/8

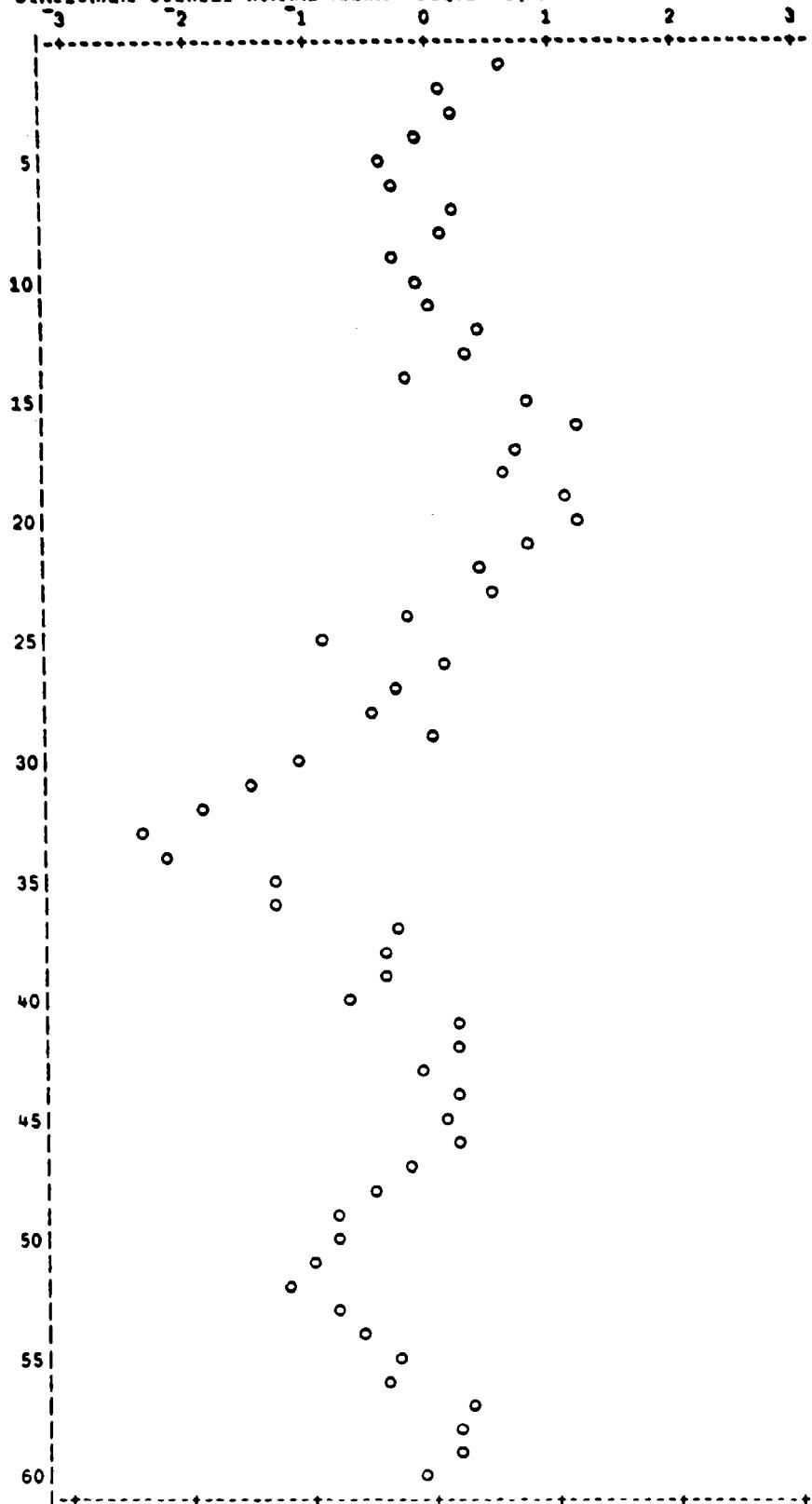


FIGURE 3

STATIONARY JUMP MARKOV SEQUENCE, $A = 15 \times 16$, $RHO = 7/8$

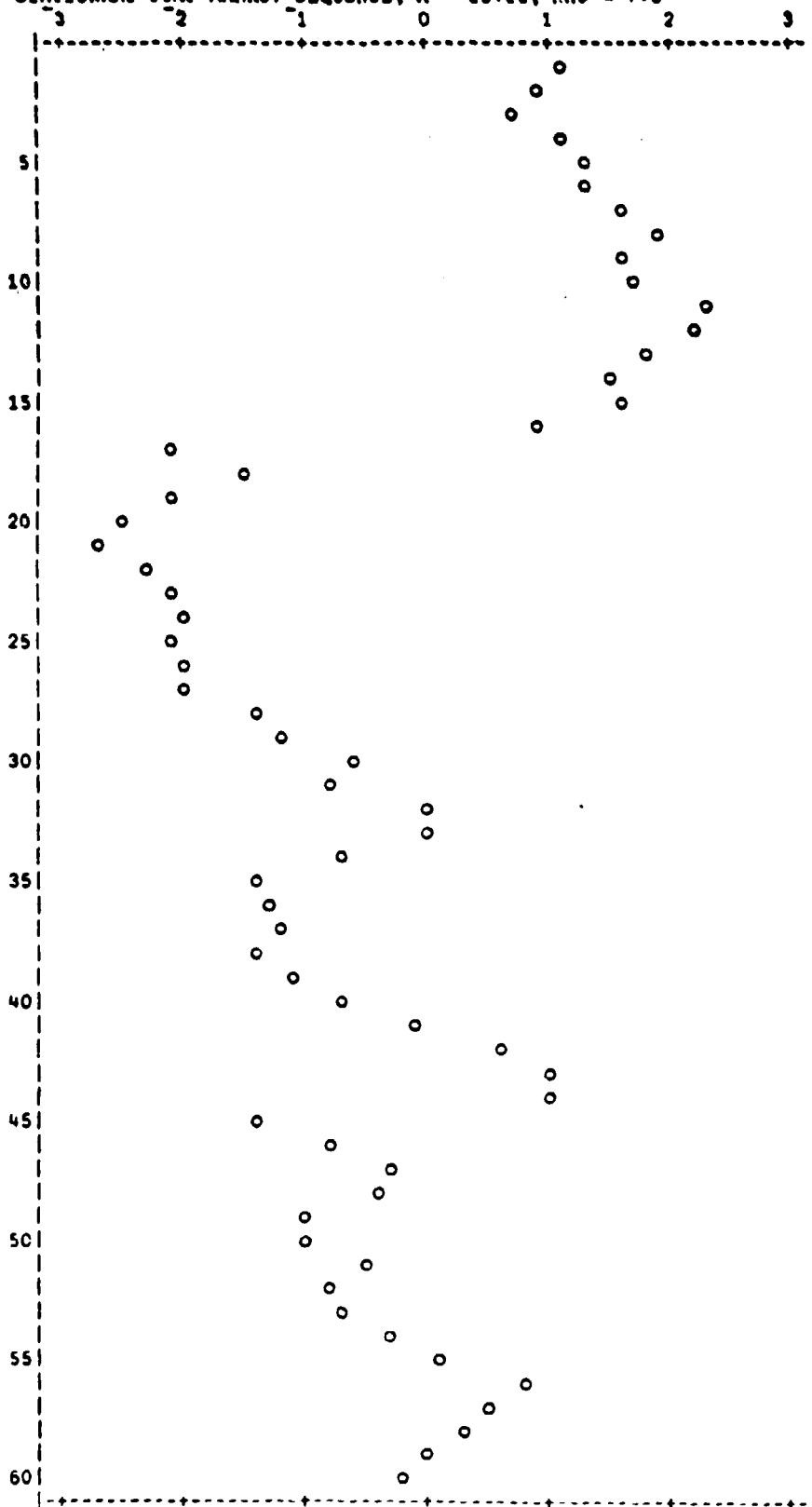


FIGURE 3

THE USE OF RIDGE REGRESSION IN TRAJECTORY ESTIMATION

William S. Agee and Robert H. Turner

Mathematical Services Branch

Data Sciences Division

US Army White Sands Missile Range

White Sands Missile Range, New Mexico 88002

ABSTRACT. Combining observations from several different trajectory measuring instruments we want to estimate the cartesian position (velocity) coordinates of a trajectory at possibly a large number of time points. Since the measuring systems are subject to systematic measurement errors as well as random measurement errors, in addition to estimating the trajectory coordinates we also want to estimate the measurement biases. The resulting estimation problem is a combined linear and nonlinear estimation problem in which the trajectory coordinates appear as nonlinear parameters in the measurements and the biases appear as linear parameters in the measurements. In practice we have found that it is often necessary to statistically constrain the measurement bias estimates by the use of Bayesian priors. These priors are assumed to be normal usually with mean zero. Thus, the specification of the priors is reduced to the specification of a prior variance for each measurement bias. There are no rules for choosing these prior variances and use of these guesses may result in rather poor bias and trajectory estimates in which the estimated bias vectors are too long and some of the biases may be of the wrong sign. We have developed the use of techniques very similar to ridge regression to treat this problem. The use of ridge regression for this problem results in significant improvements in both the trajectory and bias estimates. We demonstrate the use of this technique on several real trajectory estimation problems which have arisen at WSMR. In these problems we are estimating a trajectory and measurement biases using measurements from several radars. In some of these problems we also have measurements from optical tracking systems which, since they are more accurate and precise than radar, we use to prove the value of the ridge regression technique for obtaining improved radar trajectory and bias estimates. In using this ridge regression technique we have been successful in choosing a good value of the ridge parameter by using the ridge trace method proposed by Hoerl and Kennard. The ridge trace, although it is successful in obtaining a good value of the ridge parameter, is unsatisfactory in an automatic data processing procedure such as trajectory data reduction since a ridge trace requires human visual interpretation. We have tried some automatic methods for selection of a good ridge parameter value. So far, we have been unsuccessful in these attempts to develop an automatic method for choosing a good ridge parameter and we would like to have this considered as a clinical paper for the purpose of obtaining some new ideas for developing such a method.

1. TRAJECTORY ESTIMATION. Measurements of range, azimuth, and elevation from several different radars are used to estimate the cartesian position coordinates of a vehicle trajectory at a sequence of times, t_i , $i = 1, N$ which cover the entire trajectory. Since the measurements are subject to systematic errors as well as random measurement errors, we also want to estimate the systematic error parameters or biases in addition to the trajectory coordinates. The resulting estimation problem is a combined linear and nonlinear estimation problem in which the trajectory coordinates appear as nonlinear parameters in the measurements and the biases appear as linear parameters in the measurements.

Let $h_\alpha(\bar{x}_i)$ be a measurement function where \bar{x}_i is the cartesian position vector to the trajectory at time t_i . If we have M different radars observing the trajectory, then $\alpha = 1, 3M$. For a range measurement from the p^{th} radar

$$h_\alpha(\bar{x}_i) = [(x_i - x_p)^2 + (y_i - y_p)^2 + (z_i - z_p)^2]^{1/2} \quad (1)$$

where (x_p, y_p, z_p) are the cartesian coordinates of the origin of the local cartesian coordinate system at the p^{th} radar. For an azimuth measurement from the p^{th} radar the measurement function is

$$h_\alpha(\bar{x}_i) = \tan^{-1} \frac{x_i - x_p}{y_i - y_p} \quad (2)$$

For an elevation measurement from the p^{th} radar

$$h_\alpha(\bar{x}_i) = \tan^{-1} \frac{z_i - z_p}{[(x_i - x_p)^2 + (y_i - y_p)^2]^{1/2}} \quad (3)$$

Let $z_\alpha(t_i)$ denote the observed value of the α^{th} measurement. The observations are modeled as

$$z_\alpha(t_i) = h_\alpha(\bar{x}_i) + b_\alpha + e_\alpha(i) \quad (4)$$

where b_α is a constant measurement bias and $e_\alpha(i)$ is a zero mean, random measurement error. Let b be a $3M$ -dimensional bias vector $b^T = [b_1 \ b_2 \dots b_{3M}]$. Then the measurement model can be represented as

$$z_\alpha(t_i) = h_\alpha(\bar{x}_i) + s_\alpha b + e_\alpha(i) \quad (5)$$

where s_α is a row vector with a one in the α^{th} entry and zeros in all other entries.

$$s_\alpha = [0 \ 0 \ \dots \underset{\alpha^{\text{th}} \text{ position}}{1} \ 0 \ \dots \ 0] \quad (6)$$

Let $R_\alpha(t_i)$ be known variances of the random measurement errors, $e_\alpha(t_i)$. A normal prior with mean zero and diagonal covariance matrix P will be assumed for the bias vector b , $P = \text{diag}(P_\alpha)$, $\alpha = 1, 3M$. The estimation problem to be considered is to minimize,

$$\sum_{i=1}^N \sum_{\alpha=1}^{3M} (z_\alpha(t_i) - h_\alpha(\bar{x}_i) - S_\alpha b)^2 R_\alpha^{-1}(t_i) + b^T P^{-1} b \quad (7)$$

with respect to \bar{x}_i , $i = 1, N$ and b . Differentiating (7) with respect to \bar{x}_i and b results in the nonlinear normal equations

$$\sum_{\alpha=1}^{3M} H_\alpha^T(\bar{x}_i) R_\alpha^{-1}(t_i) (z_\alpha(t_i) - h_\alpha(\bar{x}_i) - S_\alpha \hat{b}) = 0 \quad i = 1, N \quad (8)$$

$$\sum_{i=1}^N \sum_{\alpha=1}^{3M} S_\alpha^T R_\alpha^{-1}(t_i) (z_\alpha(t_i) - h_\alpha(\bar{x}_i) - S_\alpha \hat{b}) - P^{-1} \hat{b} = 0 \quad (9)$$

where \hat{x}_i is the estimate of \bar{x}_i and \hat{b} is the estimate of b . In (8) $H_\alpha(\bar{x}_i)$ is the derivative, $\frac{\partial h_\alpha(t_i)}{\partial \bar{x}_i}$. In order to solve the normal equations, they

are linearized about a guess trajectory, $x_i(s)$. Let $x_i(s)$, $i = 1, N$ and $b(s)$ satisfy (8), i.e.,

$$\sum_{\alpha=1}^{3M} H_\alpha^T(x_i(s)) R_\alpha^{-1}(t_i) (z_\alpha(t_i) - h_\alpha(x_i(s)) - S_\alpha b(s)) = 0 \quad i = 1, N \quad (10)$$

If (8) is linearized about $x_i(s)$ and $b(s)$, we obtain

$$(\hat{x}_i - x_i(s)) = -A_i^{-1} A_{i,N+1}^T (\hat{b} - b(s)) \quad (11)$$

where

$$A_i = \sum_{\alpha=1}^{3M} H_\alpha^T(x_i(s)) R_\alpha^{-1}(t_i) H_\alpha(x_i(s)) \quad (12)$$

and

$$A_{i,N+1}^T = \sum_{\alpha=1}^{3M} S_\alpha^T R_\alpha^{-1}(t_i) H_\alpha(x_i(s)) \quad (13)$$

A_i is 3×3 and $A_{i,N+1}^T$ is $3 \times 3M$. Linearizing the second normal equation, (9), about $x_i(s)$ and solving for \hat{b} gives the result,

$$\hat{b} - b(s) = \left(P^{-1} + \sum_{i=1}^N \sum_{\alpha=1}^{3M} S_{\alpha}^T R_{\alpha}^{-1}(t_i) S_{\alpha} - \sum_{i=1}^N A_{i,N+1}^T A_{i,N+1}^{-1} A_{i,N+1} \right)^{-1} \left[\begin{aligned} & \sum_{i=1}^N \sum_{\alpha=1}^{3M} S_{\alpha}^T R_{\alpha}^{-1}(t_i) (z_{\alpha}(t_i) - h_{\alpha}(x_i(s)) - S_{\alpha} b(s)) \\ & - (P^{-1} + \sum_{i=1}^N \sum_{\alpha=1}^{3M} S_{\alpha}^T R_{\alpha}^{-1}(t_i) S_{\alpha}) b(s) \end{aligned} \right] \quad (14)$$

(11) and (14) for $\hat{x}_i - x_i(s)$ and $\hat{b} - b(s)$ are the basic equations for trajectory estimation. The solution to the normal equations are obtained by successive relinearization and solution of (8) and (9).

2. APPLICATION OF RIDGE REGRESSION. Although there are no convergence problems in solving the normal equations iteratively for the N-station radar case, another problem which is fairly common in the solution of linear least squares problem occurs quite frequently in trajectory estimation. Very often, the estimate of the bias vector, \hat{b} , converges to a solution which several of the components are too large. Sometimes the bias solution is obviously erroneous. One obviously erroneous case which frequently arises is that the elevation bias components will all be large and of the same sign. This problem of the estimated bias vector being too long is usually attributed to multicollinearity among the predictor variables in the linear least squares. The problem in the linear estimation case is often successfully treated by some method of biased estimation. This problem has not been properly recognized or successfully treated when it arises in trajectory estimation. Although the existence of these erroneous bias estimates have been recognized in trajectory estimation, the source of the difficulty was not properly identified. Some workers in trajectory estimation have stated that the existence of the problem demonstrated the need to specify a prior distribution in order to "tie down" or statistically constrain the bias estimates. Hence, the reason we have included the prior in (7). It does not take much experience in using these priors for radar trajectory estimation to realize that the problem of inflated bias estimates is usually as much present with as without the prior. There are no rules for choosing a good prior. We at first attempted to treat this problem by introducing a ridge parameter λ . Instead of minimizing (7), we would minimize

$$\sum_{i=1}^N \sum_{\alpha=1}^{3M} (z_{\alpha}(t_i) - h_{\alpha}(\bar{x}_i) - S_{\alpha} b)^2 R_{\alpha}^{-1}(t_i) + (1+\lambda) b^T P^{-1} L \quad (15)$$

Minimization of (15) merely introduces a factor of $(1 + \lambda)$ in (14) wherever P^{-1} appears. We denote the ridge solution as $\hat{b}(\lambda)$. The ridge solution reduces to

$$\hat{b}(\lambda) - \hat{b}(0) = -\lambda (Q + \lambda P^{-1})^{-1} P^{-1} \hat{b}(0) \quad (16)$$

where

$$Q = P^{-1} + \sum_{i=1}^N \sum_{\alpha=1}^{3M} S_\alpha^T R_\alpha^{-1} (t_i) S_\alpha - \sum_{i=1}^N A_{i,N+1}^T A_{i,N+1}^{-1} A_{i,N+1} \quad (17)$$

3. CHOOSING THE RIDGE PARAMETER-THE RIDGE TRACE. How should the ridge parameter λ be chosen? The graphical ridge trace method proposed by Hoerl and Kennard [1] is a standard method for choosing λ . Consider the following example from WSMR data. We have three radars, R122, R123, and R395 tracking a level flying target flying at about 30000 ft. The graph of Figure 1 shows the relative geometry of the target trajectory and radars. The diagonal elements of the prior covariance matrix used in this example were $P_{11} = P_{44} = P_{77} = 1$ (Range bias elements), $P_{22} = P_{33} = (1/\bar{R}_{122})$, $P_{55} = P_{66} = (1/\bar{R}_{123})$, and $P_{88} = P_{99} = (1/\bar{R}_{395})$, where R_j is the average range from the j^{th} radar to the trajectory. Figure 2 is a ridge trace for this example. Note $\lambda = -1$ corresponds to the least squares solution. We quickly learn two things by examination of this ridge trace. One, the range bias estimates do not stabilize from which we conclude that a ridge parameter should not be used on the range bias elements. Two, we would also want to conclude from the ridge trace that there is no benefit to be derived from the use of a prior on any of the range bias terms. Note that stability of the estimates occurs near the least squares solution which corresponds to $\lambda = -1$. The Bayesian solution with the prior specified which corresponds to $\lambda = 0$ is not plotted on the graph. This solution has the bias estimates,

	<u>R122</u>	<u>R123</u>	<u>R395</u>
Range (ft)	27	23	-20
Azimuth (mr)	-.02	-.06	.12
Elevation (mr)	.09	-.06	-.05

These estimates are not in the stability region of the ridge trace. Figure 3 is a ridge trace for this same example but without a prior on the range bias terms. The range biases are now only indirectly affected, i.e., through the angle bias estimate, by the ridge parameter λ . The value $\lambda = -.99$ appears to be a good choice of the ridge parameter. The following table confirms that large errors are present in the least squares bias estimates and that the ridge estimates with $\lambda = -.99$ provides a much better solution. The optics solution is derived from the azimuth and elevation measurements from several optical tracking cameras. The camera measurements are inherently much more precise than the radar measurements; hence, we often use an optically derived solution as a standard against which we compare radar performance.

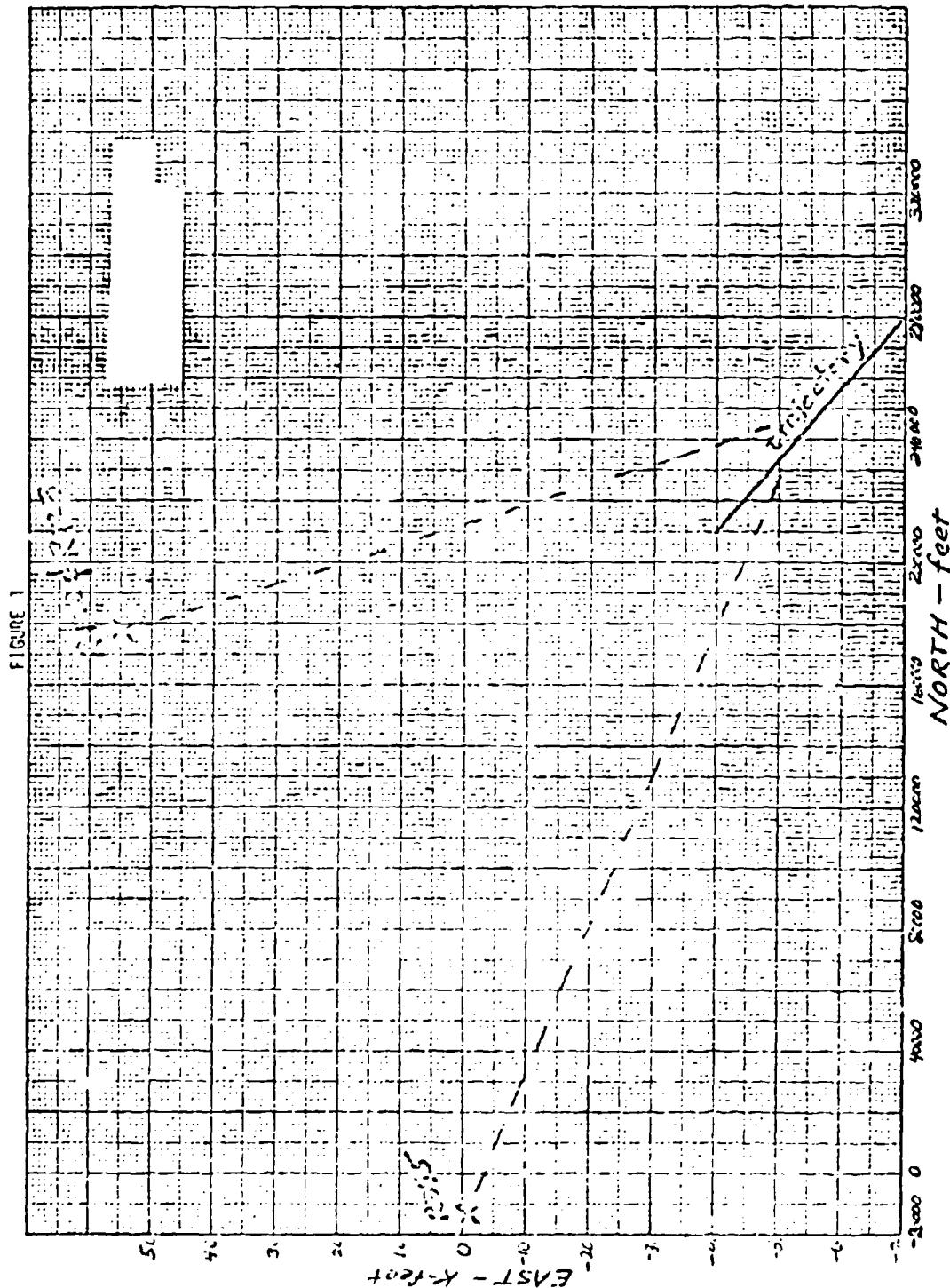


FIGURE 2

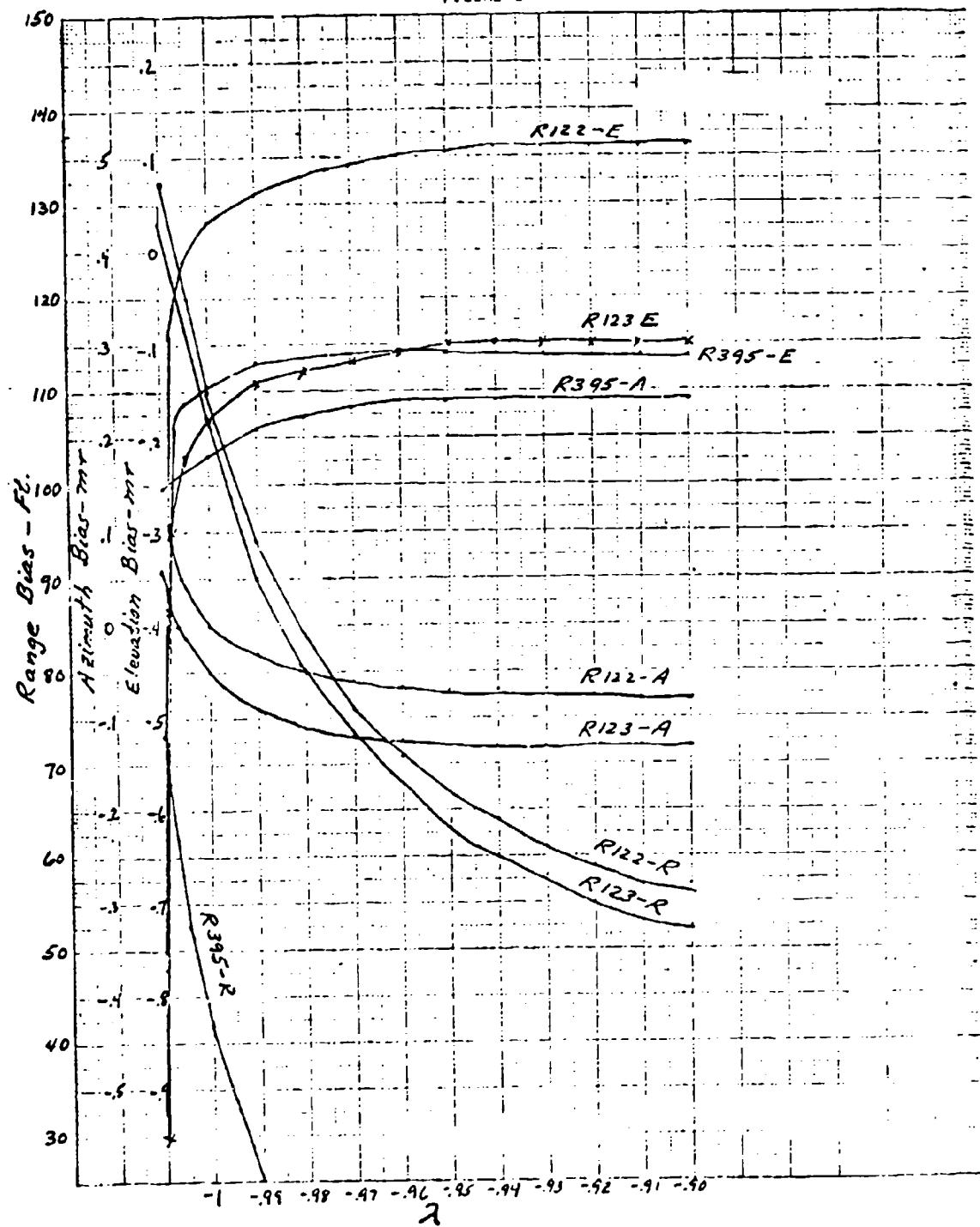
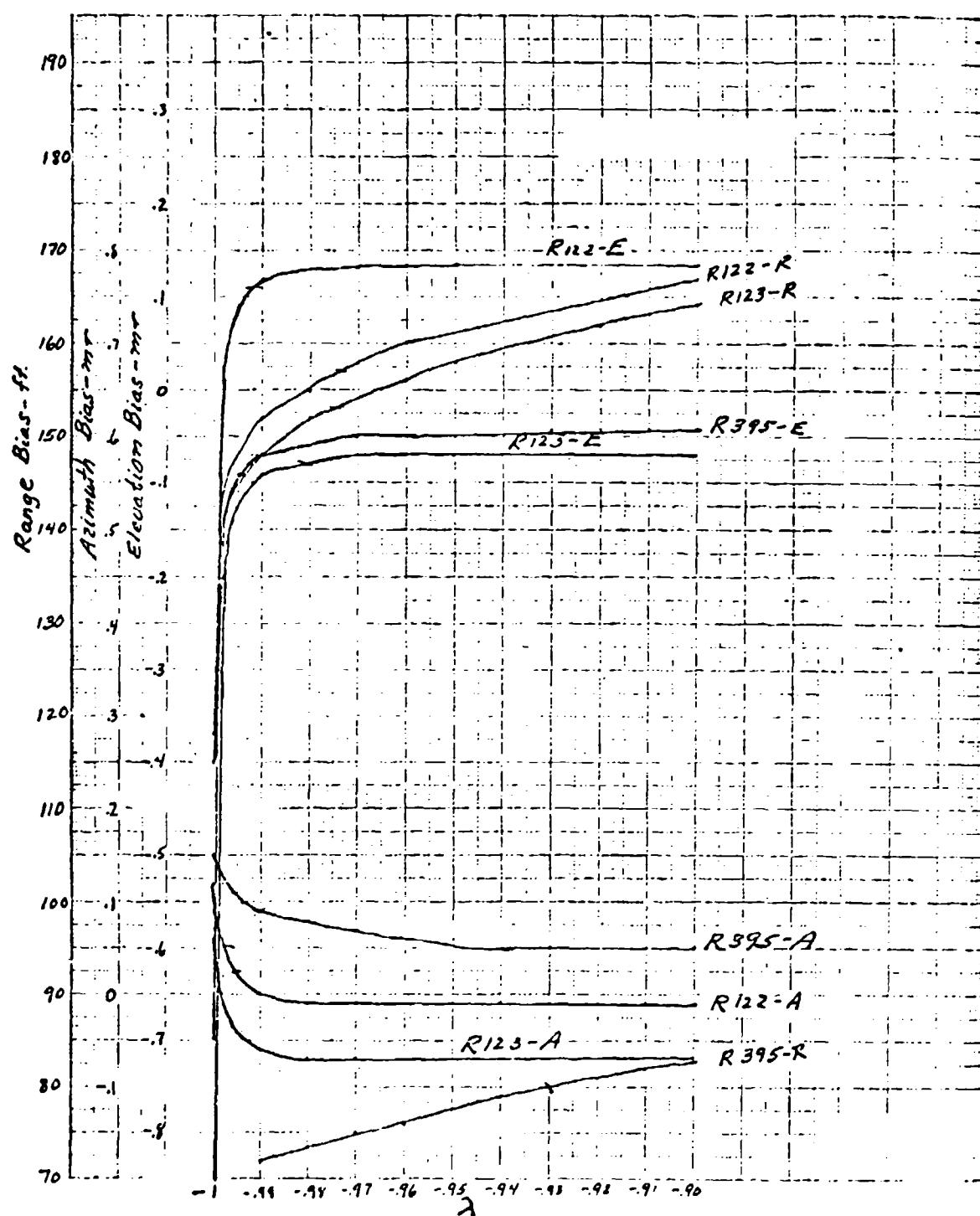


FIGURE 3



	<u>R122-R</u>	<u>R123-R</u>	<u>R395-R</u>	<u>R122-A</u>	<u>R123-A</u>	<u>R395-A</u>	<u>R122-E</u>	<u>R123-E</u>	<u>R395-E</u>
Optics	157.3	152.9	80.3	.05x10 ⁻³	.02x10 ⁻³	.09x10 ⁻³	.11x10 ⁻³	-.08x10 ⁻³	-.09x10 ⁻³
Ridge ($\lambda = -.99$)	151.8	148.2	72.0	0	-.06x10 ⁻³	.09x10 ⁻³	.12x10 ⁻³	-.09x10 ⁻³	-.07x10 ⁻³
Least squares	118.3	114.8	63.7	.116x10 ⁻³	.058x10 ⁻³	.148x10 ⁻³	.737x10 ⁻³	-.947x10 ⁻³	-.538x10 ⁻³

Even though we believe that the ridge solution provides much better trajectory and bias estimates from an N-station radar solution than does the least squares solution, we have tried and are still trying to develop a practical method of using ridge regression in trajectory data reduction. The graphical ridge trace method of Hoerl and Kennard for choosing λ is unsatisfactory for trajectory data reduction. The quick turnaround time required by range users for trajectory data products makes the ridge trace method, which requires human intervention, impractical for routine use. An automatic method for choosing a good value of λ is required in order that the ridge method be practical for trajectory data reduction.

4. CHOOSING THE RIDGE PARAMETER-MINIMIZING THE MEAN SQUARE ERROR.

One automatic method for choosing ridge parameters is an iterative method proposed by Hoerl and Kennard [2]. This method is developed for choosing the ridge parameter vector in a generalized ridge regression problem. For the generalized ridge regression in the radar bias estimation application, we want to choose a parameter vector λ having components λ_i , $i = 1, 2M$.

Thus, we will have a separate ridge parameter for each of the angle biases but none for the range bias terms. Suppose we arrange the order of the biases so that the angle biases appear in the first 2M positions of b and the range biases in the last M positions of b. We partition the bias estimation equation, $Q\hat{b} = U$ as

$$\begin{bmatrix} Q_a & R^T \\ R & Q_r \end{bmatrix} \begin{bmatrix} \hat{b}_a \\ \hat{b}_r \end{bmatrix} = \begin{bmatrix} U_a \\ U_r \end{bmatrix} \quad (18)$$

In the above we have replaced the matrix Q defined in (17) by the slightly different definition,

$$Q = \sum_{i=1}^N \sum_{\alpha=1}^{3M} S_{\alpha}^T R_a^{-1} (t_i) S_{\alpha} - \sum_{i=1}^N A_{i,N+1}^T A_{i,N+1} \quad (19)$$

Note that we have abandoned the use of priors in the definition of Q.

Suppose we transform the bias vector \hat{b}_a as $\hat{b}_a = T\hat{\beta}$ where T is orthogonal. We can eliminate \hat{b}_r from (18) obtaining,

$$\hat{b}_r = Q_r^{-1} (U_r - RT\hat{\beta}) \quad (20)$$

and

$$T^T (Q_a - R^T Q_r^{-1} R) T \hat{\beta} = T^T (U_a - R^T Q_r^{-1} U_r) \quad (21)$$

we choose T so that

$$T^T (Q_a - R^T Q_r^{-1} R) T = I \quad (22)$$

where $I = \text{diag } (\gamma_i)$. Then we have

$$T \hat{\beta} = T^T (U_a - R^T Q_r^{-1} U_r) \quad (23)$$

We now form the generalized ridge regression as

$$(I + D(\lambda))(\hat{\beta}(\lambda) - \hat{\beta}(0)) = -D(\lambda) \hat{\beta}(0), \quad (24)$$

where $D(\lambda) = \text{diag } (\lambda_i)$ and $\hat{\beta}(0)$ is the least squares solution.

The iterative method of Hoerl and Kennard attempts to choose the λ_i to minimize the mean square error in $\hat{\beta}(\lambda)$. Thus, we want to minimize

$$E[\hat{\beta}(\lambda) - \beta]^T (\hat{\beta}(\lambda) - \beta) \quad (25)$$

where β is the true parameter value. Minimizing (25) results in the choice of the optimal λ as

$$\lambda_i = \frac{1}{\beta_i^2} \quad (26)$$

Since the true values, β_i , are unknown, (26) cannot be implemented exactly. Hoerl and Kennards approximate implementation of (25) uses the iteration,

$$\lambda_i^{(K+1)} = (1/\hat{\beta}_i^{(K)})^2 \quad (27)$$

where $\beta^{(K)}$ is obtained from (24) as

$$\hat{\beta}_i^{(K)} = \gamma_i \hat{\beta}_i(0) / (\gamma_i + \lambda_i^{(K)}) \quad (28)$$

with $\hat{\beta}_i^{(0)} = \hat{\beta}_i(0)$.

Hemmerle [3] has shown that a closed form solution of this iterative scheme is easily obtained. Hemmerle shows that the iteration converges to e_* for $0 < (1/\gamma_i \hat{\beta}_i^2(0)) \leq 1/4$ where

$$e_* = \frac{(1-2e_0) - \sqrt{1-4e_0}}{2e_0} \quad (29)$$

and $e_0 = 1/\gamma_i \hat{\beta}_i(0)$. For $e_0 > 1/4$, $\hat{\beta}_i^{(K)} \rightarrow 0$.

For our three radar example applications, we obtain the following data for the application of the above method.

$\hat{\beta}_a(0)$	$\hat{\beta}_i(0)$	γ_i	$\gamma_i \hat{\beta}_i^2(0)$	β_*
$.116 \times 10^{-3}$	$-.43181174 \times 10^{-4}$	4.1857985×10^{10}	78.04	$-.42616219 \times 10^{-4}$
$-.737 \times 10^{-3}$	$-.54950324 \times 10^{-4}$	5.7866501×10^{10}	174.73	$-.54634011 \times 10^{-4}$
$.058 \times 10^{-3}$	$.14776931 \times 10^{-3}$	5.8775591×10^{10}	1283.41	$.14554624 \times 10^{-3}$
$-.947 \times 10^{-3}$	$-.12048667 \times 10^{-3}$	8.3670572×10^9	121.46	$-.11947337 \times 10^{-3}$
$.148 \times 10^{-3}$	$-.17659027 \times 10^{-3}$	8.3156874×10^7	2.59	0
$-.538 \times 10^{-3}$	$-.13022000 \times 10^{-2}$	1.2606770×10^7	21.37	$-.1237882 \times 10^{-2}$

where β_* is the limit of the iterative scheme.

This method is clearly inadequate for dealing with this example, since the bias vector has been shrunken only a very small amount whereas a considerable amount of shrinking is necessary in order that the estimated bias agree reasonably well with the biases calculated from the optics solution. The difficulty arises from those least squares bias estimates which are far from their true values. When the true values are replaced by these estimated values in the Hoerl and Kennard method, the iteration converges to the wrong values which are quite near to the least squares estimates. We have not been able to find an automatic method for choosing ridge parameters which works well for any of our examples.

5. APPLICATION OF PRINCIPAL COMPONENTS REGRESSION. In absence of a method for choosing ridge parameters we have recently begun to work with principal component regression as an alternative to the use of ridge regression for shrinking the erroneous least squares solution. We have had a fair amount of success with this method on several examples. In the principal components regression we set $\beta_i = 0$ for $i = 2M-r$, $2M$ where r is the smallest integer for which

$$\frac{\sum_{i=2M-r}^{2M} \gamma_i}{\sum_{i=1}^{2M} \gamma_i} \leq 10^{-2} \quad (30)$$

We are assuming in (30) that the eigenvalues, γ , have been ordered from largest to smallest. The use of 10^{-2} as a cutoff for the eigenvalues in (30) is arbitrary, but we have found it to work well in several examples. For the example application described above, the principal component estimator sets $\beta_5 = \beta_6 = 0$, corresponding to the two smallest eigenvalues. The results of the principal components estimator for this example is compared below with optics, least squares, and ridge regression.

	R122-R	R123-R	R395-R	R122-A	R123-A	R395-A	R122-E	R123-E	R395-E
Optics	157.3	152.9	80.3	.05x10 ⁻³	.02x10 ⁻³	.09x10 ⁻³	.11x10 ⁻³	-.08x10 ⁻³	-.09x10 ⁻³
Ridge ($\lambda = -0.95$)	151.8	143.9	72.0	0	-.06x10 ⁻³	.09x10 ⁻³	.12x10 ⁻³	-.09x10 ⁻³	-.07x10 ⁻³
Least squares	118.3	114.8	63.7	.116x10 ⁻³	.058x10 ⁻³	.148x10 ⁻³	.737x10 ⁻³	-.947x10 ⁻³	-.538x10 ⁻³
Principal components	147.9	144.3	69.4	0	-.055x10 ⁻³	.105x10 ⁻³	.123x10 ⁻³	-.089x10 ⁻³	-.067x10 ⁻³

The principal component method has been criticized as being too restrictive. Marquardt has suggested that a fractional rank procedure be used instead of the principal components integral rank procedure. The fractional rank estimator takes a linear combination of principal components estimators,

$$\hat{\beta}_f = c\hat{\beta}_r + (1-c)\hat{\beta}_{r+1}, \quad 0 \leq c \leq 1 \quad (31)$$

where $\hat{\beta}_r$ is the principal components estimator of rank r . The difficulty with the fractional rank procedure is in the choice of the parameter, c . Marquardt [4] suggests using a graphical method like a ridge trace where each component of the vector $\hat{\beta}_f$ is plotted against the fractional rank, $f = cr + (1-c)(r+1)$.

Another procedure, suggested by Hocking, Speed, and Lynn [5], for choosing c is to minimize the mean square estimation error. This method is implemented in a way similar to the iterative method of Hoerl and Kennard for choosing the generalized ridge parameters and suffers the same difficulties in application.

6. CONCLUSIONS. It has been demonstrated by simulation studies [6] that ridge regression offers potentially a better estimator than the principal component technique, but that a better estimator of the ridge parameter is necessary before that potential can be realized. I definitely believe that this conclusion is correct for our trajectory estimation applications. I also feel that a similar conclusion could be demonstrated for the fractional rank procedure. We are unable to implement these potentially better methods in a routine trajectory data reduction because we are presently unable to develop a good estimator for the parameters in either method. We are greatly in need of some fresh ideas for choosing the parameters in these biased estimation methods.

REFERENCES

1. Hoerl, A., and Kennard, R., (Feb 1970, p. 55-67), "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*.
2. Hoerl, A., and Kennard, R., (1976, p. 77-88), "Ridge Regression: Iterative Estimation of the Biasing Parameters", *Communications in Statistics*.
3. Hemmerle, W., (Aug 1975, P. 309-314), "An Explicit Solution for Generalized Ridge Regression", *Technometrics*.
4. Marquardt, D., (Aug 1970, p. 591-601), "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation", *Technometrics*.
5. Hocking, R., Speed, F., and Lynn, M., (Nov 1976, p. 425-437), "A Class of Biased Estimators in Linear Regression", *Technometrics*.
6. Gunst, R., and Mason, R., (Sep 1977, p. 616-627), "Biased Estimation in Regression: An Evaluation Using Mean Squared Error", *JASA*.

AN EIGHT VARIABLE COMPOSITE DESIGN FOR
FITTING A SECOND ORDER RESPONSE SURFACE

Carl B. Bates

US Army Concepts Analysis Agency

Bethesda, Maryland 20014

ABSTRACT. An electronic warfare study has been initiated at the Concepts Analysis Agency. The study is to provide justification for procurement and force structuring decisions concerning new electronic warfare/signal intelligence (EW/SIGINT) systems. A communications model, Communications Electronics Warfare Combat Simulation Model (COMMEL II.5), will be used to investigate the effectiveness of US EW/SIGINT systems against enemy command, control, and communications (C^3) systems. COMMEL II.5 is estimated to require four hours' running time to simulate an eight-hour battle. Eight model input variables having 6, 6, 4, 4, 4, 3, 2, and 2 levels, respectively, were selected for the investigation. The objective is to fit a second order response surface using as small a number of computer runs as possible. A $1/4 \times 2^3$ fractional factorial design is augmented with the addition of axial points. The resulting variation of a central composite design contains 80 design points. The experimental design is presented and discussed.

1. INTRODUCTION. The Concepts Analysis Agency has been tasked to analyze the relative contribution of US electronic warfare systems to the outcome of ground combat. A wide variety of new electronic warfare/

signal intelligence (EW/SIGINT) and weapon systems has been introduced for the post-1980 timeframe. Recent assessments of US and Soviet command, control, and communications (C³) show the need for improving the quantified analysis of EW assets to counter threat C³. An analytical basis is needed to provide justification for procurement and force structuring decisions with respect to EW/SIGINT systems. The analysis should provide detail sufficient to assess the potential of selected electronic countermeasures (ECM), electronic warfare support measures (ESM), and tactical SIGINT systems.

To accomplish the task, the Concepts Analysis Agency initiated the Force Electronic Warfare/Tactical SIGINT (FEWTS) Study. The purpose of the study is to analyze the relative contribution to combat potential of the denial, destruction, and exploitation of threat C³ through the application of US EW/tactical SIGINT means.

2. PROBLEM DESCRIPTION. An enhanced version of the Communications Electronic Warfare Combat Simulation Model (COMMEL II.5) was selected for use in the study. COMMEL II.5 is a fully computerized, dynamic, two-sided, division level ground combat model. It will be used to investigate the effectiveness of US EW/SIGINT systems against enemy C³ systems. The model permits detailed observation of communication events in a combat environment, and provides a tool for measuring, in terms of combat outcome, the merit of selected EW/tactical SIGINT capabilities against a threat.

Eight model input variables were selected for the investigation. The variables represent equipments and operating characteristics. The number of levels selected for the eight variables ranged from two to six. The eight variables and their levels are given in Table 1. The full $6^2 \times 4^3 \times 3 \times 2^2$ design has over 27,000 variable level combinations.

Table 1. Design Variables

Variable	Levels
x_1 - TACJAM	0, 1, 2, 3, 6, 9
x_2 - TLQ 17/A	0, 1, 2, 3, 6, 9
x_3 - QUICK FIX	0, 1, 3, 6
x_4 - TRLBLZR	0, 1, 2, 3
x_5 - CRIT NODES	1, 2, 3, 4
x_6 - JAM vs L/K	1, 2, 3
x_7 - E/W EMPL CON	1, 2
x_8 - ARTLY EMPL CON	1, 2

Nine tentative measures of effectiveness (MOE) were identified. They consisted of Red and Blue materiel and personnel losses and forward edge of the battle area (FEBA) loss. All nine MOE are continuous variables. The study members desired a second order response surface for each of the nine MOE in terms of the eight model input variables. The second order model,

$$\begin{aligned}
y = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \\
& + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 + \beta_{55} x_5^2 + \beta_{66} x_6^2 \\
& + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \dots + \beta_{17} x_1 x_7 + \beta_{18} x_1 x_8 \\
& + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \dots + \beta_{27} x_2 x_7 + \beta_{28} x_2 x_8 \\
& \vdots \\
& + \beta_{67} x_6 x_7 + \beta_{68} x_6 x_8 \\
& + \beta_{78} x_7 x_8,
\end{aligned}$$

has 43 terms (8 linear, 6 quadratic, 28 cross-products, and the intercept term). The eight independent variables (x 's) represent the eight COMMEL II.5 model input variables and the dependent variable (y) denotes a particular COMMEL II.5 model output variable, MOE.

3. BACKGROUND. Box and Wilson (1951) introduced the concept of composite designs; Box and Hunter (1957) introduced the concept of rotatability; and Box and Draper (1959) developed criteria for selecting response surface designs. Hill and Hunter (1965) and Mead and Pike (1975) give reviews of the developments of response surface methodology. More readily available sources on response surface methodology can be found in Cochran and Cox (1957), Davies (1960), Myers (1971), and Anderson and McLean (1974).

Composite designs are full or fractional 2^k factorial designs augmented with additional points which permit estimation of the quadratic

coefficients of a second order surface. The augmentation consists of $2k$ plus one (or more) center points as illustrated in Table 2. Therefore, the composite design consists of $(2^k + 2k + 1)$ design points.

Table 2. $2k + 1$ Augmentation

x_1	x_2	x_3	...	x_k
$-\alpha$	0	0	...	0
$+\alpha$	0	0	...	0
0	$-\alpha$	0	...	0
0	$+\alpha$	0	...	0
0	0	$-\alpha$...	0
0	0	$+\alpha$...	0
.....				
0	0	0	...	$-\alpha$
0	0	0	...	$+\alpha$
0	0	0	...	0

If $k=3$ and the 2^3 coded x -values are $+1$ and -1 , the design matrix for a central composite design is as shown in Table 3 and illustrated in Figure 1.

Table 3. Three Variable Central Composite Design Matrix

x_1	x_2	x_3	
-1	-1	-1	
-1	-1	1	
-1	1	-1	
-1	1	1	
1	-1	-1	
1	-1	1	
1	1	-1	
1	1	1	
- α	0	0	
$+\alpha$	0	0	
0	$-\alpha$	0	
0	$+\alpha$	0	
0	0	$-\alpha$	
0	0	$+\alpha$	
0	0	0	center point

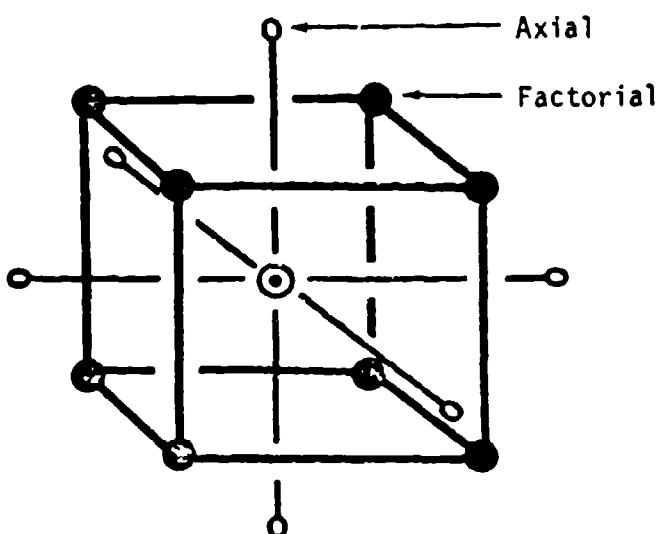


Figure 1. Central Composite Design

The literature on central composite designs contains discussions on the determination of α to yield orthogonal designs. However, no information was found applicable to the above described problem in which the x -values are prescribed and fixed. The following section discusses the attempts to develop the eight variable composite design.

4. CANDIDATE DESIGNS

a. Design A. A "Base Case" situation was defined early in the study planning phase. The Base Case combination of the levels of the eight variables is shown by the circled values in Table 4.

Table 4. Base Case

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	0	0	1	1	①	①
1	1	1	①	②	②	2	2
2	2	③	2	3	3		
③	③	6	3	4			
6	6						
9	9						

First, a fractional factorial design was developed. The lowest and the highest values of each of the eight variables were considered as the "low" and "high" values of a 2^8 factorial. A resolution V 2^{8-2} fractional factorial design was developed using $I = ABCEG = ABDFH = CDEFGH$ as the defining contrast. This gave 64 design points. The resolution V design permits fitting the 8 linear terms and the 28 cross-product terms. Table 5 repeats the Base Case values given in Table 4 and illustrates the high and low values used in the fractional factorial. Augmentation of the fractional factorial then consisted of using "inside" values as the axial points. This gave the 16 design points shown in the lower portion of Table 5. Note that for x_1 , x_2 , x_3 , and x_6 , Base Case values were treated as center points, but x_4 and x_5 were balanced over the two inside values, as were x_7 and x_8 . This gave a total of 80 design points. The composite design matrix A, however, was singular.

Table 5. Design A

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	0	0	0	1	1	1	1
1	1	1	1	1	2	2	2	2
2	2	2	3	2	3	3	3	3
3	3	3	6	3	4			
6	6	6						
9	9	9						
<u>Fractional Factorial</u>								
Low	0	0	0	0	1	1	1	1
High	9	9	6	3	4	3	2	2
<u>Augmentation</u>								
1								
2	1							
3	6	3	3	1	2	2	2	2
4	3	3	3	2	1	2	1	2
5	3	3	3	1	2	2	1	1
6	3	3	3	2	3	2	2	2
7	3	3	3	1	2	2	1	1
8	3	3	3	2	1	2	2	2
9	3	3	3	1	2	2	1	1
10	3	3	3	2	3	2	2	2
11	3	3	3	1	2	2	1	1
12	3	3	3	2	1	2	2	2
13	2	3	3	1	2	2	1	1
14	3	3	3	2	1	2	2	2
15	3	2	3	3	2	2	1	1
16	3	3	3	2	3	2	2	2

b. Design B. For the second attempt to develop a composite design, the same fractional factorial design was retained, but the 16 augmentation points were changed. The inside values for all variables except x_6 were balanced over the 16 design points. Variable x_6 was set at its Base Case value (center point). The augmentation part of the design is shown in the lower portion of Table 6. This composite design matrix was also singular.

c. Design C. The third attempt involved the same fractional factorial design, but the high and low values were changed. This time two adjacent inside values were used as the high and low values. One of the adjacent values was the Base Case value (center point). The high and low values used are shown in the center portion of Table 7. For variable x_6 , which has only one inside value, one outside value (3) was used.

For the augmentation portion of the composite design, outside values were used as the axial points for x_1 through x_6 . The next two inside values (1 and 6) were used as another pair of axial points for variables x_1 and x_2 (points 13 and 14, and points 15 and 16). All other variables were held fixed at their center points. Variables x_7 and x_8 were not varied; all 16 points were held fixed at their center points.

Table 6. Design B

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	0	0	1	1	1	1
1	1	1	1	2	2	2	2
2	2	3	2	3	3	3	3
3	3	6	3	4			
6	6	9					

Fractional Factorial

Low	0	0	0	0	1	1	1
High	9	9	6	3	4	3	2

Augmentation

1	1	2	1	1	2	1	1
2	6	3	3	2	3	2	2
3	2	1	1	2	2	1	2
6	3	6	3	1	3	2	1
5	1	2	1	2	2	1	2
6	6	3	3	1	3	2	1
7	2	1	1	2	2	1	2
8	3	6	3	1	2	1	2
9	1	2	1	2	3	2	1
10	6	3	3	1	2	1	2
11	2	1	3	1	2	1	2
12	3	6	2	3	1	2	1
13	1	6	3	1	2	1	2
14	6	2	1	3	1	2	1
15	2	1	6	3	1	2	1
16	3	6	1	3	2	1	2

Table 7. Design C

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	0	0	0	1	1	1	1
1	1	1	1	1	2	2	2	2
2	2	2	3	2	3	3	3	3
3	3	3	6	3	4			
6	6	6						
9	9	9						

<u>Fractional Factorial</u>								
Low	2	2	1	1	2	2	1	1
High	3	3	3	2	3	3	2	2

<u>Augmentation</u>								
1	0	3	3	1	2	2	1	1
2	9	3	3	1	2	2	1	1
3	3	0	3	1	2	2	1	1
4	3	9	3	1	2	2	1	1
5	3	3	0	1	2	2	1	1
6	3	3	6	1	2	2	1	1
7	3	3	3	0	2	2	1	1
8	3	3	3	3	2	2	1	1
9	3	3	3	1	1	2	1	1
10	3	3	3	1	1	2	1	1
11	3	3	3	1	1	2	1	1
12	3	3	3	1	1	2	1	1
13	1	6	3	3	1	2	1	1
14	6	3	1	6	3	1	1	1
15	3	3	1	6	3	1	1	1
16	3	3	1	1	2	2	1	1

Design matrix C was nonsingular; its rank was 43. Consequently, the full 43-term second order response surface model given above in the second section can be fitted. To date, other nonsingular design matrices have not been developed. Therefore, the design matrix C has not been compared against other nonsingular design matrices. The determinant of $(C'C)$ was evaluated, however, and $|C'C| = (3.94) \times 10^{48}$. Also, the variances of the 43 regression coefficients were obtained and are tabulated in Table 8. The variance of b_0 is 125.2. The variances of the regression coefficients of the 8 linear terms range from 1.9 to 16.4; the variances of the regression coefficients of the 6 squared terms range from 0.002 to 0.456. The variances of the regression coefficients of the cross-product terms range from 0.051 to 0.512.

Table 8. Variances of Regression Coefficients

Regression		Regression		Regression	
Coefficient	Variance	Coefficient	Variance	Coefficient	Variance
b_0	125.200	b_{12}	0.194	b_{35}	0.053
b_1	5.172	b_{13}	0.512	b_{36}	0.057
b_2	5.172	b_{14}	0.212	b_{37}	0.057
b_3	1.900	b_{15}	0.212	b_{38}	0.057
b_4	10.350	b_{16}	0.215	b_{45}	0.218
b_5	9.855	b_{17}	0.220	b_{46}	0.230
b_6	16.400	b_{18}	0.220	b_{47}	0.234
b_7	6.838	b_{23}	0.051	b_{48}	0.234
b_8	6.838	b_{24}	0.220	b_{56}	0.220
b_{11}	0.002	b_{25}	0.202	b_{57}	0.225
b_{22}	0.002	b_{26}	0.215	b_{58}	0.225
b_{33}	0.008	b_{27}	0.220	b_{67}	0.239
b_{44}	0.382	b_{28}	0.220	b_{68}	0.239
b_{55}	0.173	b_{34}	0.023	b_{78}	0.239
b_{66}	0.456				

5. CONCLUSIONS. The procedure employed above can be applied to develop second order response surface designs for situations in which the variable levels are prescribed and fixed. However, a systematic method for development of the composite design is needed. The designs attempted suggest that inside variable values should be used in the factorial or fractional factorial portion of the composite design and that the axial points should lie on or outside the k-dimensional cube of the factorial.

REFERENCES

1. Anderson, V. L. and McLean, R. A. (1974), Design of Experiments, Marcel Dekker, Inc., New York.
2. Box, G. E. P. and Draper, N. R. (1959), A Basis for the Selection of a Response Surface Design, Journal of the American Statistical Association, Vol. 54, pp. 622-654.
3. Box, G. E. P. and Hunter, J. S. (1957), Multifactor Experimental Designs for Exploring Response Surfaces, Annals of Mathematical Statistics, Vol. 28, pp. 195-241.
4. Box, G. E. P. and Wilson, K. B. (1951), On the Experimental Attainment of Optimum Conditions, Journal of the Royal Statistical Association, Series B, Vol. 13, pp. 1-45.
5. Cochran, W. G. and Cox, G. M. (1957), Experimental Designs, John Wiley and Sons, Inc., New York.
6. Davies, O. L. (Ed.) (1960), Design and Analysis of Industrial Experiments, Hafner Publishing Company, New York.
7. Hill, W. J. and Hunter, W. G. (1966), A Preview of Response Surface Methodology: A Literature Survey, Technometrics, Vol. 8, pp. 571-590.
8. Mead, R. and Pike, D. J. (1975), A Review of Response Surface Methodology from a Biometric Viewpoint, Biometrics, Vol. 31, pp. 803-851.
9. Myers, R. H. (1971), Response Surface Methodology, Allyn and Bacon, Inc., Boston.

AN APPLICATION OF ORDER STATISTICS TO TIME-SEQUENCE LOGIC

William E. Baker and Malcolm S. Taylor
Probability and Statistics Branch
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

ABSTRACT. If X_1, X_2, \dots, X_n are independent, identically-distributed random variables, then $Y_1 < Y_2 < \dots < Y_n$, where the Y_i 's are the X_i 's rearranged in order of increasing magnitudes, are defined to be the order statistics corresponding to the original random sample. Order statistics have been applied to the solution of a problem involving the determination of time windows for firing impulses in a fuzing system. A computer program has been written which provides the probability of a warhead fusing as a function of the parameters which characterize the detonators. Conversely, given a required probability of fusing, the program will determine the necessary detonator characteristics. Although motivated by this specific problem, the work is general in nature and should have additional applications in the armament research and development community.

1. INTRODUCTION. Let X_1, X_2, \dots, X_n be independent, identically-distributed random variables. Then $Y_1 < Y_2 < \dots < Y_n$, where the Y_i 's are the X_i 's rearranged in order of increasing magnitudes, are defined to be the order statistics corresponding to the original random sample. Order statistics find immediate application in the design and evaluation of logical structures which make decisions based on the relative values assumed by a set of n random variables. One particularly interesting application involves the determination of time windows for firing impulses in a fuzing system. This is the problem which motivated the work on which we are reporting. However, the work is general in nature and may prove useful for other applications in the armament research and development community.

2. STATEMENT OF THE PROBLEM. In the particular problem which we addressed, a fuze contains N detonators, K of which must function within a specific time span. Furthermore, the second detonator (which functions at time Y_2) partitions the time span into two subintervals. The first subinterval $[Y_2 - \delta_1, Y_2]$ is examined to determine if the first detonator functioned within that time segment, and the second subinterval $[Y_2, Y_2 + \delta_2]$ is monitored to count the number of additional detonators activated during that period of time. If, within the time interval $[Y_2 - \delta_1, Y_2 + \delta_2]$, K detonators have functioned, then the command to fire will be initiated;

otherwise, it will not. The times to function for the detonators are random variables and, as such, can be characterized by a cumulative distribution function F . Assuming that the time to function of each detonator is identically distributed, then the problem consists of expressing the probability of fusing as a function of K , N , δ_1 , δ_2 , and F .

3. SOLUTION. Let X_i be the time to function of detonator i in its operating environment. Then X_1, X_2, \dots, X_N are independent, identically-distributed random variables; and we can define Y_1, Y_2, \dots, Y_N to be the order statistics corresponding to the X_i 's. For our problem, if $Y_2 - Y_1 < \delta_1$, we are interested in the probability that $Y_K - Y_2 < \delta_2$; however, if $Y_2 - Y_1 > \delta_1$, we need to determine the probability that $Y_{K+1} - Y_2 < \delta_2$ assuming $K + 1 \leq N$. Therefore, we need to evaluate

$$\begin{aligned} \Pr \{\text{warhead fusing}\} &= \Pr \{Y_2 - Y_1 < \delta_1\} \Pr \{Y_K - Y_2 < \delta_2 \mid Y_2 - Y_1 < \delta_1\} \\ &\quad + \Pr \{Y_2 - Y_1 > \delta_1\} \Pr \{Y_{K+1} - Y_2 < \delta_2 \mid Y_2 - Y_1 > \delta_1\}. \end{aligned} \quad (1)$$

Applying the definition of conditional probability we obtain

$$\begin{aligned} \Pr \{\text{warhead fusing}\} &= \Pr \{Y_2 - Y_1 < \delta_1\} \frac{\Pr \{Y_K - Y_2 < \delta_2 \text{ and } Y_2 - Y_1 < \delta_1\}}{\Pr \{Y_2 - Y_1 < \delta_1\}} \\ &\quad + \Pr \{Y_2 - Y_1 > \delta_1\} \frac{\Pr \{Y_{K+1} - Y_2 < \delta_2 \text{ and } Y_2 - Y_1 > \delta_1\}}{\Pr \{Y_2 - Y_1 > \delta_1\}}, \end{aligned} \quad (2)$$

which upon simplifying yields

$$\begin{aligned} \Pr \{\text{warhead fusing}\} &= \Pr \{Y_K - Y_2 < \delta_2 \text{ and } Y_2 - Y_1 < \delta_1\} \\ &\quad + \Pr \{Y_{K+1} - Y_2 < \delta_2 \text{ and } Y_2 - Y_1 > \delta_1\}. \end{aligned} \quad (3)$$

Defining

$$F(a) = \int_{-\infty}^a f(x)dx, \quad (4)$$

we can proceed to evaluate the first term on the right-hand side of Equation 3. As shown in Appendix A we can obtain the joint probability density function of the 1st, 2nd, and K th order statistics,

$$f(y_1, y_2, y_K) = \frac{N!}{(K-3)!(N-K)!} f(y_1) f(y_2) [F(y_K) - F(y_2)]^{K-3} \\ \cdot f(y_K) [1 - F(y_K)]^{N-K} \quad (5)$$

If we let $u = y_K - y_2$, $v = y_2 - y_1$, and $w = y_1$, then we can rewrite Equation 5,

$$f(u, v, w) = \frac{N!}{(K-3)!(N-K)!} f(w) f(v+w) [F(u+v+w) - F(v+w)]^{K-3} \\ \cdot f(u+v+w) [1 - F(u+v+w)]^{N-K}, \quad (6)$$

and the desired probability is

$$\int_{-\infty}^{+\infty} \int_0^{\delta_1} \int_0^{\delta_2} f(u, v, w) du dv dw \quad (7)$$

which is equal to

$$\frac{N!}{(K-3)!(N-K)!} \int_{-\infty}^{+\infty} f(w) \int_0^{\delta_1} f(v+w) \int_0^{\delta_2} [F(u+v+w) - F(v+w)]^{K-3} \\ \cdot f(u+v+w) [1 - F(u+v+w)]^{N-K} du dv dw. \quad (8)$$

In an analogous manner we can obtain the joint probability density function of the 1st, 2nd, and $K+1$ st order statistics; and, letting $u = y_{K+1} - y_2$, $v = y_2 - y_1$ and $w = y_1$, we can obtain the necessary probability for the second term on the right-hand side of Equation 3. That probability is equal to

$$\frac{N!}{(K-2)!(N-K-1)!} \int_{-\infty}^{+\infty} f(w) \int_{\delta_1}^{+\infty} f(v+w) \int_0^{\delta_2} [F(u+v+w) - F(v+w)]^{K-2} \\ \cdot f(u+v+w) [1 - F(u+v+w)]^{N-K-1} du dv dw. \quad (9)$$

The probability of fuzing is then just the sum of Equation 8 and Equation 9.

Appendix B contains a computer program which evaluates these integrals. In its current form it assumes that the distribution of the times to function of the detonators is normal; however, it can be easily modified to change this distribution. The program requires as input N , K , δ_1 , δ_2 , and σ (standard deviation of the assumed distribution).

4. RESULTS. For the problem we addressed, the value for σ was specified to be 10^{-5} seconds. Figure 1 presents the results of changing δ_1 and δ_2 for a fusing system in which K is equal to $N-1$. In Figure 2 we considered a fuze with eight detonators; and, keeping σ equal to 10^{-5} , we varied K as well as δ_1 and δ_2 . Of course, given any four of the five input variables (σ , K , N , δ_1 , and δ_2), we can obtain a specific probability of warhead fuzing by parametrically varying the remaining variable.

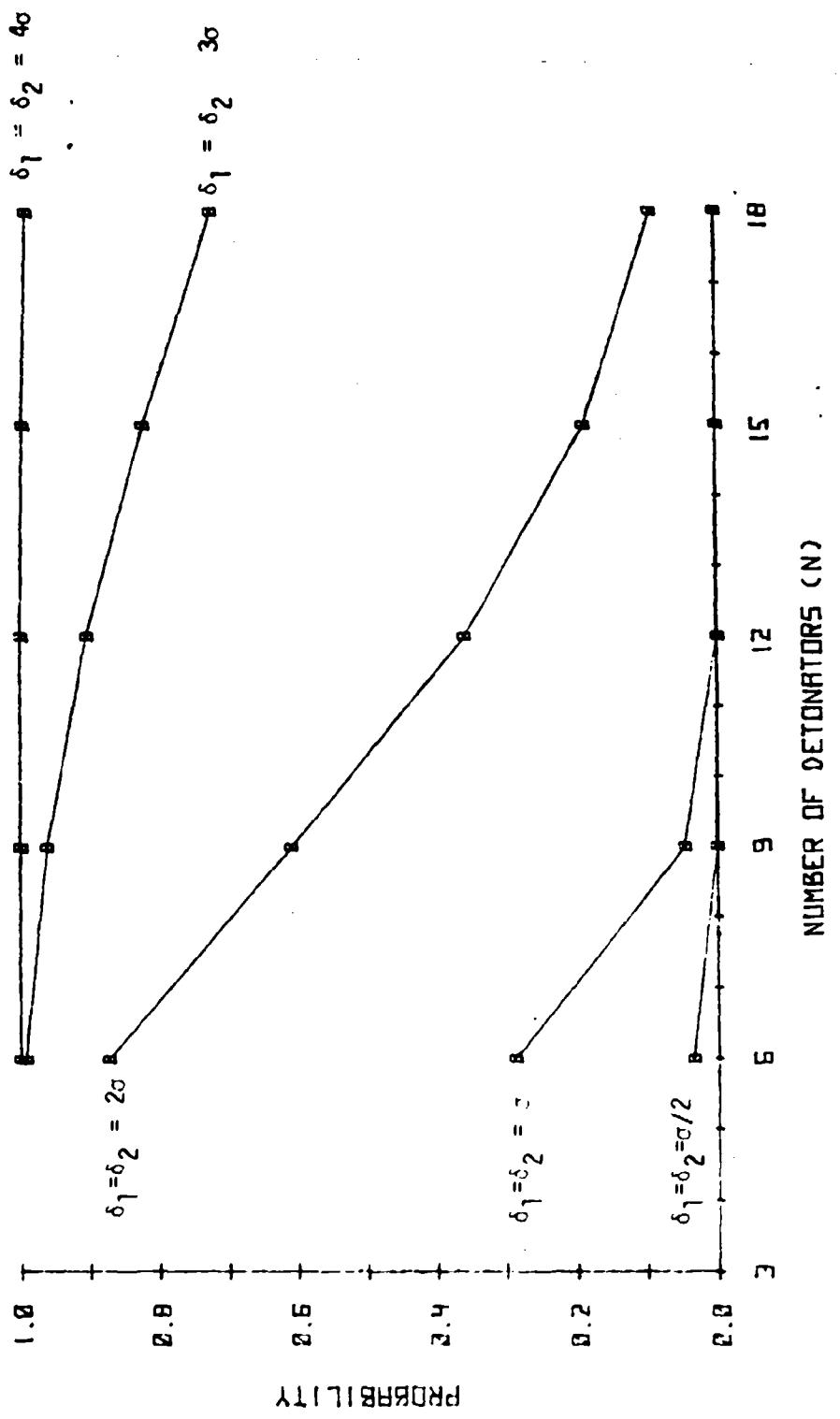


FIGURE 1. PROBABILITY OF WARHEAD FUZING IN AN $N-1$ OUT OF N FUZING SYSTEM ($\sigma = .00001$).

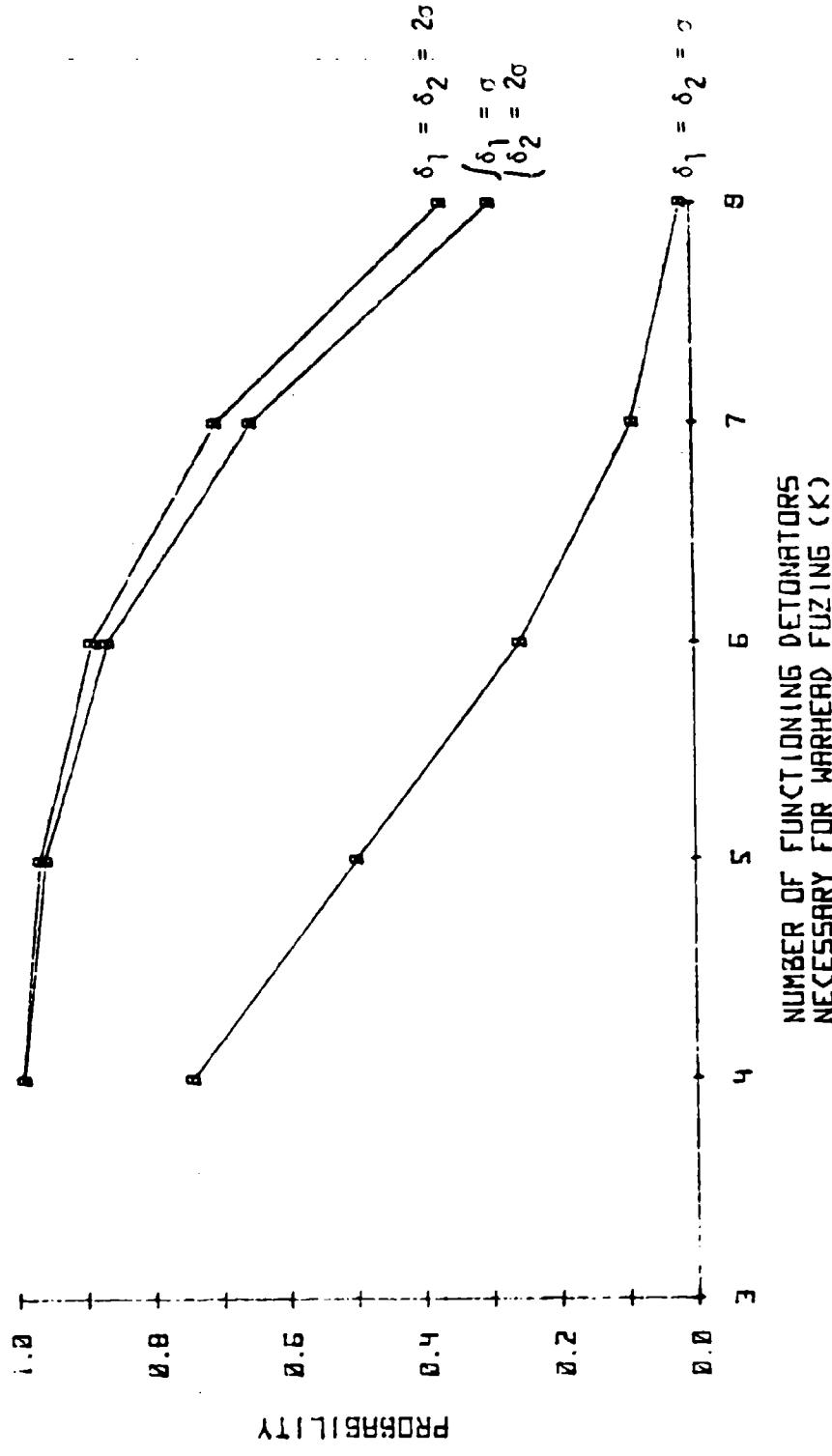
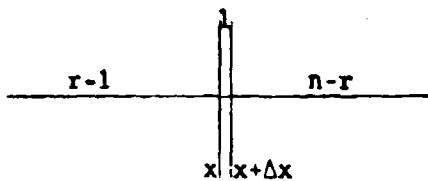


FIGURE 2. PROBABILITY OF WARHEAD FUZING IN A K OUT OF 8 FUZING SYSTEM ($\sigma = .00001$).

APPENDIX A. As shown in Reference 1 we can obtain the joint probability density function for Y_1, Y_2, \dots, Y_n . Let X_1, X_2, \dots, X_n be a random sample from a population with density $f(x) = F'(x)$; and let Y_1, Y_2, \dots, Y_n be the corresponding order statistics. Then the probability density function for the r th order statistic may be derived by considering the following configuration:



That is, $X_i < x$ for $r-1$ of the X_i , $x < X_i < x + \Delta x$ for one X_i , and $X_i > x + \Delta x$ for the remaining $n-r$ of the X_i . The number of ways this combination of events can occur is

$$\frac{n!}{(r-1)! 1! (n-r)!}$$

and each such way has probability

$$[F(x)]^{r-1} [F(x+\Delta x) - F(x)]^1 [1 - F(x+\Delta x)]^{n-r}.$$

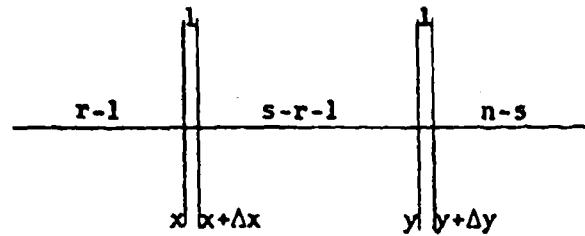
Therefore, we have

$$\Pr \{x < Y_r \leq x + \Delta x\} = \frac{n!}{(r-1)! (n-r)!} \cdot [F(x)]^{r-1} [F(x+\Delta x) - F(x)] [1 - F(x+\Delta x)]^{n-r} + O(\Delta x^2) \quad (\text{A1})$$

where $O(\Delta x^2)$ means terms of order $(\Delta x)^2$ and includes the probability of realizations of $x < Y_r \leq x + \Delta x$ in which more than one X_i is in $(x, x + \Delta x)$. Dividing both sides of Equation A1 by Δx and then letting $\Delta x \rightarrow 0$, we obtain

$$f_{Y_r}(x) = \frac{n!}{(r-1)! (n-r)!} [F(x)]^{r-1} f(x) [1 - F(x)]^{n-r}. \quad (\text{A2})$$

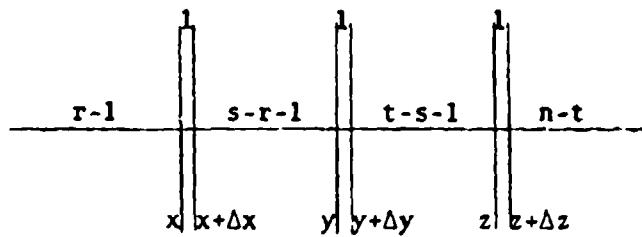
In a similar manner we can derive the joint probability density function of Y_r and Y_s :



and through an analogous argument we obtain

$$f_{Y_r, Y_s}(x, y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} [F(x)]^{r-1} f(x) \\ \cdot [F(y) - F(x)]^{s-r-1} f(y) [1 - F(y)]^{n-s}. \quad (A3)$$

Finally, for the joint probability density function of Y_r , Y_s , and Y_t :



and

$$f_{Y_r, Y_s, Y_t}(x, y, z) = \frac{n!}{(r-1)!(s-r-1)!(t-s-1)!(n-t)!} [F(x)]^{r-1} f(x) \\ \cdot [F(y) - F(x)]^{s-r-1} f(y) [F(z) - F(y)]^{t-s-1} f(z) [1 - F(z)]^{n-t}. \quad (A4)$$

For the case $r=1$, $s=2$, and $t=k$, we obtain

$$f_{Y_1, Y_2, Y_k}(x, y, z) = \frac{n!}{(k-3)!(n-k)!} f(x) f(y) \\ \cdot [F(z) - F(y)]^{k-3} f(z) [1 - F(z)]^{n-k}. \quad (A5)$$

APPENDIX B. We are presenting here the computer program which evaluates the desired probability function. As noted in the program comments, we have assumed that the times to function of the detonators are normally distributed with mean zero and variance σ^2 . However, with just a few changes to the subroutines, a different distribution may be assumed. To do this, all cards containing "NORMAL" in columns 74 through 79 must be replaced by others with the appropriate probability density function or cumulative distribution function.

PROGRAM DETUN (INPUT,OUTHUT,TAPES=INPUT,TAPE6=OUTPUT)

THIS PROGRAM DETERMINES THE PROBABILITY THAT K OUT OF N
DETONATORS WILL FUNCTION WITHIN A GIVEN TIMESSPAN. K MUST
BE GREATER THAN OR EQUAL TO 3, AND K MUST BE LESS THAN OR
EQUAL TO N. THE DISTRIBUTION OF THEIR FUNCTIONING IS ASSUMED
TO BE GAUSSIAN WITH A MEAN EQUAL TO 0.0 AND A STANDARD
DEVIATION EQUAL TO SIGMA. THE SOLUTION IS DERIVED THROUGH
THE USE OF ORDER STATISTICS AND IS OBTAINED BY EVALUATING
A TRIPLE INTEGRAL.

IF THE NUMBER OF DETONATORS IS GREATER THAN THIRTY,
THEN THE TOLERANCE LIMITS OF THE INTEGRALS MUST BE REDEFINED.
THAT IS, THE VARIABLE 'ERROR' IN THE MAIN ROUTINE AND THE
VARIABLES 'TOLX,TOLY' IN SUBROUTINE 'DIST' SHOULD BE ADJUSTED.
WITH THE TOLERANCE LIMITS CURRENTLY IN THE PROGRAM,
THE RESULTING PROBABILITIES ARE CORRECT TO TWO DECIMAL PLACES.

INPUT IS AS FOLLOWS

N NUMBER OF DETONATORS
K NUMBER OF DETONATORS TO FUNCTION
SIGMA STANDARD DEVIATION OF DISTRIBUTION
DELT1 TIMESPAN (FIRST TO SECOND DETONATORS)
DELT2 TIMESPAN (SECOND TO LAST DETONATORS)

REAL NFACT,K2FACT,K3FACT,NKFACT,NK1FACT

DIMENSION IERR(6)

COMMON /COM1/ N,K,SIGMA,DELT1,DELT2
COMMON /COM2/ PI,W
COMMON /COM6/ IND

EXTERNAL DIST

DATA IERR /3*(-0),0,2*(-0)/

CALL SYSTEMC (34,IERR)
CALL SYSTEMC (115,IERR)
WRITE (6,200)

READ INPUT
INITIALIZE VARIABLES

5 READ (5,100) N,K,SIGMA,DELT1,DELT2
IF (K.LT.3 .OR. K.GT.N) GO TO 15
IF (N .GT. 30) WRITE (6,500) N
IF (DELT1 .GT. 10.*SIGMA) DELT1=10.*SIGMA
IF (DELT2 .GT. 10.*SIGMA) DELT2=10.*SIGMA
RES1=0.
RES2=0.
PI=3.1415926536
NN=N/5+6

```

IF (N .EQ. 50) NN=12
ERR0H=10.**(-NN)
A=-10.*SIGMA
B=+10.*SIGMA

C C C
      EVALUATE TRIPLE INTEGRAL
C C C

IN0=1
RES1=SQUANK (A,B,ERROR,RUM,DIST)
IF (N .EQ. K) GO TO 7
IN0=2
RES2=SQUANK (A,B,ERROR,RUM,DIST)

C C C
      DETERMINE FACTORIALS

7 CONTINUE
NFACT=1
K2FACT=1
K3FACT=1
NKFACT=1
NK1FACT=1
DO 10 I=1,N
NFACT=NFACT*I
IF (I .LE. (K-2)) K2FACT=K2FACT*I
IF (I .LE. (K-3)) K3FACT=K3FACT*I
IF (I .LE. (N-K)) NKFACT=NKFACT*I
IF (I .LE. (N-K-1)) NK1FACT=NK1FACT*I
10 CONTINUE

C C C
      COMPUTE AND WRITE PROBABILITIES
C C C

PROB1=NFACT/(NKFACT*K3FACT)*RES1
PROB2=NFACT/(NK1FACT*K2FACT)*RES2
PROB=PROB1+PROB2
WRITE (6,300) N,K,SIGMA,DELT1,DELT2,PROB1,PROB2,PROB
GO TO 5
15 WRITE (6,400) N,K
STOP

C
100 FORMAT (2I5,3F20.7)
200 FORMAT (1H1//)
300 FORMAT (1H ,4MN = ,I2.3X,6MK = ,I2.3X,8MSIGMA = ,F10.7,3X,
          *     8MDELT1 = ,F15.8,3X,8MDELT2 = ,F15.8,3X/1M0,
          *     37MPROBABILITY (1 THROUGH K FUNCTION) = ,F10.6/1M ,
          *     39MPROBABILITY (2 THROUGH K+1 FUNCTION) = ,F8.6/1M ,
          *     22MPROBABILITY (TOTAL) = ,F25.6/////
400 FORMAT (1H ,35H***** INVALID VALUE OF N OR K ***** ,3X,
          *     4MN = ,I2.3X,6MK = ,I2)
500 FORMAT (21H ***** WARNING - N = ,I2,
          *     51H, CHECK THE TOLERANCE LIMITS OF THE INTEGRALS *****//)

C C C
      END

```

```

C FUNCTION DIST (W)
C -----
C
C      THIS ROUTINE PROVIDES THE INTEGRAND FOR THE THIRD INTEGRAL
C      AS WELL AS THE LIMITS OF INTEGRATION FOR THE SECOND INTEGRAL.
C
C COMMON /COM1/ N,K,SIGMA,DELT1,DELT2
C COMMON /COM2/ PI,W
C COMMON /COM4/ IND
C
C      EXTERNAL FXX,FXY
C
C      WWWW
C      NN=N/5+6
C      IF (N .EQ. 50) NN=12
C      TOLX=10.**(-NN)
C      TOLY=10.**(-NN)
C      PHIW=1./(SQRT(2.*PI)*SIGMA)*EXP(-1.*W**2/(2.*SIGMA**2))      NORMAL
C
C      IF (IND .EQ. 2) GO TO 5
C      DOWN=0.
C      UP=DELT1
C      GO TO 10
C 5  DOWN=DELT1
C      UP=2.*DELT1
C      IF (DELT1 .LT. 10.*SIGMA) UP=10.*SIGMA
C
C 10 CALL DBLINT (DOWN,UP,FXX,FXY,TOLX,TOLY,ANS,RUM,RUMYH)
C      DIST=PHIW*ANS
C      RETURN
C      END
C
C
C      SUBROUTINE FXX (V,Y1,Y2,Y3)
C -----
C
C      THIS ROUTINE PROVIDES THE INTEGRAND FOR THE SECOND INTEGRAL
C      AS WELL AS THE LIMITS OF INTEGRATION FOR THE FIRST INTEGRAL.
C
C COMMON /COM1/ N,K,SIGMA,DELT1,DELT2
C COMMON /COM2/ PI,W
C      PHIW=1./(SQRT(2.*PI)*SIGMA)*EXP(-1.* (V-W)**2/(2.*SIGMA**2))      NORMAL
C
C      Y1=PHIW
C      Y2=0.
C      Y3=DELT2
C      RETURN
C      END
C
C

```

```

FUNCTION FAY (V,U)
-----
C THIS ROUTINE PROVIDES THE INTEGRAND FOR THE FIRST INTEGRAL.
C
COMMON /COM1/ N,K,SIGMA,DELT1,DELT2
COMMON /COM2/ PI,W
COMMON /COM4/ IND
C
PHIUVW=1.0/(SQRT(2.0*PI)*SIGMA)*EXP(-1.0*(U+V+W)**2/(2.0*SIGMA**2))      NORMAL
CAPVW=FND ((V+W)/SIGMA)                                              NORMAL
CAPUVW=FND ((U+V+W)/SIGMA)                                              NORMAL
IF (IND .EQ. 2) GO TO 15
C
IF (IN .EQ. 3) GO TO 2
IF (IN .EQ. K) GO TO 5
IF (K .EQ. 3) GO TO 10
FXY=(CAPUVW-CAPVW)**(K-3)*PHIUVW*(1.0-CAPUVW)**(N-K)
GO TO 25
2 FXY=PHIUVW
GO TO 25
5 FXY=(CAPUVW-CAPVW)**(K-3)*PHIUVW
GO TO 25
10 FXY=PHIUVW*(1.0-CAPUVW)**(N-K)
GO TO 25
C
15 CONTINUE
C
IF (IN .EQ. (K+1)) GO TO 20
FXY=(CAPUVW-CAPVW)**(K-2)*PHIUVW*(1.0-CAPUVW)**(N-K-1)
GO TO 25
20 FXY=(CAPUVW-CAPVW)**(K-2)*PHIUVW
C
25 RETURN
END
C
C

```

ACKNOWLEDGEMENTS. We would like to acknowledge Prof. H. A. David of Iowa State University for his personal correspondence directed toward the evaluation of the joint distribution of order statistics; also, Mr. Denis Silvia of the Ballistic Research Laboratory for providing the problem which prompted this work.

REFERENCES

1. David, H. A.; Order Statistics; John Wiley and Sons; New York; 1970.
2. Mood, A. M., Graybill, F. A., Boes, D. C.; Introduction to the Theory of Statistics; McGraw-Hill, Inc.; New York; 1974.

RADAR ERROR SIGNAL IMPROVEMENT

Robert E. Green
Programs Management Office
Instrumentation Directorate
US Army White Sands Missile Range
White Sands Missile Range, New Mexico

ABSTRACT. Monopulse tracking radars are subject to pointing errors that are induced by target-caused signal fluctuations. These signal fluctuations introduce non-Gaussian noise into the angle tracking servo error signals of the radar. This paper raises the question of the efficacy of making corrections for these errors based on other radar measurements that are significantly corrupted by noise. Actual radar error signals are displayed along with the results of spectral analysis of the signals. The spectral analysis confirms the presence of non-Gaussian components in the error signals.

I. INTRODUCTION. One of the types of devices used at White Sands Missile Range to keep track of missiles in flight is an instrumentation radar. These devices transmit a burst of energy which is reflected off the target and back to the radar. Automatic control devices in the radar keep the antenna pointed at the target. The measurement of the time interval from burst transmission to echo reception permits the determination of target range; and the position of the antenna pedestal in azimuth and elevation permit the designation of target position in polar coordinates. The radars transmit these bursts of energy at rates of 160, 320, or 640 times per second. The process of tracking a target automatically requires that the device (radar) sense how far it is off from the target and make the necessary corrections. The purpose of this paper is to define a problem that has been detected in this error sensing circuitry and solicit suggestions for improving radar system performance.

II. ERROR SIGNALS. Instrumentation radars of the type used at White Sands Missile Range utilize a monopulse feed system to generate the error signals that are used to direct the tracking mount in azimuth and elevation. This monopulse feed system uses a quadrangle of four sensors (Figure 1) to determine the necessary direction to drive the antenna so that the target is centered in the quadrangle.

A ° ° E

60 ° E

Figure 1. Sensor Quadrangle.

The directional error is measured by observing the difference of the following signal levels:

$$\text{Azimuth error} = (A + C) - (B + D) \quad (1)$$

$$\text{Elevation error} = (A + B) - (C + D) \quad (2)$$

These sums and differences are formed at the operating radio frequencies (5.5 GHz) of the radar. This means that the individual signals from A, B, C and D are not available for processing. In the radar equipment, these quantities (1) and (2) are sensed as voltages instead of digital numbers. These quantities are sensed each time the radar receives an echo. The signals differenced to form the quantities, (1) and (2) are of almost equal magnitude. This results in a very weak, low frequency signal, embedded in a large amount of noise. The radar equipment uses a very narrow band filtering process to extract the signal from the noise. This filtering process is accomplished in two stages. The data is first processed through a low pass filter of approximately 10 Hz bandwidth. The filtered signal is then applied to the servo which typically has a bandwidth of approximately 3 Hz. It is not possible to significantly reduce the bandwidth of the 10 Hz filter and maintain the stability of the servomechanism.

The noise in the error signals has two major contributing sources. One of these is "thermal noise" and is naturally occurring in the environment. It is assumed to have a Gaussian distribution with zero mean. The other major noise source is contributed by the response of the radar system to changes in the reflectivity pattern of the target. The reflectivity pattern of the target is a very complex function that changes rapidly with changes in target aspect. This means that signal amplitude can change drastically between two successive echos from the same target. These rapid signal fluctuations are the origin of the second major noise source.

III. AUTOMATIC GAIN CONTROL. The gain of the radar receiver performs the same function as the volume control on a radio. As the signal gets weaker the volume or gain is increased to keep the output at a constant level. In a radar system, an automatic control system is used to sense the average received signal level and adjust the receiver gain accordingly. If the signal level does not change too rapidly, the automatic gain control system maintains the receiver gain at the correct level so that the radar system functions normally. If the received signal level changes faster than the automatic gain control can adjust, then the noise is introduced into the error signals. This noise is deterministic in the sense that if the receiver gain setting and the received signal level are known, then the incorrect value for the error signal can be predicted, using a known deterministic function. The previous paragraph indicated that the radar system is designed to respond only to low frequencies. This would make it appear that high frequencies would have no effect on the system. It should be noted that the radar is a sampled data system and that frequencies near the sampling frequency will appear to be low frequencies due to aliasing.

IV. ERROR GRADIENT CURVES. The error gradient curve is used to relate the magnitude of the error to the voltage sensed (Figure 2). This function is established as a part of the radar system set-up procedure. It is also used in the tracking function to sense how far the target is off center. A common set-up would be to assign a deviation of one milliradian a value of one volt. Notice that the curve is linear in the region where the deviation does not exceed ± 1 milliradian. This is the region where the radar system would normally be expected to operate. In the radar set-up procedure, the deviations shown in Figure 2 are assigned for the voltages sensed in the azimuth and elevation error signal detection circuitry. These values are correct, as long as the receiver gain is set at the appropriate level for the received signal.

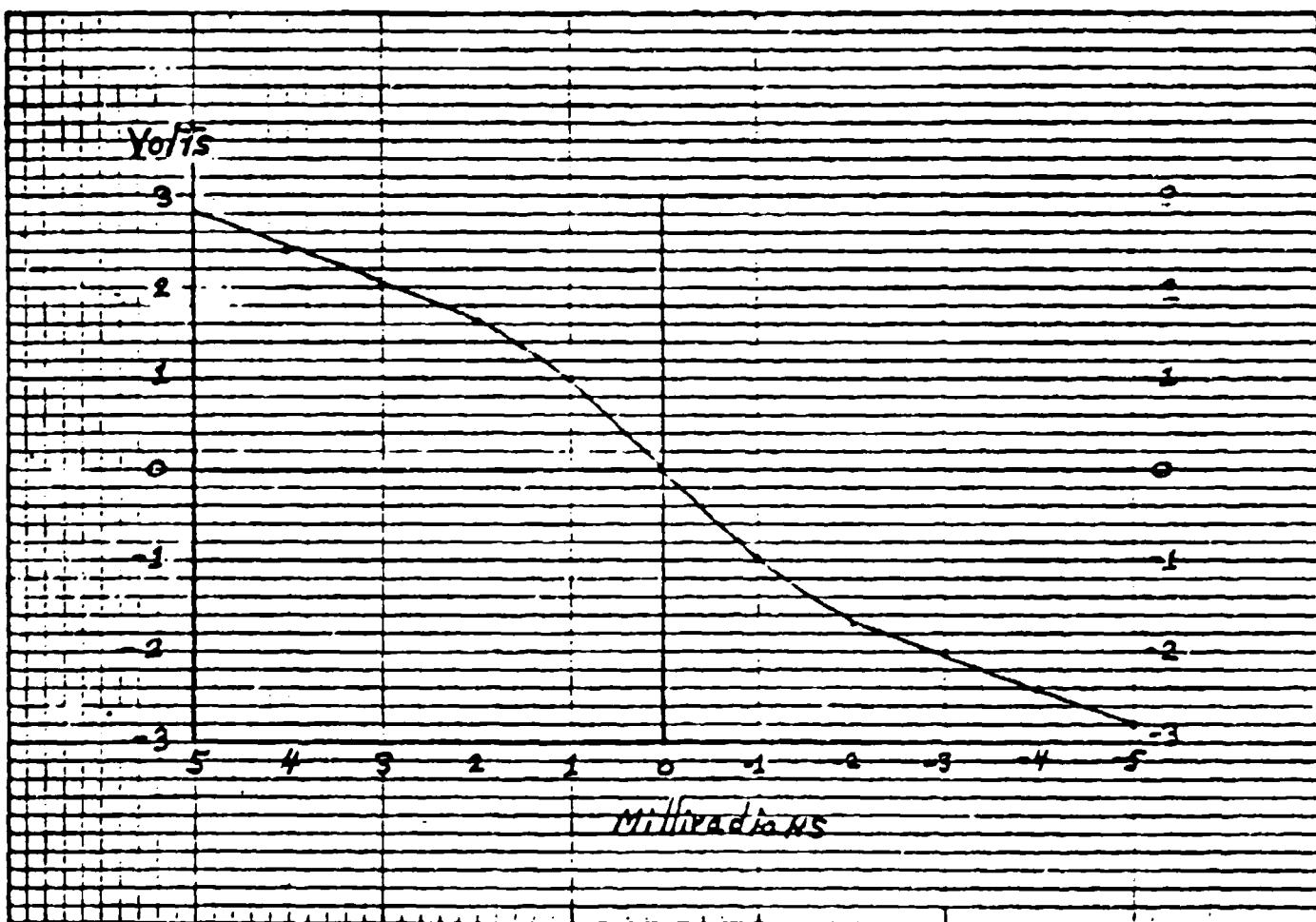


Figure 2. Error gradient curve

The automatic gain control system sets the receiver gain based on the average value of the last few signals received. When a received signal has a significantly different amplitude from this average, it has a profound effect on the error sensing circuitry of the radar system. Figure 3 illustrates how the error gradient curve appears to the error-sensing circuitry when the received signal is either significantly stronger or significantly weaker than the average value over the last few samples. Curve A illustrates how the error sensor reacts if the received signal is much stronger than expected. In this case, the voltage sensed for a given angular deviation is much larger than that shown in Figure 2. Curve B illustrates how the sensor reacts if the received signal is much weaker than expected. For this condition, the voltage sensed for a given angular deviation is much smaller than shown in Figure 2. Since Figure 2 was used to calibrate the system, these conditions introduce errors in the sensed angular deviations. In actual experience this type of signal fluctuation occurs frequently in tracking targets such as missiles and aircraft.

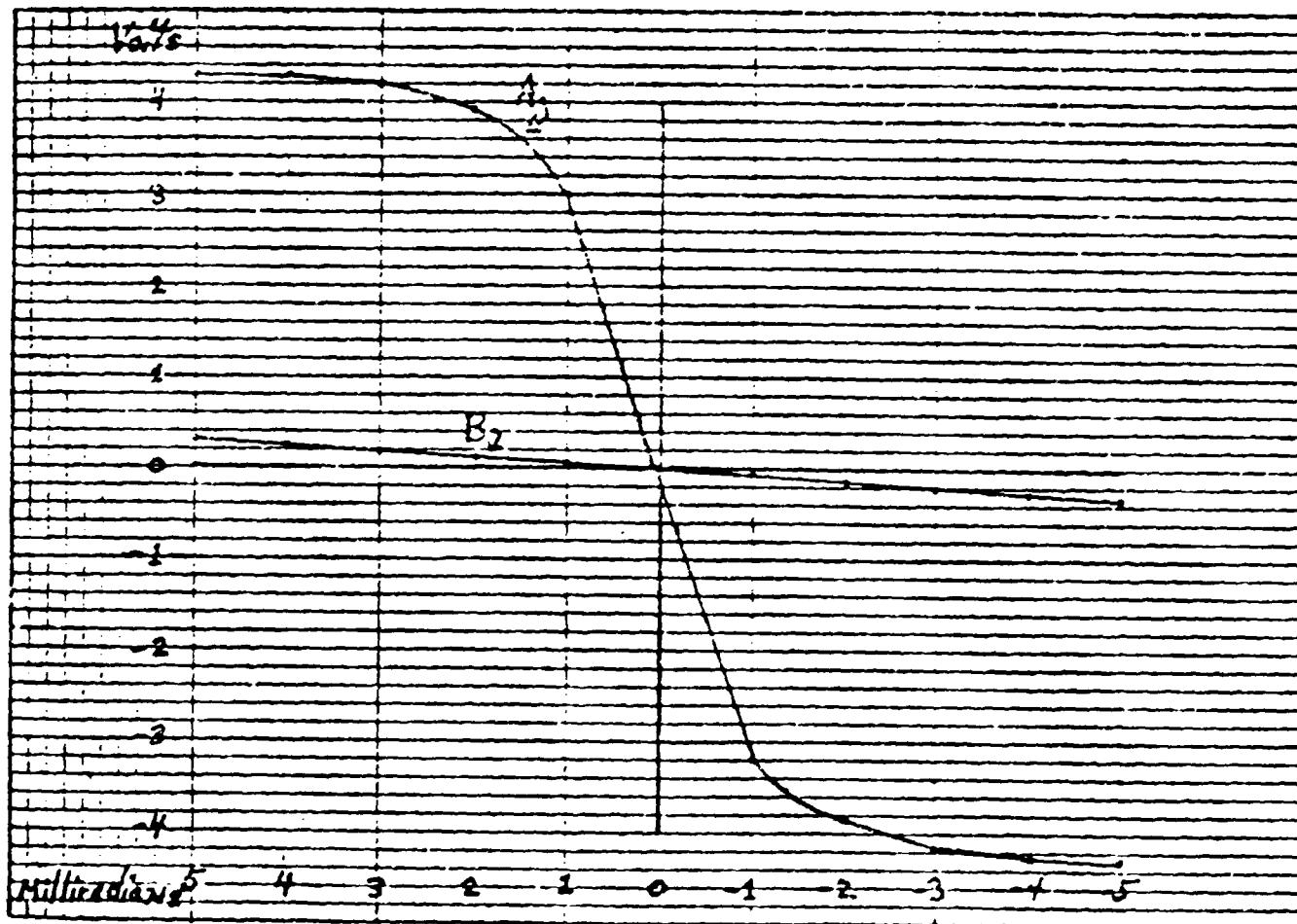


Figure 3. Mismatched error gradient curves.

V. SPECTRAL ANALYSIS. The introduction of a large signal fluctuation into the error sensing circuitry of the radar affects the system for a period of time. This is due to the action of the filtering process. The introduction of an impulse into the system lasting only one sample period would at first appear to be at too high of a frequency to affect the system. If the amplitude of the impulse is very large its effect will be spread over several samples by system filtering action resulting in antenna pointing errors. Figure 4 shows the power spectral density of a radar error signal when the radar was tracking a spinning missile. Note there is power at low frequencies near the servo bandwidth of the instrument. Such non-Gaussian noise will create incorrect responses in the radar angle tracking system. Figure 5 shows similar spectral analysis for a roll stable missile. Notice that the non-Gaussian error signals occur much less frequently in this case. This indicates that rapid changes in reflectivity pattern are much less common for non-spinning missiles. Figure 6 is a recording of radar error signals where large signal fluctuations were known to occur. The large angle errors resulting from these signal variations are evident throughout the period shown. The data presented indicate that the presence of large short-term signal fluctuations do affect the radar angle error signals.

VI. THE PROBLEM OF CORRECTION. The composite of the Gaussian and deterministic noise sources results in a function such as the one shown in Figure 6. Observe that this data is so noisy that no trend can be discerned by inspection. The available measurements of received signal level are also very noisy. In spite of this, it is possible to generate the required set of error gradient curves. The necessary values can be generated by fixing the radar parameters and then observing the data over a few hundred samples. The averages of these samples have the expected characteristics. The individual samples of received signal level are so corrupted by noise that the correction of individual error signal samples may not result in an improved value.

The information presented raises the following questions:

- ° Will applying corrections, based on functions derived from average values, produce a better behaved sequence of error signals?
- ° Is spectral analysis an adequate method of measuring the improvement resulting from the correction process?
- ° How noisy must the measurement of signal level become in order to make the correction process ineffective?

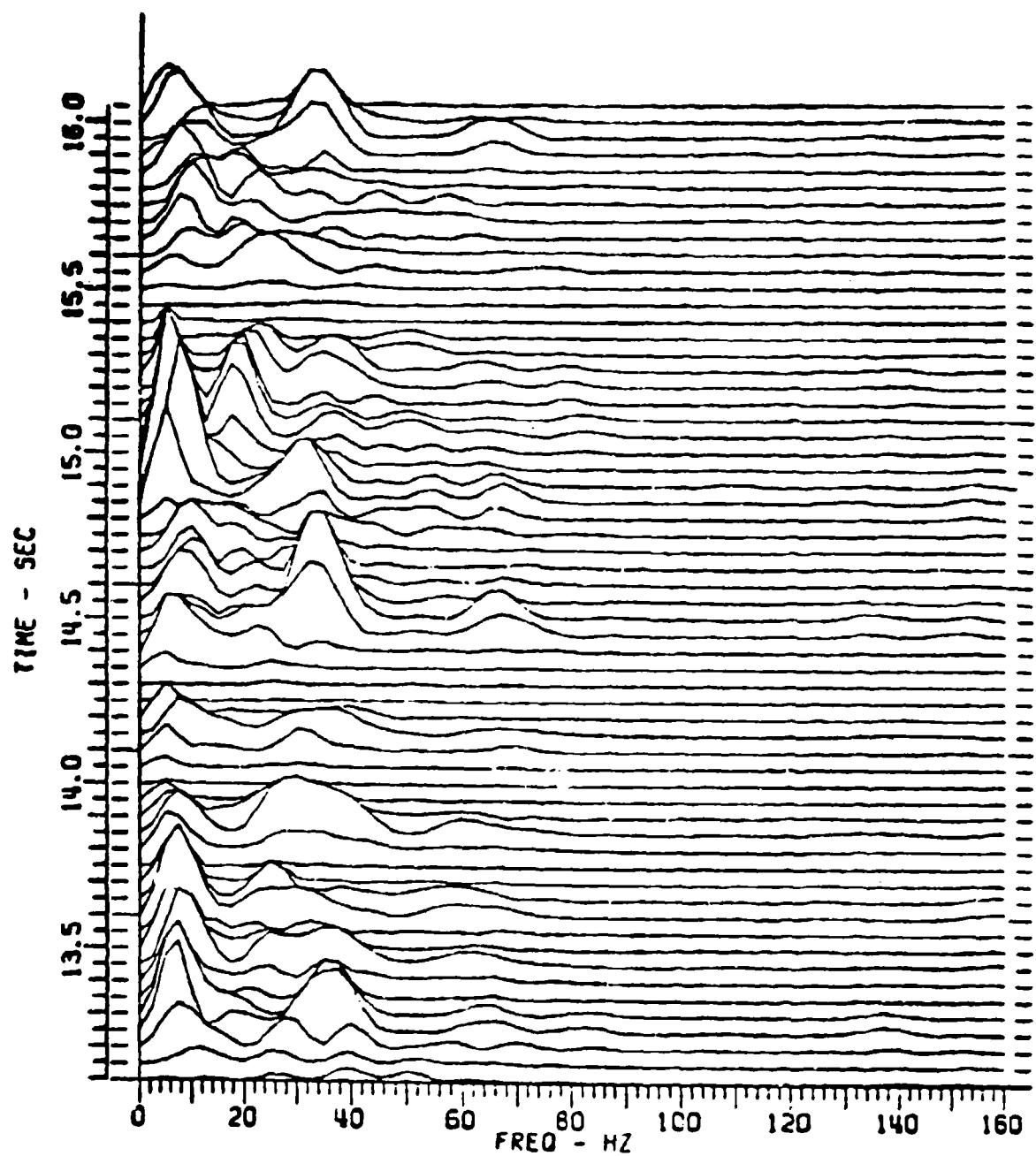


Figure 4. Spinning missile error signal power spectral density.

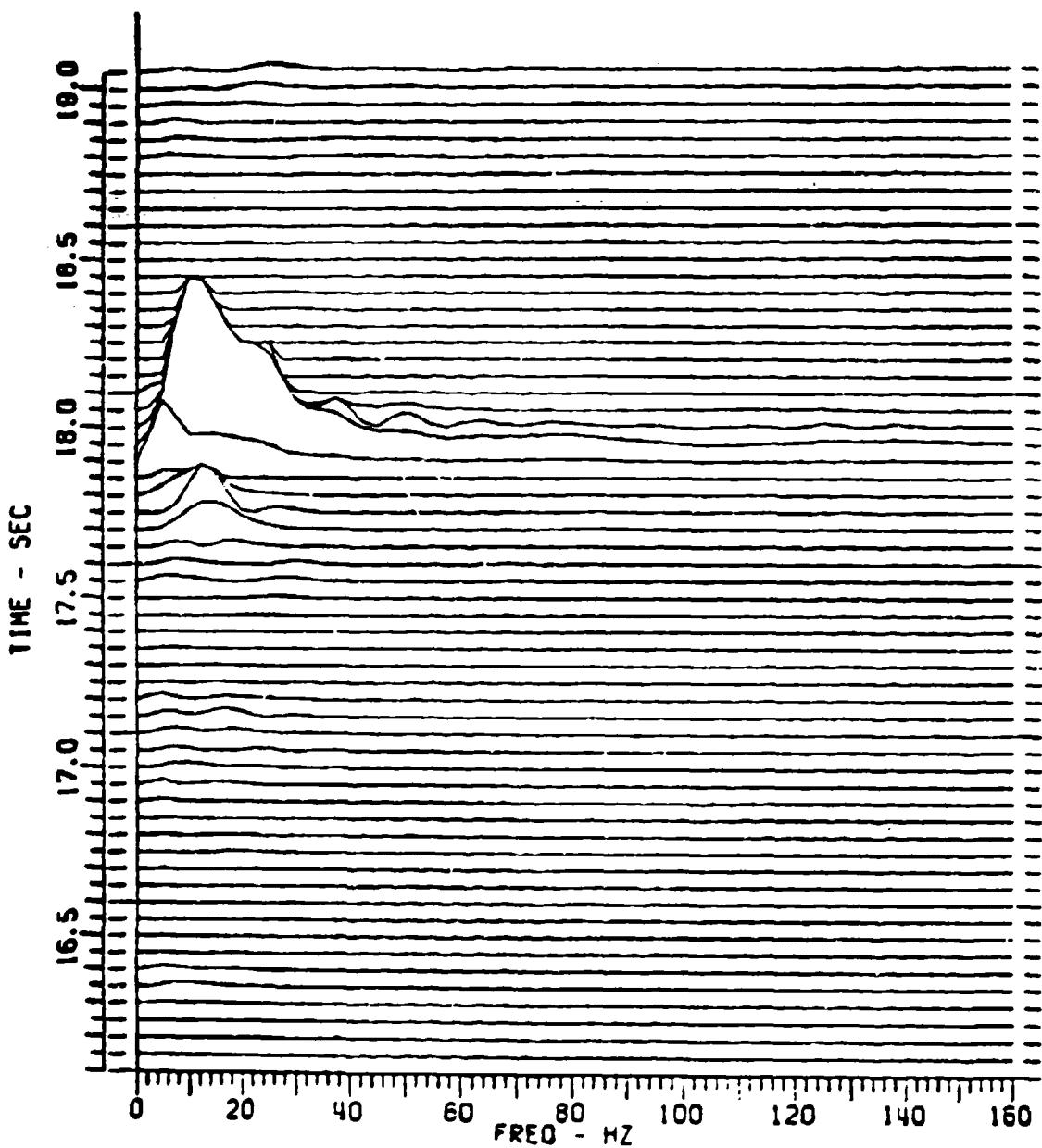


Figure 5. Roll stable missile error signal power spectral density.

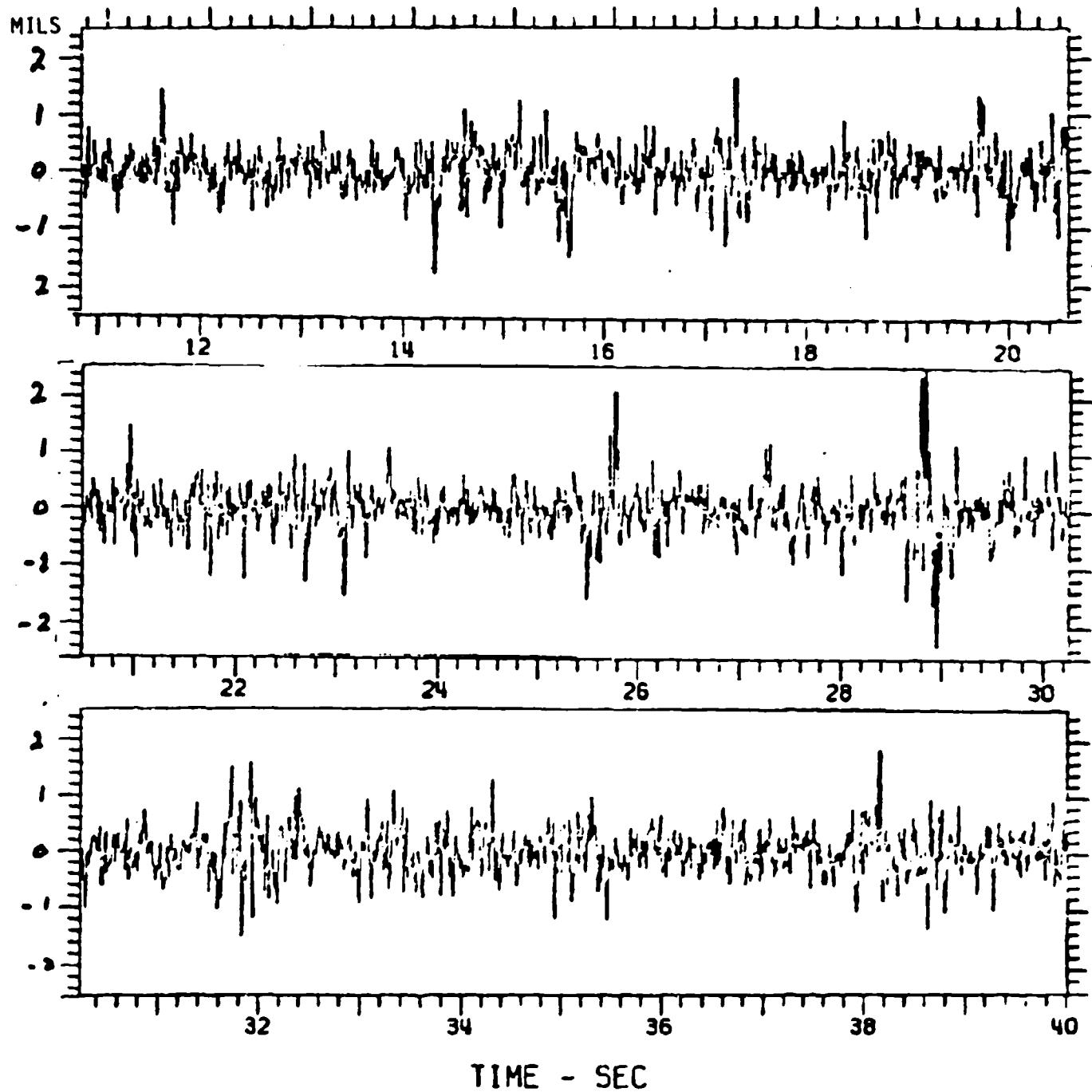


Figure 6. Radar Error signals.

USE OF THE BILINEAR Z-TRANSFORM IN IMPLEMENTING DIGITAL FILTERS

Donald W. Rankin
Army Materiel Test and Evaluation Directorate
US Army White Sands Missile Range
White Sands Missile Range, NM 88002

ABSTRACT

The Laplace transform is an extremely versatile tool for solving differential equations. The s-plane transfer function converts the problem in integration to an algebraic one. But when a control system employs digital filters in an embedded computer, the variables are necessarily discrete, and the problem is better stated by means of a difference equation. The operator equivalence relationships are

$$Z = 1 + \Delta = e^D$$

and show that the s-plane transfer functions will be quite complicated.

To circumvent this difficulty, we employ the Bilinear Z-transform which has the form

$$\frac{1}{A} \cdot \frac{F}{\phi} = \frac{1 + BZ^{-1}}{1 + CZ^{-1}}$$

and maps into what we shall call the r-plane, where the transfer functions are well behaved. It is a most useful tool, admirably performing its mission of simplifying calculations, but seems to be seldom used --- rarely correctly.

This paper re-examines the theory of the Bilinear Z-transform utilizing two new parameters (essentially the reciprocals of those traditionally used). It is felt that the method results in considerable simplification and clarification.

I. BACKGROUND. During the test of a ballistic missile control system, an analysis was performed of the digital filters used in the attitude compensation channels. The filters were implemented by cascading two, three or four stages of the type

$$\frac{F}{\phi} = \frac{(a + \beta Z^{-1})x}{(1 + \gamma Z^{-1})x} \quad (1)$$

This filter is implemented in a digital computer by the following two successive steps:

$$x_1 = \phi_1 - \gamma x_0 \quad (2)$$

$$F_1 = \alpha x_1 + \beta x_0 \quad (3)$$

Subscripts refer to increments of time; hence the sampling interval is given by

$$t_1 - t_0 = \Delta t$$

χ_1 is merely a convenient computational parameter.

In the actual case, only real coefficients were encountered. However, the treatment which follows can be extended without much difficulty to include the case of complex coefficients.

Several shortcomings were found. The proximate cause is easy to state --- incorrectly computed coefficients and too many stages are examples. The ultimate cause, however, cannot be determined with any certainty. Perhaps it is a lack of understanding of the principles of digital filters on the part of both design and test engineers.

The theory of digital filters is the theory of electrical networks. It is not surprising, then, that the bulk of the literature on the subject has been written by electrical engineers, and is couched in engineering terms. The frequency domain is the vehicle of thought.

But when a linear control system is operated by an embedded digital computer, the signal to be sampled often is virtually aperiodic in nature, exhibiting a frequency spectrum that is quite primitive when compared with that of even a common electrical phenomenon. Under these conditions, restating matters in the time domain results in worthwhile simplification and clarification.

Accordingly, the theory is re-investigated, utilizing the terminology of the time domain. Two variables are identified there, and expressed in units of time (e.g., seconds or sampling intervals).

II. THE PROBLEM OF FREQUENCY FOLDING.* Nearly everyone has seen in the movies the spoked wheel which, starting from rest, turns faster and faster until a speed is reached (the Nyquist frequency) where the spokes appear to slow down, eventually (at twice the Nyquist frequency) coming to an apparent halt. As the actual wheel speed continues to increase, the spokes seem to turn backward, and the phenomenon is repeated in mirror image. In fact, it is repeated indefinitely, like the images in a barber shop mirror. The frequencies are said to be "folded," the folds occurring at odd multiples of the Nyquist frequency.

The Nyquist frequency is equal to half the sampling frequency. Thus, if the movie projector operates at 24 frames per second, the Nyquist frequency is 12 spokes per second. It is convenient for theoretical purposes to express frequencies in radians per second, rather than in hertz. For a finite sampling interval of Δt seconds, then, the radian sampling frequency is $2\pi/\Delta t$ and the Nyquist frequency $\pi/\Delta t$.

*J. W. Tukey employed the term "aliasing."

If the spectrum of a signal to be sampled contains frequencies (of sufficient amplitude to be detected) greater than $\pi/\Delta t$, frequency folding will surely occur. Having occurred, it can neither be detected in nor removed from the sample. Necessarily, steps to be taken are limited to preventive ones. Three cases are discussed:

Case 1. It may be possible to decrease Δt , thereby increasing the Nyquist frequency until it spans the troublesome frequencies.

Case 2. Unwanted frequencies can be removed by processing the signal with a suitable band pass filter before sampling. This will result in a loss of "power," which perhaps can be partially compensated for in the subsequent digital filter.

Case 3. Analysis of the output of the plant (signal) may reveal that it is aperiodic. The principal frequency thus is zero.

In all three cases, the function (signal) is said to be band limited, since its frequency spectrum contains no frequencies outside the band defined by the Nyquist frequency; i.e., $-\pi/\Delta t < \omega < \pi/\Delta t$.

This paper will treat only band limited functions. However, this does not mean that the specter of frequency folding can be ignored, since either a poor filter design or a faulty feed-back mechanism can induce periodicity.

III. THE BILINEAR TRANSFORMATION. In its most general form, the bilinear transformation is given by

$$Awz + Bw + Cz + D = 0 \quad (4)$$

where A, B, C and D are constants while w and z are variables, any of which might be complex. Avoiding the trivial case $A = 0$, division by A obviously does not disturb the equality. Simplifying thus, let

$$\alpha = -\frac{C}{A}; \beta = -\frac{D}{A}; \gamma = \frac{B}{A}$$

Then

$$wz + \gamma w = \alpha z + \beta \quad (5)$$

$$w + \gamma wz^{-1} = \alpha + \beta z^{-1}$$

$$w = \frac{\alpha + \beta z^{-1}}{1 + \gamma z^{-1}} \quad (6)$$

It is equally easy to solve for either z or z^{-1} in terms of w. This demonstrates that the inverse of a bilinear transformation also is a bilinear transformation.

The operators $\frac{d}{dt}$, $Z = 1 + \Delta$, and $Z^{-1} = 1 - \nabla$ submit to algebraic manipulation, so that

$$w = \frac{(a + \beta Z^{-1})x}{(1 + \gamma Z^{-1})x} \quad (1)$$

becomes a useful digital filter. The variable x is ANY parameter that can be sampled in the computer.

As will be seen, the filter provides an approximate solution to the differential equation

$$\frac{dx}{dt} = f(x, t)$$

since

$$Z = 1 + \Delta = \exp\left(\frac{d}{dt}\right)$$

from which

$$\frac{d}{dt} = \ln_e Z = -\ln_e Z^{-1} = -\ln_e(1 - \nabla)$$

Expanding in ascending powers of ∇ , then substituting $1 - Z^{-1} = \nabla$, the infinite series in Z^{-1} can be approximated by a rational function in Z^{-1} . The set of Padé approximants is a convenient source. If the degree of numerator and denominator are chosen to be the same, the rational function can be decomposed into factors, each of which has the form of filter w .

If the variable w is used to define an output/input ratio, then the filter is of exactly the form encountered during the test.

IV. FILTER CONSTRAINTS. To be realizable, the filter (rational function) must be bounded. That is, there must be no poles at infinity. Obviously, the degree of the numerator must not exceed that of the denominator.

For a stable filter, under a conformal mapping into the r -plane (described later), all poles must be found in the left half-plane. A sufficient condition seems to be that the real parts of all denominator coefficients be of like sign.

V. DEFINITION OF TERMS.

t_1 present time

t_0 previous time

$\Delta t = t_1 - t_0$ the sampling interval.

Sometimes referred to as one Real Time Interrupt (RTI). In the case at hand, $\Delta t = 0.008192$ sec.

ϕ or ϕ_i the input at t_i (see Figure 1.)

F or F_i the output of the filter at t_i

F_i/ϕ_i momentary gain of the filter

G steady state gain of the filter

$$G = \lim_{i \rightarrow \infty} \frac{F_i}{\phi_i}$$

For this discussion, $G = 1$, with no loss of generality.

It is profitable to investigate the response of the filter to unit step function at t_1 ; i.e.,

$$\phi_n = 0 \quad (n < 0)$$

$$\phi_n = 1 \quad (n > 0)$$

where n admits only of integral values.

Thus

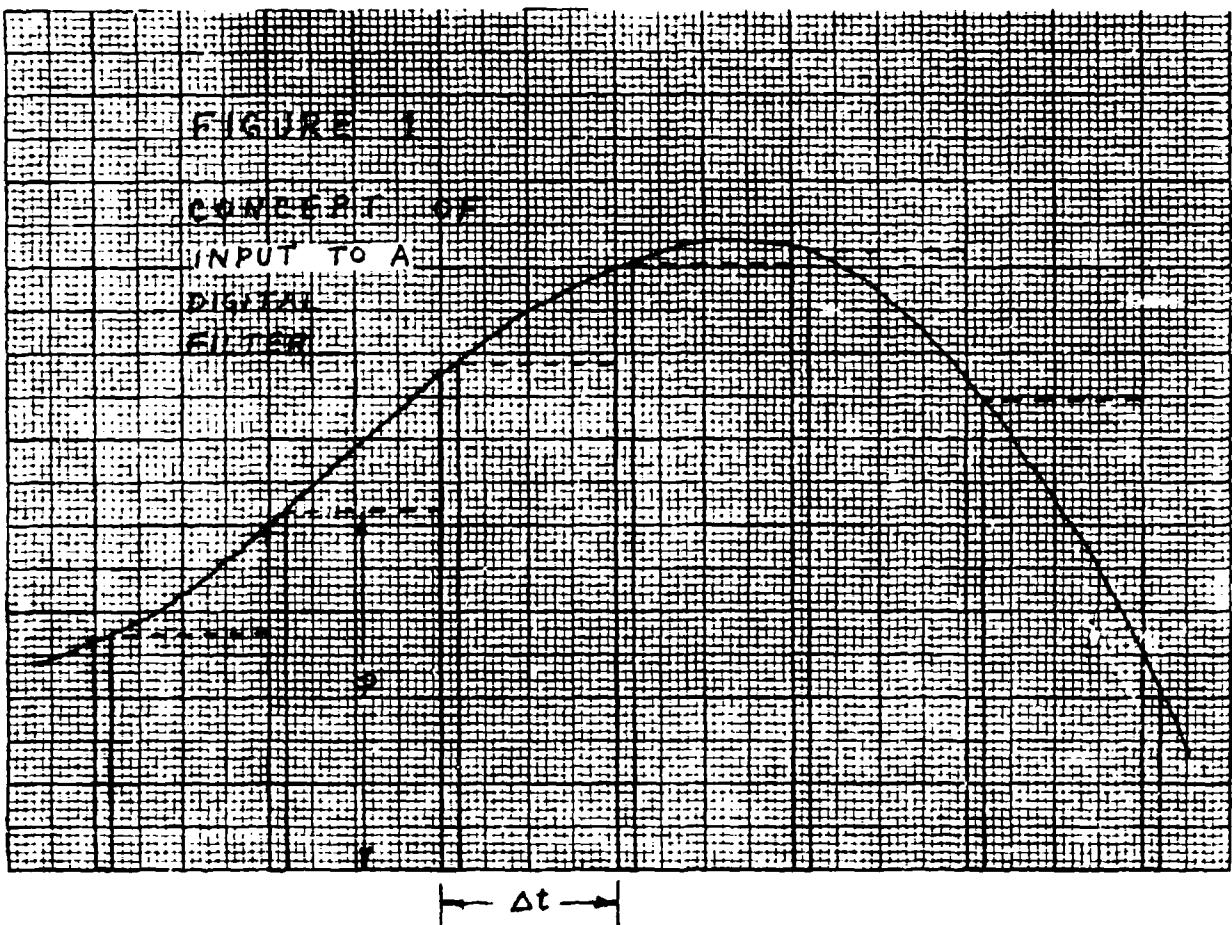
$$\lim_{i \rightarrow \infty} F_i = 1$$

We shall now define I , the total impulse of the filter, as

$$I = \sum_{i=1}^{\infty} \Delta t (F_i - G\phi_i)$$

which, upon substitution, reduces to

$$I = \Delta t \sum_{i=1}^{\infty} (F_i - 1) \quad (7)$$



The partial impulse of the filter is defined as

$$I_n = \Delta t \sum_1^n (F_i - 1) \quad (8)$$

When n is small, the ratio $|I_n|$ is useful.

When $|I_n| = \frac{1}{2}I$, that time can be called the half-life of the filter.

It is a measure of the filter's responsiveness (or sluggishness).

When $|I| = 0$, there is no net feedback. The filter is a smoothing filter only (possibly a very good one). It is useless for control.

When $|I| > 0$, there is an excess of output over input, which is available to the system for control. The amount of this excess defines "total impulse" in a useful filter.

When $|I| < 0$, there is excessive "power" loss. The filter allows the system to drift toward instability.

VI. OSCILLATING FILTERS. Repeating for convenience

$$x_1 = \phi_1 - \gamma x_0 \quad (2)$$

$$F_1 = ax_1 + bx_0 \quad (3)$$

and continuing to investigate the response to unit step function, it is seen that

$$x_0 = 0$$

$$x_1 = 1$$

$$x_2 = 1 - \gamma$$

$$x_3 = 1 - \gamma + \gamma^2$$

$$x_4 = 1 - \gamma + \gamma^2 - \gamma^3$$

In the limit, there will be generated an infinite series which converges to

$$\sum_{i=0}^{\infty} (-\gamma)^i = \frac{1}{1 + \gamma} \quad (9)$$

provided $|Y| < 1$. When $0 < Y < 1$, the series will have terms of alternating sign. The successive values of x_i will oscillate about some value, as will the output of the filter. As a general rule, an oscillating filter is not desirable for control. So much so that an oscillating filter can be viewed as evidence of poor design.

When $-1 < Y < 0$, the series for $(1 + Y)^{-1}$ will have terms of like sign, and the filter will be relatively smooth.

Notice the behavior of x for various real values of Y .

$Y < -1$. The successive values of x diverge. The filter is unstable.

$-1 < Y < 0$. x converges to the value $\frac{1}{1+Y}$. The filter is stable and relatively smooth.

$Y = 0$. $x = 1$ (constant). The output of the filter therefore is constant $(\alpha + \beta)$. An exception occurs at t_1 , where $F_1 = \alpha$.

$0 < Y < 1$. x converges, but oscillates about the value $\frac{1}{1+Y}$. The filter oscillates with period $2\Delta t$. This is equivalent to exactly the Nyquist frequency. (See Figure 2.)

$Y = 1$. x alternates between the two values 0 and 1. The filter output oscillates between α and β .

$Y > 1$. x and the filter output are both oscillatory and divergent.

For a filter to be both stable and non-oscillatory, the coefficient Y must fall within the range

$$-1 < Y < 0$$

VII. THE FUNDAMENTAL IMPULSE FORMULA. Repeating for convenience

$$w = \frac{F}{\phi} = \frac{\alpha Z + \beta}{Z + Y} \quad (1.1)$$

Let $\kappa = \beta/\alpha$. Then

$$\frac{F}{\phi} = \alpha \left(\frac{1 + \kappa Z^{-1}}{1 + Y Z^{-1}} \right) \quad (1.2)$$

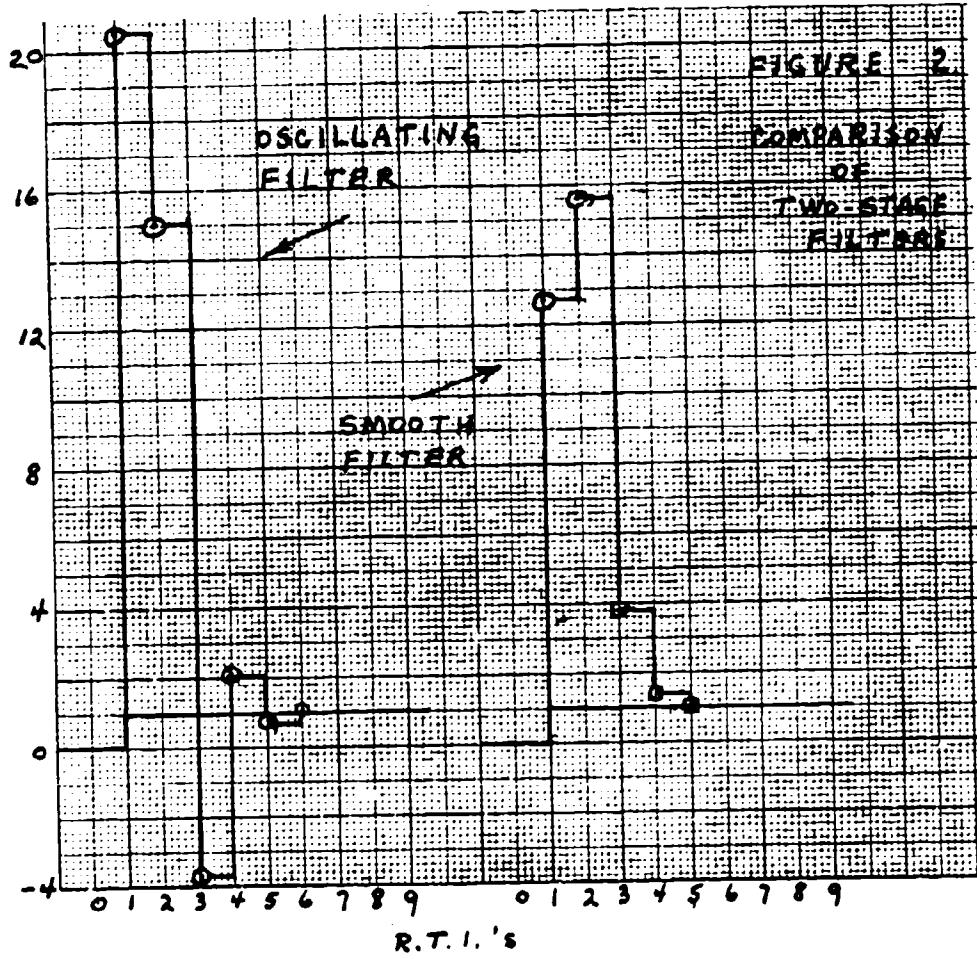
Clearing of fractions,

$$(1 + Y Z^{-1})F = \alpha(1 + \kappa Z^{-1})\phi$$

This difference equation can be written

$$F_n + Y F_{n-1} = \alpha[\phi_n + \kappa \phi_{n-1}]$$

FIGURE 17



But

$$\phi_n = 1 \quad (n > 0)$$

and

$$\lim_{i \rightarrow \infty} F_i = 1$$

Hence

$$1 + \gamma = \alpha [1 + \kappa] \quad (10)$$

From the difference equation is derived

$$\sum_{i=1}^n F_i + \gamma \sum_{i=0}^{n-1} F_i = \alpha [\sum_{i=1}^n \phi_i + \kappa \sum_{i=0}^{n-1} \phi_i]$$

Since $F_0 = \phi_0 = 0$, this can be written as

$$\sum_{i=1}^{n-1} F_i + F_n + \gamma \sum_{i=1}^{n-1} F_i = \alpha [\sum_{i=1}^{n-1} \phi_i + \phi_n + \kappa \sum_{i=1}^{n-1} \phi_i]$$

$$(1 + \gamma) \sum_{i=1}^{n-1} F_i - \alpha (1 + \kappa) \sum_{i=1}^{n-1} \phi_i = \alpha \phi_n - F_n$$

But

$$1 + \gamma = \alpha (1 + \kappa),$$

hence

$$(1 + \gamma) \sum_{i=1}^{n-1} (F_i - \phi_i) = \alpha \phi_n - F_n$$

Remembering that $\phi_i = 1$ and letting $n \rightarrow \infty$,

$$\sum_{i=1}^{\infty} (F_i - 1) = \frac{\alpha - 1}{1 + \gamma} = \frac{1}{\Delta t} I \quad (11)$$

Making the substitution

$$\alpha = \frac{1 + \gamma}{1 + \kappa} \quad (10)$$

$$\sum_1^{\infty} (F_i - 1) = \frac{1}{1 + \kappa} - \frac{1}{1 + \gamma} = \frac{\gamma}{1 + \gamma} - \frac{\kappa}{1 + \kappa} \quad (12)$$

For two cascaded stages

$$\frac{F}{\phi} = \alpha_1 \alpha_2 \frac{(1 + \kappa_1 Z^{-1})(1 + \kappa_2 Z^{-1})}{(1 + \gamma_1 Z^{-1})(1 + \gamma_2 Z^{-1})} \quad (13)$$

An exactly similar development yields

$$\begin{aligned} \sum_1^{\infty} (F_i - 1) &= \frac{1}{1 + \kappa_1} + \frac{1}{1 + \kappa_2} - \frac{1}{1 + \gamma_1} - \frac{1}{1 + \gamma_2} \\ &= \frac{\gamma_1}{1 + \gamma_1} + \frac{\gamma_2}{1 + \gamma_2} - \frac{\kappa_1}{1 + \kappa_1} - \frac{\kappa_2}{1 + \kappa_2} \end{aligned} \quad (14)$$

The subscripts on the right refer to filter stages.

It is apparent that the process can be extended to any number of stages --- say j . Thus can be stated in general form (for j stages) the FUNDAMENTAL IMPULSE FORMULA

$$\begin{aligned} I &= \Delta t \sum_1^{\infty} (F_i - 1) \\ &= \Delta t \left(\frac{1}{1 + \kappa_1} + \frac{1}{1 + \kappa_2} + \dots + \frac{1}{1 + \kappa_j} - \frac{1}{1 + \gamma_1} \right. \\ &\quad \left. - \frac{1}{1 + \gamma_2} - \dots - \frac{1}{1 + \gamma_j} \right) \\ &= \Delta t \left(\frac{\gamma_1}{1 + \gamma_1} + \frac{\gamma_2}{1 + \gamma_2} + \dots + \frac{\gamma_j}{1 + \gamma_j} - \frac{\kappa_1}{1 + \kappa_1} \right. \\ &\quad \left. - \frac{\kappa_2}{1 + \kappa_2} - \dots - \frac{\kappa_j}{1 + \kappa_j} \right) \end{aligned} \quad (15)$$

Or, since $1 + \gamma = \alpha (1 + \kappa)$, (10)

$$I = \Delta t \left(\frac{\alpha_1 - 1}{1 + \gamma_1} + \frac{\alpha_2 - 1}{1 + \gamma_2} + \dots + \frac{\alpha_j - 1}{1 + \gamma_j} \right) \quad (16)$$

$$\text{For a single stage, } I = \Delta t \left(\frac{\alpha - 1}{1 + \gamma} \right) \quad (11)$$

The constraints upon γ imposed by the requirement for filter stability ensure that the denominator is positive. Hence $\alpha > 1$ is the necessary condition for a useful control filter ($|I| > 0$). It follows immediately that $1 + \kappa$ also is positive and that $\gamma > \kappa$.

For a given value of κ , $\alpha/(1 + \gamma)$ remains constant, but $-1/(1 + \gamma)$ increases with increasing γ . The function is maximized (for the allowable range) at $\gamma = 1$. We choose the notation

$$I_{\max} = \frac{\Delta t}{2} (\alpha - 1)$$

For j stages,

$$I_{\max} = \frac{\Delta t}{2} (\alpha_1 + \alpha_2 + \dots + \alpha_j - j)$$

$$\text{But if } \gamma = 1, \alpha - 1 = \frac{1 - \kappa}{1 + \kappa}, \text{ and}$$

$$I_{\max} = \frac{\Delta t}{2} \left(\frac{1 - \kappa_1}{1 + \kappa_1} + \frac{1 - \kappa_2}{1 + \kappa_2} + \dots + \frac{1 - \kappa_j}{1 + \kappa_j} \right) \quad (17)$$

It is proper to think of I_{\max} as a boundary condition. But I_{\max} is clearly unattainable, since the boundary is not included in the (open) region. It will be found that I_{\max} always is reduced by an amount which shall be called the attenuation and which is defined by

$$A = I_{\max} - I \quad (18)$$

The proper dimension is some unit of time; e.g., seconds or RTI's.

I_{\max} can be called the "desired total impulse." It appears in the numerator of the r-plane transfer function, as will be seen later. Since it is a function of only the κ_j 's, the latter can be called the "impulse coefficients."

VIII. THE ATTENUATION. A. Attenuation is defined by

$$A = I_{\max} - I \quad (18)$$

Two forms of the Fundamental Impulse Formula can be combined by simple addition to yield

$$\begin{aligned} I &= \frac{\Delta t}{2} \left(\frac{1 - \kappa_1}{1 + \kappa_1} + \frac{1 - \kappa_2}{1 + \kappa_2} + \dots + \frac{1 - \kappa_j}{1 + \kappa_j} - \frac{1 - \gamma_1}{1 + \gamma_1} \right. \\ &\quad \left. - \frac{1 - \gamma_2}{1 + \gamma_2} - \dots - \frac{1 - \gamma_j}{1 + \gamma_j} \right) \end{aligned} \quad (14)$$

Thus

$$A = \frac{\Delta t}{2} \left(\frac{1 - \gamma_1}{1 + \gamma_1} + \frac{1 - \gamma_2}{1 + \gamma_2} + \dots + \frac{1 - \gamma_j}{1 + \gamma_j} \right) \quad (19)$$

If we define

$$A_j = \frac{\Delta t}{2} \left(\frac{1 - \gamma_j}{1 + \gamma_j} \right)$$

then each $A_j > 0$, due to the limits imposed upon γ . Further, since $I > 0$ (for a control filter), $A < I_{\max}$.

Therefore

$$I_{\max} > A = A_1 + A_2 + \dots + A_j > 0$$

The attenuation can be computed separately for each stage and the parts added.

Notice that each A_j is a function of γ_j alone. It is therefore proper to call the γ_j 's "coefficients of attenuation."

For a single stage, let us observe the effect upon A_j of various values of γ .

$$-1 < \gamma < 0 \text{ implies } A_j > \frac{\Delta t}{2}$$

$$\gamma = 0 \text{ implies } A_j = \frac{\Delta t}{2}$$

$$0 < \gamma < 1 \text{ implies } A_j < \frac{\Delta t}{2}$$

The filter characteristics, previously stated in terms of γ , can now be stated in terms of A . For j stages, then

$$0 < A < \frac{1}{2}j\Delta t \quad \text{filter oscillates}$$

$$A = \frac{1}{2}j\Delta t \quad \text{filter finite of duration } j\Delta t$$

$$\frac{1}{2}j\Delta t < A < I_{\max} \quad \text{filter smooth}$$

$$A = I_{\max} \quad \text{zero total impulse}$$

As A approaches I_{\max} , the filter becomes more sluggish.

Particularly note that a requirement that the filter be non-oscillatory places a finite limit upon the number of filter stages. This limit is, of course

$$j < \frac{2I_{\max}}{\Delta t}$$

IX. COMPUTING THE COEFFICIENTS. For a single stage

$$A = \frac{\Delta t}{2} \left(\frac{1 - \gamma}{1 + \gamma} \right) \quad \text{or} \quad \gamma = \frac{\Delta t - 2A}{\Delta t + 2A} \quad (19.1)$$

$$I_{\max} = \frac{\Delta t}{2} \left(\frac{1 - \kappa}{1 + \kappa} \right) \quad \text{or} \quad \kappa = \frac{\Delta t - 2I_{\max}}{\Delta t + 2I_{\max}} \quad (17.1)$$

Now A is a function of frequency response and may be amenable to some adjustment. Not so I_{\max} . It is the "power" demanded of the filter, and we expect to deliver only I of it. Can we recover I_{\max} completely? The answer is yes. Instead of

$$A = I_{\max} - I \quad (18)$$

we write

$$A = (I_{\max} + A) - (I + A)$$

or

$$A = (I_{\max} + A) - I_{\max} \quad (20)$$

We now find that

$$I_{\max} + A = \frac{\Delta t}{2} \frac{1 - \kappa'}{1 + \kappa'} \quad (21)$$

or

$$\kappa' = \frac{\Delta t - 2I_{\max} - 2A}{\Delta t + 2I_{\max} + 2A}$$

In other words, we enter $I_{\max} + A$ into the formula for the impulse coefficient and let the filter attenuate it back to I_{\max} . It is possible to do this because, in the time domain, so many of the terms are additive. Particularly note that A must be determined first.

Now I_{\max} is a design requirement imposed upon the filter. Arbitrarily augmenting it by some amount (say A) alters nothing in principle. The formulae can be stated

$$w' = \frac{F'}{\phi} = \frac{\alpha' Z + \beta'}{Z + \gamma} \quad (22)$$

$$\kappa' = \beta'/\alpha' \quad (23)$$

$$\frac{F'}{\phi} = \alpha' \left(\frac{1 + \kappa' Z^{-1}}{1 + \gamma Z^{-1}} \right) \quad (24)$$

$$1 + \gamma = \alpha' [1 + \kappa'] \quad (25)$$

where the primes denote the new values resulting from the augmentation of I_{\max} . Note that ϕ and γ are independent of I_{\max} (as is A) and hence are not primed.

X. COMPUTING THE COEFFICIENTS FOR A TWO-STAGE FILTER. If I_{\max} and A are known, a single-stage filter is uniquely determined, since it contains, in essence, only two coefficients (γ and κ').

Suppose the task is to design a two-stage filter. Other means must be found for determining two of the four coefficients.

The first step is easy. For a non-oscillatory filter,

$$I_{\max} > A > \frac{1}{2} j \Delta t$$

If we set $\gamma = \gamma_1 = \gamma_2$, then $A_1 = A_2 = \frac{1}{2} \Delta t$ and it is ensured that neither stage will be oscillatory.

$$I_{\max} + A = \frac{\Delta t}{2} \left(\frac{1 - \kappa'_1}{1 + \kappa'_1} + \frac{1 - \kappa'_2}{1 + \kappa'_2} \right) \quad (26)$$

requires only that the sum within parentheses be constant. Solving for either κ' in terms of the other

$$\kappa'_1 = \frac{\Delta t - (1 + \kappa'_2)(I_{\max} + A)}{\kappa'_2 \Delta t + (1 + \kappa'_2)(I_{\max} + A)} \quad (27)$$

Allowing κ'_2 to increase without bound

$$\lim_{\kappa'_2 \rightarrow \infty} \kappa'_1 = \frac{-(I_{\max} + A)}{\Delta t + I_{\max} + A} = P \quad (28)$$

If $\kappa'_1 = \kappa'_2 = \kappa'$, then

$$\kappa' = \frac{\Delta t - (I_{\max} + A)}{\Delta t + I_{\max} + A} = 1 + 2P \quad (29)$$

We thus establish limits for the κ' 's.

$$P < \kappa'_1 < 1 + 2P < \kappa'_2$$

It is observed that as the κ' 's approach $1 + 2P$, more and more "power" is delivered at the first RTI. In some cases, the half-life can be less than $\Delta t/2$, causing the filter to over-correct and generate unwanted noise. This effect is most marked when the two κ' 's are equal. (At $1 + 2P$, of course.)

Little case can be made for a half-life less than Δt . Using this as a restriction, and noting that at the first RTI

$$\frac{E'}{\phi} = \alpha'_1 \alpha'_2 \quad (30)$$

it develops that

$$(\alpha'_1 \alpha'_2 - 1) \Delta t < \frac{1}{2} I_{\max} \quad (31)$$

Now

$$1 + \gamma = \alpha' (1 + \kappa') = \frac{2\Delta t}{\Delta t + A}$$

allowing us to develop a second equation in the κ' 's. The solution, provided it exceeds $1 + 2P$, will furnish a practical lower limit for κ'_2 . To avoid negative total impulse (for the stage), use an upper limit of $\kappa'_2 < 1$.

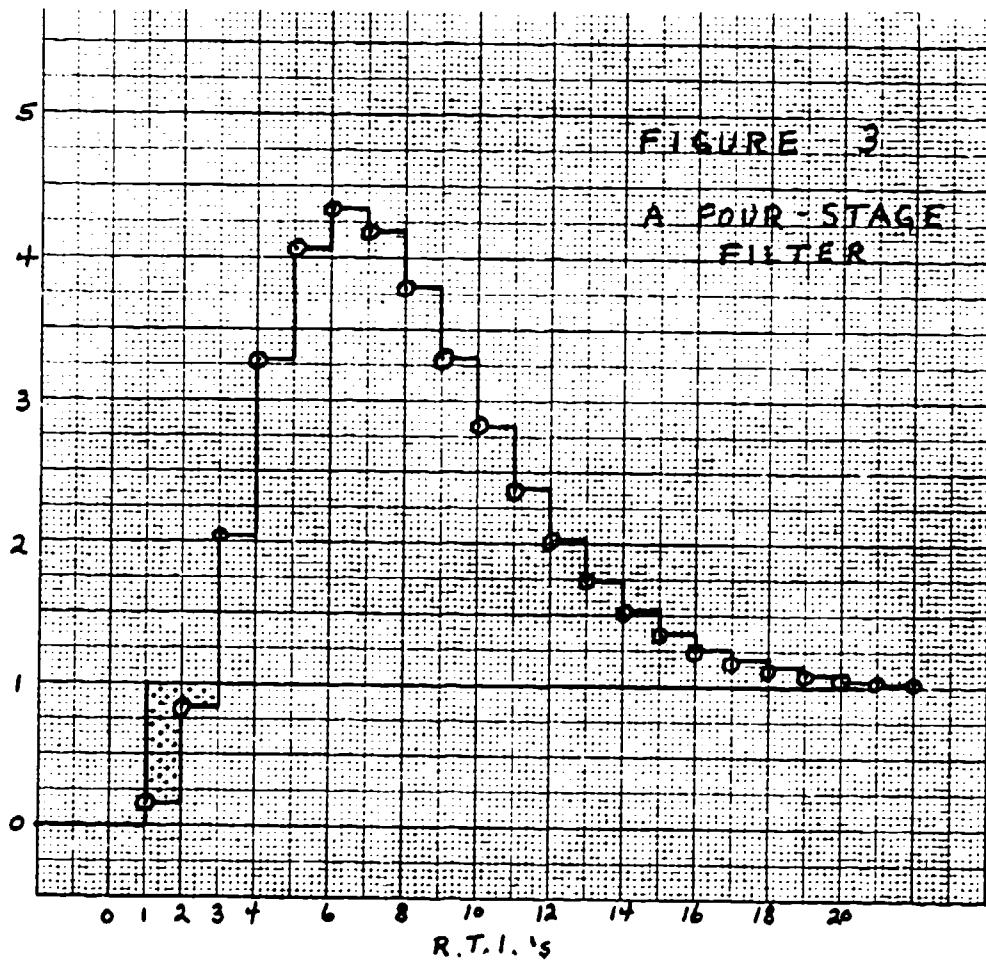
Often it is effective to set κ'_2 equal to unity. A useful side effect is that one term drops out of the I_{\max} equation, since

$$I_{\max} + A = \frac{\Delta t}{2} \left(\frac{1 - \kappa'}{1 + \kappa'} + \frac{1 - 1}{1 + 1} \right)$$

and thus the formula for κ' is the same as the single-stage formula. The filter is now implemented by

$$\frac{F'}{\phi} = \alpha'_1 \alpha'_2 \left[\frac{(1 + \kappa' z^{-1})(1 + z^{-1})}{(1 + \gamma z^{-1})(1 + \gamma z^{-1})} \right] \quad (32)$$

Extending the method ($1 = \kappa'_2 = \kappa'_3 = \text{etc.}$) to filters of still more stages may not be warranted. The additional stages will be smoothing stages only, and the resulting filter can be very sluggish. In fact, the ratio $|I_n|/|I|$ may actually take on negative values for the first few RTI's, a most undesirable characteristic for a control filter. (See Figure 3.)



In order to illustrate the effect of varying κ'_2 , three compensated two-stage filters were synthesized to the requirements

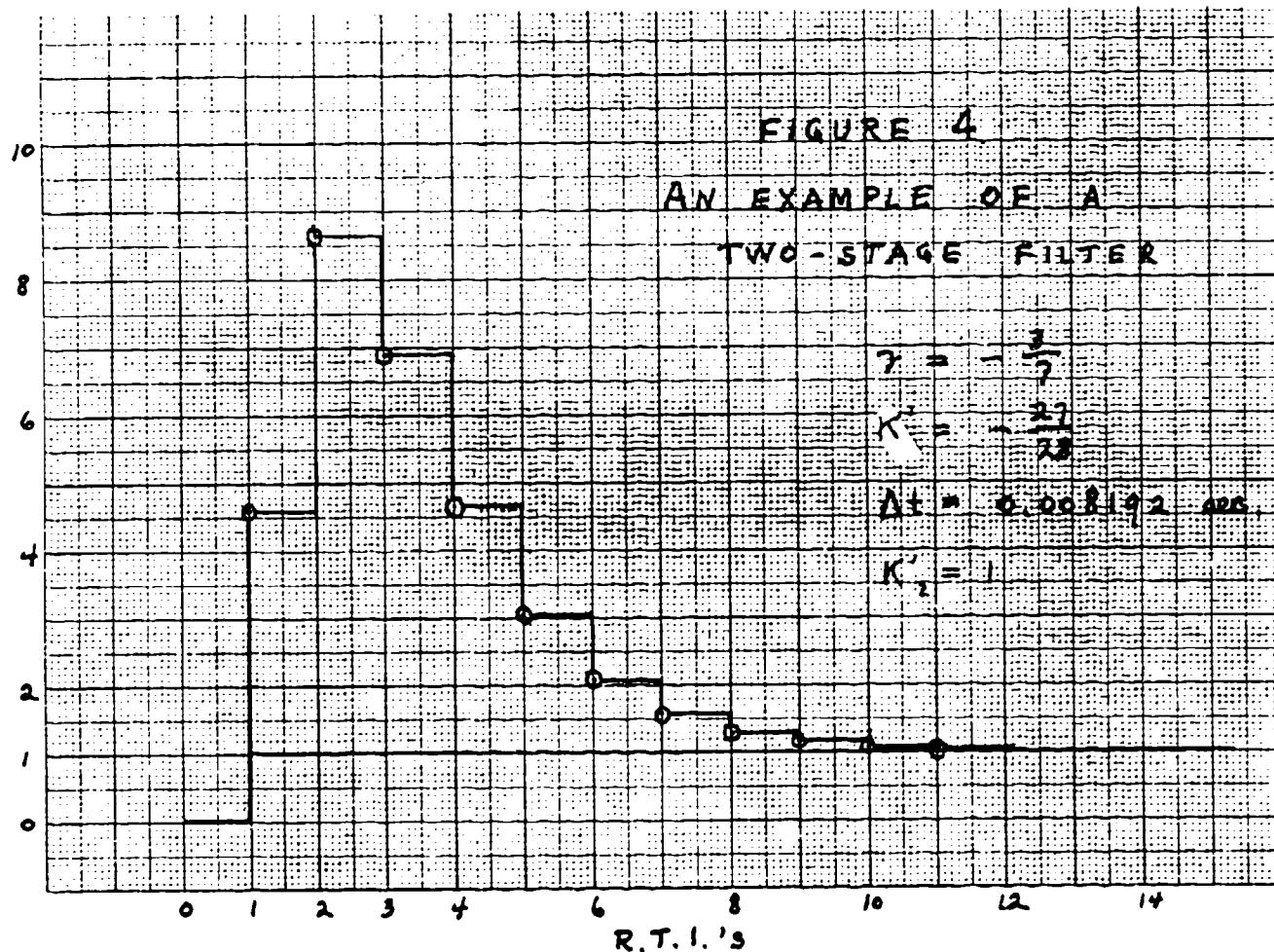
$$I_{\max} = 25 \Delta t$$

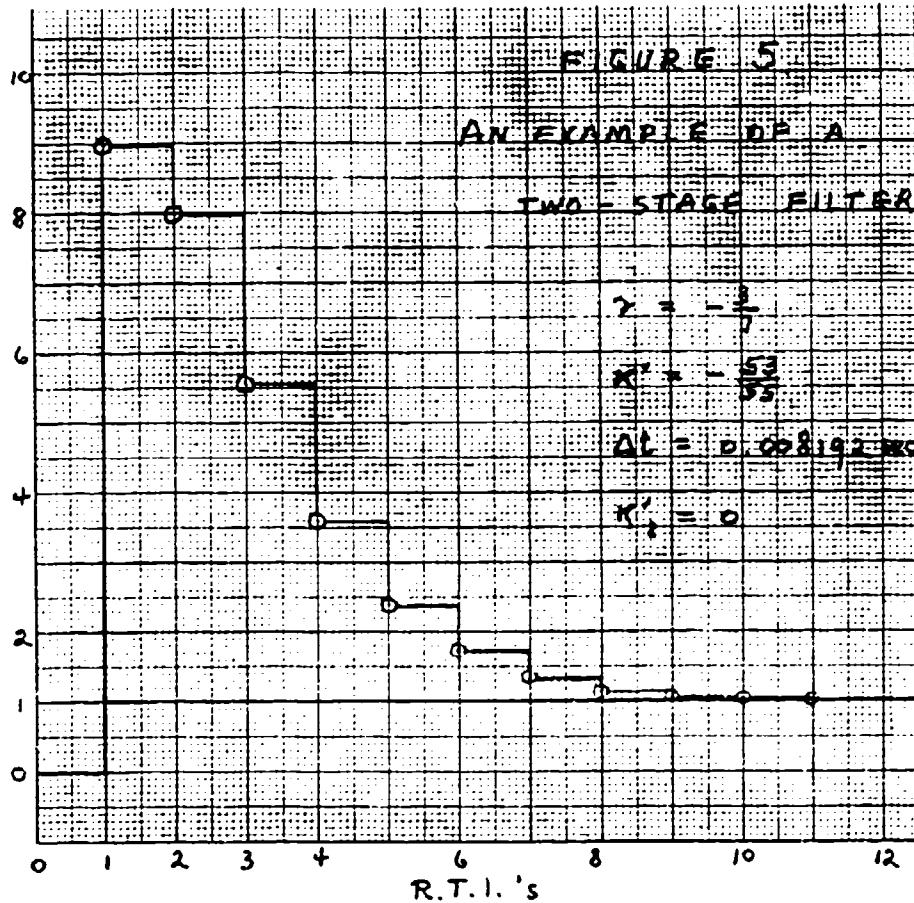
and

$$A = \frac{5}{2} \Delta t$$

κ'_2 was arbitrarily set in turn to the values 1, 0, and $-\frac{1}{3}$.

Results are depicted graphically in Figures 4, 5 and 6.



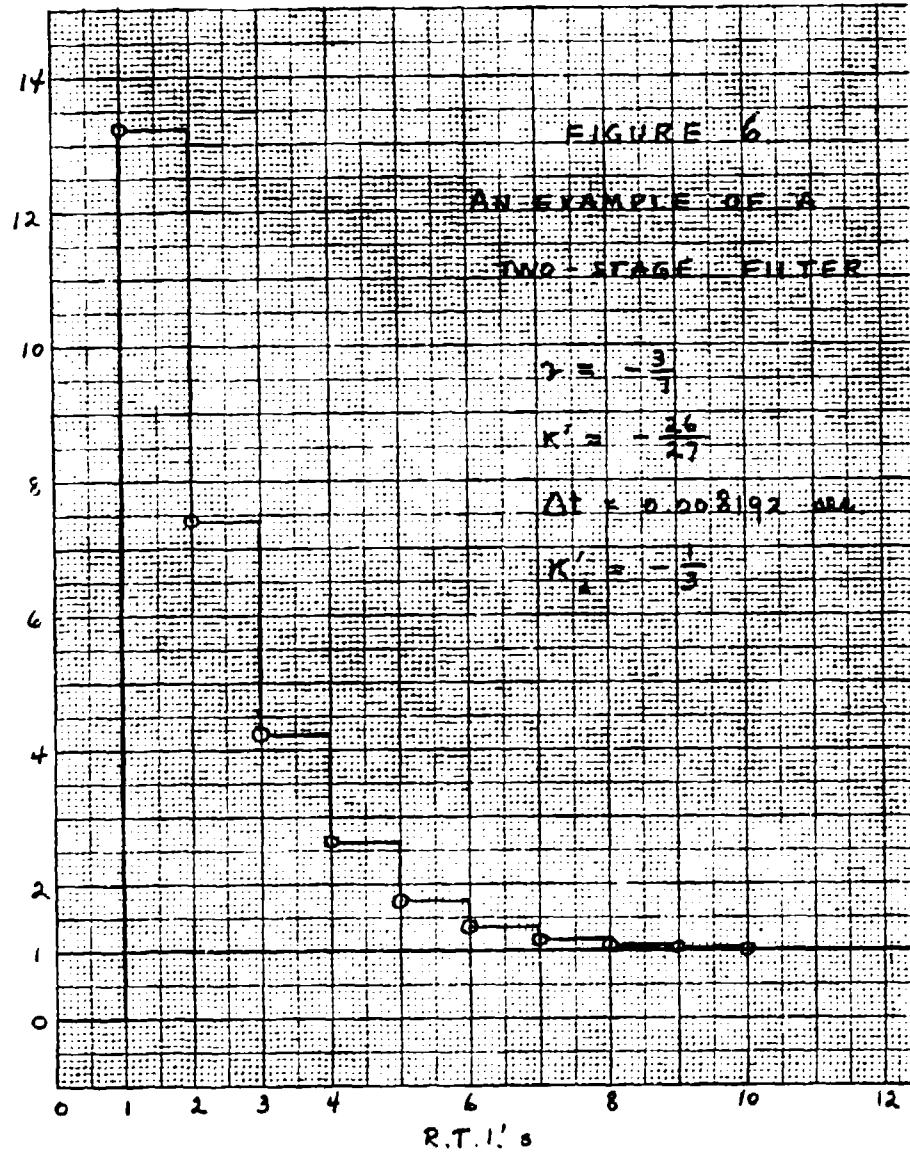


XI. TRANSFORMATIONS AND CONFORMAL MAPPINGS.* A useful tool developed to simplify the solution of certain differential equations is the Laplace Transform, defined by

$$L [F(x)] = \int_0^{\infty} e^{-sx} F(x) dx = f(s)$$

provided the integral exists. Therefore, the Laplace Transform is an integral operator.

*In this paragraph, $j = \sqrt{-1}$



The complex variable s is of the form $s = \sigma + j\omega$, σ and ω being real. The function $f(s)$ is many-valued, being periodic in $4jn\omega_0$, where n is any integer and ω_0 is the Nyquist frequency (often expressed in radians per second).

For band limited functions, it becomes expedient to define another complex variable

$$r = \rho + jv$$

by

$$\frac{r\Delta t}{2} = \tanh \frac{s\Delta t}{2} \quad (33)$$

The frequency response in both r - and s -planes is given by setting $\rho = \sigma = 0$, from which

$$\frac{jv\Delta t}{2} = \tanh \frac{j\omega\Delta t}{2} = -j \tan \left(\frac{\omega\Delta t}{2} \right)$$

Thus $\frac{v\Delta t}{2} = \tan \frac{\omega\Delta t}{2}$. It follows that for $\omega_0 = \frac{1}{\Delta t}$, the s -plane zero-strip maps into the entire r -plane. As a result, the new variable r is single-valued.

ω is called the NATURAL frequency.

v is called the WARPED frequency. It is found to be related to the attenuation by

$$A = \frac{2}{v} \quad (34)$$

It is very easy to demonstrate that the transformation $\frac{r\Delta t}{2} = \tanh \frac{s\Delta t}{2}$ is bilinear in Z^{-1} and r , since $Z = e^{s\Delta t}$ and $\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$, whence

$$\frac{r\Delta t}{2} = \tanh \frac{s\Delta t}{2} = \frac{1 - e^{-s\Delta t}}{1 + e^{-s\Delta t}} = \frac{1 - Z^{-1}}{1 + Z^{-1}} \quad (35)$$

The inverse transformation is, of course

$$Z^{-1} = \frac{1 - \frac{r\Delta t}{2}}{1 + \frac{r\Delta t}{2}} \quad (36)$$

In the present case, the form of the filter is known, and is expressed in powers of Z^{-1} . The r -plane transfer function is recoverable immediately by direct substitution.

For a single-stage filter, the desired difference equation is of the form

$$\frac{F'}{\phi} = a' \left(\frac{1 + \kappa' Z^{-1}}{1 + \gamma Z^{-1}} \right) \quad (24)$$

Substituting for Z^{-1} , it is found that

$$\frac{F'}{\phi} = a' \left(\frac{1 + \kappa' + \frac{r\Delta t}{2}[1 - \kappa']}{1 + \gamma + \frac{r\Delta t}{2}[1 - \gamma]} \right) \quad (37)$$

Repeating for convenience

$$A = \frac{\Delta t}{2} \left(\frac{1 - \gamma}{1 + \gamma} \right) \quad (19.1)$$

$$I_{max} + A = \frac{\Delta t}{2} \left(\frac{1 - \kappa'}{1 + \kappa'} \right) \quad (21)$$

$$1 + \gamma = a'(1 + \kappa') \quad (25)$$

it is seen that the r-plane transfer function can be written

$$\frac{F'}{\phi} = \frac{1 + r(I_{max} + A)}{1 + rA} \quad (38)$$

It should be clear that solution in either the r-plane or the Z^{-1} -plane (which two are connected by the stated bilinear transformation) is easily implemented in a digital computer, since only straightforward arithmetic is involved. Not so in the s-plane, where logarithms (e.g., $\tanh^{-1}[\frac{r\Delta t}{2}]$), exponentials, and the like will be required.

For a filter of j stages, if $\gamma = \gamma_1 = \gamma_2 = \dots = \gamma_j$ and $\kappa'_2 = \kappa'_3 = \dots = \kappa'_j = 1$, the r-plane transfer function is

$$\frac{F'}{\phi} = \frac{1 + r(I_{max} + A)}{(1 + rA)^j} \quad (39)$$

BIBLIOGRAPHY

- (1) Boole, G., THE CALCULUS OF FINITE DIFFERENCES; 5th Ed., Chelsea Publishing Co., New York, 1970.
- (2) Churchill, R. V., INTRODUCTION TO COMPLEX VARIABLES AND APPLICATIONS; McGraw-Hill Book Co., New York, 1948.
- (3) Hamming, R. W., NUMERICAL METHODS FOR SCIENTISTS AND ENGINEERS; McGraw-Hill Book Co., New York, 1962.
- (4) Kaiser, J. F., "Digital Filters" in SYSTEMS ANALYSIS BY DIGITAL COMPUTER; John Wiley & Sons, New York, 1966.
- (5) Lindorff, D. P., THEORY OF SAMPLED-DATA CONTROL SYSTEMS; John Wiley & Sons, New York, 1965.

INFERENCE PROCEDURES FOR DETERMINING LIFE TIME ESTIMATES
OF ADVANCED MATERIALS

Donald Neal
Edward M. Lenoe
Donald Mason

Army Materials and Mechanics Research Center
Watertown, Massachusetts 02172

EXTENDED ABSTRACT

An improved procedure for treatment of so-called censored data has been developed and life-time estimates made for proof tested ceramic rotor hubs, in addition to development of quality assurance control of powder metallurgically produced turbine engine discs. These represent situations for structures to perform under extreme environmental conditions and analytical procedures to aid in achieving required component capability.

Two and three parameter Lognormal and Weibull functions represent the candidate statistic models. These functions are examined for best representation of data in order to provide flexibility in the fitting process. The functional parameters are obtained from the maximum likelihood (M.L.) method. This method provides a superior representation of the cyclic fatigue data as compared to the more conventional procedures. The M.L. method can also provide the desired confidence limits for the parameter and reliability determinations associated with the given data set. The inadequacies associated with the method of moments, graphical procedures, etc., in obtaining the functional parameters is recognized from the arbitrariness of the functional representation of the data. The acceptability of these methods is acutely data dependent.

The need for considering all data including censored data is established. Both lower and upper bound censored data are considered as they relate to proof testing and run-outs respectively. An improved probability of failure computation can be obtained when the total data set is represented. Partial probability ranking procedures tend to introduce substantial errors in the extrapolation process necessary in obtaining minimum life-time estimates. By including censored data, one can provide a more complete understanding of the materials capabilities.

The results of combining the M.L. method with the inclusion of censored data are compared with conventional procedures in obtaining both structural reliability and material probability of failure computations. The comparison indicates a substantial nonconservative estimate of failure probabilities can occur if threshold stress values are obtained from proof testing without consideration of the censored data. Application of the M.L. procedure provided an improvement in the functional representation of data.

ANALYSIS OF CENSORED DATA

Introduction

Oftentimes in procurement of structural ceramic components, screening tests are employed to attempt to verify component (part) quality. Typical of such tests are room temperature (cold) spin tests of rotor hubs. Usual practice calls for delivery of successful spin tested parts and these are then treated as if guaranteed strengths were existent. We desired to more fully exploit the information gained during such screening tests. As an example, during the conduct of screening experiments, failures were observed. Ordinarily these failures, or the failure data is not reported. It is obvious that the failure rate data provides useful information in planning for component reliability levels and is necessary to establish a rational quality assurance plan. Thus this study explores the use of censored statistics to provide reliability estimates incorporating minimum screening strength levels, and also the failure rates (standard deviations ind.) associated with spin tests. The influence, for instance, of 5% versus 10% failure rate during screening tests are documented. Monte Carlo techniques are employed to establish desired screening test procedures.

Suppose, for instance, fast fracture probability of failure estimates were made using conventional Weibull statistics. In this instance, the screening level is treated as a lower bound. However, censored data techniques allow taking into account the likely component failure rates, based on the observed screening test data. The purpose of the following calculation is to compare the degree of conservatism of the two reliability estimates. (It was observed that the censored data technique provides the more conservative results.)

These comparisons provide confidence in using the screening test data itself for the estimates of production reliability. The implication of these results is that continued local mechanical strength determination testing can be minimized and lot component sampling, coupled with spin tests can be adopted to insure hardware reliability. Treated random sample selection from lots can predict corresponding failure of total lot. It is, therefore, important to consider failure below minimum load level.

In representing fatigue data were run-outs (non-failed specimens tested at predetermined number of cycles) exists, it is usually necessary to apply graphical methods in determining prescribed probabilities and their corresponding cycles to failure. The graphical approach requires representation of only the failed results. The remaining data is included only in representing the ranking of the data. This method is often susceptible to error because of the arbitrariness that exists in interpreting an acceptable regression line for ranked data. Optimum coefficient methods, for example, will introduce sizable variation in slopes such that a unique threshold value for function becomes very difficult to determine. Since the extrapolation of the regression line provides the necessary probability

number it is therefore critical that the slope of this line be properly determined. If all data is considered including run-outs then a censored data analysis procedure must be applied. The present analysis outlined in this text applies this analysis including the appropriate M.L. procedures. The Weibull and Lognormal distribution were candidate functions since they are usually the most acceptable representation of fatigue data. A comparison between graphical results and that of the M.L. censored data method indicate substantial differences. The graphical results showed highly non-conservative estimates. In this instance the component material would have been rejected instead of accepted as indicated by the censored data method.

Following is a general description of the analytical technique developed in treating the problems.

Weibull Function

The Weibull function has been commonly used in failure prediction of ceramic and high strength fatigued materials. It was determined from the analysis of the rotor disc and helicopter component data that the best representative function was also Weibull, therefore, the M.L. analysis of censored data for this function will be developed in this paper. The Weibull probability density function of the random variable X is

$$f(X|\sigma_u, \sigma_o, m) = [m(\lambda-\sigma_u)^{m-1}/\sigma_o^m] \exp\{-[(\lambda-\sigma_u)/\sigma_o]^m\} \quad (1)$$

where $\sigma_o, m > 0$ and $X > \sigma_u > 0$

σ_u , σ_o and m are the location, scale and shape parameters respectively. The log of the Weibull likelihood function for dual censoring can be written as [2]

$$\begin{aligned} \ln L &= \ln N! - \ln r! + (N_o - r)(\ln m - \ln \sigma_o) - \ln(N - N_o) \\ &+ (m-1) \sum_{i=r+1}^{N_o} \ln(X_i - \sigma_u) - \sum_{i=r+1}^{N_o} \{ (X_i - \sigma_u)/\sigma_o \}^m \\ &+ r \ln(1 - \exp[-(X_{r+1} - \sigma_u)^m/\sigma_o^m]) - (N - N_o) \{ (X_{N_o} - \sigma_u)/\sigma_o \}^m \end{aligned} \quad (2)$$

where

N = total number of data points including the censored values,
 N_o = number of values prior to run-outs (fatigue data)

and

r = number of data values less than the proof tested value

The M.L. equations are determined from the partial derivative of $\ln L$ with respect to the three parameters set equal zero*. That is,

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma_o} &= 0 \\ \frac{\partial \ln L}{\partial m} &= 0 \\ \frac{\partial \ln L}{\partial \sigma_u} &= 0\end{aligned}\tag{3}$$

where

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma_o} &= -m(N-r)/\sigma_o^{m+1} \sum_{i=r+1}^N (x_i - \sigma_u)^m / \sigma_o^{m+1} \\ &\quad - mr(x_{r+1} - \sigma_u)^m \exp[-(x_{r+1} - \sigma_u)^m / \sigma_o^m] / [\sigma_o^{m+1} \{1 - \exp[-(x_{r+1} - \sigma_u)^m / \sigma_o^m]\}] \\ \frac{\partial \ln L}{\partial m} &= (N-r)(\frac{1}{m} - \ln \sigma_o) + \sum_{i=r+1}^N \ln(x_i - \sigma_u) - \sum_{i=r+1}^N [(x_i - \sigma_u) / \sigma_u]^m \\ &\quad \cdot \ln[(x_i - \sigma_u) / \sigma_o] + r(x_{r+1} - \sigma_u)^m \ln[(x_{r+1} - \sigma_u) / \sigma_o] \exp[-(x_{r+1} - \sigma_u)^m / \sigma_o^m] / \\ &\quad [\sigma_o^m \{1 - \exp[-(x_{r+1} - \sigma_u)^m / \sigma_o^m]\}]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma_u} &= (1-m) \sum_{i=r+1}^N (x_i - \sigma_u)^{-1} + m \sigma_o^{-m} \sum_{i=r+1}^N (x_i - \sigma_o)^{m-1} - mr(x_{r+1} - \sigma_u)^{m-1} \\ &\quad \cdot \exp[-(x_{r+1} - \sigma_u)^m / \sigma_o^m] / [\sigma_o^m \{1 - \exp[-(x_{r+1} - \sigma_u)^m / \sigma_o^m]\}]\end{aligned}$$

*Note: When $N = N_o$ and $r \neq 0$ then lower censoring is applied as required for ceramic disc analysis. If $r = 0$ and $N \neq N_o$ then upper censoring is used in evaluating fatigue results for the helicopter component. The resultant equations above are for the case where $N = N_o$.

An iterative procedure has been developed for determining the M.L. parameters $\hat{\sigma}_o$, $\hat{\mu}_u$ and \hat{m} . Initial estimates are obtained from the moment method without censoring. From these estimates each parameter is determined one at a time in cyclic order in equation 3 until reasonable convergence is obtained. At each step, the rule of false position is used to determine values which satisfied the likelihood equation with prior estimates of the other parameters remaining constant. Note, if a large percent (greater than 50) of censored values exist, then it is necessary to gradually increase the amount of censoring in order to obtain desired convergence. If a two parameter Weibull function is desired then omit σ_u in the computation process.

Lognormal Function

Although the Weibull extreme value function is commonly applied in representing ceramic strength data, the lognormal function can provide an option if the Weibull function is not acceptable. The lognormal function has been included in the evaluation procedure. The likelihood function is defined as:^[3]

$$f(X_{r+1}, \dots, X_N | \mu, \sigma, \tau) = \frac{N!}{(N-N_0)!r!} \prod_{i=r+1}^{N_0} \frac{1}{\sigma\sqrt{2\pi}(X_i - \tau)} \exp \left\{ -\sum_{i=r+1}^{N_0} \frac{[\ln(X_i - \tau) - \mu]^2}{2\sigma^2} \right\} \cdot (1 - F[z_{N_0}])^{N-N_0} \cdot \{F[z_{r+1}]\}^r \quad (4)$$

where τ is the location or threshold parameter and μ , σ are mean and standard deviations respectively. N , r and N_0 are defined in the Weibull analysis.

The function F is defined as:

$$F = \int_{-\infty}^{z_i} f(t) dt$$

where

$$z_i = [\ln(X_i - \tau) - \mu]/\sigma$$

and

$$f(z_i) = (2\pi)^{-\frac{1}{2}} \exp(-z_i^2/2)$$

The complete development of the M.L. function for the log normal is omitted since it is similar to that developed for the Weibull function. A much more severe convergence problem exists in determination of Lognormal parameters, particularly for 50% or more censoring. Therefore, it is important to obtain reasonable initial estimates in addition to introducing small increments of censoring until the desired amount is obtained.

Quality Assurance of Rotor Discs

The failure prediction procedures described previously were applied to data obtained from both spin and flexure tests of rotor discs. Tests were made in order to establish quality assurance of the disc material prior to manufacture into ceramic engine rotors. The spin test is applied initially in order to guarantee a minimum strength level for the disk. This lower bound (threshold strength) was obtained from spinning the disk at an angular velocity of 60,000 rpm. The equivalent forth point flexure test results are 350 N/mm^2 at rim section (R_1 , Figure 3) of disc, this stress value is obtained as shown in Figure 4.

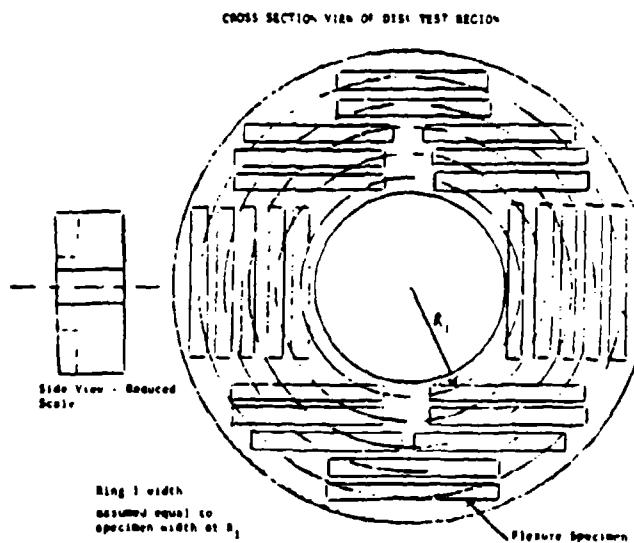
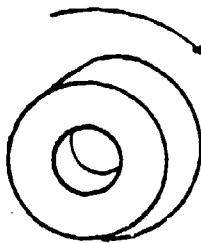


Figure 3

QUALITY ASSURANCE OF ROTOR DISC



QUALIFICATION SPIN TEST (60,000 RPM)

σ_R -STRESS AT RIM REGION : 350 N/mm²

$$\text{WHERE } \sigma_R = \sigma_T \frac{K \cdot V}{M}$$

K = LOADING CHARACTERISTICS

V = VALUE OF ELEMENTS

M = MEASURE OF DISPERSION

σ_T = TANGENTIAL STRESS FROM ROTATION

FIGURE 4

Subsequent flexure tests were conducted on the surviving disks. The test specimens are selected from locations outlined in Figure 3.

The Results of the Censored Data Analysis Procedures

Data is analyzed for flexure specimens obtained from ring 1 of disk (see Figure 4) in order to be consistent with the threshold strength level obtained from spin test at this same location. Since there was a limited amount of data from available disks, it was necessary to generate additional data in order to demonstrate the effects of the censoring process as related to failure predictions of the material. The censored data relates to the number of failed discs resulting from spin tests. The remaining data was obtained from selecting randomly, values generated from the Weibull functional representation of flexure results in ring 1.

Initially it was assumed that flexure data was obtained from 100 rings without consideration or knowledge of the number of failed discs in spin test. A plot of the ranked data (flexure strength) and the corresponding Weibull functional representation is shown in Figure 5. The RMS error tabulation determines the best functional representation and is defined as:

$$\text{RMS} = \left[\sum_{i=1}^N (R_j(x_i) - F_x(z_i))^2 / N \right]^{1/2} \quad (5)$$

where

$$R_1 = \frac{i-.5}{N}, \quad R_2 = \frac{i-.3}{N+.4}, \quad \text{and} \quad R_3 = \frac{i}{N+1},$$

$$i = 1, 2, 3, \dots, N$$

and N = sample size

F_x is the cumulative density function selected from the four candidate functions (normal, lognormal, Weibull and the radical function). See Reference 4 for details regarding radical function. In Figure 5, the radical function was the best fit with Weibull the next best representation. The mean and standard deviation is also tabulated with their corresponding 90% confidence intervals. In box the labelled Weibull parameters, the dispersion scale (char. value) and threshold value σ_u^u (origin) are displayed with 90% confidence intervals. σ_u^u intervals are not as precise estimates as those for the other parameters.

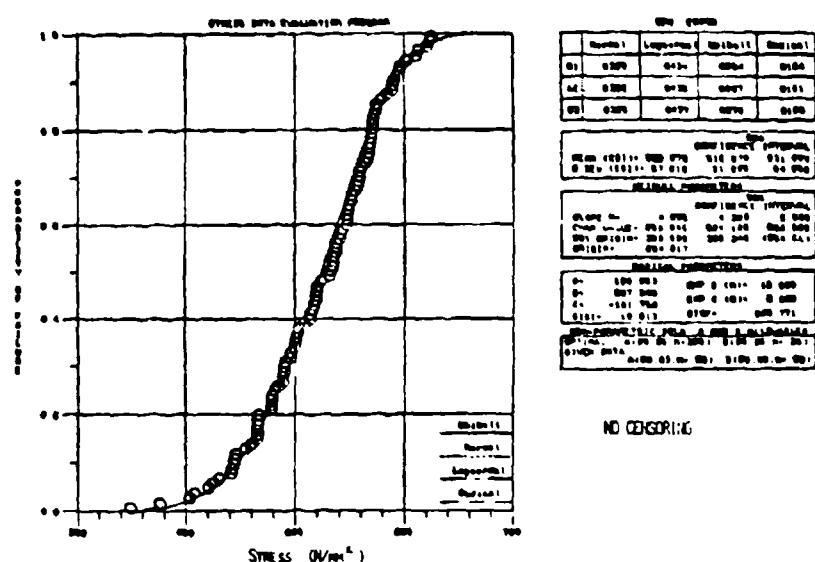


Figure 5

The 99% origin number is the 99% probability of survival value with adjacent number being the lower 95% confidence limit. The latter number represents the so-called design A - allowable. In this instance it is 368.35 N/mm². The radical parameters are also tabulated but will not be described in this text (see Reference 4). The box for non-parametric solution^[5] provides for design A and B allowables when parametric representation of the data is not acceptable. The optimal tabulation indicates 300 and 30 data points are necessary for the A and B allowables. Instances where 99 data points are available, a 99% survivability point (smallest stress value) has a 63% chance of being correct. The B allowable (90% survivability) has at least a 95% guarantee of being correct. The design A allowable determined from Weibull function will be of primary consideration in evaluating the effects of censoring data as it relates to hypothetical failure rates of the discs.

The results from Figure 5 essentially describe the failure prediction of the disc when with additional data (spin test) omitted. There can be a serious problem existing if this spin test result is omitted since the test only prescribes material strength guarantee for small regions of the disc. Therefore, it is essential to recognize that both spin test and flexure tests are material strength characterization tests. Figure 6 presents the results from a 10% failure rate, that is, assuming spin test resulted in 10% of total number of discs failed. In this case, the A-allowable is 297.59 N/mm^2 , a sizeable reduction from the case where no spin-test failure existed. Lognormal was the other candidate function but did not provide the best representation of data. Figure 7 shows the results for 20% failure with a resultant A-allowable of 230.42 N/mm^2 . If a failure rate of 30% existed then A-allowable would be 194.57 N/mm^2 as noted in Figure 8. The data is not well represented in this case, but if no alternative is available then these results will provide the necessary conservatism in contrast to omission of the censored results. The effects of ignoring spin test failures as they relate to censored disc data is obvious, therefore, it is important to consider all test results; both flexure and spin tests. In Figure 9, a plot of probability of survival versus RPM is shown. Note the reduction in allowable RPM when spin test failures have been considered. For example, if 30% failure rate existed, then 95% survivability of additional discs would limit the maximum speed to 40,000 RPM.

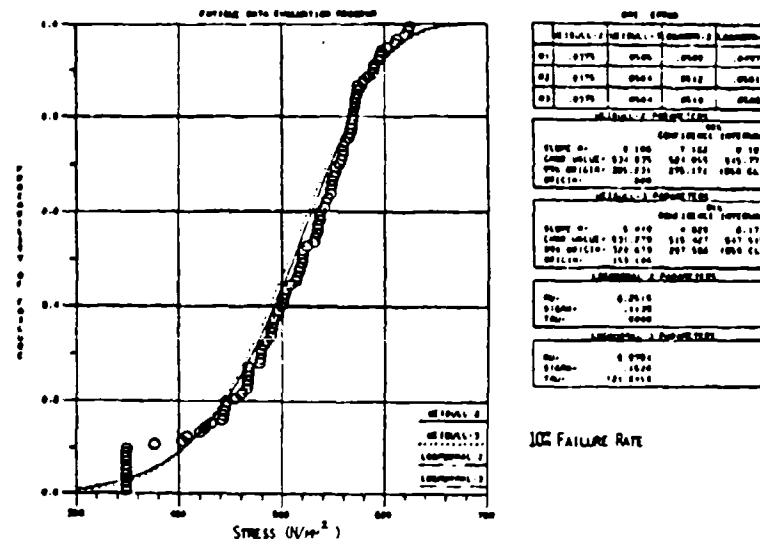


Figure 6

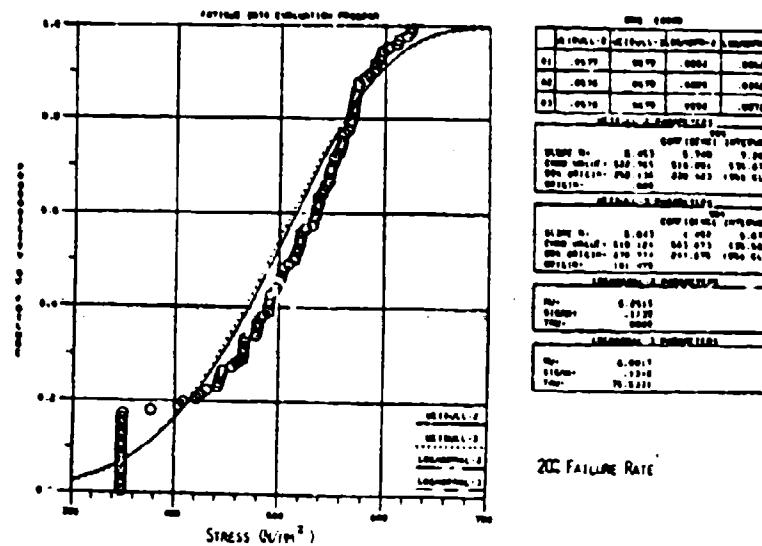


Figure 7

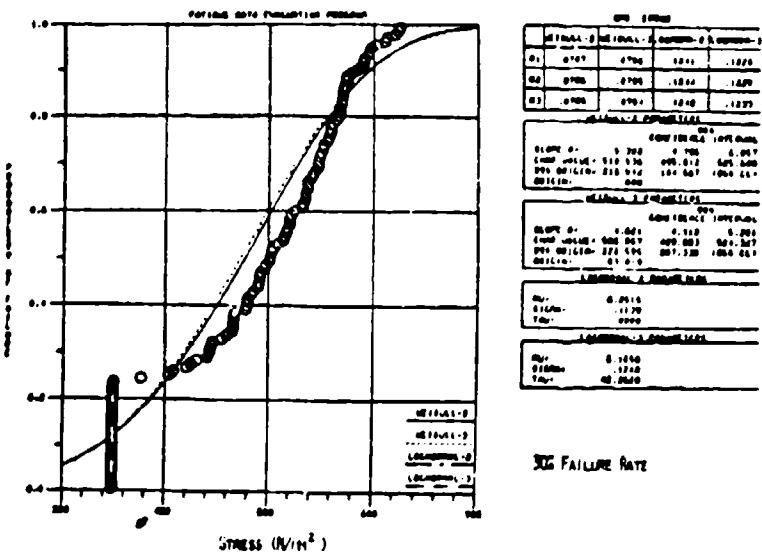


Figure 8

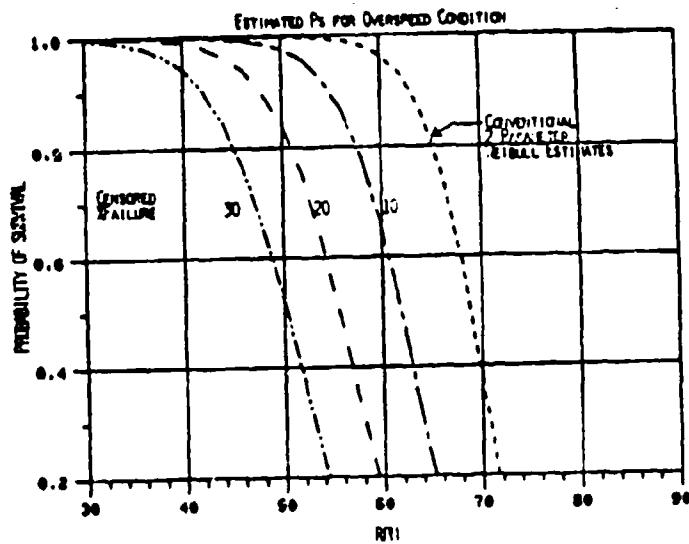


Figure 9

Fatigue Data Evaluation (Helicopter Engine Component)

The low cycle fatigue data with upper censoring (run-outs) shown in Table I was obtained by strain control mode of testing where total axial strain range is the controlled parameter being held constant. The material is a HIP René 95 powder metal used in helicopter engine components.

Run (i/m ²)	FATIGUE TEST RESULTS René 95 Powder Metal Censored Data		
	Cycles (Failure)	Log Cycles	
1	.019	3358	3.923
2	.038	1161	4.056
3	.058	10201	4.212
4	.077	16408	4.216
5	.096	17013	4.241
6	.115	17634	4.241
7	.135	17761	4.246
8	.154	17888	4.254
9	.173	18056	4.469
10	.192	18164	4.468
11	.212	20103	4.507
12	.231	23355	4.589
13	.250	41982	4.623
14	.269	47377	4.674
15	.288	47309	4.675
16	.307	47300	4.680
17	.327	47978	4.681
21	.981	40978	4.681

TABLE I

Initially the ranked data tabulated in Table I was plotted on Weibull probability paper (see Figure 10). A regression line was obtained from an optimum condition coefficient results. The 95% confidence limits for the line are shown in the figure. The 3.54 cycles designation describes the 99% probability for a larger log cycle to failure. That is, there is a one percent chance for the log-cycles to failure to be less than 3.54. The 95% confidence limit was determined in conjunction with the 99% greater cycles to failure in order to describe an A-allowable for fatigue strength.

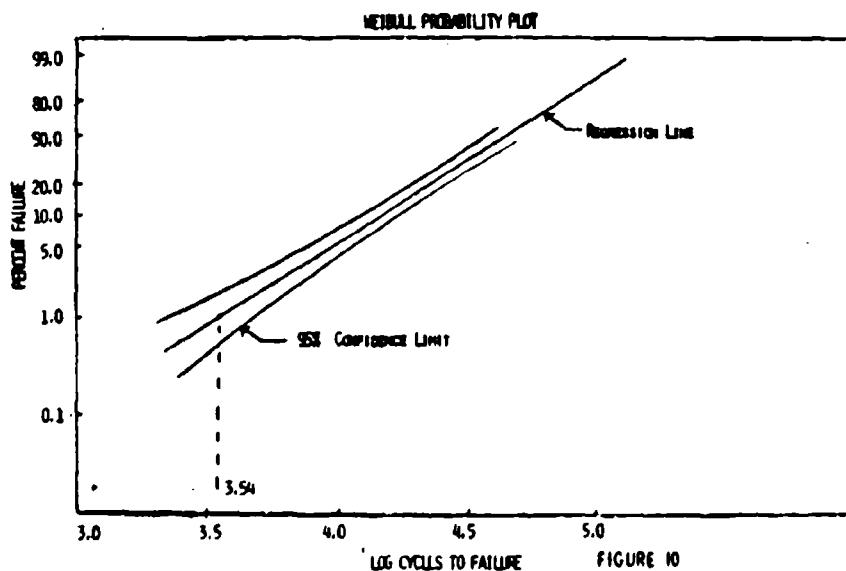


FIGURE 10

Figure 10

In Figure 11, a plot of the ranked data with the corresponding functional relation are shown. Functional parameters were obtained from the censored data analysis described previously. The failure to represent lognormal results was due to excessively large percent of censoring. This could have been corrected if partial censoring was implemented. The two parameter Weibull functions excellent representation discouraged the need for this modification. The relatively small RMS values for the Weibull function are consistent with results for the small residuals noted in the graph (see Figure 11). The broken line representing the graphical method results shows relatively poor representation of the data. This was not noticeable from the graphical plot of data. Although this poor representation occurs in this instance, other sets of censored data were well represented by the graphical procedure. The problem exists, in that the graphical method is not consistent in providing a good representation. The M.L. method applied to the censored case invariably results in a desirable data representation. A tabulation of the A and B allowables are

shown in the figure. Note the relatively large differences in the results from the two methods. The consequence of this overly conservative estimate from the graphical procedure can result in unnecessary rejection of a very expensive engine component. Figure 12 shows the results from another set of data where a considerable large amount of censoring exists. Note, the excellent representations of this data by the M.L. method.

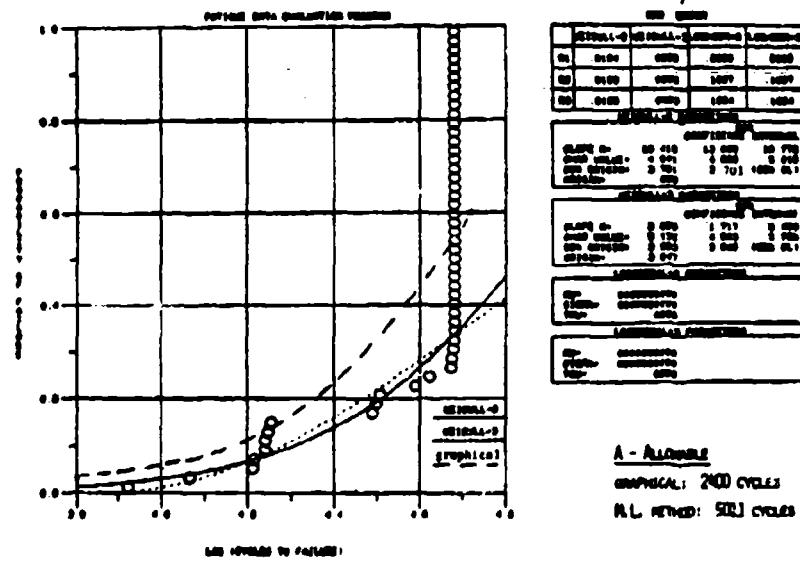


FIGURE 11

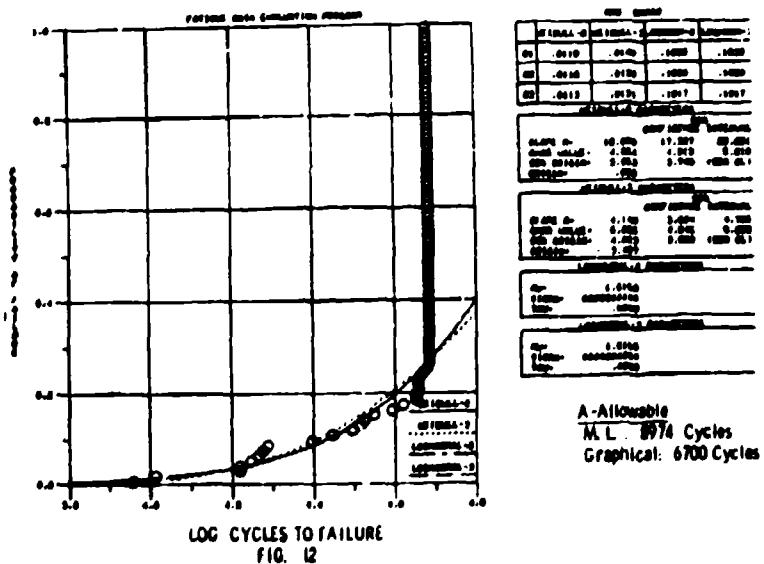


FIG. 12

CONCLUSIONS

The analytical treatment of truncated data obtained from the ceramic rotor has been discussed in some detail since the results have important implications regarding use of proof testing and qualification data. Furthermore, calculations of the type presented herein are of importance for establishing meaningful production lot sampling procedures which rely on limited quality assurance and spin test data. It was evident that neglect of statistical information contained in the spin test failure rate data, leads to non-conservative mathematical representations for material behavior.

The previous discussion obviously represents one narrow facet of analyzing failure of components. Thus far nothing has been said concerning time dependent failure response of structural ceramics. It is worthwhile commenting on studies directed towards the objective of understanding such phenomena. While fairly extensive data is available to the designer which permits materials choices for particular applications, it is worth noting that many of the inherent mechanisms which produce microstructural and physical and chemical changes are not fully understood. It is not possible at this time to present a comprehensive mathematical model for all ceramics which accurately accounts for all controlling materials phenomena, such as the physical changes induced under severe environmental limits, as well as creep, slow crack growth and other aspects of time dependent behavior.

An accurate determination of a prescribed probability for specific minimum of cycles to failure of the René 95 material can be realized if all data is censored, represented by the Weibull function where corresponding parameters are described by M.L. methods. The uncertainty involved in the graphical approach should be avoided. The argument that it is easier to use is not valid at the present time. The simplest programmable calculator can provide the necessary computation for the M.L. method. Although probabilistic life estimates have been made for the previously mentioned material, the M.L. method can also be applied to most other fatigue data.

REFERENCES

1. J. R. Benjamin and C. A. Cornell (1970), Probability, Statistics and Decision for Civil Engineers, McGraw Hill, Inc., pp. 396-402.
2. H. Leon Harter and Albert H. Moore (1965), "Maximum Likelihood Estimation of the Parameters of the Gamma and Weibull Populations from Complete and from Censored Samples", *Technometrics*, Vol. 7, pp. 639-643.
3. H. Leon Harter and Albert H. Moore (1966), "Local Maximum Likelihood Estimation of the Parameters of Three-Parameter Lognormal Populations from Complete and Censored Samples", *Journal of the American Statistical Association*, Vol. 61, pp. 842-851.
4. R. Beeuwkes, Jr. and D. M. Neal (1978), "A Simple Density Function with Finite Distribution Limits", presented at 25th Conference of Army Mathematicians, Johns Hopkins University, Baltimore, Maryland.
5. James Knaub, Jr., "Small Sample Size Effects on Tolerance Limits, Exceedance", presented at 25th Conference on the Design of Experiments in Army Research, Development and Testing", U. S. Army Natick R&D Laboratories, Natick, MA.

RISKS TO NEIGHBORING FACILITIES*

Paul C. Cox
2930 Huntington Drive
Las Cruces, New Mexico
(505) 522-1756

ABSTRACT: Many military installations, as well as industrial facilities, conduct operations which can present a safety hazard to personnel, property, vehicles, industrial facilities, and communities that lie in the neighborhood of the installation. This report considers a military installation which tests missiles and rockets, as an example, and discusses procedures for estimating risks from these operations to neighboring facilities. The procedures of this report should also be applicable to many other types of operations that are found on a variety of military installations and with a number of industries. Estimates of risk will be provided for certain critical points and also in the form of contour maps, which will show the risks for the entire region. Methods for obtaining confidence limits for these risks will also be discussed. Finally, some suggestions are offered regarding comparisons of risks from operations to every day life; and from these comparisons, it may be possible to decide whether the risks from an operation are sufficiently small to be accepted.

1. INTRODUCTION:

a. Operations conducted by many military installations may cause safety hazards to personnel, property, vehicles, industrial facilities, and communities in the neighborhood of the installation. These operations include the testing of rockets, missiles, airborne targets, aircraft, explosive devices, materiel emitting radiation, etc. It is the purpose of this report to discuss a few methods for estimating the risks created by military operations to neighboring facilities.

b. Specifically, consider a military installation with the primary mission of testing rockets and missiles, and the risks that may occur as a result of a malfunctioning round flying off-course, impacting in an undesired location, and causing serious damage to an industrial facility located at the unplanned impact point. The methods of this report are easily extendable to other types of military installations, various types of industrial operations, and a variety of possible targets.

*This is a condensation of the original report. A copy of the complete report may be obtained by writing to the author at the above address. 99

c. The purpose of this study may be to consider the risks upon one (or possibly two or three) specific target. This target may be an industrial plant, a town, a highway, etc. On the other hand, it may be desired to learn the risks at every point lying within the region surrounding the military installation. If it is the latter, the end product of the study may be a contour map of the area, with contour lines indicating the probability that during any 12 month period, a malfunctioning object may strike at a point along the line and do damage greater than at some specified level. The reasons for studying the risks for an entire neighborhood include: (1) The entire region around this installation may be covered with industrial plants, farms and ranches, communities, highways, and other points of concern; or (2) a company may want to locate a plant somewhere in the region around the installation and will want to know which areas are safe enough.

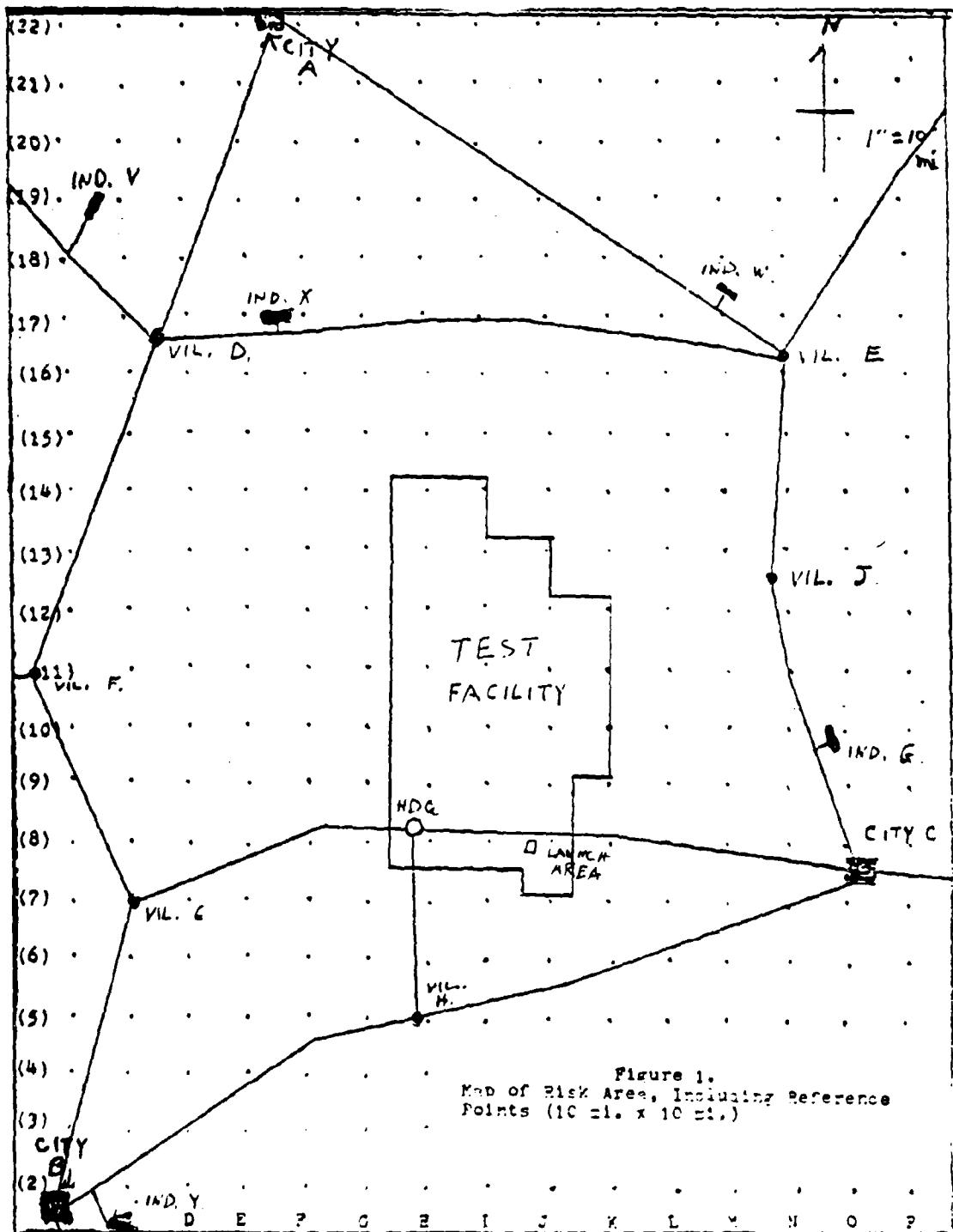
d. A primary working tool for this study is a set of maps. These maps will cover the entire region of concern. They will be in black and white, will show very little detail, but will show the boundaries of the test facility, major highways, larger communities and points of interest. These maps will also contain reference points, which may be thought of as the points of intersection of equally spaced vertical and horizontal lines. The closeness of these reference points will depend upon the accuracies desired and the amount of work that one wishes to do when evaluating the data. Figure one is an example of such a map, with reference points located 10 mi. x 10 mi. apart. Actual working maps should be two to three times as long and wide as figure one.

e. The target that will be used as the example for this report will be an industrial complex covering 100 acres = .1563 sq. mi. It will be assumed that if an object tested under the project under consideration impacts within this 100 acre complex, damage at an unacceptable level has a 100% probability of occurring. (Note references 2 and 3, or Appendix E of the original report, for some techniques for the extent of damage to expect if an impact occurs.) Finally, risks will be computed over a 12 month period of time. It is then believed that the risks obtained from a study such as this can easily be extended to other types of targets and for different periods of time.

2. Project Classification:

a. Record all test programs that are presently assigned to the installation as well as those expected to be assigned within the next few years. Also, review some of the programs that were previously assigned to the installation, because some of these might provide information that can be useful in the evaluation of present or future systems.

b. Determine which projects present no risk to targets of concern and remove them from further study. These projects may involve testing objects with insufficient range to reach the boundaries of the test installation; the test objects may be of such material that they will do no serious damage if they do impact in a critical area; or the system may be so reliable that an unplanned impact is virtually impossible.



c. Collect the following information for all of the remaining projects:

- (1) System design and performance characteristics.
- (2) Expected reliability and accuracy of the system.
- (3) Expected number of future tests.
- (4) Expert opinion on the reliability and performance characteristics of the system.
- (5) Mass, shape, penetration capabilities, and other destructive characteristics of the test object.
- (6) If test data exists, collect the following:
 - (a) Total number of tests that have been conducted.
 - (b) Number of rounds that have malfunctioned, resulting in unplanned impacts.
 - (c) The location of unplanned impacts.

d. Order the projects on the basis of the amount of test data that is available. Analyze those projects with a good deal of data first, because the results from these projects may be useful when evaluating those projects with little or no test data.

3. Constructing a Footprint:

a. A footprint must be constructed for each project which can provide a threat to any target under consideration. This footprint, when constructed, will be superimposed upon figure one, as illustrated by figure 2. (See page 7). It should be pointed out that some projects will require more than one footprint, one for each test configuration.

b. The footprint that is being used to illustrate these procedures consists of a set of concentric ellipses. The procedures used to construct this footprint are discussed in detail in appendices A, C, and D of the original report. Examples of other types of footprints are discussed in Appendix B of the original report. The footprint used in this study may be desirable if it can be determined that:

- (1) The unplanned impacts appear to be distributed approximately as the bivariate normal.
- (2) The center of impact, the angle of rotation, and the variances can all be estimated by one procedure or another.

c. It is assumed that in this illustrated example, there exists a considerable amount of test data which can be used to estimate the required parameters. The coordinates of the unplanned impacts are listed in table 1 (Page 6). There will be many instances in which it will be necessary to estimate these parameters by methods that do not depend upon test data.

d. The project used to illustrate these procedures will be referred to as Project A. The following information will be used to construct this footprint, in addition to the information listed in paragraph 1e:

- (1) 240 relevant tests have been conducted under project A. Of this number, 18 rounds were unreliable and unplanned impacts resulted. Impact data for the 18 unplanned impacts may be found in columns 2 and 3 of table 1.
 - (2) It has been estimated that there will be about 32 tests each year for several years to come. Thus, we can expect about 2.40 unplanned impacts per. year.
 - (3) Assume that for this type of test, a flight surveillance system has a 90% capability of destroying or diverting malfunctioning rounds so that no damage will occur to a target.
- e. The x and y coordinates of the 18 data points (see col. 2 & 3 of table 1) must be transformed as follows before constructing the footprint:
- (1) The (x,y) coordinates must be translated to provide an (x',y') coordinate system such that \bar{x}' and \bar{y}' equal zero. The (x',y') coordinates are listed in columns 4 & 5 of table 1, and the values of $x' \cdot y'$ are listed in column 6.
 - (2) A rotation of axes, providing (x'',y'') coordinates, is necessary so that $r_{x''y''}$ will equal zero. The rotation formulas are: $x'' = x' \cdot \cos \theta + y' \cdot \sin \theta$ and $y'' = -x' \cdot \sin \theta + y' \cdot \cos \theta$
- Where θ may be obtained as follows:

$$\tan 2\theta = \frac{2 \sum x' \cdot y'}{\sum x'^2 - \sum y'^2}$$

The derivation of the above formulas may be found in appendix C of the original report.

The (x'',y'') coordinates for the 18 unplanned impact points may be found in columns 7 and 8 of table 1, page 6. The values of $x'' \cdot y''$ are listed in column 8. It may be seen from table 1 that $\bar{x}'' = \bar{y}'' = r_{x''y''} = 0$, which is exactly what the translation and rotation was expected to accomplish. Using (x',y') data from table 1, $\theta = 37.7^\circ$.

- f. The footprint will now be constructed and superimposed upon figure one, using the following procedures. This will be illustrated by figure two.

- (1) The footprint will consist of 9 arbitrarily chosen, concentric ellipses. These ellipses will be constructed to contain 20%, 40%, 60%, 80%, 90%, 95%, 99%, 99.5%, and 99.9% of the expected unplanned impacts. If the number of ellipses is increased, it should result in improving the precision of the estimates of risk. It will, however, increase the labor required to obtain the estimates.

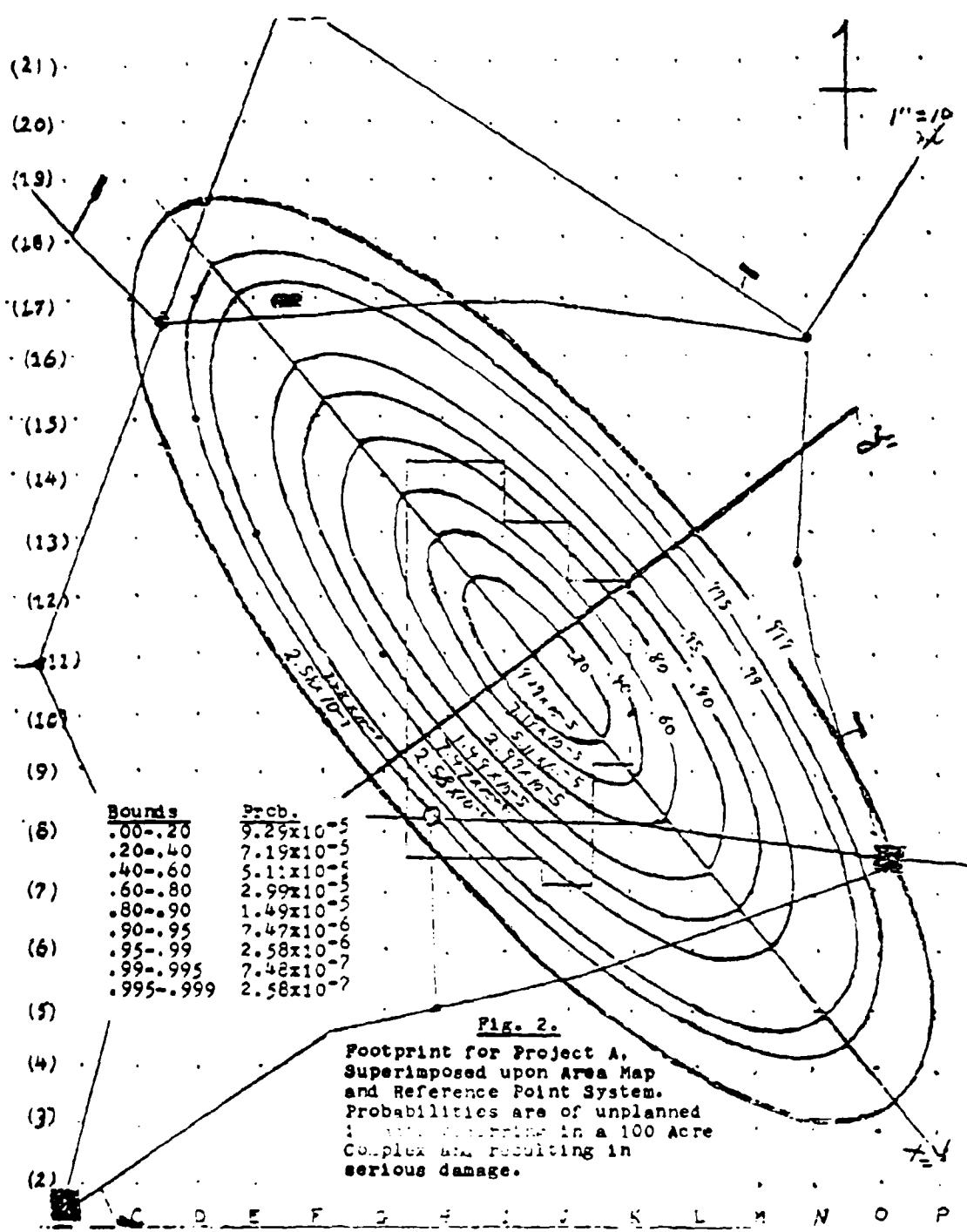
(2) The 9 ellipses will be of the form: $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$,
 since $r_{x''y''} = 0$. $a = k \cdot s_x$; $b = k \cdot s_y$; $k = \sqrt{-2 \cdot \ln(1 - p)}$.
 The formula for k is developed from the x' distribution
 with 2 degrees of freedom, in appendix D of the original
 report.

(3) Table 1, showing the (x, y) ; (x', y') ; and (x'', y'') coordinates
 for the 18 unplanned impact points for Project A, is shown
 below.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
n	x	y	x'	y'	$x' \cdot y'$	x''	y''	$x'' \cdot y''$
1	-36	-4	-19	-14	250	-23.6	0.5	-11.8
2	-30	-4	-13	-14	162	-18.8	-3.1	58.3
3	-23	-1	-6	-11	60	-11.5	-5.0	57.5
4	-30	4	-13	-6	78	-14.0	3.2	-44.8
5	-28	2	-11	-3	33	-10.5	4.4	-46.2
6	-22	5	-5	-5	25	-7.0	-0.9	6.3
7	-22	11	-5	1	-5	-3.3	3.8	-12.5
8	-23	16	-6	6	-36	-1.1	8.4	-9.2
9	-18	9	-1	-1	1	-1.4	-0.2	0.3
10	-12	1	5	-9	-45	-1.5	-10.2	15.3
11	-13	8	4	-2	-8	1.9	-4.0	-7.6
12	-15	13	2	8	16	6.5	5.1	33.2
13	-11	16	6	0	36	8.4	1.1	9.2
14	-10	21	7	11	77	12.3	4.4	54.1
15	-6	13	11	3	33	10.5	-4.3	-45.2
16	-6	19	11	9	99	14.2	0.4	5.7
17	-1	18	16	8	128	17.6	-3.4	-59.8
18	0	23	17	13	221	21.4	-0.1	-2.1
Sum:	-306	180	0	0	1167	0.1	0.1	0.7
Means:	-17	10	0	0	0	0	0	0
Sum of Sq.			1860	1250		2761.5	347.2	
St.Dev.			10.4	8.6	0.77	12.745	4.519	0.0
r_{xy}								

TABLE 1
 (x, y) ; (x', y') ; and (x'', y'') Coordinates of 18 Unplanned Impact Points.

(4) Figure 2, shows the footprint for Project A,
 superimposed upon the area map of figure one.



4. Assignment of Probabilities to Reference Points.

a. Reference point E-14 will be used as an example to illustrate how risks are assigned. Note that point E-14 is located within the band which is bound by the .95 and .99 ellipses. (see the footprint, fig. 2.)

b. Begin by computing the probability that over a period of 12 months, a test object from Project A will impact within a target as described in para. 1a, and do damage at an unacceptable level. The computation will be for a target lying within the band bound by the .95 and .99 ellipses, the band which contains the reference point E-14. Therefore, the probabilities will apply to such a target lying at or near point E-14.

- (1) The risk from a single test assigned to Project A is as follows:

$$R = .1563 \times 1.00 \times 18/240 \times .10 \times .04/582.674 = \\ 8.0474 \times 10^{-8}, \text{ where:}$$

- (a) .1563 sq. mi. = 100 acres, which is the size of the industrial complex under consideration.
- (b) 1.00 is the probability that damage at an unacceptable level will occur if there is an impact within the industrial complex. (note appendix E of the original report for further discussion).
- (c) 18/240 = .0750 comes from 18 unplanned impacts (unreliable rounds) from a total of 240 tests.
- (d) .10 is the estimate of the probability that the flight surveillance system will fail to destroy or divert the unreliable round in such a way as to avert an unacceptable level of damage. Assume this was based upon 380 attempts in which 38 were not successful.
- (e) .04 is the probability of falling in the band that is bound by the .95 and .99 ellipses.
- (f) 582.674 sq. mi. is the area of this band.

- (2) Since it is estimated that there will be 32 tests per. year under Project A, the risk to the target from Project A, over a 12 month period is as follows:

$$P(\text{Risks for 32 tests}) = 1 - (1 - R)^{32}. \text{ However, for small } R, \\ P = 32 \cdot R = 32(8.0474 \times 10^{-8}) = 2.5752 \times 10^{-6}$$

Also, for small R, this can be extended to any number of years by multiplication. For example, for 25 years,

$$P = 25(2.5752 \times 10^{-6}) = 6.438 \times 10^{-5}.$$

Note that 2.5752×10^{-6} is the risk assigned to the band bound by the .95 and .99 ellipses. Risks from Project A to the other bands are computed by a similar method. (Note the footprint on figure 2.)

c. Finally, the risks will be computed for a target as described in para. 4e, located at point E-14, from all projects located at the installation, and for a period of 12 months.

- (1) The first step is to review all footprints from all projects and observe which contain point E-14.
- (2) Then, using the methods of para. 4b, determine risks for point E-14, from all relevant footprints. These risks will be added together to give the overall risk from the entire military installation. It was observed that there were 10 footprints that covered the point E-14, and the risks associated with each is listed below.

PROJECT	RISK	PROJECT	RISK
A	2.58×10^{-6}	F	0.55×10^{-6}
B	$1.32 \times "$	G	$2.11 \times "$
C	$0.32 \times "$	H	$1.05 \times "$
D	.00	I	$0.08 \times "$
E	$1.18 \times "$	J	$1.12 \times "$
		Sum:	10.32×10^{-6}
			$= 1.032 \times 10^{-5}$

- (3) From these results, it can be seen that the probability is about one in 100,000 that during any 12 month period an object from the installation may impact at the target site, located near point (E,14), and do damage at an unacceptable level.

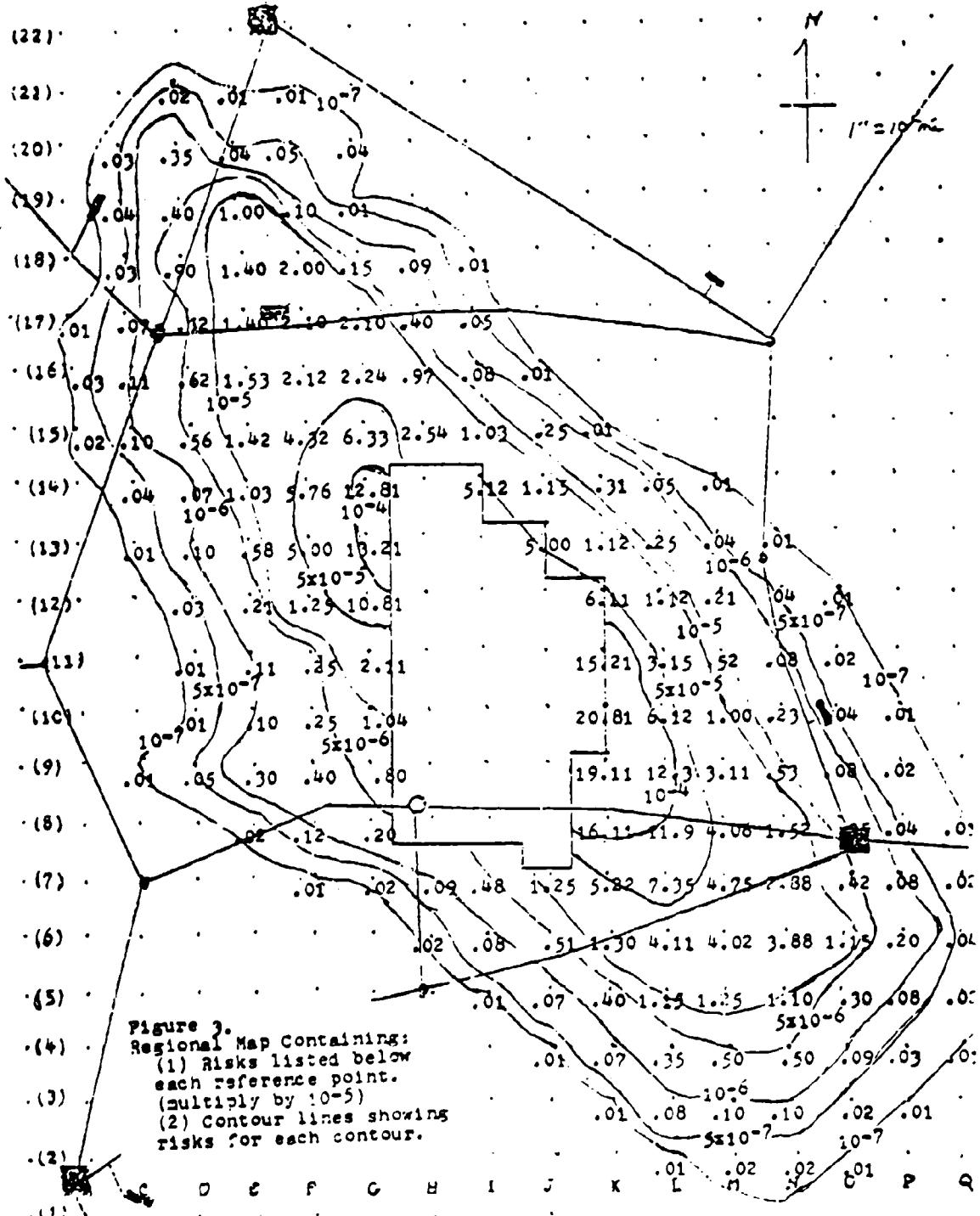
5. Constructing the Reference Point and Contour Maps

a. By using the procedures of section 5, it is possible to assign a probability to every reference point on the map, and it is quite possible that a map with the level of risk recorded at each reference point is all that a user will want. Such a map is illustrated by figure 3.

b. If a contour map is desired, proceed as follows:

- (1) Determine the probabilities desired to assign to each contour line.
- (2) Locate the adjacent reference points, with probabilities just greater than and just less than that of the contour lines being considered.
- (3) By appropriate interpolation (probably logarithmic), determine where the contour lines should lie between these ref. points.
- (4) Connect these points by ruler and french curve and thus construct the contour lines.

c. Figure 3 illustrates how the reference points can be labeled, showing the risks of impact and serious damage at each of these points. Then, using the reference points as a guide, a contour map has been superimposed upon the map of the region. If the labeled reference points appear to be most useful, it will be unnecessary to construct the contour lines.



6. Confidence Limits.

A great deal of additional study is needed to develop improved procedures for determining confidence limits for risks at desired points throughout the region surrounding the military installation. Pages 9, 11, and 12 of the original report discuss how a modification of the procedures of "propagation of errors" may be used to obtain estimates of the upper confidence limits for the risks. References 4, 5, and 6 discuss the methods of propagation of errors. It is hoped, however, that better methods than propagation of errors will someday be found.

7. How Safe is Safe Enough?

a. When it is indicated from a study that the risk at some point is some figure (one in 100,000 for example), the user is likely to ask several questions about this figure, including:

- (1) What does this level of risk mean?
- (2) Is this level safe enough?
- (3) Why should we accept any risk?

b. The following suggestions may help in answering some of the questions of the users.

- (1) Risks from a military installation may be compared to risks that occur in every day life. To make these comparisons, such publications as "Accident Facts" (see reference 12) may provide some useful information. For example, about one out of every 400 Americans will die as a result of a motor vehicle accident within the next 12 months, and one of every 100 Americans can expect to receive a disabling injury from the same source. Thus, if it can be shown that the operations from the installation are very much less than the risks from the daily use of the American automobile, this may be convincing.
- (2) Perhaps risks can be reduced to economic terms. For example, if a plant is worth 10 million dollars, and the annual risk is one in 100,000. This would indicate a risk of only \$100 per year. Thus, insurance may be available to cover such a risk.
- (3) Finally, when users are reluctant to accept any risk, it should be pointed out that some risk is associated with every activity carried on by the human race. Thus, it is necessary to find ways to reduce the risk of all activities to an acceptable level.

9. References:

1. Owen, D.B. Handbook of Statistical Tables 1962, Addison Wesley Publishing Co., Inc. Reading, Mass. Note:
 - P. 184, a function for computing bivariate normal probabilities. (this table is especially useful if $r_{xy} \neq 0$).
 - P. 170, Critical values of the Circular Normal Distribution. (This table can be useful for computing circular or elliptical footprints).

The next two reports can be useful in determining the capability of targets to withstand impact from an object. Note Appendix E.

2. Kennedy, R.P. A Review of Procedures for the Analysis and Design of Concrete Structures to Resist Missile Impact Effects., Nuclear Engineering and Design, vol. 37, 1976, p. 183-203 North Holland Publishing Co.
3. Structural Analysis and Design of Nuclear Plant Facilities. Prepared by the Editing Board and Task Groups of the Committee of Structures and Materials of the Structural Division of the American Society of Civil Engineers, J.D. Stevenson, Chairman of Editing Board and Task Group. Note especially the chapter written by R.P. Kennedy.

The next three references can be useful in getting a background in the theory of propagation of errors.

4. Scarbrough, James B., Numerical Mathematical Analysis, 6th Edition, 1966, Johns Hopkins Press, Baltimore, Md.
5. Ku, Harry H., Notes on the Use of Propagation of Error Formulas, 1965, National Bureau of Standards Report no. 9011.
6. Hahn, Gerald J. and Shapiro, Samuel S., Statistical Models in Engineering, 1967, John Wiley and Sons, New York (note P. 252-255)

The next report can be useful in obtaining daily risks to life and health in many occupations and in just ordinary living. This information may be useful in comparing with risks obtained from special studies. Note section 8, p. 12, of this report

7. Accident Facts, published yearly by the National Safety Council, 425 N. Michigan Ave., Chicago, Ill. 60611

THE 1980 SAMUEL S. WILKS MEMORIAL MEDAL

Frank E. Grubbs

The Samuel S. Wilks Memorial Medal Award was initiated in 1964 by the US Army and the American Statistical Association, and has been administered for the Army by the American Statistical Association, a non-profit, educational and scientific society founded 140 years ago in 1839. The Wilks Medal and Award is given each year to a statistician - and a top-notch one! - and is based primarily on his contributions to the advancement of scientific or technical knowledge in Army statistics, ingenious application of such knowledge, or successful activity in the fostering of cooperative scientific matters which coincidentally benefit the Army, the Department of Defense, the US Government, and our country generally. The Award consists of a medal, with a profile of Professor Wilks and the name of the Award on one side, the seal of the American Statistical Association and the name of the recipient on the reverse side, and a citation and honorarium related to the magnitude of the Award funds, which were generously donated by Phillip G. Rust of the Winnstead Plantation, Thomasville, Georgia. Mr. Rust originally stimulated the interest of Sam Wilks in distributional properties of the "extreme spread" (bivariate range), a measure of the "accuracy" of rifle shot on a target.

These annual Army Design of Experiments Conference, at which the Wilks Medal is awarded each year, are sponsored by the Army Mathematics Steering Committee on behalf of the Office of the Chief of Research, Development and Acquisition, Department of the Army.

BIOGRAPHY OF THE RECIPIENT OF THE 1980 SAMUEL S. WILKS MEMORIAL MEDAL

by

Churchill Eisenhart

The 1980 Samuel S. Wilks Memorial Medalist is an internationally recognized authority on statistics whose leadership has contributed greatly to the adoption, acceptance, and effective use of statistical thinking and statistical methods in many areas of research and human affairs, in both the governmental and private sectors.

He was born in Philadelphia, Pennsylvania, on November 5, 1912. When three years old, his family moved to the West Coast and lived successively in Berkeley, Fresno, and Los Angeles, California, and Portland, Oregon, while his father, primarily a physical anthropologist trained at Oxford under a Rhodes Scholarship, with Ph.D. in Philosophy from the University of Pennsylvania, taught at the University of California at Fresno State College, served with the California Commission on Emigration and Housing, and taught at Reed College.

At the age of 9, he had a newspaper route in Portland and won a Thanksgiving turkey for an unusually large increase in circulation. Recalling this at the time (1962) of his appointment as President of the University of Rochester, he said: "I remember this partly because it was my first lesson in pitfalls of statistical measurement. The base set for measuring the growth of my route was the month of August, and it was no feat at all to triple circulation when the Reed College faculty returned from their vacations and especially when hundreds of students arrived at the college. Most of all, I remember that turkey because my father and I received it--alive--in downtown Portland and took it home by streetcar."

The family moved to Minneapolis, Minnesota in 1923 when his father joined the faculty of the Department of Anthropology at the University of Minnesota. As a boy in Minneapolis, our Wilks Medalist became an ardent stamp collector, tennis player and photographer. He organized a small stamp company, selling stamps partly by mail but mostly to boys in the neighborhood. One of his best customers was Richard M. Scammon, who was later to become Director (1961-1965) of the U. S. Bureau of the Census and to serve with our Medalist on the President's Commission on Federal Statistics.

Our 1980 Wilks Medalist entered the University of Minnesota in the fall of 1928; majored in psychology with a minor in sociology, took nearly as much work in mathematics and in philosophy, and gained valuable writing experience as an editorial writer for the college paper, the Minnesota Daily, then known as "the world's largest college daily". His high scholarship led to his election to Phi Beta Kappa, and to receipt of his A.B. degree magna cum laude in 1932 at the age of 19. A paper on "The Influence of Color on Apparent Size" which he wrote during his junior year was published in the Journal of General Psychology, Vol. 13 (1935) and has been reprinted in books of readings in psychology.

Shortly before graduation he decided on a career in economics, with emphasis on mathematical economics and statistics. He remained at the University of Minnesota for the academic year 1932-1933 to continue his study of mathematics and to study economics and then moved on to the University of Chicago, where he held a University Fellowship in the Department of Economics from 1933 to 1935. It was at Chicago that he began life-long friendships with Milton Friedman and George J. Stigler (father of the present Theory and Methods Editor of the Journal of the American Statistical Association). They (and two other fellow students) selected, arranged, and saw through to publication, the first book of Professor Frank H. Knight's essays, The Ethics of Competition (Harper & Brothers, 1935)--an early instance of our Medalist's drive to see worthwhile material formally published in the open literature.

He spent the academic year 1935-1936 as Granville W. Garth Fellow in Political Economy at Columbia University, where he studied with, among others, Wesley C. Mitchell, one of the most eminent of American economists, doyen of American business cycle analysts, the 1918 President of the American Statistical Association, co-founder (1920) and director of research of the National Bureau of Economic Research (an independent non-profit organization), etc., and especially with Harold Hotelling, a pioneer in mathematical economics, econometrics, and multivariate statistical analysis, and the person in the United States then most versed in R. A. Fisher's theory of small samples and statistical inference. Furthermore, this was the year in which Hotelling revealed the potential of systematic treatment of functions of the relative ranks of sample observations as a basis for what are now termed "distribution-free" or "non-parametric" methods of statistical inference, through his joint paper with Margaret Richards Pabst, "Rank correlation and tests of significance involving no assumption of normality". Presented at the New York meeting of the American Mathematical Society on October 26, 1935, and published in the March 1936 issue of the Annals of Mathematical Statistics (Vol. 7, 29-43), this paper marked "the true beginning" of research on such methods as "an important special field of statistics" (I. Richard Savage, JASA, Vol. 58, p. 844, December 1953). Hotelling himself, his teaching, and his research exerted a far reaching influence on our 1980 Wilks Medalist's career in statistics as will become evident as we proceed. For the moment we may note simply that our Medalist's first statistical paper, "The Poisson Distribution and the Supreme Court", published in the June 1936 issue of JASA (Vol. 31, 376-380) was written as a course paper for Hotelling.

During the summer of 1935 and the academic year 1936-37, our 1980 Wilks Medalist was an economist for the National Resources Committee, a New Deal agency in Washington, D.C. Milton Friedman was there also, and while there, wrote his paper, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance" (JASA 32, 675-701, December 1937) in which he thanks our Medalist for bringing to his attention a more informative method of handling tied ranks. While serving with this Committee, our Medalist co-authored a book on estimates of consumer expenditures in the United States for 1935-1936, and worked on an article on the temporal stability of consumption that saw publication in 1942.

His first teaching appointment was as Instructor in Political Economy at Yale University in 1937-38. "The high-point of that year," he has said, "was getting to know Irving Fisher who was especially hospitable because of my interest in mathematical economics, a field in which he had pioneered forty years earlier."

In the fall of 1938 our Wilks Medalist joined the Department of Economics at Stanford University, an association that continued, with extensive interruptions, until 1946. At Stanford he taught courses in economic theory, mathematical economics, and advanced statistics. His first contribution to statistical methodology, "The correlation ratio for ranked data", published in the September 1939 issue of JASA (Vol. 34, 533-538), grew out of consulting he did with a psychologist during his first year at Stanford, and anticipated, or perhaps we should say paralleled, independent work by M. G. Kendall and B. Babington Smith, published in the September 1938 issue of the Annals of Mathematical Statistics (Vol. 10, No. 3, 275-287--our Medalist's "rank correlation ratio", r_s , is exactly

their "coefficient of concordance", W. A dozen years later, at the University of Chicago, he returned to consideration of methods of analysis of ranked data as means of avoiding the implications of the normality assumption underlying many common statistical tests, and with William H. Kruskal prepared a comprehensive treatment of the "Use of ranks in one-criterion variance analysis", published in the December 1952 issue of JASA (Vol. 47, 583-621), in which they introduced their now widely used H test, thus designated in honor of Hotelling.

During 1939-40 and the last half of 1941, our 1980 Wilks Medalist was a Carnegie Research Associate at the National Bureau of Economic Research (NBER) in New York City, on leave of absence from Stanford; and took advantage of the proximity of Columbia University to attend the lectures there of Abraham Wald, newly arrived (1939) from Austria. At the NBER he was closely associated with Arthur F. Burns (later Chairman, President's Council of Economic Advisors; Chairman, Board of Governors of the Federal Reserve System, etc.), Wesley C. Mitchell (mentioned above), Frederick C. Mills (1934 President of the ASA; and author of a statistical methods text, the second edition of which in 1938 incorporated many of the new ideas and methods of R. A. Fisher and was used widely by students in economics, business and other fields), and Geoffrey H. Moore (who later became the 1968 President of the American Statistical Association and Commissioner 1969-1973, of Labor Statistics, U. S. Department of Labor.) Analysis and interpretation of economic time series occupied center stage at the NBER. With Moore he published A Test of Significance for Time Series and Other Ordered Observations (National Bureau of Economic Research Technical Paper 1, September 1941, 59 pp) in which they developed a test for randomness, relative to either a monotonic or oscillatory trend, based on the distribution of length of runs up and down; and two articles (with Moore) in JASA, "A significance test for time series" (Vol. 36, 401-409, September 1941), and "Time series significance tests based on signs of differences" (Vol. 38, 1953-1964, June 1943). The first provided a brief summary of the Technical Paper, with examples of the application of the test developed therein; the second, an alternative but not independent test based on the total number of runs up and down. These two tests are today standard tools of nonparametric statistics.

At NBER, he also carried out much of the research embodied in his paper "Compounding probabilities from independent significant tests", published in the July-Oct. 1942 issue of Econometrica (Vol. 10, Nos. 3&4, 229-248), in which he gave a clear mathematical exposition of the basis of R. A. Fisher's procedure for combining "significance probabilities" yielded by independent statistical tests having continuous probability distributions (Fisher, Statistical Methods for Research Workers, Fourth Edition (1932), Sec. 21.1)--the basis of which was a mystery to many individuals and incorrectly explained by others--and provided the requisite mathematical extension to cases in which at least one of the "significance probabilities" is obtained from a statistical test having a discrete probability distribution such as, for example, a rank or run test. At that time, too, he was co-author with Milton Friedman of a paper on the empirical derivation of indifference functions which saw publication in 1942.

Our 1980 Wilks Medalist returned to Stanford University for the first half of 1942. The United States was then at war with both Germany and Japan, so rationing and price control were matters of paramount concern. Our Medalist responded by writing a paper, "How to ration consumers' goods and control their prices", published in the American Economic Review later that year. Then, on April 17, 1942 our 1980 Wilks Medalist wrote to W. Edwards Deming (then Head Mathematician, Mathematical Advisor, U.S. Bureau of the Census) stating that he and several of the others teaching statistics in various departments of Stanford considered it "probable that a good many students with research training might by training in statistics become more useful than in their present work, or might increase their usefulness within their present fields" and asked for Deming's advice on the development of "a curriculum adapted to the immediate statistical requirements of the war". Deming responded by April 24, 1942, on the letterhead of the Chief of Ordnance, War Department, suggesting a concentrated effort--a "short" course followed by a "long" course on Shewhart methods of quality control, the short to be "for executive and industrial people who want to find out some of the main principles and advantages of a statistical program in industry"; the long course for "people who actually intend to use statistical methods on the job; "both courses [to] be thrown open to engineers, inspectors, and industrial people with or without mathematical or statistical training". (Portions of both letters are reproduced on pages 320-321 of the June 1980 issue of JASA.) In one of the paragraphs not reproduced, Deming points out the relevance of Wallis and Moore's work to statistical quality control, adding: "The theory of runs and patterns is destined to receive a great deal of attention from now on, and it is a pleasure to see the superb effort that you and Mr. Moore have put forth."

The impact of Deming's suggestions was such that by May 1st, Holbrook Working (Statistician and Economist at the Stanford Food Research Institute, and Chairman of the University Committee on Statistics) had arranged a general meeting of everyone in statistics; a first letter about the course went out on May 21 to firms in the Western states that were supplying Army ordnance; and in July 1942 the first course was given at Stanford, by Working and Eugene L. Grant (of the Engineering School.) Our Medalist had been scheduled to teach this course (with Grant), and had been "beginning to wonder how to learn what [he] was supposed to teach", when he was asked "to head up an economic research unit in the Office of Price Administration" in Washington. So he dropped out and was replaced by Working. The course was such a success that early in 1943 Working was chosen

to head the now famous major national program that put on intensive 8-day courses in statistical quality control throughout the country, under the auspices of the Office of Production Research and Development of the United States Office of Education. By March 1945 these had been attended by more than 1900 persons from 678 industrial concerns in the United States and 13 in Canada. Many of the "students" in the earlier of these courses went out to serve as "instructors" in part-time courses that brought the message to an additional 3,100 persons in American and Canadian industry. This program had an enormously beneficial effect on the quality and volume of American and Canadian war production; and "prepared the soil" for the establishment of the American Society for Quality Control, in February 1946. That our 1980 Wilks Medalist played a role in initiating, and came so close to participating in this very effective venture was a secret well kept from many of us until we saw mention of it in the June 1980 issue of JASA.

Our 1980 Wilks Medalist never made it to the OPA position in Washington. Before his appointment to that position became official, he received a telegram from Warren Weaver, Director (Natural Sciences) of the Rockefeller Foundation (1932-1955) then up to his ears in support of the war effort as Chairman (1940-42) of Section D-2 of the National Defense Research Committee (NDRC) of the Office of Scientific Research and Development (OSRD). Weaver, whom our Medalist has described as "one of the most remarkable, admirable, brilliant, sagacious and civilized human beings on the American scene in the past half-century" had perceived an urgent need for a concentrated effort focused on resolution of the various mathematical and statistical problems that were arising in the several armed services and suppliers of their material, and especially those problems that were arising more or less simultaneously in different places with, as he put it, "the same verbs but different nouns"; and was engaged in setting up several mathematical and statistical groups to do the "spade work" of the soon to be established Applied Mathematics Panel of the NDRC, of which he was to be the Chief (1943-46), and Thornton C. Fry (of the Bell Telephone Laboratories), the Deputy Chief. Wilks had suggested to Weaver the establishment of a statistical group at Columbia with Hotelling as Principal Investigator, and Hotelling had brought our 1980 Wilks Medalist to Weaver's attention. Thus it came to pass that on July 1, 1942 our Medalist assumed his first administrative post, Director of Research of the Statistical Research Group (SRG) at Columbia University in New York City.

SRG got off to a start with just three experienced researchers or "principals" as he terms them in his article "The Statistical Research Group, 1942-45" in June 1980 issue of JASA (Vol. 75, 320-330): Hotelling, our Medalist, and Jacob Wolfowitz, another former student of Hotelling. Before its dissolution on September 30, 1945 the number of "principals" had risen to 17--or to 18, if Frederick Mosteller is included, who though actually on the payroll of another group, worked closely and extensively with this SRG for essentially one full year and co-authored two of its books and co-edited one of these. (The names of all 18, with their respective lengths of service with SRG, are listed on page 324 of our Medalist's aforementioned article.) These "principals" were supported at one time or another by about 60 others: typists, secretaries, a switchboard operator, an administrative assistant, a librarian, a messenger, and about 30 young women, mostly mathematics graduates of Hunter or Vassar, who did the necessary computing under the direction of Albert Bowker. The "principals"

worked on problems of tactics, equipment, and operations for the Army, Navy, the Air Force (which was a branch of the Army in World War II), and other units of OSRD. Many of these activities stemmed from AMP studies assigned to SRG, but a large number stemmed from consultation with and informal assistance to Army, Navy, or NDRC groups. Sometimes one problem would lead to a related problem in another setting, or experience with a particular technique would lead to another application of the technique to an unrelated problem.

Our Medalist mentions in his article a great number of the military problems on which SRG worked, so there is no need to give such details here. The most famous, and probably the most influential and lasting contribution was, of course Abraham Wald's development of sequential analysis, full instructions and tables for the practical application of which saw open publication in Sequential Analysis of Statistical Data: Applications (Columbia University Press, 1945); and the theoretical development, in Wald's book Sequential Analysis (Wiley, 1947). Our Medalist in his article gives two accounts of the history of sequential analysis, one written in April 1943, soon after the development; and the other written from memory in March 1950, when the 1943 memorandum could not be located. Both accounts bring out clearly the essential roles of our Medalist and Milton Friedman in getting the development "off the ground" after they had recognized its possibility of achievement.

Although SRG was formally dissolved on September 30, 1945, our Medalist stayed on until March 31, 1946 to make sure that some of SRG's other wartime contributions achieved open publication in a unified form creditable to both the individuals concerned and SRG. Although he listed himself alphabetically as the third editor of Selected Techniques of Statistical Analysis (McGraw-Hill, 1947), and alphabetically as the fourth of the editors of Sampling Inspection (McGraw-Hill, 1948) he was in fact the Editor-in-Chief for both.

As Director of Research of SRG, our 1980 Wilks Medalist brought together an absolutely extraordinary group of research workers in statistical theory and methodology, in both number and quality. The experience of working in SRG contributed significantly to the subsequent careers of a substantial number of the "principals". Many became leaders in statistics in the next three decades. Four became President of the American Statistical Association: Bowker (1964), our Medalist (1965), Mosteller (1967), and Eisenhart (1971). Seven became President of the Institute of Mathematical Statistics: Wald (1948), Girshick (1952), L. J. Savage (1958), Wolfowitz (1959), Bowker (1962), Herbert Solomon (1965) and Mosteller (1975)--Hotelling had been President in 1941. Mosteller is the 1980 President of the American Association for the Advancement of Science. Friedman received the Nobel Prize in Economics for 1976. At least nine subsequently became chairmen of university departments of statistics: K. A. Arnold, Bowker, Girshick, Hotelling, Mosteller, Savage, Solomon, Wald, and our Medalist. Two became heads of major universities: Bowker (of two: City University of New York and University of California at Berkeley), and our Medalist (University of Rochester). Three received the Samuel S. Wilks Medal: Solomon (1975), Eisenhart (1977), now this year's Medalist.

The influence of SRG continues through the work of its "principals" alive and deceased (Girshick, Hotelling, Savage, Wald) and through the statistical tools developed at SRG, theoretical and practical, which have become established parts of statistics. Several effective statistical consulting groups have since been modeled on SRG, notably those at the Bell Telephone Laboratories and at the National Bureau of Standards. But even today it is probably not saying too much to say that SRG was the best statistical research and consulting group ever. Those who worked there know this to be the consequence of the high standards of excellence established, maintained, and insisted upon by its Director of Research, our 1980 Wilks Medalist.

After all that, whatever more is said is bound to be anticlimatic, but needs to be said nonetheless to round out the record and give the full picture of this champion of statistical theory and methodology.

In the Spring of 1946 our Medalist returned to Stanford University as Associate Professor of Economics, and immediately instituted steps toward the establishment of a Statistics Department there. However, before that department came into being, he had left in the Fall of 1946 to join the University of Chicago as Professor of Statistics and Economics in the Graduate School of Business. (Later he was also named a Professor in the Department of Economics, in the Division of Social Sciences). In 1949, he became Chairman of the newly formed Department of Statistics in the Division of the Physical Sciences, a post he held until 1957. During his chairmanship, the Department of Statistics at the University of Chicago became one of the outstanding departments in its field in the world a position that it still retains. (In addition, he played behind-the-scenes roles in the establishment of Department of Statistics at Columbia University, Harvard University, and the University of Rochester, making a total of five whose establishment he influenced in minor to major ways.) In 1956 our Wilks Medalist was appointed Dean of the University of Chicago's Graduate School of Business, became financially self-supporting while tripling its annual expenditures, and came to be widely recognized as one of the Nation's very best.

There is an amusing side to our Medalist's appointment to his first tenured professorship at the University of Chicago. He has no so-called "earned degrees" beyond his 1932 A.B. from the University of Minnesota. He had satisfied nearly all the requirements for a Ph.D., some at the University of Chicago, others at Columbia University; had had two thesis accepted; but before he had completed the remaining requirements at the University of Chicago, he was appointed a professor with permanent tenure there and became ineligible under that university's rules to receive an advanced degree from it. He has, however, received three honorary degrees, Doctor of Science from Hobart and William Smith Colleges (1973), Doctor of Laws from Roberts Wesleyan College (1973), and Doctor of Humane Letters from Grove City College (1975).

While at the University of Chicago, our 1980 Wilks Medalist published a number of noteworthy papers on statistical methodology. The first was a long paper, "Standard sampling-inspection procedures", presented at the 25th Meeting of the International Statistical Institute at Washington, D.C., in 1947 and published subsequently in its Proceedings, Vol. 3, 331-350. This was essentially an exposition of the basic principles and state of the art of acceptance sampling procedures as spelled out in more detail in the SRG book, Sampling Inspection (1948), then in press. Next was a basic paper, "Tolerance Intervals for Linear regression", presented at the Second Berkeley Symposium on Mathematical Statistics

and Probability at Berkeley, California, in the summer of 1950, and published in its Proceedings, 43-51. "Rough-and-ready statistical tests", published in the March 1952 Issue of Industrial Quality Control (Vol. 8, No. 5, 35-40) was a composite and updated version of some notes on these matters made available to SRG staff during WW II, updated to include some of the material to be published in the forthcoming (1952) joint paper with W. H. Kruskal, "Use of ranks in one-criterion analysis of variance", mentioned earlier. With Harry V. Roberts (an Associate Professor of Statistics in the School of Business) he co-authored Statistics: A New Approach (The Free Press, 1956), an 84-page work that became a widely used text in English speaking countries and saw translation into German (1959) and Portuguese (1964). A paperback version of the first quarter was issued by Collier Books (1962) under the title, The Nature of Statistics, and has been translated into Swedish, Danish, Norwegian, and Japanese.

In addition, while at Chicago, our 1980 Wilks Medalist served as the Editor of the Journal of the American Statistical Association for nearly a decade, 1950-1959. During 1955 he chaired an Inner-University study group formed under the aegis of the University of Chicago, and funded by the Ford Foundation, to reach a decision on the desirability of a new or revised edition of the Encyclopedia of the Social Sciences (that had been published in 15 volumes by the Macmillan Company between 1930 and 1935. (The study group included members from the University of California at Berkeley, Harvard University, University of Illinois, Reed College, and Princeton University--see David L. Sills, "Editing a Scientific Encyclopedia", Science, Vol. 163, 1169-1175, 14 March 1969.) The project layed dormant for five years, until late 1960, when the Macmillan Company decided to publish a new encyclopedia of the social sciences, The International Encyclopedia of the Social Sciences (IESS), which saw publication in 17 volumes by the Macmillan Company and The Free Press in April 1968. Our 1980 Medalist served as Chairman of the Editorial Advisory Board, and as Chairman of the Executive Committee, for this vast undertaking. Unlike its predecessor and other encyclopedias, the IESS contains a great many articles on statistical concepts, theory, and methodology, together with biographies of a host of individuals who made significant contributions to statistical thinking and practice, excluding those still alive in the 1960's. Consequently, anyone wishing to know something about the contemporary state of statistics--concepts, theory or methodology--or about their historical development, found this encyclopedia a most useful source. It proved so useful in this regard that the articles on statistics and articles relevant to statistics published in the IESS were brought up to date by the addition of Postscripts or by revision in whole or in part, and republished together with a few additional articles and biographies, as the International Encyclopedia of Statistics, two volumes, by The Free Press in 1978.

Our 1980 Wilks Medalist became President, Professor of Economics and Statistics, and Trustee of the University of Rochester in 1962. His title was changed to Chancellor in 1970. In 1975, in anticipation of retirement, he turned over the chief executive responsibilities but remained in administration. In 1978, he retired as an officer of the university but continues there with the same title, Chancellor.

He has been a director of Bausch & Lomb, Inc., since 1963, Eastman Kodak Company since 1965, Lincoln First Banks, Inc., since 1967, Macmillan, Inc., since 1964, Metropolitan Life since 1973, Rochester Telephone Corporation since 1964, Standard Oil Company (Ohio) since 1977, and Trans Union Corporation since 1962; and was for fourteen years a director of Esmark, Inc., (1963-1977).

He was a consultant to the RAND Corporation from 1948 until 1966; a member (1952-1953) of an advisory panel to the Secretary of Army on operations research; a member (1969-1970) of the President's Commission on an All Volunteer Armed Force; Chairman (1969-1978) of the Commission on Presidential Scholars, and Chairman (1970-1971) of the President's Commission on Federal Statistics, as well as a member of chairman of various other Presidential or national commissions and councils. The present-day Committee on National Statistics of the National Research Council was established on the recommendation of "his" President's Commission on Federal Statistics to grapple with and help resolve conflicts over statistical aspects of such national problems as environmental monitoring, presentation of statistical evidence in court, effect of changes in stratospheric ozone on incidence of skin cancer, and recently the 1980 Census undercount.

Even more could be said about this Wilks-like individual, but the foregoing is more than sufficient to explain why the 1980 Samuel S. Wilks Memorial Medal is awarded

To W. Allen Wallis in recognition of his extraordinary contributions to the effective use of statistical theory and methodology by the armed services during World War II, for his outstanding contributions to clear statistical thinking and effective statistical practice through the publications he authored or edited, for his leadership of statisticians, and for his service to the nation through chairmanship of, or membership on numerous high-level Governmental and non-Governmental commissions and councils.

OPTIMAL ESTIMATION TECHNIQUES FOR FORECASTING PROPAGATION PARAMETERS

K. E. Kunkel and D. L. Walters
Atmospheric Sciences Laboratory
White Sands Missile Range, New Mexico

ABSTRACT

The prediction of stochastic atmospheric variables such as wind and temperature on time scales of 6 hours down to a few seconds is being attempted. The statistical problem is to find the optimal estimation technique that combines the existing climatological data base with current measurements to provide the best real time forecast. An autoregressive technique has been attempted but yields results which reduce the error by only 10% over persistence forecasts. Panel recommendations indicate that mixed autoregressive-moving average techniques and Kalman filter techniques are the most promising to attempt.

I. INTRODUCTION

An important goal of meteorology is to accurately predict the state of the atmosphere at some time in the future. For the large scale features of the atmosphere, this can be done in a quasi-deterministic (if somewhat inaccurate) sense by predicting the movement of large scale weather systems. However for some purposes, such as predicting the optical propagation characteristics of the atmosphere, it is necessary to know the small scale, or turbulent, nature of the wind and temperature fields. The problem that we wish to address here is the prediction of stochastic atmospheric variables for time scales of six hours down to a few seconds.

II. DATA DESCRIPTION

An example of the kind of parameter that needs to be predicted is the temperature near the surface of the earth. Fig. 1 shows a time series of temperature for a 15 min period at heights of 3 and 33m. This figure illustrates the stochastic nature of the variable. In general, the short time scale variations cannot be predicted deterministically. However, the data can be characterized in a statistical sense. The following are typical characteristics of the data:

- 1) non-zero, non-stationary mean
- 2) The power spectral density follows a $k^{-5/3}$ behavior (k = wave-number) for $k > k_{\min}$ where k_{\min} is some wavenumber scale of the flow.

For $k < k_{\min}$, the power spectral density has no definite shape. Fig. 2 shows the power spectral density for temperature at 3 and 33m above ground. The $k^{-5/3}$ behavior is exhibited for $\log k > -1.5$. Assuming ergodicity, k can be related to the frequency domain by $k = 2\pi f / \bar{U}$ where f = frequency and \bar{U} = mean wind speed.

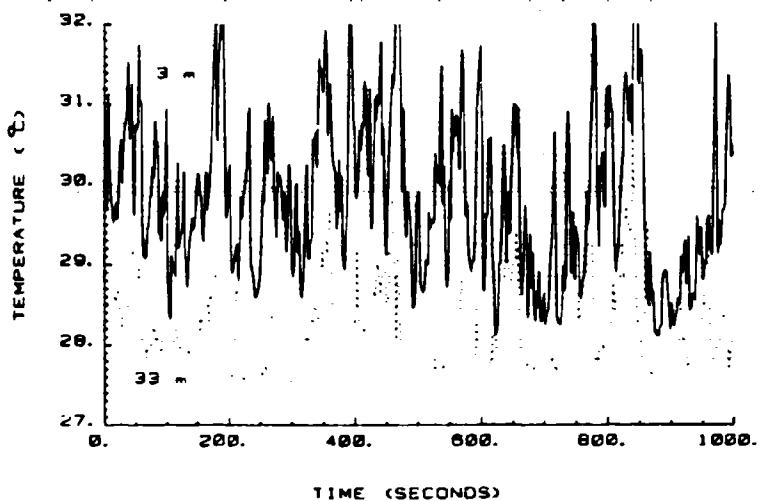


Figure 1. Time series of temperature ($^{\circ}\text{C}$) at heights of 3m (solid) and 33m (dotted).

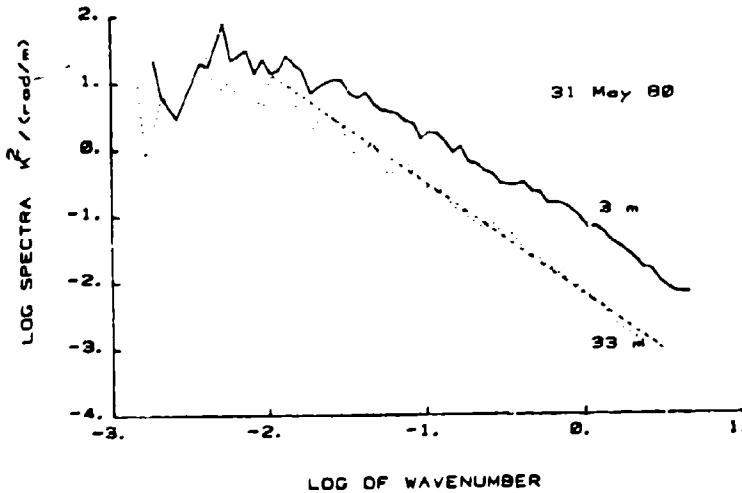


Figure 2. The \log_{10} of power spectral density of temperature vs. \log of wavenumber at heights of 3m (solid) and 33m (dotted). Units of wavenumber are rad/m. Dashed line show a power spectral density $\propto k^{-5/3}$ dependence.

3) Low frequency trends are often present. These are usually tied to the daily cycle of heating and cooling and occur at frequencies in the range $10^{-4} - 10^{-5} \text{ s}^{-1}$.

There are two sources of data available that can be used as the basis for making a prediction. These are:

1) Climatological data base. These data provide general characteristics of the data in the past. These include low frequency trends

tied to the daily cycle and typical expected power spectral levels as a function of time of day and other external factors.

2) Measurements taken on the current day.

The problem then becomes to predict a stochastic variable given climatological data and present measurements.

III. RESULTS

One attempt has been made to solve this problem by using an adaptive linear prediction filter similar to one described by Keeler and Griffiths (1977). This is an autoregressive approach and can be briefly summarized as follows. If x is the variable (with mean removed) to be predicted, then the prediction at time t , $x_p(t)$, is given by

$$x_p(t) = \sum_{i=1}^I g_i x(t - i\Delta t)$$

where $x(t - i\Delta t)$ are measured values, Δt is the time interval, and g_i are weighting coefficients. The prediction error $E(t)$ is given by

$$E(t) = x(t) - x_p(t)$$

Since in general we don't know how to calculate a priori the coefficients g_i , we allow the coefficients to be changed as data is collected in order to provide the minimum mean square error. An algorithm is used which updates the coefficients as each measurement sample is collected by using the method of steepest descent. This is given by

$$g_i(t + \Delta t) = g_i(t) + \mu E(t) x(t - i\Delta t)$$

where

$$\mu = \frac{\alpha}{I \sigma_x^2}$$

σ_x^2 = variance of x

α = constant which determines rate of convergence

This type of algorithm was applied to a number of data sets with Δt ranging from 10 secs to 15 min. The predictions were compared with predictions based on persistence, i.e.,

$$x_p(t) = x(t - \Delta t)$$

By using a wide range of α and I values, the best we could do was to decrease the mean square error by 0-10% over persistence. This is not very encouraging.

IV. QUESTIONS

Our questions to the panel are:

- 1) Given the nature of the data, can we predict these quantities significantly more accurately than by simply using persistence or the climatological mean?
- 2) What are the limits of predictability? Can these limits be calculated?
- 3) What prediction technique is likely to be most successful for this problem? Possible techniques that have been discovered in the literature are:
 - a) Autoregressive (all-pole)
 - b) Moving average (all-zero)
 - c) Mixed pole-zero (Box-Jenkins)
 - d) Kalman type filter

V. PANEL RECOMMENDATIONS

The problem is a difficult one and may not be amenable to solution. However, two techniques should be attempted. One is the mixed autoregressive moving average technique (Box-Jenkins) for which software packages exist. The other is the Kalman filter technique.

REFERENCES

Keeler, R. J. and L. J. Griffiths, 1977: Acoustic doppler extraction by adaptive linear-prediction filtering. J. Acous. Soc. Am., 61, 1218-1227.

SOME RESULTS FOR THE UNIVARIATE NORMAL RANDOM
LINEAR REGRESSION MODEL PREDICTION THEORY

D. G. Kabe

New Mexico State University, Las Cruces, New Mexico

and

St. Mary's University, Halifax, N.S., Canada

ABSTRACT. Optimal prediction, within the normal theory framework, of one vector variable by the linear functions of another correlated random vector variable, when certain values on the predicted variables are missing is considered. The optimal predictors are derived by using both the conditional expectation minimization theory and the canonical correlation theory. However, the maximum likelihood estimators of the unknown parameters are derived only for the canonical correlation theory.

I. INTRODUCTION. This paper presents some of the author's discussion (as one of the panelists) on the following two papers presented at the twenty-sixth United States Army conference, on Design of Experiments, held at New Mexico State University, 22-24, October 1980. The first paper, "Optimal Estimation Techniques for Forecasting Propagation Parameters," was presented by K. E. Kunkel and D. L. Walters of the United States Army, White Sands Missile Range, New Mexico; and the second paper, "A Stochastic Mesoscale Meteorologic Model," was presented by E. P. Avara of the United States Army, White Sands Missile Range, New Mexico. Both the papers studied the optimal prediction theory of one vector variable by the linear functions of another correlated random vector

variable, and the second paper considered such theory when certain sample values on the predicted variable were (missing) unobservable. In his discussion the author pointed out some known results (the derivations given here are different), within the normal theory framework, to the above problems.

The prediction of one vector random variable by the linear functions of another correlated random vector variable, within the normal distribution theory framework, is a very well known problem in statistical literature, and an exhaustive paper on this topic is published by Scobey and Kabe (1980). The two classical techniques often used for this purpose are: 1) the conditional expectation minimization theory (CEMT), and 2) canonical correlation theory. In brief we shall discuss these two techniques. Both are based on the fact that for two correlated vectors x, y

$$E(y-f(x))'(y-f(x)) , \quad (1)$$

is minimized when $E(y|x) = f(x)$, and $f(x)$ is the optimal predictor of y , for a given x .

The known results of CEMT and canonical correlation theory (CCT) are presented in the next section, prediction intervals are derived in section 3, and missing values are considered in section 4.

Sometimes the same symbol denotes different quantities, however, its meaning is made explicit in the context.

II. SOME USEFUL RESULTS. We first present the results for CEMT. Let y be a q component vector (all vectors are column vector and all matrices are full rank matrices) with $E(y) = 0$, $E(yy') = \Sigma$, $E(x) = 0$, $E(xx') = \Delta$, $E(yx') = B$, B $q \times p$, x $p \times 1$, $p \leq q$, and then consider the minimum value problem

$$\begin{aligned}
 & \underset{\Lambda}{\text{Min}} \left\{ \underset{x}{\text{Min}} E \left[(y - \Lambda x)' (y - \Lambda x) | x \right] \right\} \\
 &= \underset{\Lambda}{\text{Min}} E \left\{ \underset{x}{\text{Min}} \left[(y - E(y|x))' (y - E(y|x)) \right. \right. \\
 &\quad \left. \left. + ((\Lambda x - E(y|x))' (\Lambda x - E(y|x)) | x \right] \right\},
 \end{aligned} \tag{2}$$

which by using (1) may be written as

$$\underset{\Lambda}{\text{Min}} E \left\{ E(y - \Lambda x)' (y - \Lambda x) | x \right\}, \quad \Lambda x = E(y|x). \tag{3}$$

However, (3) reduces to

$$\begin{aligned}
 & \underset{\Lambda}{\text{Min}} E (y - \Lambda x)' (y - \Lambda x) = \underset{\Lambda}{\text{Min}} \text{tr} [\Sigma - 2AB' + \Lambda \Delta \Lambda'] \\
 &= \underset{\Lambda}{\text{Min}} \text{tr} [\Sigma - B\Delta^{-1}B' + (\Lambda - B\Delta^{-1}) \Delta (\Lambda - B\Delta^{-1})'],
 \end{aligned} \tag{4}$$

where Λ is $q \times p$. From (4) obviously

$$\Lambda = B\Delta^{-1}, \quad \text{i.e.,} \quad \Lambda \Delta = B, \tag{5}$$

is a necessary condition for our minimization problem.

Now we have to find that Λ , which yields $(\Sigma - \Lambda \Delta \Lambda')$ singular. We now consider

$$\underset{\Lambda}{\text{Min}} \text{tr} [\Sigma - \Lambda \Delta \Lambda'], \tag{6}$$

to find the minimum of (2). By our assumption Λ satisfies

$$0 = |\Sigma - \Lambda \Delta \Lambda'| = |\Sigma - B\Delta^{-1}B'|. \tag{7}$$

whence a solution Λ to (7) is

$$\Delta^{\frac{1}{2}} = T \Lambda^{\frac{1}{2}} (Q' 0)', \quad \Delta \Delta' = \lambda_1 t_1 t_1' + \dots + \lambda_q t_q t_q', \quad (8)$$

where

$$\begin{aligned} \Sigma &= T \Delta T' = (t_1, \dots, t_q) \Lambda (t_1, \dots, t_q)' \\ &= \lambda_1 t_1 t_1' + \dots + \lambda_q t_q t_q', \end{aligned} \quad (9)$$

and $\lambda_1 > \dots > \lambda_q$ are the roots of Σ and $q \times q T$ is the matrix of the latent vectors of Σ , $q \times p$ is any arbitrary orthogonal matrix. With A given by (8), we find from (6) that

$$\min_A \text{tr} [\Sigma - A \Delta A'] = \lambda_{p+1} + \dots + \lambda_q. \quad (10)$$

Now with A given by (8), we find it convenient to denote Ax by \hat{y} , and say that \hat{y} optimally predicts y , for a given x (actually \hat{y} estimates $E(y|x)$), i.e.,

$$\hat{y} = Ax = B \Delta^{-\frac{1}{2}} x, \quad B \Delta^{-\frac{1}{2}} = A \Delta^{-\frac{1}{2}} = T \Lambda^{-\frac{1}{2}} (Q' 0)'. \quad (11)$$

When A is given by (8), equation (11) predicts y for a given x by CEMT. Thus (8) implies that CEMT predicts y optimally by linear functions of x by assuming all canonical correlations between y and x are unity, i.e., from (8) the generalized variance of $(x' y')'$ vanishes. CEMT exactly follows CCT except that in CCT the equation (8) is not satisfied. Thus CEMT deals with the singular normal distribution theory and CCT deals with the nonsingular normal distribution theory. Since the results for the singular normal distribution follow exactly on the same lines as for the nonsingular case, we consider parameter estimation for CCT only.

In CCT A satisfies (5), and also (11), except that the second member of (11) is now discarded. The p canonical correlations between y and x are the positive square roots of

$$|\rho^2 \Sigma - B \Delta^{-1} B'| = 0 , \quad (12)$$

and the canonical variates of y and x corresponding to a particular ρ are $\xi'y$ and $n'x$, where ξ and n satisfy

$$\begin{bmatrix} \rho^2 \Sigma & -B \\ -B' & \Delta \end{bmatrix} \begin{bmatrix} \xi \\ n \end{bmatrix} = 0 . \quad (13)$$

If all the canonical correlations are zero, then y cannot be predicted by linear functions of x. When all canonical correlations are unity, then y is a linear function of x, given by (11). When some canonical correlations are unity, and others between zero and unity, then the prediction is to be carried on partly by JEMT and partly by CCT. An example of such a case is given by Kshirsagar (1962), and discussed by Scobey and Kabe (1980). If all canonical correlations are between zero and unity, then the first member of (11) holds and A satisfies (5), but not (7). When (12) holds $a'y$ is optimally predicted by a linear function of x only if a is proportional to ξ , where ξ satisfies

$$(\rho^2 \Sigma - B \Delta^{-1} B')\xi = 0, \quad \text{and} \quad \rho\xi = \Sigma^{-1} B n . \quad (14)$$

In this case

$$|\Sigma - B \Delta^{-1} B'| = |\Sigma| \left| I - \frac{1}{2} B' \Sigma^{-1} B \frac{1}{2} \right|$$

If the means of y and x are not zero, then (11) changes to

$$\hat{y} = E(y) + A(x - E(x)) = E(y) + B\Delta^{-1}(x - E(x)) . \quad (16)$$

Thus to predict y optimally by linear functions of x , the population parameters must be known. If the population parameters are unknown, then they are replaced by their maximum likelihood estimators, when a sample of size N on $(x', y')'$ is available. Thus e.g., in the usual notation

$$\begin{aligned} |S_{12} - S_{21}S_{11}^{-1}S_{12}| &= |S_{22}| |I - S_{11}^{-\frac{1}{2}}S_{12}S_{22}^{-1}S_{21}S_{11}^{-\frac{1}{2}}| \\ &= |S_{22}| \prod_{i=1}^p (1 - \hat{\rho}_i^2) , \end{aligned} \quad (17)$$

is the sample counterpart of (15), and

$$U = Y - S_{21}S_{11}^{-1}X, \quad UU' = (S_{22} - S_{21}S_{11}^{-1}S_{12})/N , \quad (18)$$

is the maximum likelihood estimator of

$$V(y - \hat{y}) = V(y - Ax) = \Sigma - B\Delta^{-1}B' . \quad (19)$$

However, the optimal properties of such sample counterparts are not as yet fully investigated in statistical literature.

III. PREDICTION INTERVALS. We obtain prediction intervals for a single future observation. These prediction intervals are based on Rao's U statistic, see Kabe (1968).

Let a $(p+q)$ component vector $(x', y')'$ have a $(p+q)$ variate normal distribution with mean value μ , and covariance matrix Σ . Then assuming μ, Σ to be partitioned correspondingly we have

$$E(y|x) = u_2 - \Sigma_{21} \Sigma_{11}^{-1} u_1 + \Sigma_{21} \Sigma_{11}^{-1} x = n + \beta' x , \quad (20)$$

as the linear regression of y on x . The equation (20) is known as the univariate random linear regression model. We take $(\bar{y} - S_{21} \bar{S}_{11}' \bar{x})$, and $\beta = \bar{S}_{11}' S_{12}$ as the maximum likelihood estimates of n and β respectively, where $(\bar{x}', \bar{y}')'$ is the sample mean vector and S is the sample dispersion matrix based on a sample of size N on $(x', y')'$. If $(x', y')'$ is any future observation, then from (16) the predictor of y is

$$\begin{aligned}\hat{y} &= \bar{y} - S_{21} S_{11}^{-1} x + S_{21} S_{11}^{-1} x \\ &= \bar{y} + S_{21} S_{11}^{-1} (x - \bar{x}) .\end{aligned}\quad (21)$$

The joint density of S and $v = (x' - \bar{x}', y' - \bar{y}')'$ is

$$g(S, v) = K \exp \left(-\frac{1}{2} \text{tr} \Sigma^{-1} (S + hvv') \right) |S|^{\frac{1}{2}(N-p-q-2)} , \quad (22)$$

where $h = N/(N+1)$, and K as a generic letter denotes the normalizing constants of density functions in this paper.

Now partition S , v , Σ^{-1} as,

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \quad \Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix}, \quad (23)$$

and by setting

$$n = S_{22} - S_{21} S_{11}^{-1} S_{12}, \quad u = v_2 - S_{21} S_{11}^{-1} v_1, \quad \beta' = S_{21} S_{11}^{-1} , \quad (24)$$

write (22) as

$$\begin{aligned}
g(S_{11}, D, B, v_1, u) &= K \exp \left\{ -\frac{1}{2} \text{tr} \Sigma_{11}^{-1} S_{11} - \frac{1}{2} \text{tr} \Sigma^{22} D \right. \\
&\quad \left. - \frac{1}{2} \text{tr} \Sigma^{22} (B-B)' S_{11} (B-B) - \frac{1}{2} h v_1' \Sigma_{11}^{-1} v_1 - \frac{1}{2} h (u + (B-B)' v_1)' \right. \\
&\quad \left. \Sigma^{22} (u + (B-B)' v_1) \right\} \\
&\quad |v_1|^{\frac{1}{2}(N-p-q-2)} |S_{11}|^{\frac{1}{2}(N-p+q-2)} \tag{25}
\end{aligned}$$

Now integrate out B and find the density of u , v_1 , S_{11} , and D to be

$$\begin{aligned}
g(S_{11}, D, u, v_1) &= K \exp \left\{ -\frac{1}{2} \text{tr} \Sigma_{11}^{-1} S_{11} - \frac{1}{2} \text{tr} \Sigma^{22} D \right. \\
&\quad \left. - \frac{1}{2} h v_1' \Sigma_{11}^{-1} v_1 - \frac{1}{2} h u' \Sigma^{22} u / (1 + h v_1' S_{11}^{-1} v_1) \right\} \\
&\quad (1 + h v_1' S_{11}^{-1} v_1)^{-\frac{1}{2}q} |v_1|^{\frac{1}{2}(N-p-q-2)} |S_{11}|^{\frac{1}{2}(N-p-2)} \tag{26}
\end{aligned}$$

It now follows from (26) that

$$u = h u' D^{-1} u / (1 + h v_1' S_{11}^{-1} v_1), \tag{27}$$

has the density

$$g(u) = K u^{\frac{1}{2}(q-2)} / (1+u)^{\frac{1}{2}(N-p)}, \tag{28}$$

and hence the prediction intervals for y , for a given x , can be based on (28).

IV. MISSING OBSERVATIONS. If a sample of size k $(X' Y')'$ is now available on $(x', y')'$, and a sample of size $(N-k)$, $N \geq k \geq q+1$ is later available on $(x', y')'$, then the relations between the maximum likelihood estimators of

the original sample and the total sample are available in the literature, see e.g., Kabe (1967). The results given by Kabe (1967) can be easily modified to suit the missing value prediction theory.

We now state the problem proposed by E. P. Avara and its possible solution outlined by the author. The first sample is $(X', Y')'$ and the second sample is $(X_2', 0)'$, as no observations on y were recorded in the second sample. The problem is how does the entire theory of optimal prediction be carried on under such circumstances.

The fact that the problem of missing values forms a significant line of research is known in the statistical literature. However, the extreme difficulty involved in its mathematical treatment is the cause of its not being thoroughly investigated as yet.

Let \bar{x}_N denote the mean of the entire sample and \bar{x}_1 the mean of the first K observations. If S_{11} , S_{12} , S_{22} , and S_{11}^* , S_{12}^* , S_{22}^* are the old and new maximum likelihood estimates of Σ_{11} , Σ_{12} , Σ_{22} respectively, then a relation between old and new estimates is desired. Obviously $S_{11}^* = (X_N X'_N - N \bar{x}_N \bar{x}'_N)/N$ is the maximum likelihood estimate of Σ_{11} . Further from (21)

$$\bar{y} + h S_{21} S_{11}^{-1} (\bar{x}_N - \bar{x}_1) , \quad (29)$$

is the predictor of \bar{y}_N , and hence is the maximum likelihood estimate of $E(y)$. The constant $h = K/(N-K)$ from Kabe (1967). To derive the relations between the old and new maximum likelihood estimates we first consider the old data representation. If $U = [I, 0]$, $V = [0, I]$, then the sample is represented by

$$x = S_{11}^{-\frac{1}{2}} u \quad (30)$$

$$y = S_{21}S_{11}^{-1}x + (S_{22} - S_{21}S_{11}^{-1}S_{12})^{\frac{1}{2}} v \quad . \quad (31)$$

However, the entire sample must be represented by

$$x_N = S_{11}^{*\frac{1}{2}} u^*$$

$$y_N = S_{21}^{*}S_{11}^{*-1}x_N + (S_{22}^{*} - S_{21}^{*}S_{11}^{*-1}S_{12})^{\frac{1}{2}} v^* \quad . \quad (32)$$

However, since there are no new observations on (32) must reduce to

$$y = S_{21}^{*}S_{11}^{*-1}x + (S_{22} - S_{21}S_{11}^{-1}S_{12})^{\frac{1}{2}} v \quad . \quad (33)$$

Note that v and its coefficient do not change because y does not change.

It follows from (31) and (33) that

$$S_{21}S_{11}^{-1} = S_{21}^{*}S_{11}^{*-1} \quad , \quad (34)$$

and hence

$$S_{21}^{*}/N = S_{21}\bar{S}_{11}^{*}S_{11}^{*/N} \quad , \quad (35)$$

is the maximum likelihood estimate of Σ_{21} . Again from (33)

$$S_{21}^{*}S_{11}^{*-1}S_{12}^{*}/N + (S_{22} - S_{21}S_{11}^{-1}S_{12})/k \quad , \quad (36)$$

is the maximum likelihood estimator of $V(y) = \Sigma_{22}$.

This research is supported by a National Research Council of Canada grant A-4018.

REFERENCES

- Kabe, D. G. (1967). On multivariate prediction intervals for sample mean and covariance based on partial observations, J. Amer. Statist. Assoc. 62, 634-637.
- Kabe, D. G. (1968). On the distribution of the regression coefficient matrix of a normal distribution, Austrl. J. Statist. 10, 21-23.
- Kshirsagar, A. M. (1962). Prediction from simultaneous equations systems and Wold's implicit causal chain model, Econometrica 30, 804-811.
- Scobey, P, and Kabe, D. G. (1980). Some correlation optimization problems of Econometrics, Statistics, and Psychology, Biometrical J. 22, No. 7, 1-20.

ADAPTIVE MEDIAN SMOOTHING

William S. Agee and Jose E. Gomez
Mathematical Services Branch
Data Sciences Division
US Army White Sands Missile Range
White Sands Missile Range, NM 88002

ABSTRACT. We have developed a robust data smoothing method which was motivated by the necessity of extracting a small but nevertheless important signal which is imbedded in a large band of noise. It is assumed that nothing is known about the signal other than that it is of significantly lower frequency than the noise in which it is imbedded. The noise variance may vary over a rather large range. The signal to noise ratio may also vary over a large range, sometimes the signal will predominate but usually the signal will be almost invisible in a large band of noise. We adapt Tukey's idea of using medians to smooth data for exploratory data analysis to develop our robust smoother for extracting this small signal from noise. If Z_k , $k=1, 2, \dots$ are the measured values of signal plus noise, the smoothed values of this time series are given by

$$\bar{X}_k = \text{median } \{Z_{k-i}, Z_{k-i+1}, \dots, Z_{k+i}\}$$

$$i = 0, N$$

where N is variable and is made data dependent by choosing N as a function of locally computed values of the signal and noise statistics. The application of the robust adaptive smoothing technique is illustrated on some WSMR data sequences.

INTRODUCTION. Median smoothing has been strongly advocated by Tukey [1] and [2] as a tool for exploratory data analysis. Suppose we have a noisy time sequence of measurements, x_i , $i=1, 2, \dots$, which we want to smooth. Median smoothing of this measurement sequence basically means to compute a smoothed value at any time t_i by the median of the measurements about t_i . More specifically, the smoothed value at time t_i , denoted by \hat{x}_i , is computed by

$$\hat{x}_i = \text{med}_{j=0, N} \{x_{i-j}, x_{i+j}\} \quad (1)$$

The smoother in (1) has a smoothing span of $2N+1$ points.

Several advantages of median smoothing are readily apparent. Median smoothing is very simple to implement since it only requires the use of a subroutine which will order the measurements. Median smoothing does not require the specification of a model of either the signal or noise, i.e., it is a nonparametric method as opposed to most other smoothing methods. Median smoothing is robust. It tends to reject spurious values or outliers in the measurements. An outlier or short burst of outliers in the measurements will not appear in the smoothed output if the length of the burst is smaller than 11 points. Median smoothing has been applied to speech processing, [3] and [4], and to image processing, [5] and [6]. Our motivation for the development of an adaptive median smoothing routine was for the smoothing of radar error signals.

SMOOTHING RADAR ERROR SIGNALS. Let $R_o(t_i)$, $A_o(t_i)$, and $E_o(t_i)$ be the range, azimuth, and elevation output values of a radar at time t_i when tracking a target. These output readings specify a point in space at which the radar is pointing. These values are usually close to the true target range, azimuth, and elevation values, which we denote as $R(t_i)$, $A(t_i)$, $E(t_i)$. The target tracking errors are defined as

$$\begin{aligned} r_T(t_i) &= R(t_i) - R_o(t_i) \\ a_T(t_i) &= A(t_i) - A_o(t_i) \\ e_T(t_i) &= E(t_i) - E_o(t_i) \end{aligned} \tag{2}$$

Measured values, $r(t_i)$, $a(t_i)$, and $e(t_i)$ of the tracking errors are available. These measured values of the tracking errors are called the radar error signals. These error signals are usually very noisy compared to their signal content. We want to obtain smoothed values, $\hat{r}(t_i)$, $\hat{a}(t_i)$, and $\hat{e}(t_i)$ of the radar error signals in order to construct improved measurements, $R_m(t_i)$, $A_m(t_i)$, $E_m(t_i)$, of the targets range, azimuth, and elevation.

$$\begin{aligned} R_m(t_i) &= R_o(t_i) + \hat{r}(t_i) \\ A_m(t_i) &= A_o(t_i) + \hat{a}(t_i) \\ E_m(t_i) &= E_o(t_i) + \hat{e}(t_i) \end{aligned} \tag{3}$$

Median smoothing appears to be an immediately applicable method which can be quickly implemented for the task of constructing smoothed values, $\hat{r}(t_i)$, $\hat{a}(t_i)$, $\hat{e}(t_i)$ from the measured error signals, $r(t_i)$, $a(t_i)$, $e(t_i)$. However, when one attempts to directly apply median smoothing to the error signals, some common characteristics of radar error signals reduce the quality of the smoothed output to such an extent that the attempt is unsuccessful. A constant span median smoother cannot be applied successfully to smoothing radar error signals because the signal vs. noise content of the measurements varies over such a wide range during a mission. Fig 1 gives an indication of this wide variation. Initially, when the radar is acquiring the target, Fig 1 shows a rather strong signal content as compared with noise. This portion of the measurement sequence would require a short span smoother in order to preserve the signal characteristics. In the later portion of Fig 1 the range of the target from the radar is increasing and the elevation angle may be quite low resulting in a very large noise content relative to any signal which may be present. This portion of the data would require a large smoothing span in order to filter out the large amount of unwanted noise. Thus, the characteristics of the radar error signals force the use of a variable span median smoother where the span at time t_i must be dependent on the relative content of signal and noise at times near t_i . We call the result an adaptive median smoother.

ADAPTIVE MEDIAN SMOOTHING. An adaptive median smoother is defined by

$$\hat{x}_i = \text{med}_{j=0, N_i} \{x_{i \pm j}\}, \quad N_{\text{MIN}} \leq N_i \leq N_{\text{MAX}} \quad (4)$$

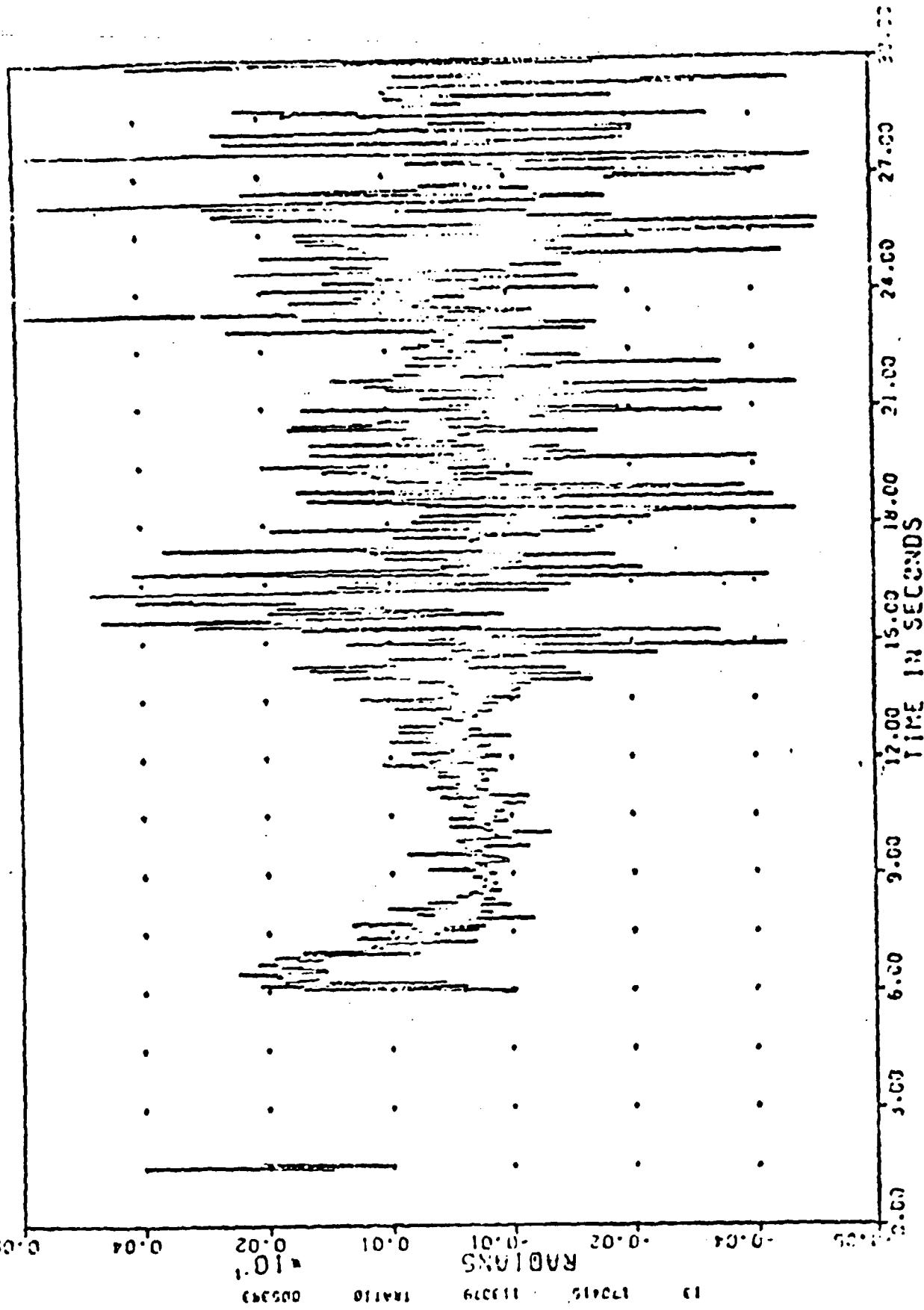
The choice of N_i is based on the measured values, x_j , where t_j is near t_i . The maximum, N_{MAX} , and the minimum, N_{MIN} , values of N_i can be specified by the general characteristics of the data and thru experience. The definition given in (4) obviously does not specify how to obtain smoothed values near the beginnings and ends of data sequences. At the start and end of data strings we use the simplest possible smoothing:

START

$$\begin{aligned} \hat{x}_1 &= x_1 \\ \hat{x}_2 &= \text{med} \{x_1, x_2, x_3\} \\ \hat{x}_i &= \text{med}_{j=0, i-1} \{x_{i \pm j}\} \quad 1 \leq N_{\text{MIN}} \end{aligned} \quad (5)$$

EL BORESIGHT

Fig. 1



END

$$\begin{aligned}\hat{x}_L &= x_L \\ \hat{x}_{L-1} &= \text{med } \{x_{L-2}, x_{L-1}, x_L\} \\ \hat{x}_{L-i} &= \text{med}_{j=0,i} \{x_{L-i+j}\} \quad i \leq NMIN\end{aligned}\tag{6}$$

In (6) the subscript L denotes the last point.

METHODS OF ADAPTATION. In order to adapt the span of the median smoother to the local characteristics of the data sequence at each time point t_i , we examine the residuals in the vicinity to determine if there is some signal remaining in the residuals indicating that we have been oversmoothing and should shorten the span or whether the residuals exhibit a random behavior indicating that we could possibly lengthen the smoothing span. We have used two different measures, the serial correlation, and the Von Neumann ratio to examine the residuals for trends. In using the serial correlation we have tried both the usual parametric definition and also a nonparametric correlation coefficient which will serve to preserve the robustness of the overall method.

Let t_i be the current time at which a smoothed value is being computed and let r_{i-j} , $j=1, N_{MAX}$ be the residuals from the smoothed values

$$r_{i-j} = x_{i-j} - \hat{x}_{i-j} \tag{7}$$

Let $\bar{r} = \text{ave}_{j=1, N_i} (r_{i-j})$ be the sample average of the residuals.

The usual definition of the serial correlation coefficient is given by

$$s = \frac{\sum_{j=1}^{N_i-1} (r_{i-j+1} - \bar{r})(r_{i-j} - \bar{r})}{\sum_{j=1}^{N_i-1} (r_{i-j} - \bar{r})^2} \tag{8}$$

If $S_L \leq S \leq S_U$ we conclude that there is no reason to believe that the residuals are serially correlated. In this case we can increase the smoothing interval, i.e., we set $NS \leftarrow NS+2$, subject to $NS \leq (2NMAX+1)$. If $S < S_L$ or $S > S_U$, we conclude that the residuals are correlated and may contain a significant signal component because of over smoothing. In this case we set $NS \leftarrow NS-2$ subject to $NS \geq (2NMIN+1)$. The upper and lower limits for large N_i are

$$S_U = \frac{-1 + 1.645\sqrt{N_i-2}}{N_i-1} \quad \text{and} \quad S_L = \frac{-1 - 1.645\sqrt{N_i-2}}{N_i-1}.$$

For $N_i < 20$ we use the tables

N	S_U	S_L
5	.253	-.753
6	.345	-.708
7	.370	-.674
8	.371	-.625
9	.366	-.593
10	.360	-.564
11	.353	-.539
12	.348	-.516
13	.341	-.497
14	.335	-.479
15	.328	-.462
16	.322	-.446
17	.316	-.432
18	.310	-.420
19	.304	-.409
20	.299	-.399

In order to preserve the robustness of the median smoother, we should use an adaptation test which is itself robust. This is easily achieved in the serial correlation case by using a rank correlation coefficient in place of the ordinary serial correlation coefficient given in (8). Let $R(j) = \text{rank}(r_{i+j})$,

$r_{i+j} \in \{r_{i+k}, k=1, N_i-1\}$ and let $R_1(j) = \text{rank}(r_{i+j+1})$, $r_{i+j+1} \in \{r_{i+k+1}, k=1, N_i-1\}$.

Then with $d_j = R_1(j) - R(j)$ we compute the Spearman rank correlation coefficient as

$$S_p = 1 - \frac{6 \sum_{j=1}^{N_i-1} d_j^2}{N_i(N_i-1)(N_i-2)} \tag{9}$$

If $S_u < S_p > S_L$ we conclude that there is no reason to believe that the residuals are serially correlated. Thus, we increase the smoothing interval by $NS \leftarrow NS + 2$, if either $S_p < S_u$ or $S_p > S_L$ we decrease the smoothing interval by $NS \leftarrow NS - 2$. For large values of N_i the upper and lower limits are

$$S_u = \frac{1.645}{\sqrt{N_i - 2}} \quad \text{and} \quad S_L = \frac{-1.645}{\sqrt{N_i - 2}}$$

For values $N_i \leq 10$ we use the following table

N_i	S_u	S_L
5	1	-1
6	.9	-.9
7	.771	-.771
8	.679	-.679
9	.643	-.543
10	.600	-.600

Another useful method for adjusting the smoothing span is the Von Neumann ratio which is the ratio of the mean square successive difference to the variance. Specifically,

$$V = \frac{\frac{1}{N_i-1} \sum_{j=1}^{N_i-1} (r_{i-j+1} - r_{i-j})^2}{\frac{1}{N_i} \sum_{j=1}^{N_i} (r_{i-j} - \bar{r})^2} \quad (10)$$

Then if $V_L \leq V \leq V_u$ we decide that there is insufficient evidence that the residuals are correlated so that we then increase the smoothing interval by $NS \leftarrow NS + 2$. If either $V < V_L$ or $V > V_u$ the residuals show evidence of being correlated so that we decrease the smoothing interval by $NS \leftarrow NS - 2$. The upper and lower limits are sample size dependent and are chosen by the following table.

N_i	S_L	S_u
5	1.0255	3.9745
6	1.0682	3.7318
7	1.0919	3.5748
8	1.1228	3.4486
9	1.1524	3.3476
10	1.1803	3.2642
11	1.2062	3.1938
12	1.2301	3.1335
13	1.2521	3.0812
14	1.2725	3.0352
15	1.2914	2.9943
16	1.3090	2.9577
17	1.3253	2.9247
18	1.3405	2.8948
19	1.3547	2.8675
20	1.3680	2.8425

EXAMPLE-SMOOTHING RADAR ERROR SIGNALS. Figs 2-19 present the application of adaptive median smoothing to smoothing of range, azimuth, and elevation tracking error signals from a WSMR radar. The minimum smoothing interval was 11 points while the maximum smoothing interval was 41 points. The method used to adapt the smoothing interval to the data was Spearman rank correlation coefficient.

Figs 2-3 present the raw range error signal. Figs 4-5 show the smoothed range error signal and Figs 6-7 show the range residuals. The range tracking error does not exhibit significant signal content so that the smoothing of this signal is quite uninteresting. Note that the smoothed range in Figs 4-5 show bumps and dips having flat tops and bottoms. These flat peaks and valleys are characteristic of median smoothing. Tukey suggests a method of removing these peaks and valleys but we have not attempted to implement this in our median smoothing routine. Note also that the smoothed range exhibits a very constant behavior. This constant, which is not zero, may be due to the granularity of range output readings.

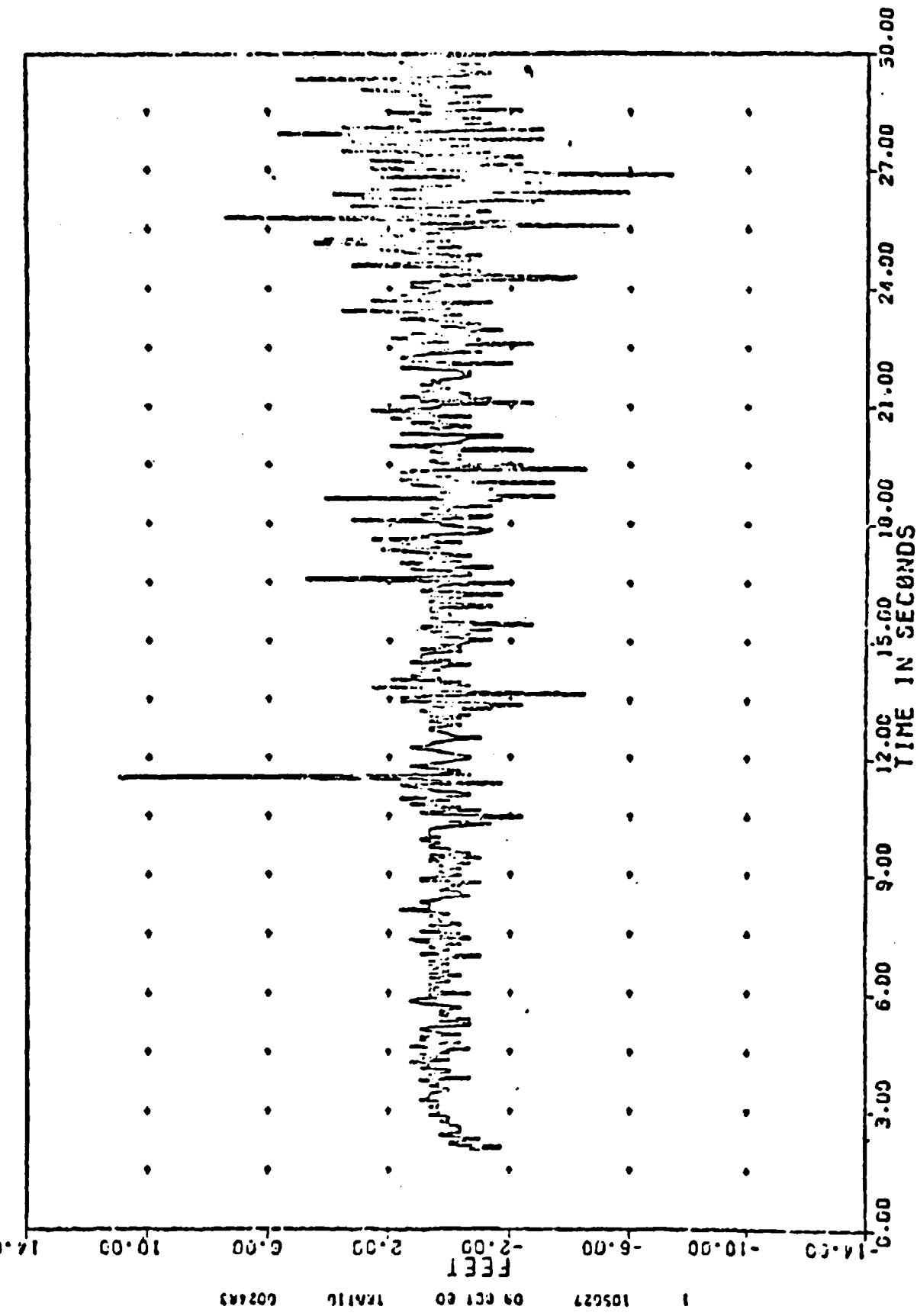
Figs 8-9 are the azimuth tracking error signal. At the beginning the radar is acquiring the target so that the error signal has a very strong signal content. After the target has been acquired the signal level decreases drastically and as the target recedes from the radar the noise content of the tracking error increases until it is virtually one large band of noise. Figs 10-11 are the smooth azimuth tracking error and Figs 12-13 exhibit the azimuth tracking residuals.

Figs 14-15 are the elevation tracking error. Again, this tracking error indicates a strong signal when the target is being acquired. Near the end of the mission the noise amplitude becomes very large. Also, near the end of the track the elevation tracking error indicates again a strong signal content. Figs 16-17 present the smooth elevation tracking error and Figs 18-19 are the elevation tracking residuals.

CONCLUSIONS. Adaptive median smoothing is a very simple method which can be readily applied to smoothing almost any data sequence without modeling either the signal or noise characteristics of the sequence. Adaptive median smoothing is also robust with respect to outliers. When applied to smoothing radar tracking errors as in the example given above, adaptive median smoothing does remarkably good in extracting the signal from the noisy data sequence. Some additional features of Tukey's proposals for median smoothing remain to be implemented and tested in our adaptive median smoothing. We plan to test the use of repeated median smoothing and the technique of twicing, i.e., resmoothing the residuals in our adaptive median smoothing routine.

RANGE BORESIGHT

Fig 2



105627 09 0016 16A116 602643

Fig 3

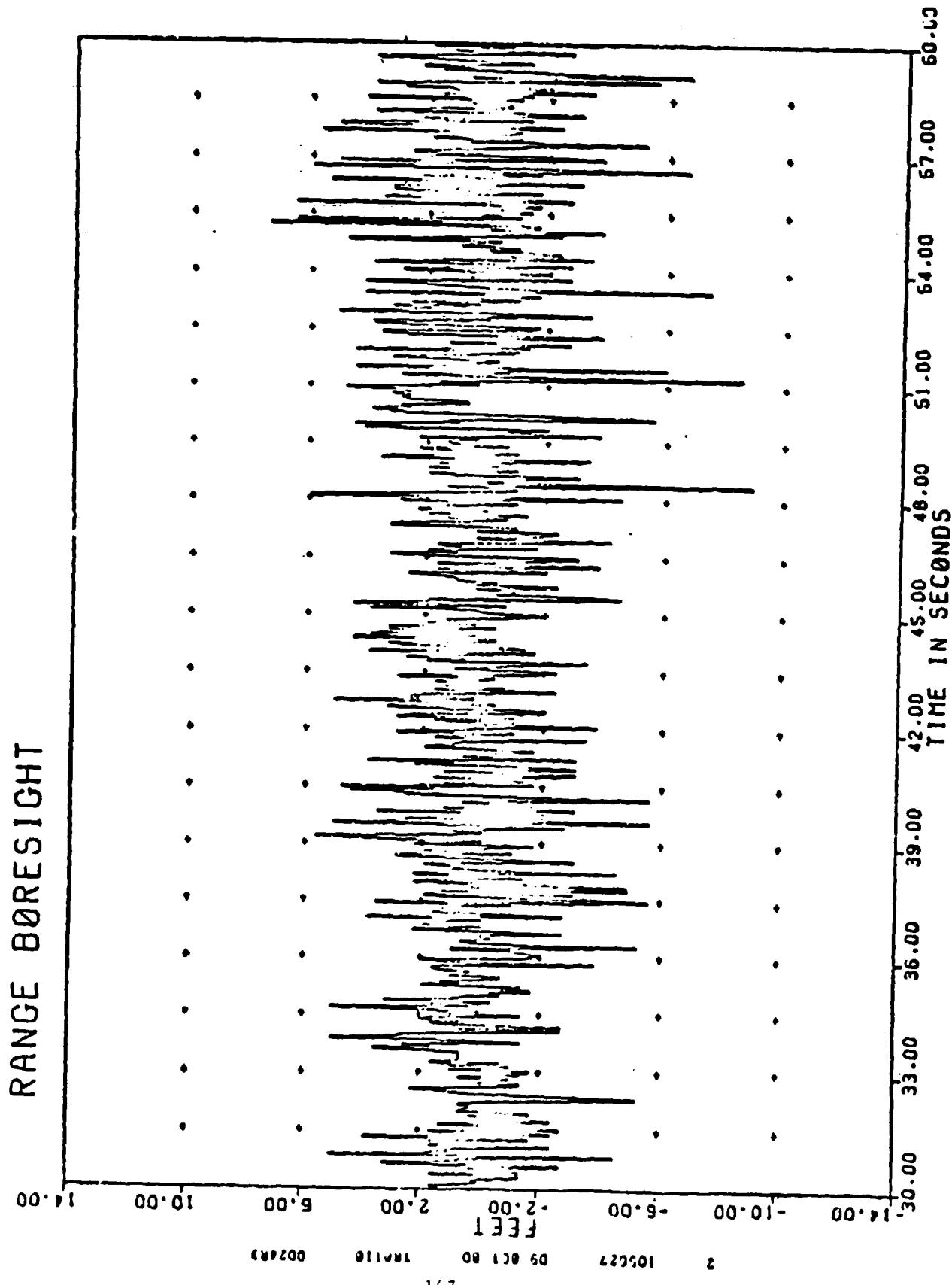


Fig 4

SMOOTHED RANGE

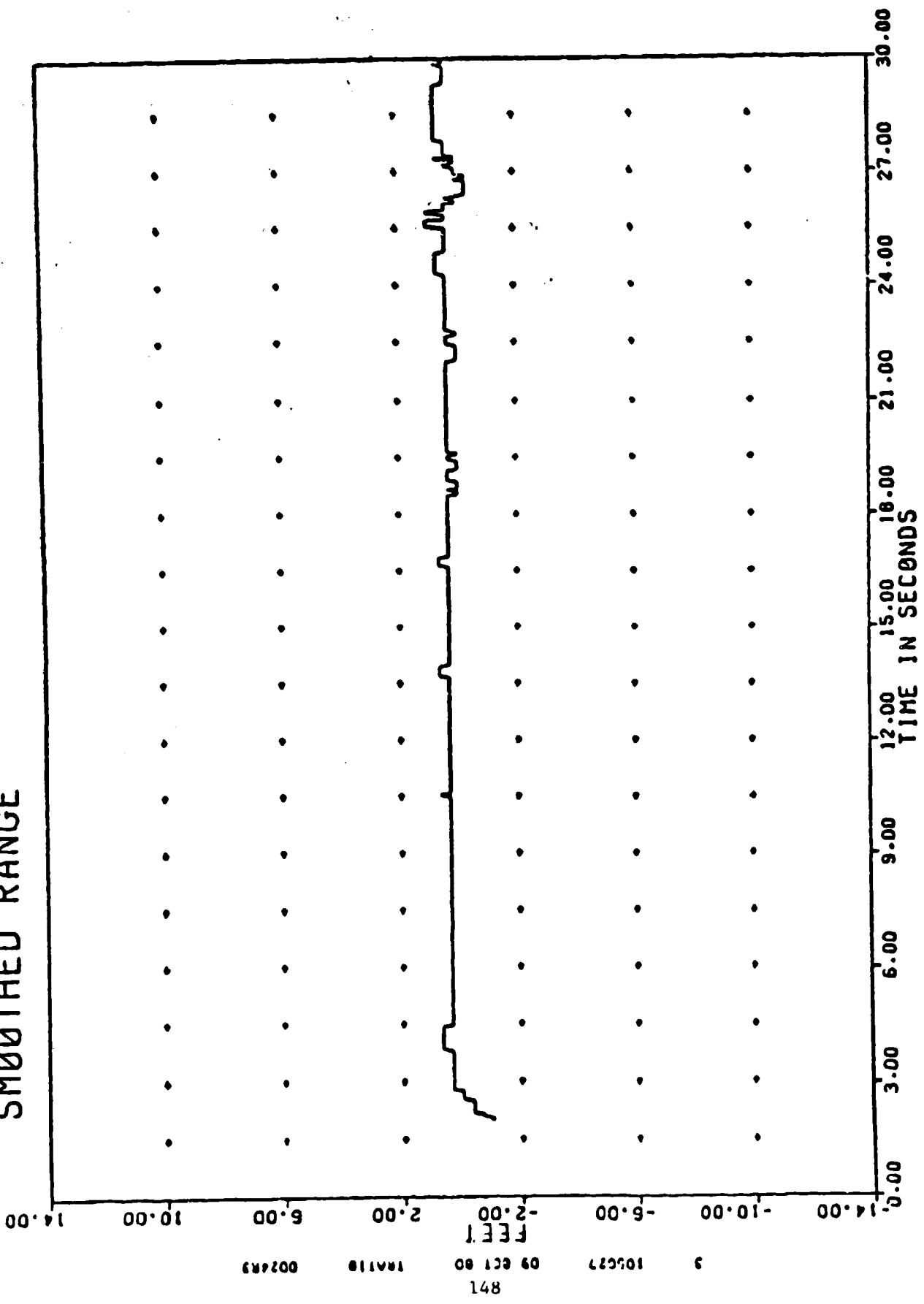


Fig 5

SMOOTHED RANGE

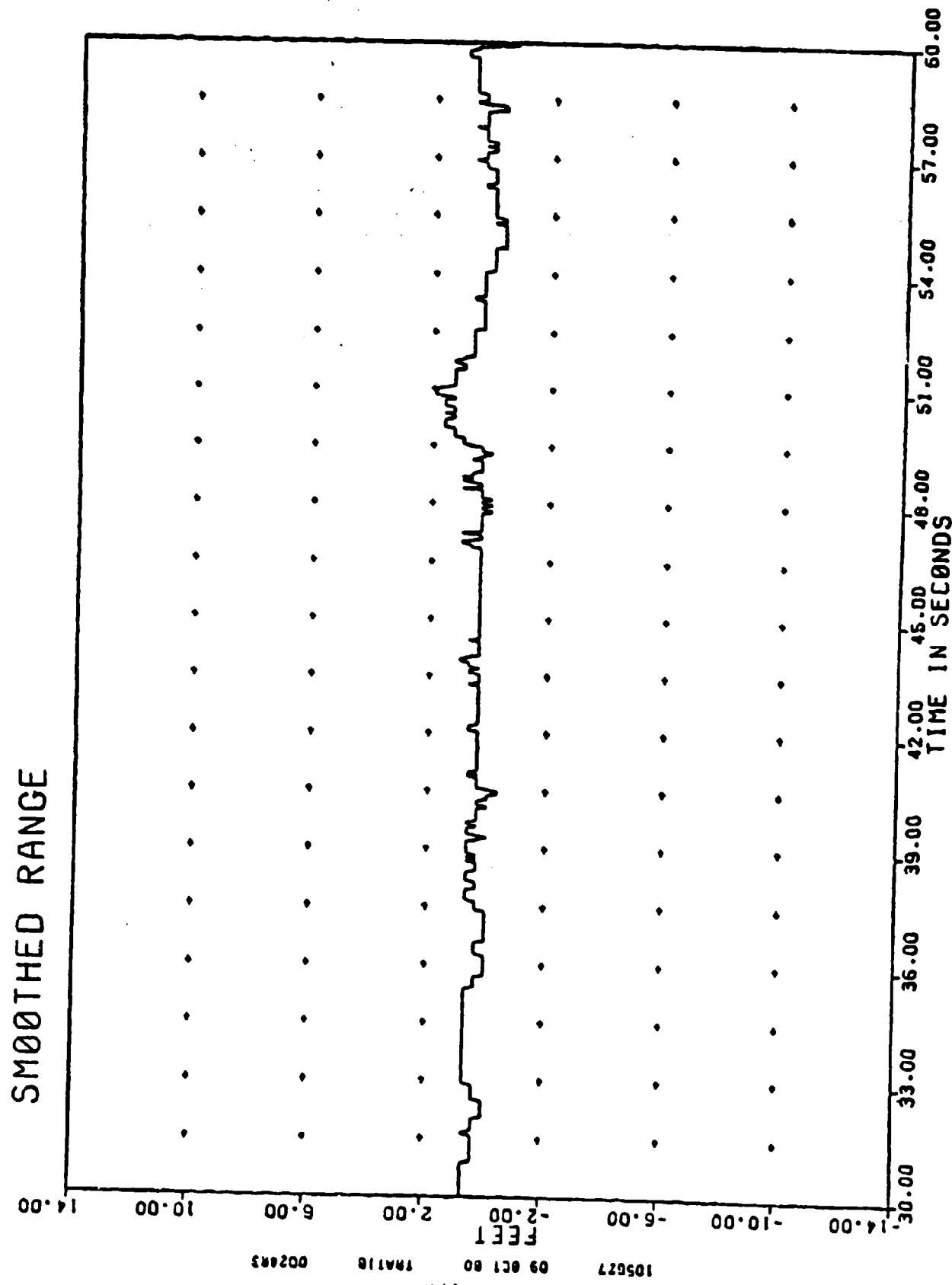


Fig 6

RANGE RESIDUALS

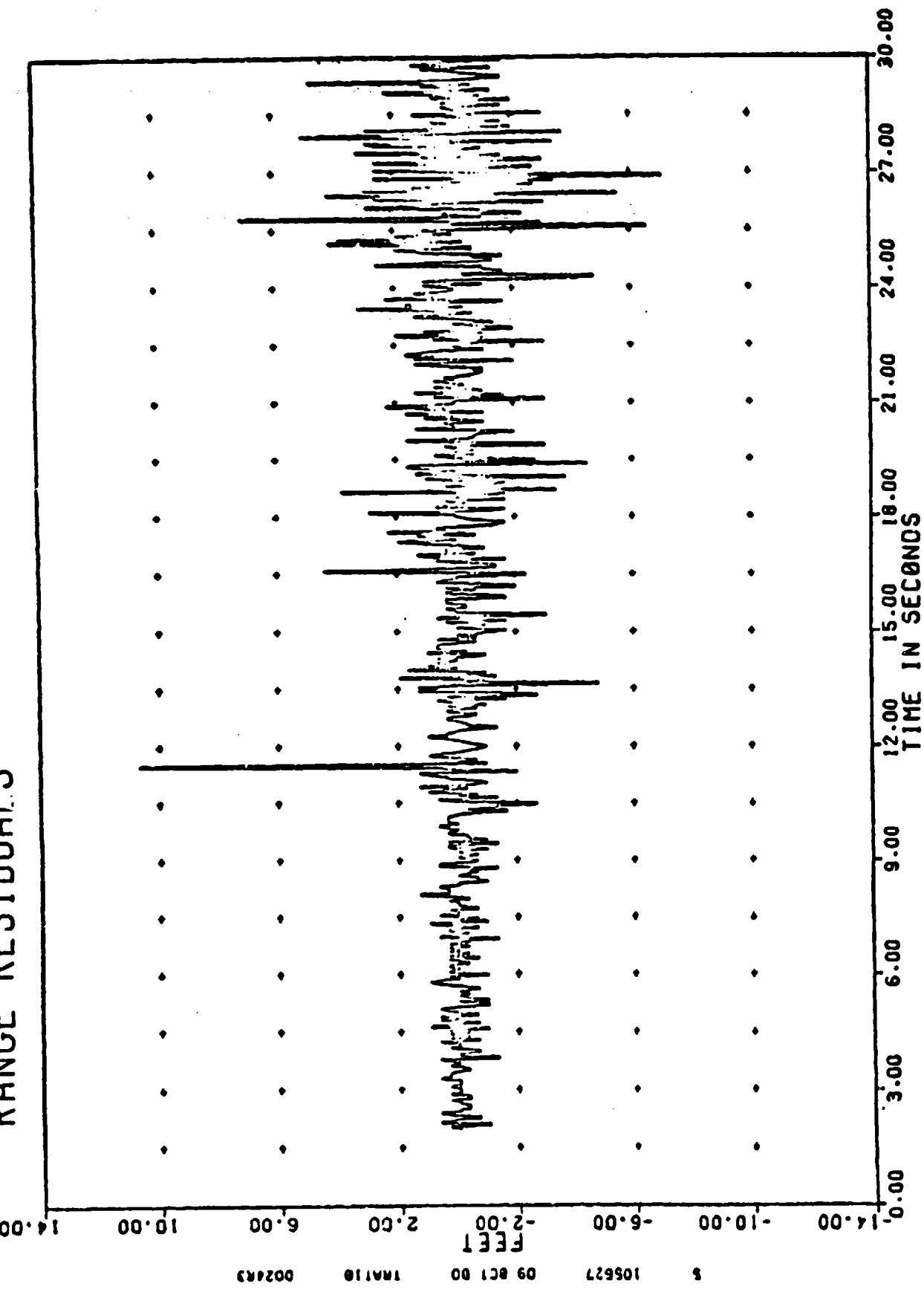
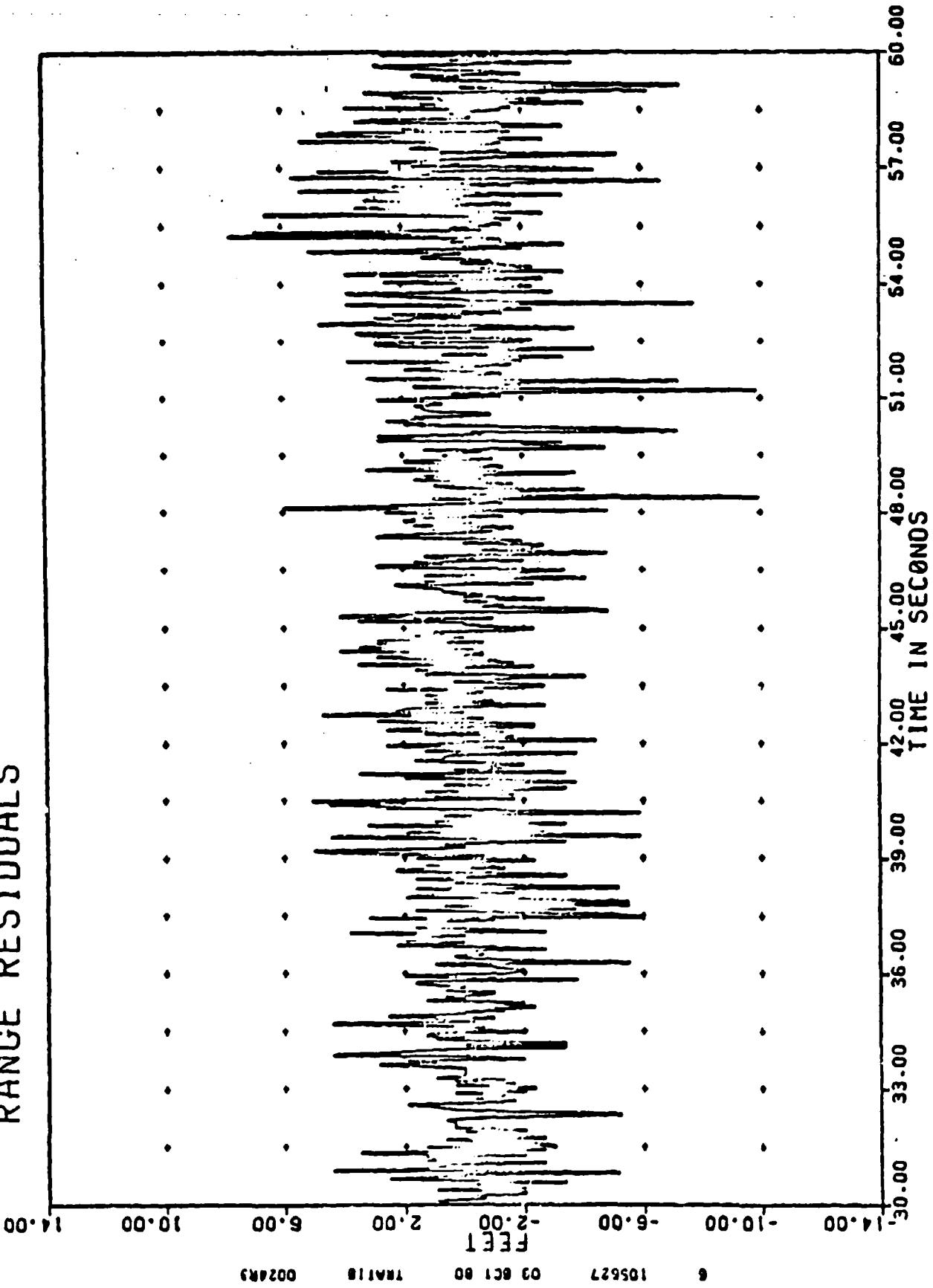


Fig 7

RANGE RESIDUALS



AZ BORESIGHT

Fig. 8

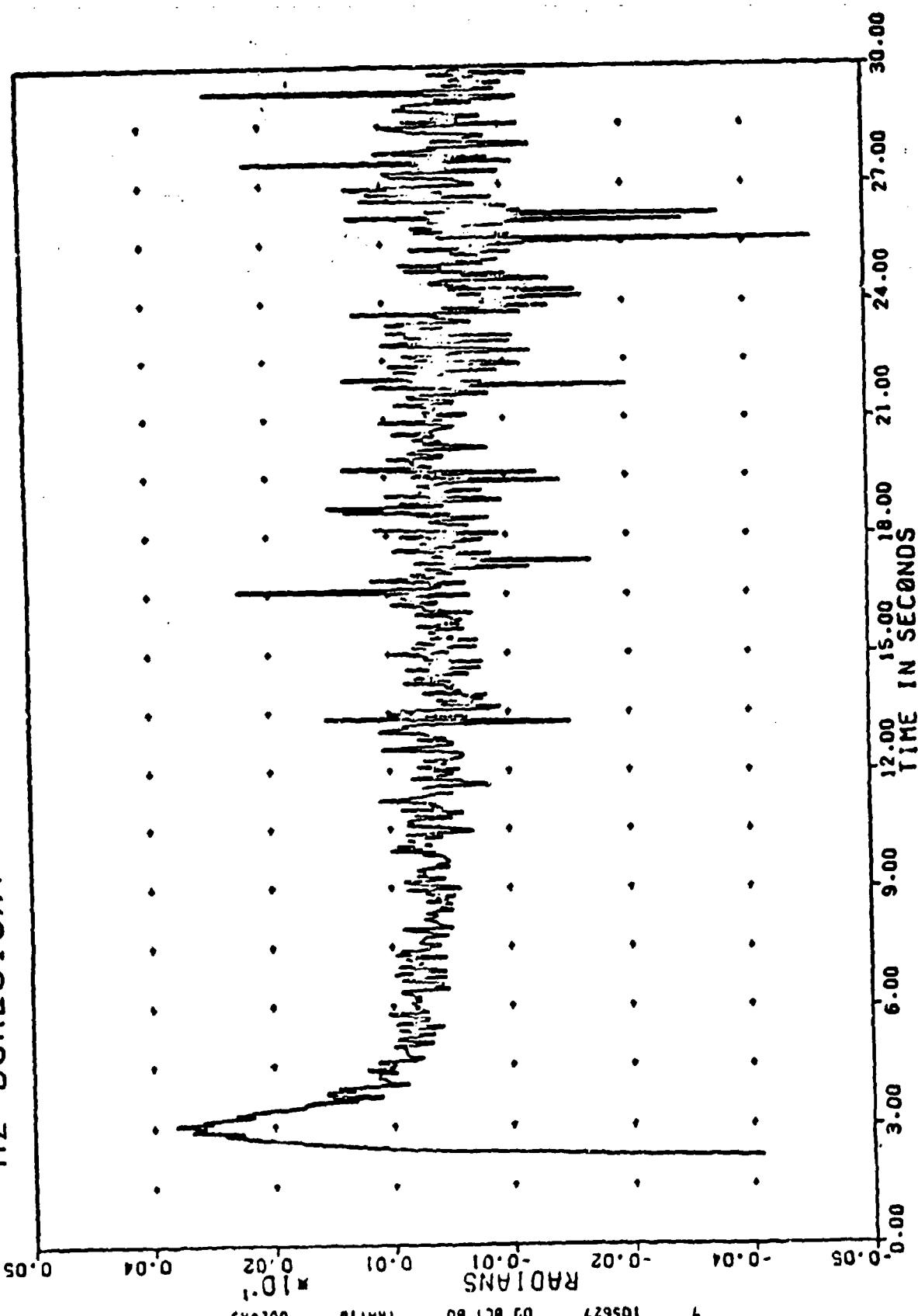


Fig 9

AZ BORESIGHT

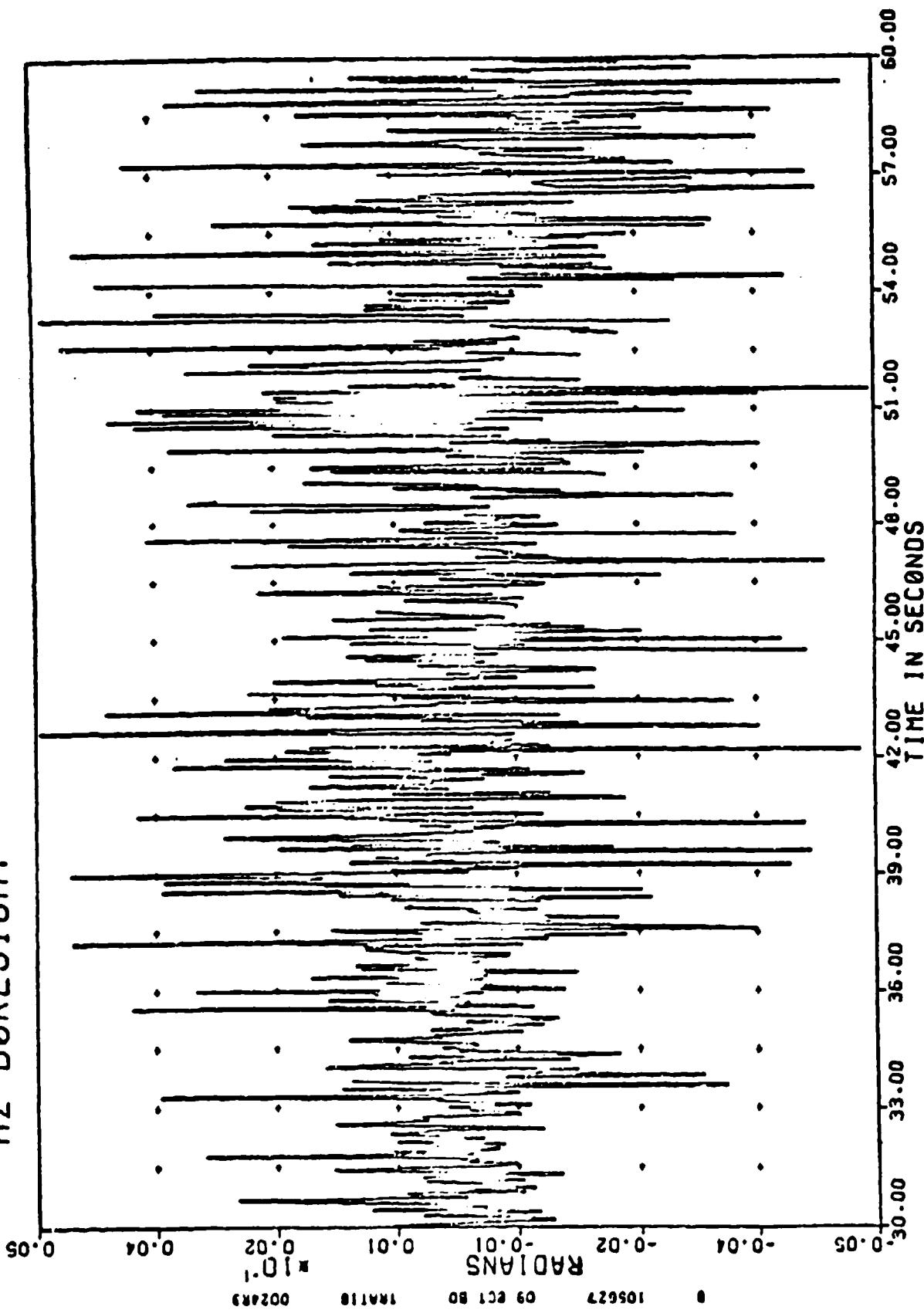


Fig 10

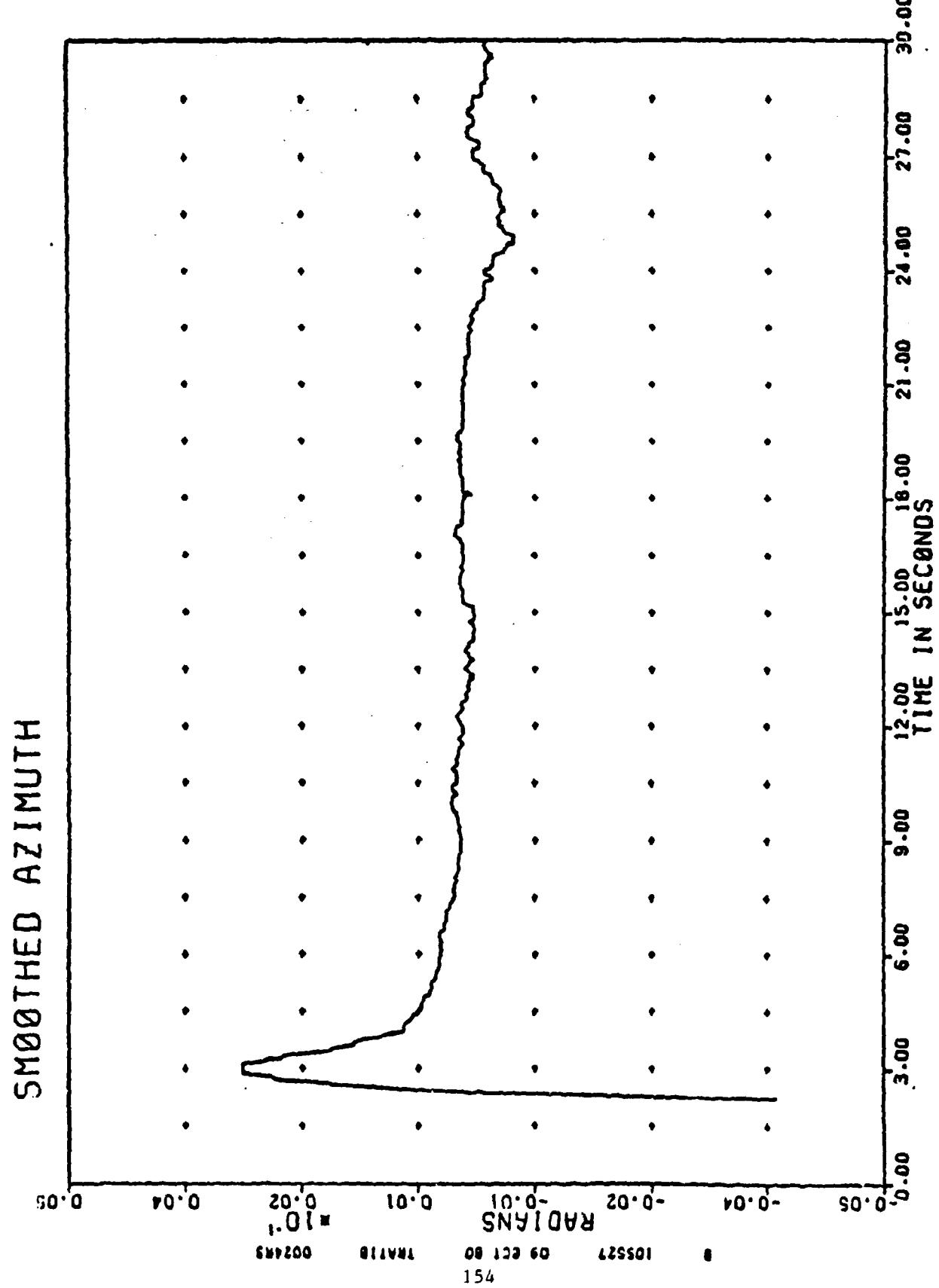


Fig 11

SMOOTHED AZIMUTH

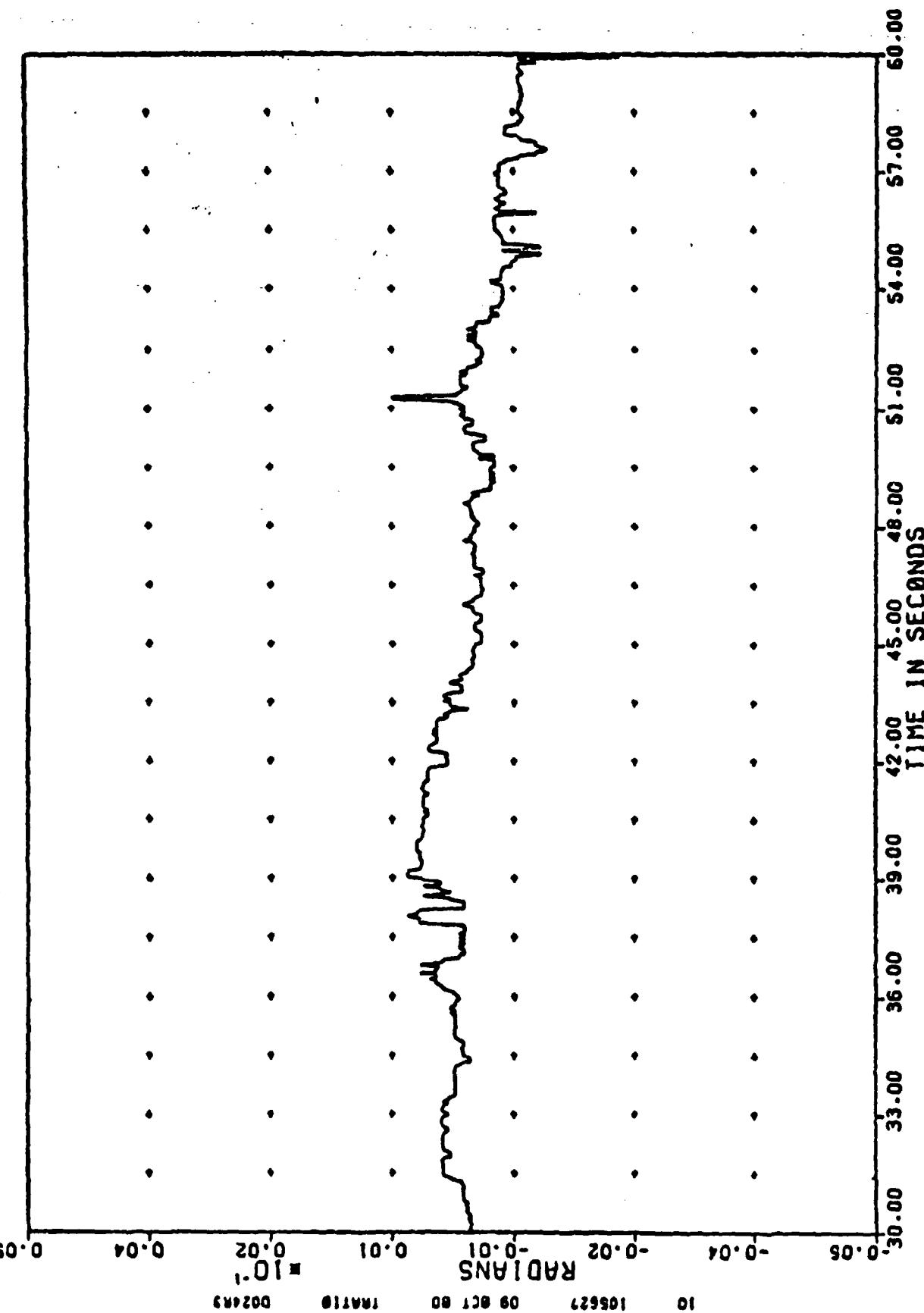
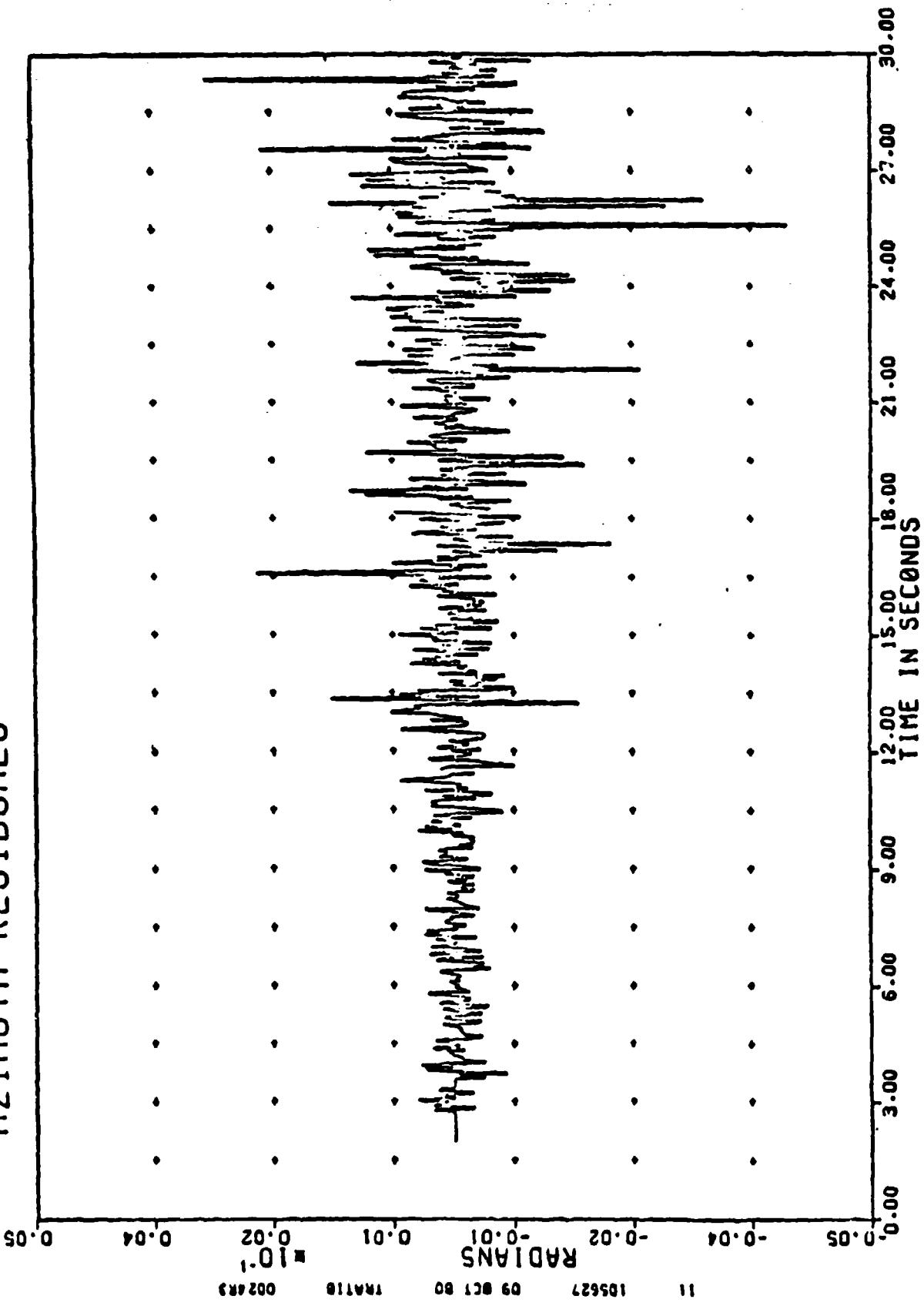


Fig 12

AZIMUTH RESIDUALS



AZIMUTH RESIDUALS

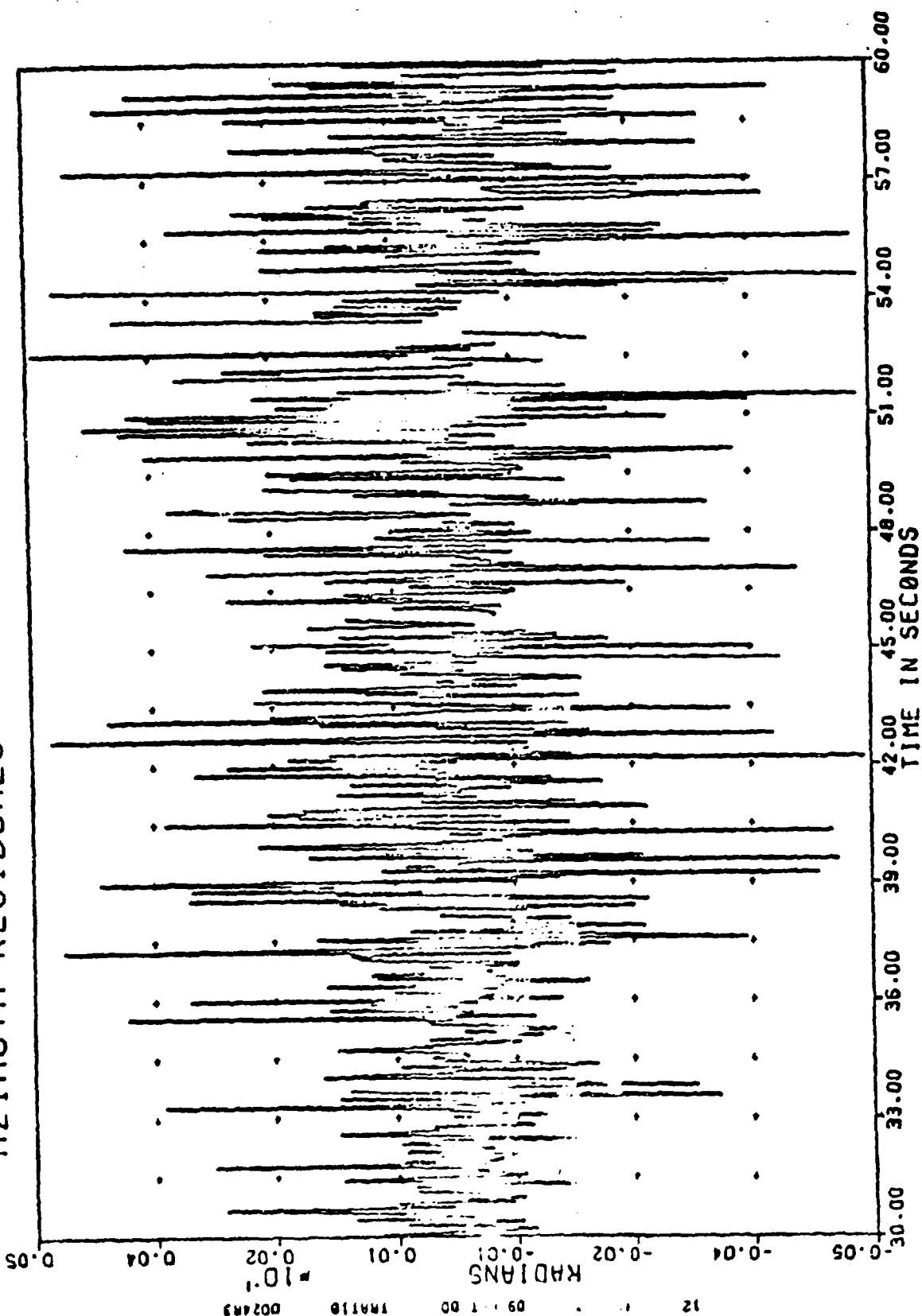


Fig 14

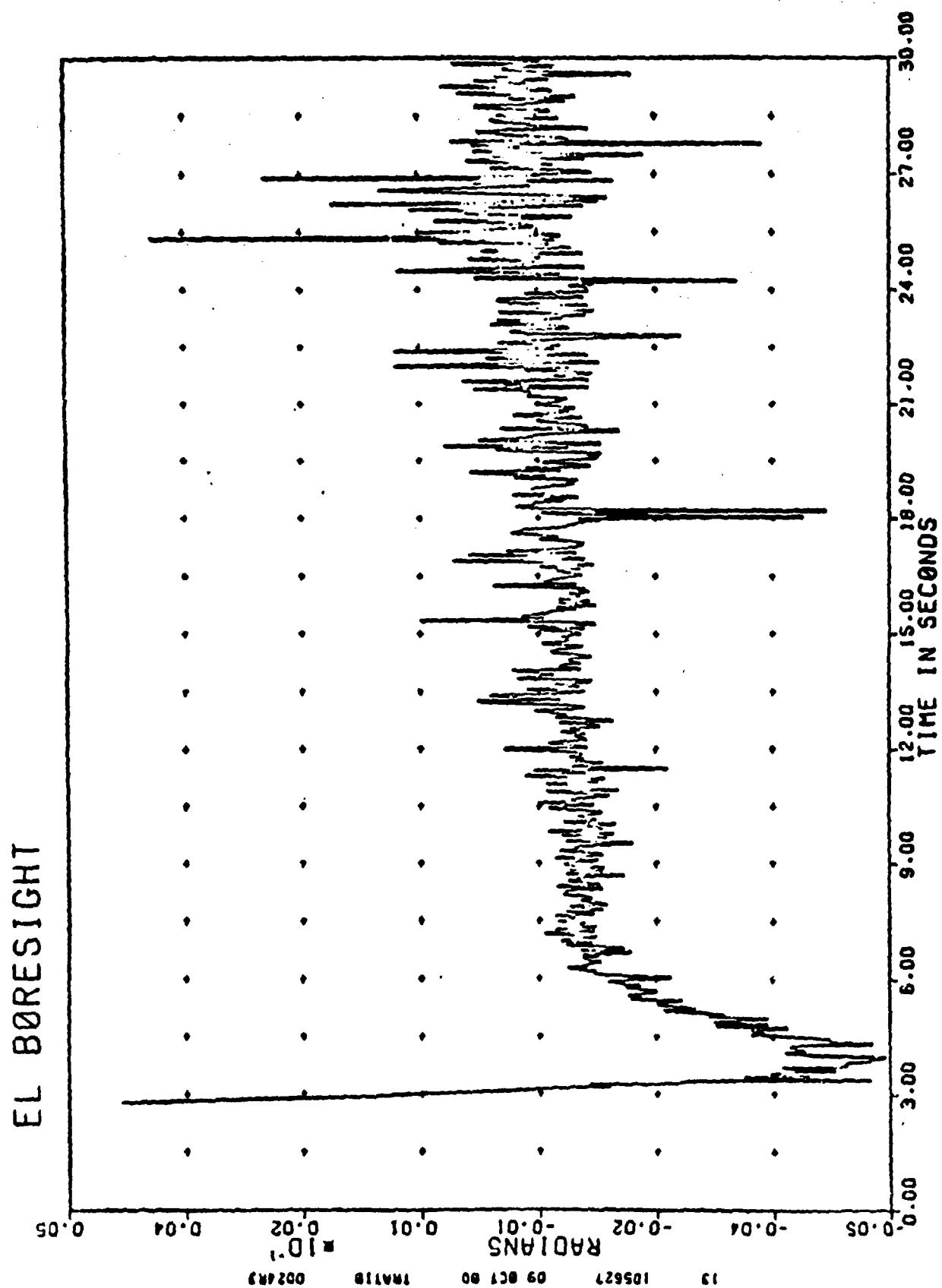


Fig 15

EL BORESIGHT

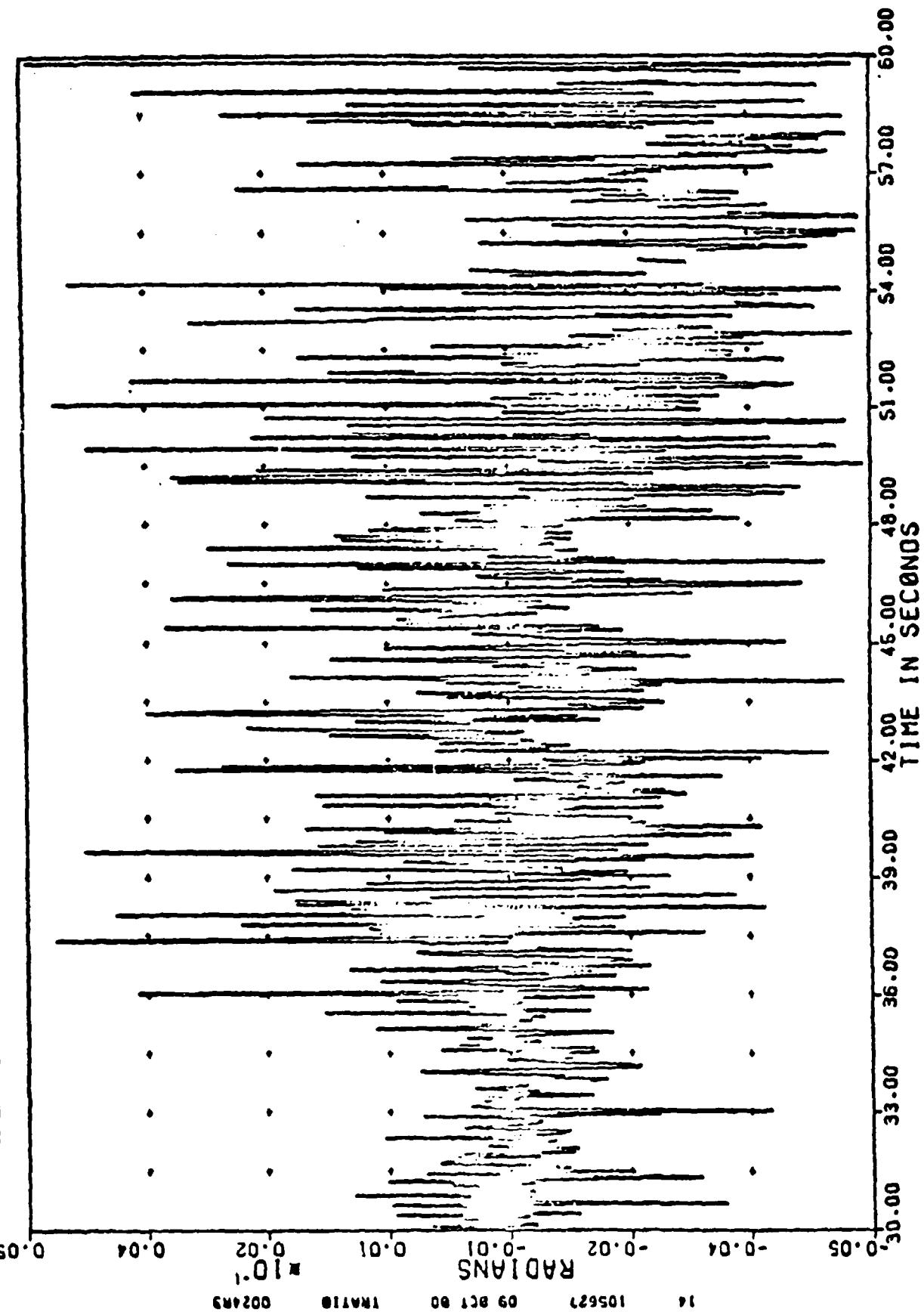


Fig 16

SMOOTHED ELEVATION

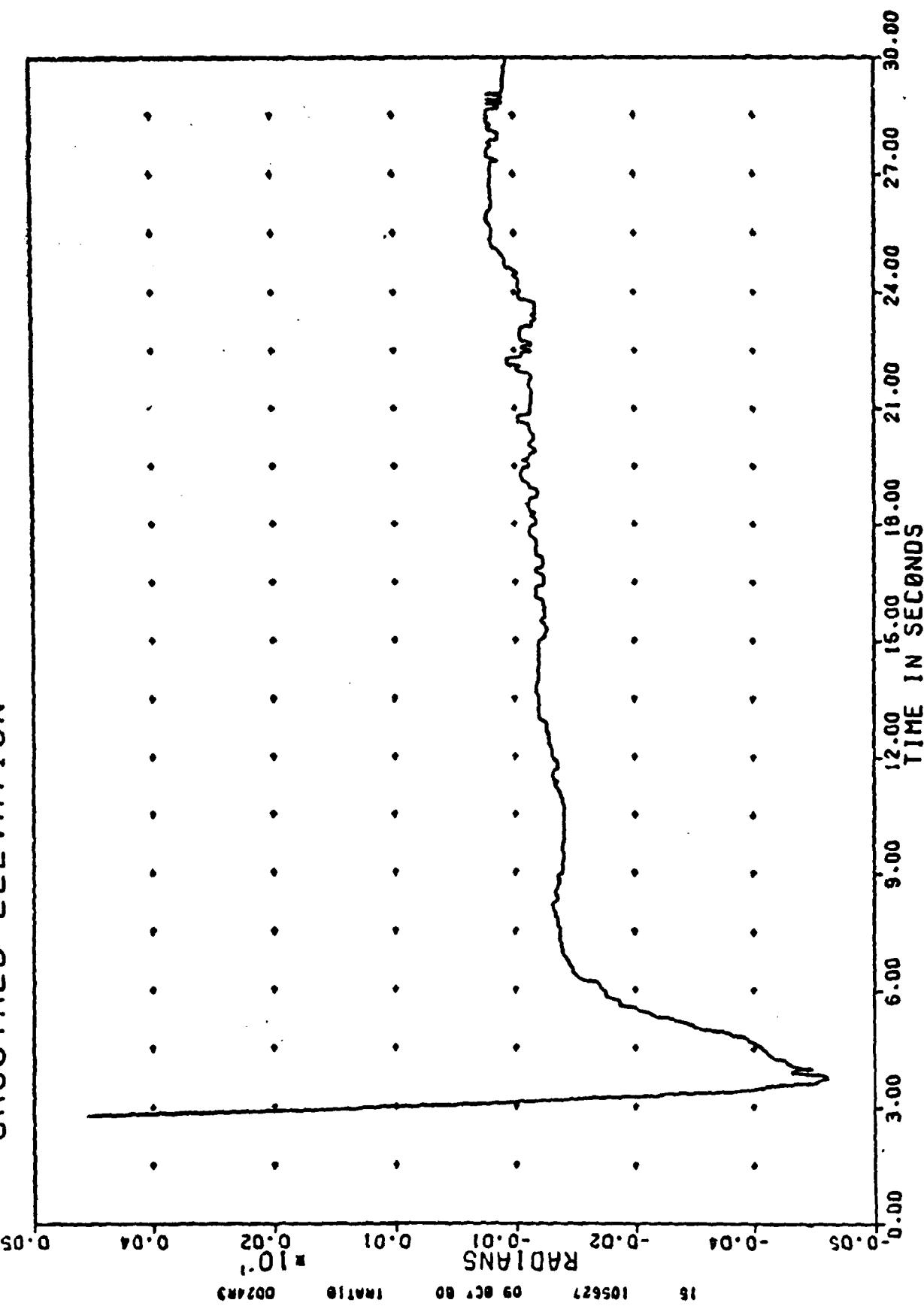


Fig 17

SMOOTHED ELEVATION

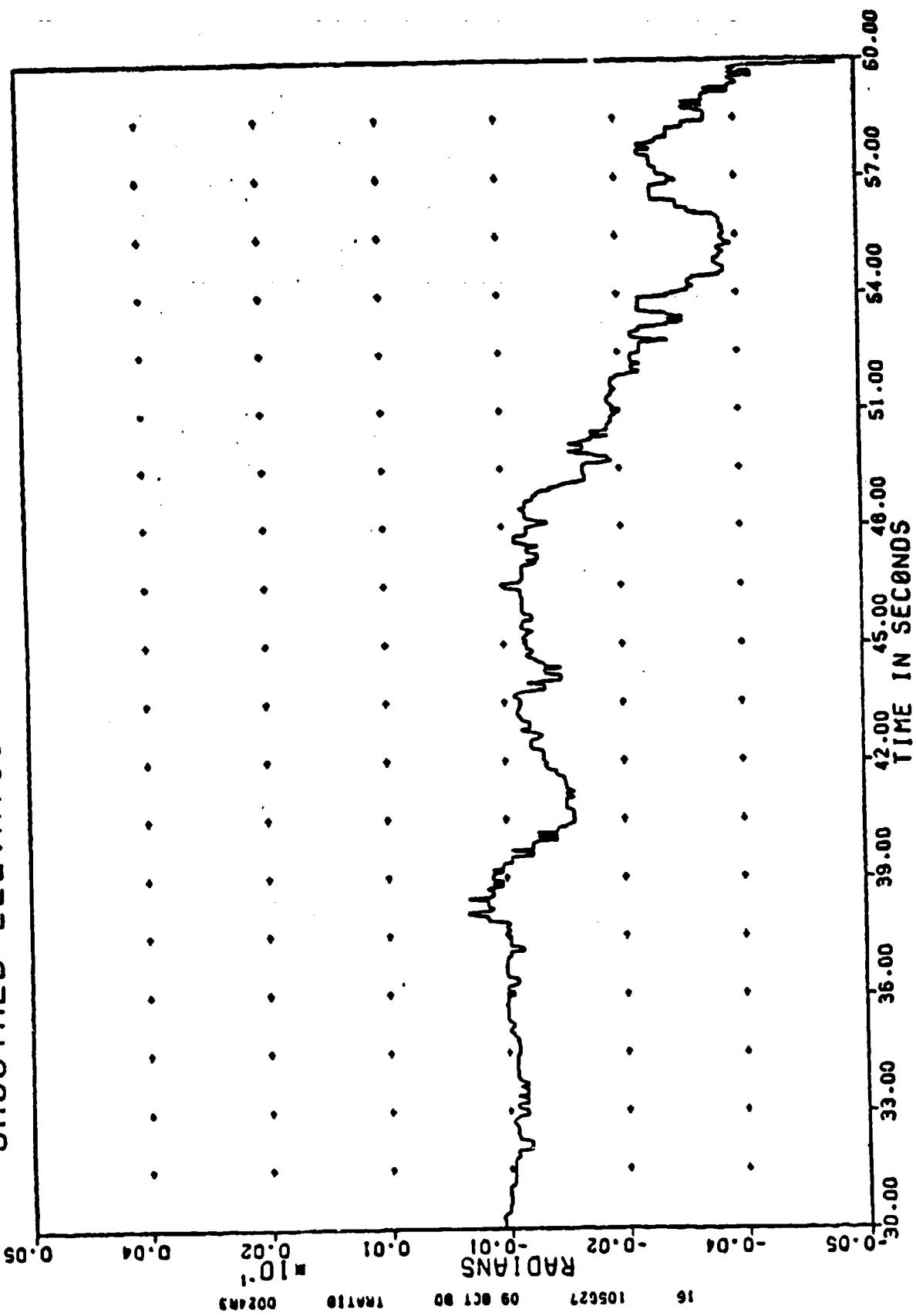
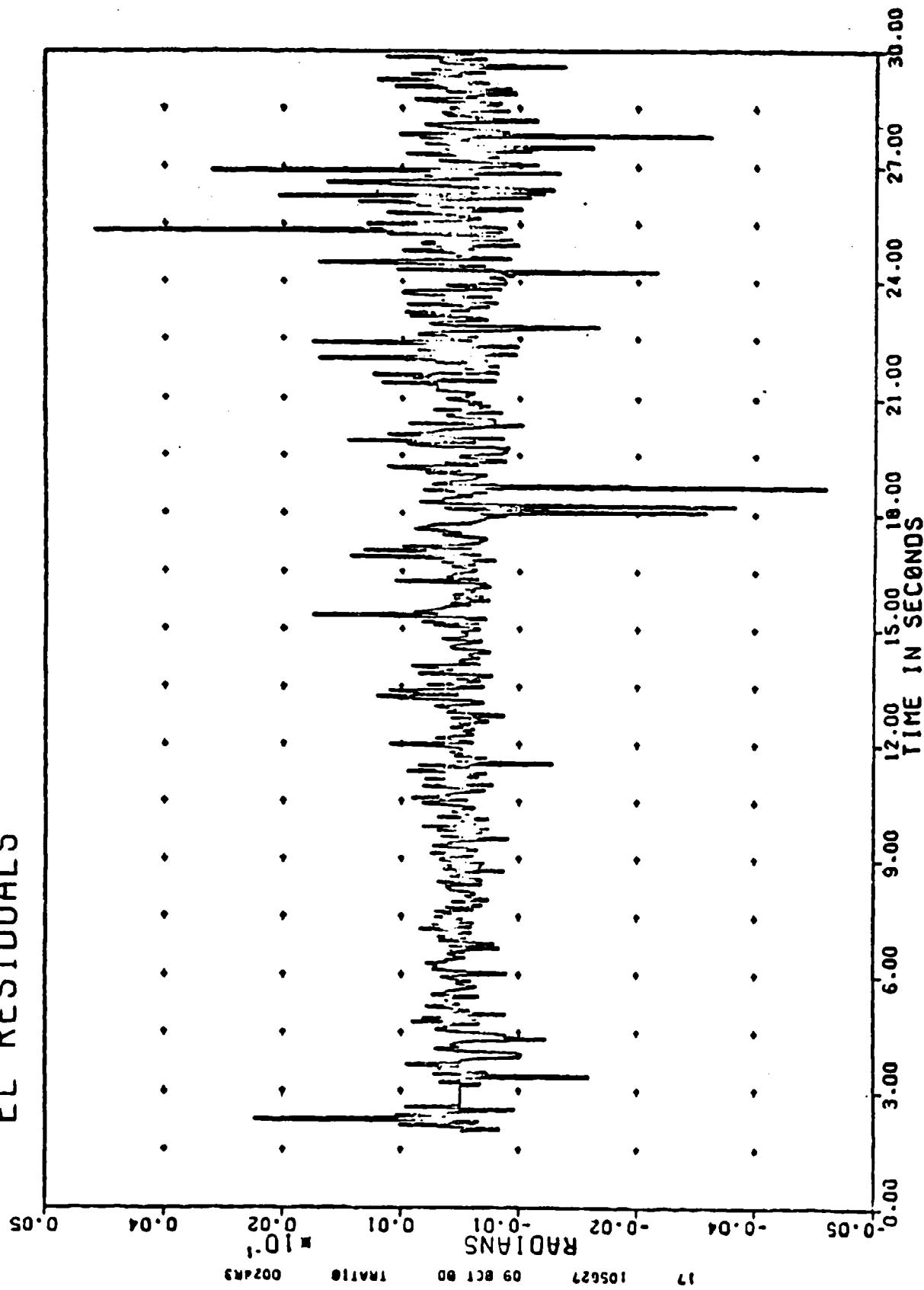


Fig 18

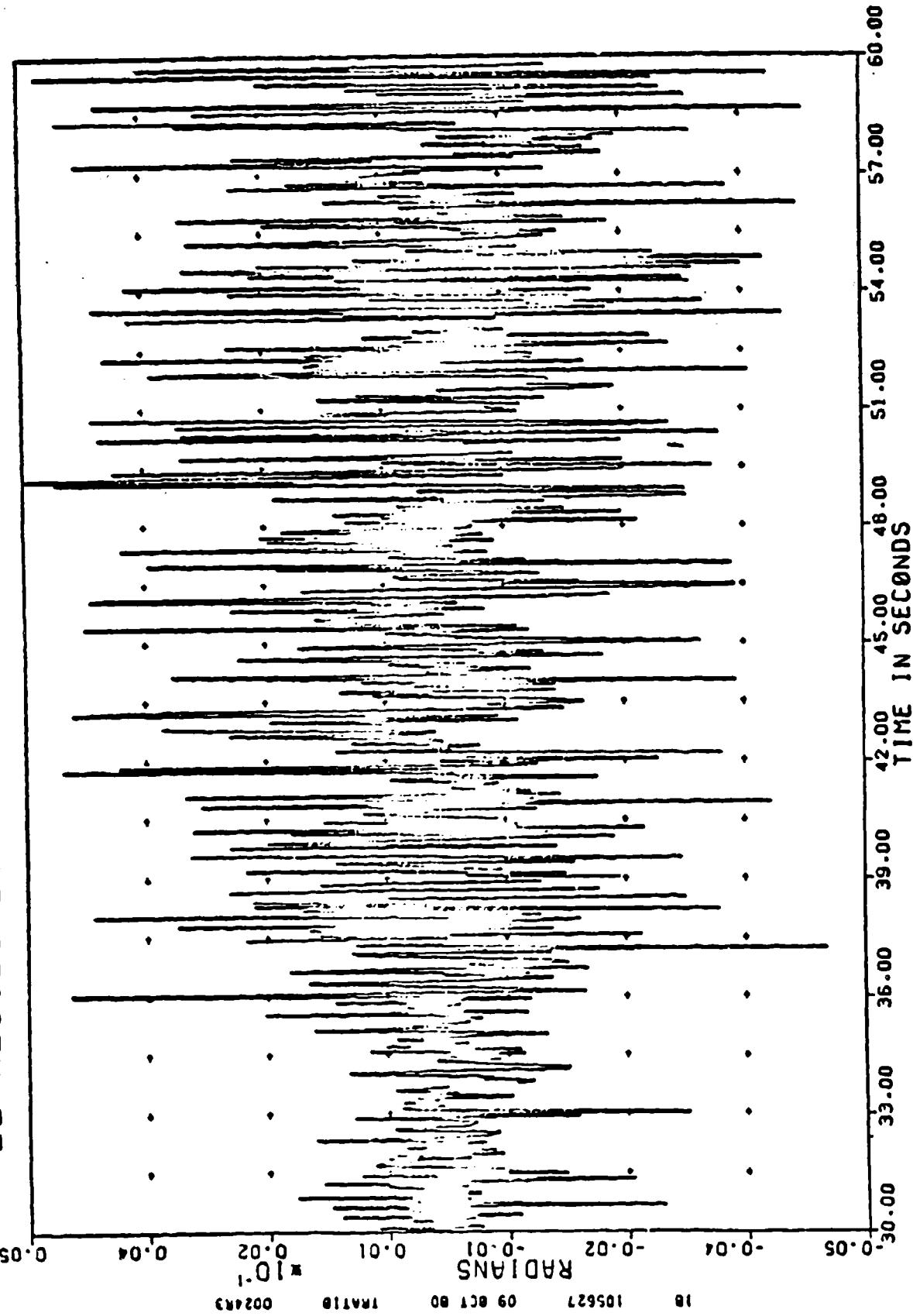
EL RESIDUALS



17 105627 09 OCT 80 TRAVIS 0024R3

Fig 19

EL RESIDUALS



REFERENCES

1. Tukey, J.W., Nonlinear (Nonsuperposable) Methods for Smoothing Data, Cong. Rec., 1974 EASCON, 673.
2. Tukey, J.W., Exploratory Data Analysis, Addison-Wesley, Reading MA, 1977.
3. Jayant, J.S., Average and Median Based Smoothing Techniques for Improving Digital Speech Quality in the Presence of Transmission Errors, IEEE Trans. Commun., COM-24, 1976, 1043-1045.
4. Rabiner, L.R., Sambur, M.R., and Schmidt, C.E., Applications of a Nonlinear Smoothing Algorithm to Speech Processing, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-23, 1975.
5. Narendra, P.M., A Separable Median Filter for Image Noise Smoothing, Proc. 1978 Conf. on Pattern Recognition and Image Processing, Chicago, Ill., May 1978, 137-141.
6. Pratt, W.K., Digital Image Processing, Wiley-Interscience, New York, NY, 1978.

FITTING AN ELLIPSE

Donald L. Buttz
US Army White Sands Missile Range
White Sands Missile Range, NM 88002

ABSTRACT

Three program procedures to fit an ellipse to x, y data constrained by two criterias are described. The first procedure is an iterative approach and the remaining two procedures are of a statistical nature, using a line of regression.

1. Introduction

Beginning with an impact pattern plot as points on an x, y graph, the problem is to evaluate an ellipse of minimum area circumscribing 95% of "activated" submunition impacts. The first procedure hereafter named PARAM is an iterative procedure. The second procedure known as SELLIPSE is a statistical procedure and the third procedure labeled ELLP3 is also statistical in nature. However, ELLP3 is more a probability approach to fit an ellipse of minimum area to a plotted impact pattern.

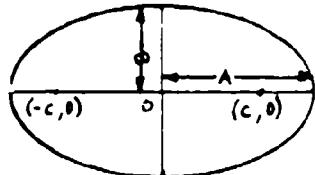
The problem of fitting an ellipse to impact pattern arose as a scoring criteria. The two criterias for that scoring are that the area of the ellipse must be the smallest while secondly containing exactly 95% of the "activated" submunition impacts.

The result of every procedure must provide the area of the ellipse, the length of the major and minor axes and the angle of axes rotation.

This paper shall describe the program procedures PARAM, SELLIPSE and ELLP3 in that order and then a summary of comparative results.

2. Program Procedure Called PARAM

The notation and relationships expressed below apply to the following discussion:

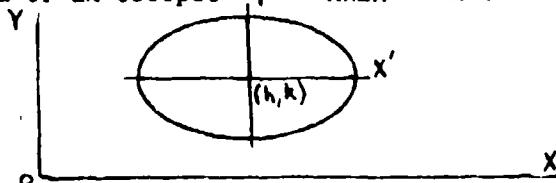


A is the semimajor axes
B is the semi minor axes
C is the ellipse focal point

ELLIPSE CENTER AT ORIGIN

Standard equation of an ellipse: $\frac{x^2}{A^2} + \frac{y^2}{B^2} = 1$.

Area of an ellipse πAB .

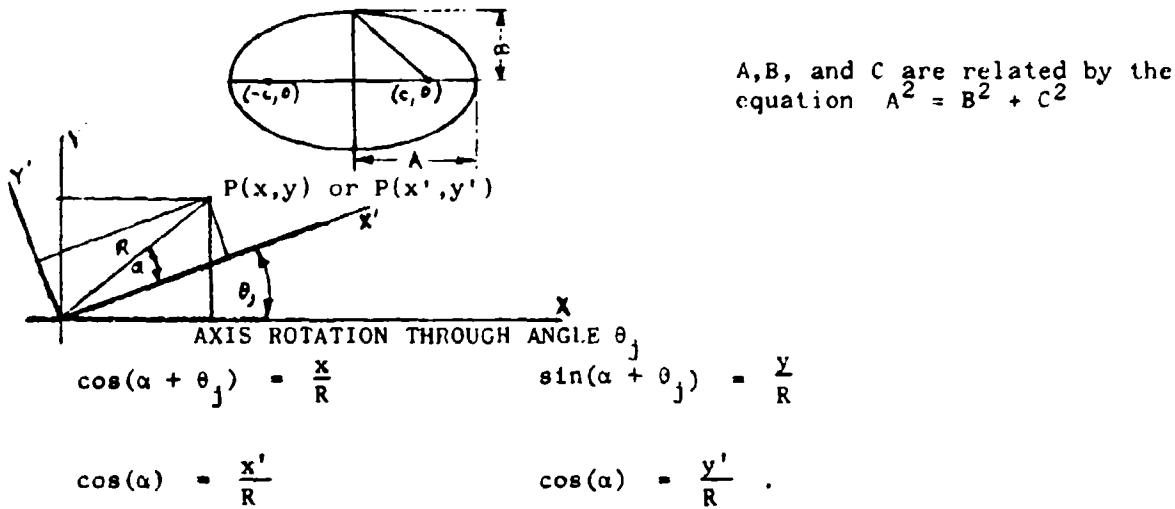


ELLIIPSE WITH CENTER AT (h, k) AND NEW AXES x' AND y'

The equation of the curve relative to these axes is $\frac{x'^2}{A^2} + \frac{y'^2}{B^2} = 1$.

The equation relative to the x and y axes, by setting $x' = x - h$ and $y' = y - k$ becomes, $\frac{(x-h)^2}{A^2} + \frac{(y-k)^2}{B^2} = 1$

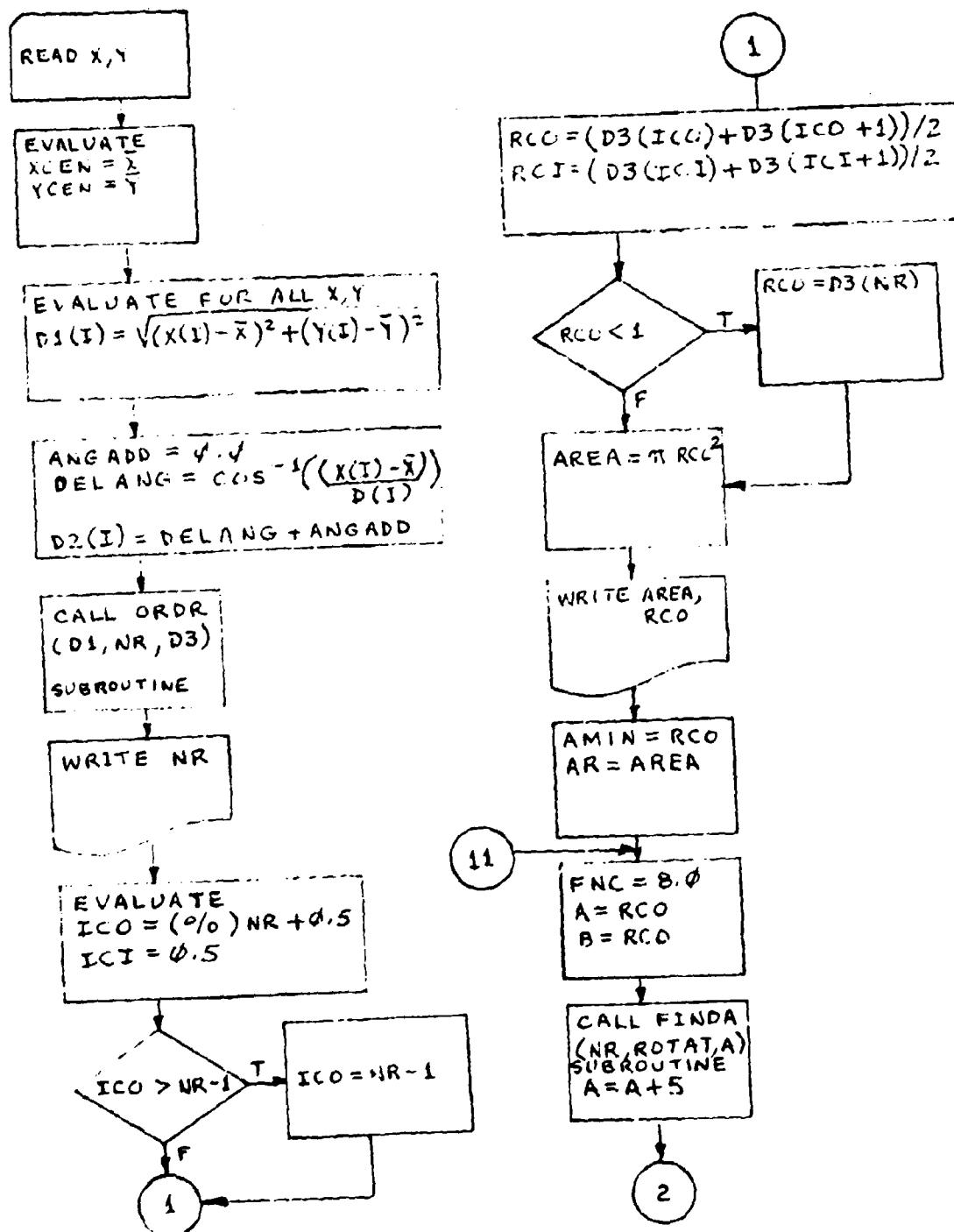
This is the standard equation of the ellipse with center at (h, k) and major axis is parallel to the x -axis.



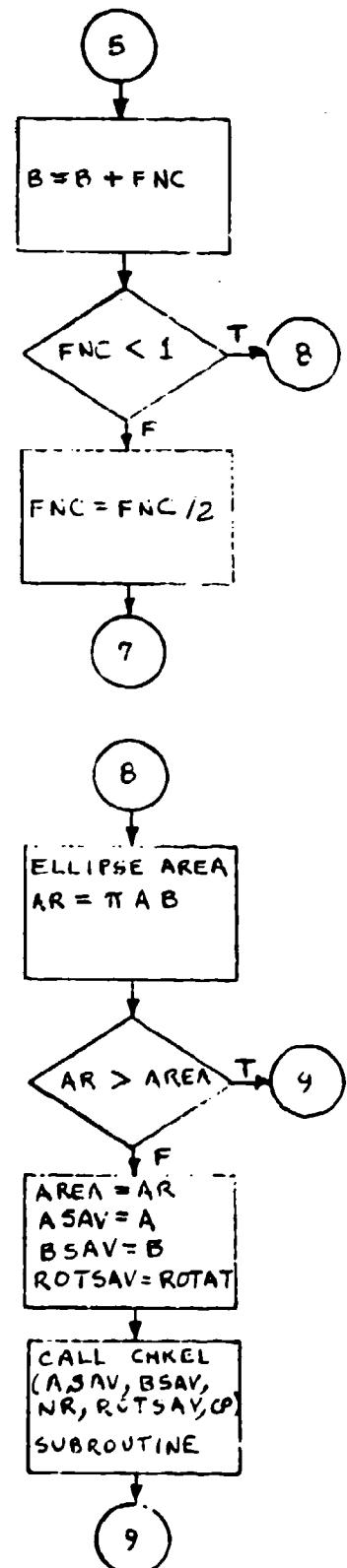
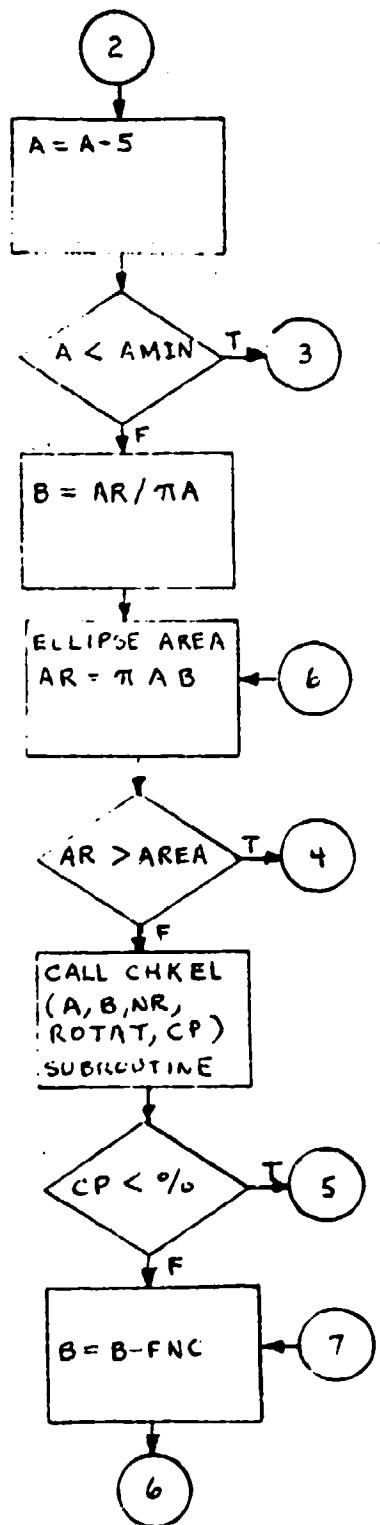
Let it be noted that for purposes of this procedure an activated bomblet will be defined to lie within the ellipse if and only if the sum of its distance from $(C, 0)$ and $(-C, 0)$ after adjustment for axis rotation is less than or equal to "2A". Secondly each fitted ellipse will be centered at the mean center of impact in the rotated coordinate system, (h, k) .

PARAM PROGRAM FLOWCHART

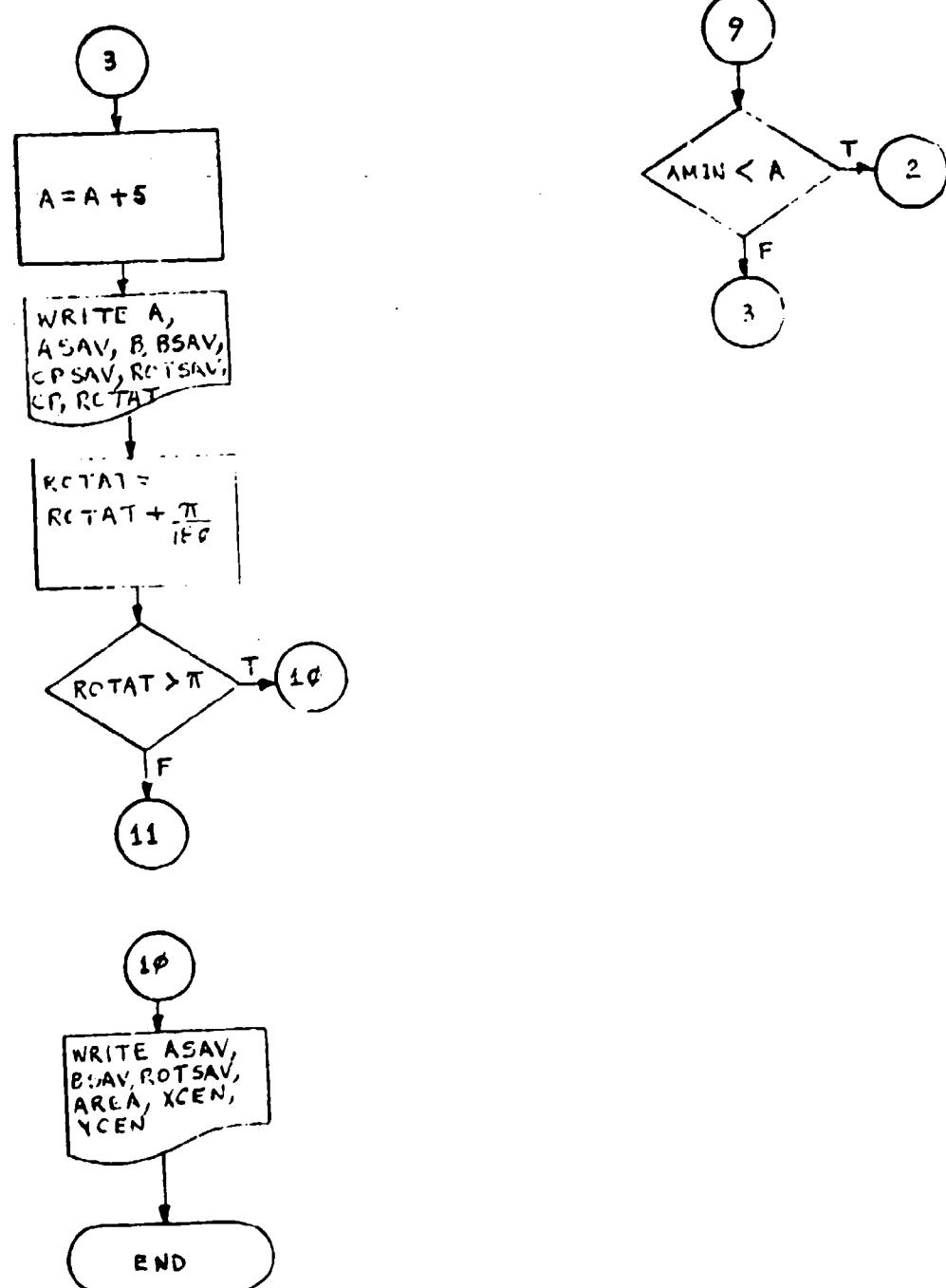
PARAM IS AN ITERATIVE PROCEDURE FOR FITTING AN ELLIPSE OF MINIMUM AREA TO PLOTTED IMPACT PATTERN DATA



PARAM FLOWCHART



PARAM FLOWCHART



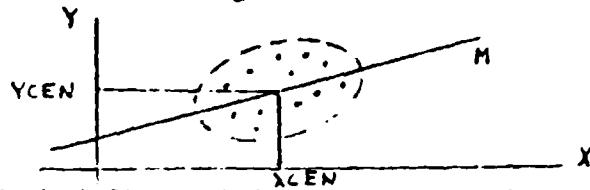
After iterating until θ_j has been adjusted through 179 degrees the ellipse of minimum area is determined. The equation of the ellipse of minimum area which contains 95% of the activated submunition impacts is

$$\frac{(x'-h)^2}{A^2} + \frac{(y'-k)^2}{B^2} = 1$$

where A and B were saved from the comparisons, and the rotation of the coordinate axis system is specified by θ_j .

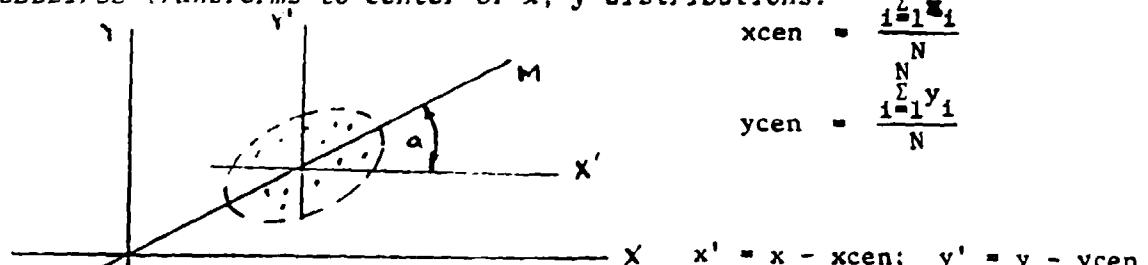
3. Program Procedure Called SELLIPSE

SELLIPSE computes the slope M and the y intercept B from the normal equations for linear regression.



The dashed figure indicates the possible ellipse to be determined.

Next, SELLIPSE transforms to center of x, y distributions.



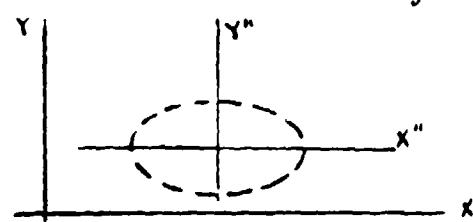
$$x_{cen} = \frac{\sum_{i=1}^N x_i}{N}$$

$$y_{cen} = \frac{\sum_{i=1}^N y_i}{N}$$

SELLIPSE rotates coordinates through an angle α to obtain the x'' and y'' coordinates.

$$y'' = -x' \sin \alpha + y' \cos \alpha$$

$$x'' = x' \cos \alpha + y' \sin \alpha$$



SELLIPSE computes the standard deviations of elliptic distribution along x'' and y'' .

$$\sigma_{x''} = \sqrt{\frac{\sum x''^2}{N}}$$

$$\sigma_{y''} = \sqrt{\frac{\sum y''^2}{N}}$$

The standard deviation in terms of x and y respectively are

$$\sigma_x = \frac{1}{N} \sqrt{\frac{1}{1+M^2} [(N\Sigma x^2 - (\Sigma x)^2) + M^2(N\Sigma y^2 - (\Sigma y)^2) + 2M(N\Sigma xy - \Sigma x\Sigma y)]}$$

$$\sigma_y = \frac{1}{N} \sqrt{\frac{1}{1+M^2} [M^2(N\Sigma x^2 - (\Sigma x)^2) + (N\Sigma y^2 - (\Sigma y)^2) - 2M(N\Sigma xy - \Sigma x\Sigma y)]}$$

The equation of the ellipse in terms of x'' and y'' is

$$\frac{x''^2}{A^2} + \frac{y''^2}{B^2} = 1$$

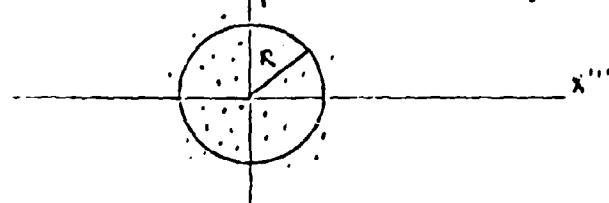
where A is the semimajor axis and B is the semiminor axis.
The ratio A/σ_x is k and the ratio of B/σ_y is k .

$$A = k\sigma_x \quad \text{and} \quad B = k\sigma_y$$

where k is the percentile radius.

SELLIPSE does a computation to circular coordinates.

Let $x''' = x''/\sigma_x$ and $y''' = y''/\sigma_y$.



The equation of a circle is

$$x'''^2 + y'''^2 = R^2$$

$$\text{then } R = \sqrt{x'''^2 + y'''^2}$$

where R is the radius.

SELLIPSE computes the mean and standard deviation of the radius.

$$R_{\text{mean}} = \frac{\sum_{i=1}^N R_i}{N}$$

$$\sigma_R = \sqrt{\frac{\sum (R - R_{\text{mean}})^2}{N}}$$

Percentile Radius is defined as follows

$$R\% = RMEAN + K_{R\%} \sigma_R$$

where $K_{R\%}$ for a normal distribution equals 1.645 for 95% ellipse.

In a normal probability distribution function, $f(x)$ is the probability density where

$$f(x) = (1/\sqrt{2\pi}) \exp[-(1/2)x^2].$$

For negative values of x , one uses the fact that $f(-x) = f(x)$.

Also, let $F(x)$ be the cumulative distribution function. Therefore:

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp[-(1/2)t^2] dt$$

For $F(x) = 95\%$ x becomes 1.645 our $K_{R\%}$. Note that $R\% = RMEAN + K_{R\%} \sigma_R$.

The semi major axis is $A = R\% \sigma_x$

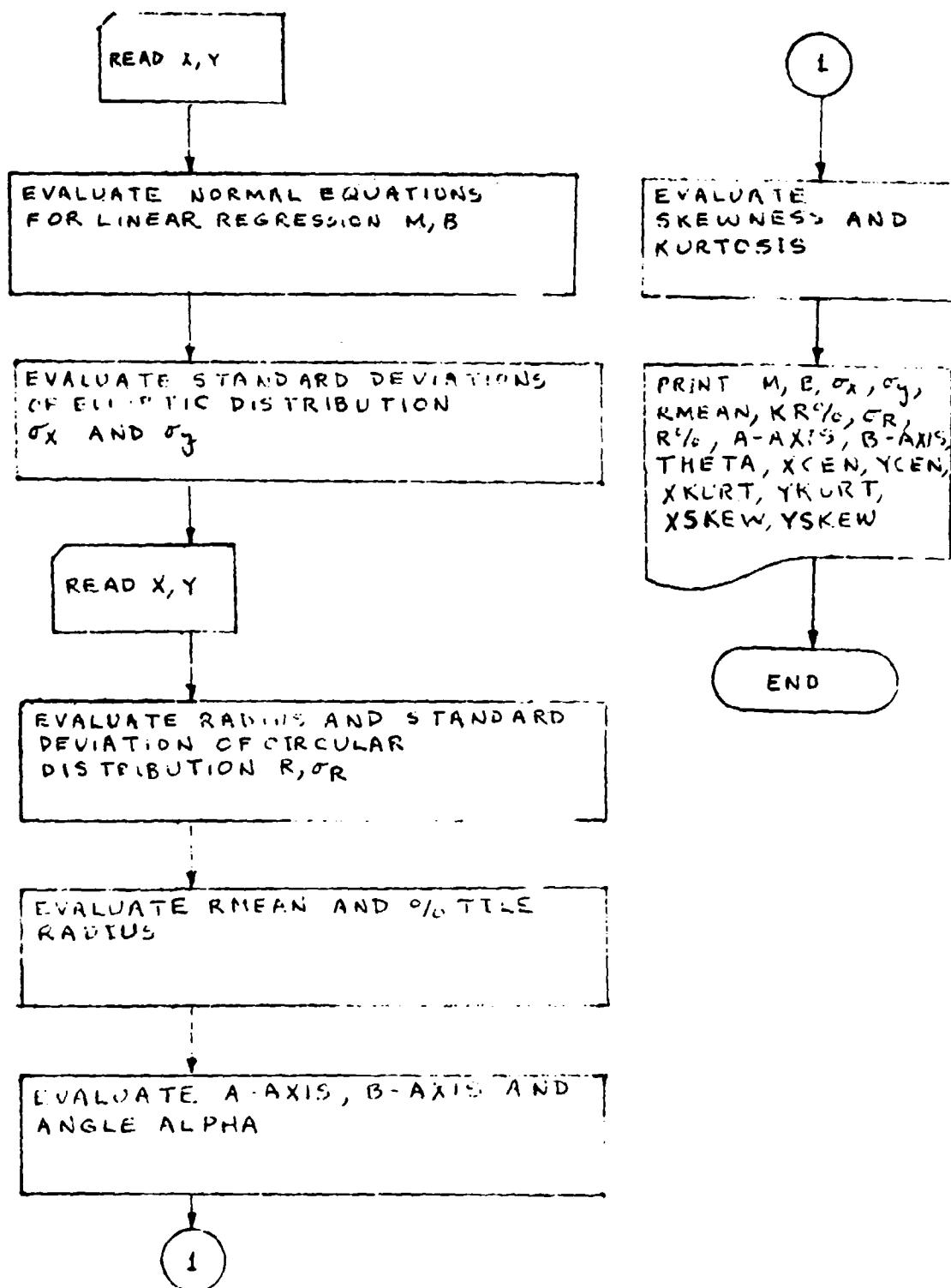
and the semi minor axis is $B = R\% \sigma_y$.

The rotated angle is $\alpha = \tan^{-1}(M)$.

The equation of the ellipse is then

$$\frac{x''^2}{A^2} + \frac{y''^2}{B^2} = 1$$

SELLIPSE PROGRAM FLOWCHART



4. Program Procedure Called ELLP31

Assumptions and derivations will be discussed first.

The data are presented as N ordered pairs (x_i, y_i) representing locations of submunition impacts in a north-south/east-west rectangular coordinate system. x and y are assumed to be jointly distributed, normal, random variables with respective means M_x and M_y and non-zero variances estimated by S_x^2 and S_y^2 , where

$$S_j^2 = \frac{\sum_{i=1}^N (j_i - \bar{j})}{N - 1}, \quad \text{where } j = x, y.$$

There may exist between x and y a non-zero correlation whose coefficient, R , is estimated by

$$R = \frac{\sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i / N}{(N - 1) S_x S_y}.$$

The joint probability density function is approximately

$$f_{xy}(x, y) = \frac{1}{2\pi S_x S_y \sqrt{1 - R^2}} e^{-Q(x, y)}$$

where

$$Q(x, y) = \frac{1}{2(1 - R^2)} \left[\left(\frac{x - M_x}{S_x} \right)^2 - 2R \left(\frac{x - M_x}{S_x} \right) \left(\frac{y - M_y}{S_y} \right) + \left(\frac{y - M_y}{S_y} \right)^2 \right].$$

"Modern Probability and its Applications", by Emanuel Parzen, (John Wiley & Sons, New York).

$$-\ln(P[Z > z]) \doteq \frac{1}{2(1-R^2)} \left[\left(\frac{x - M_x}{S_x} \right)^2 - 2R \left(\frac{x - M_x}{S_x} \right) \left(\frac{y - M_y}{S_y} \right) + \left(\frac{y - M_y}{S_y} \right)^2 \right].$$

By translating axis so that the new origin is at the joint means (M_x, M_y) and simplifying, the equation becomes:

$$-2(1-R^2)\ln(P[Z > z])S_x^2S_y^2 \doteq S_y^2x'^2 - 2RS_xS_yx'y' + S_x^2y'^2$$

By rotating the axis about the new origin and simplifying, the equation becomes:

$$\begin{aligned} -2(1-R^2)\ln(P[Z > z])S_x^2S_y^2 &\doteq x''^2(S_y^2\cos^2\theta - 2RS_xS_y\sin\theta\cos\theta + S_x^2\sin^2\theta) + \\ &\quad y''^2(S_y^2\sin^2\theta + 2RS_xS_y\sin\theta\cos\theta + S_x^2\cos^2\theta) + \\ &\quad + x''y''(-2S_y^2\sin\theta\cos\theta - 2RS_xS_y\cos^2\theta + \\ &\quad + 2RS_xS_y\sin^2\theta + 2S_x^2\sin\theta\cos\theta) \end{aligned}$$

Setting the coefficient of $x''y''$ equal to zero yields the following for θ :

$$\theta \doteq \frac{1}{2}\tan^{-1}\left[\frac{2S_xS_yR}{S_x^2-S_y^2}\right].$$

Simplifying, to derive the equation of the ellipse

$$1 \doteq \frac{x''^2}{\left[\frac{2(1-R^2)(-\ln(P[Z > z]))S_x^2S_y^2}{S_y^2\cos^2\theta - 2RS_xS_y\sin\theta\cos\theta + S_x^2\sin^2\theta} \right]} + \frac{y''^2}{\left[\frac{2(1-R^2)(-\ln(P[Z > z]))S_x^2S_y^2}{S_y^2\sin^2\theta + 2RS_xS_y\sin\theta\cos\theta + S_x^2\cos^2\theta} \right]}$$

$$\text{where } A = \sqrt{\frac{2(1 - R^2)(-\ln(P[z > z]))S_x^2 S_y^2}{S_y^2 \cos^2 \theta - 2RS_x S_y \sin \theta \cos \theta + S_x^2 \sin^2 \theta}}$$

$$B = \sqrt{\frac{2(1 - R^2)(-\ln(P[z > z]))S_x^2 S_y^2}{S_y^2 \sin^2 \theta + 2RS_x S_y \sin \theta \cos \theta + S_x^2 \cos^2 \theta}}$$

The ellipse is plotted with center (M_x, M_y) with semi major axis A, with semi minor axis B, and with rotation θ relative to the usual north-south/east-west coordinate axis system. The constants R, S_x, S_y, S_x², and S_y² are all computed from the coordinates before rotation.

The following discussion applies to deriving of PROGRAM ELLP3 demonstrating the intermediate steps.

Recall that we translated the old axis to the new origin (M_x, M_y) .

$$\begin{array}{lll} (x, y) & x = x' + M_x & (x', y') \\ \text{OLD} & y = y' + M_y & \text{NEW} \end{array}$$

then in the new axis our origin is $(0,0)'$.

As a result of translating we get:

$$-2(1 - R^2)\ln(P[z > z])S_x^2 S_y^2 = S_x'^2 - 2RS_x S_y x'y' + S_y'^2$$

setting the left hand side of the equation to K, a constant, we get:

$$K = S_x'^2 - 2RS_x S_y x'y' + S_y'^2.$$

We do a rotation of axes thru θ with respect to the old axes about (M_x, M_y) which is now $(0,0)'$

$$\begin{array}{lll} (x', y') & x' = x''\cos\theta - y''\sin\theta & (x'', y'') \\ \text{TRANSLATED} & y' = x''\sin\theta + y''\cos\theta & \text{ROTATED} \end{array}$$

$$\begin{aligned} K &= S_y^2(x''\cos\theta - y''\sin\theta)^2 - 2RS_x S_y(x''\cos\theta - y''\sin\theta)(x''\sin\theta + y''\cos\theta) + \\ &\quad + S_x^2(x''\sin\theta + y''\cos\theta)^2 \end{aligned}$$

$$\begin{aligned} K &= x''^2(S_y^2 \cos^2 \theta - 2RS_x S_y \sin \theta \cos \theta + S_x^2 \sin^2 \theta) + x''y''(-2S_y^2 \sin \theta \cos \theta - \\ &\quad - 2RS_x S_y \cos^2 \theta + 2RS_x S_y \sin^2 \theta + 2RS_x^2 \sin \theta \cos \theta) + \\ &\quad + y''^2(S_y^2 \sin^2 \theta + 2RS_x S_y \sin \theta \cos \theta + S_x^2 \cos^2 \theta). \end{aligned}$$

Take the $x''y''$ coefficient and set it equal to zero

$$-2S_y^2 \sin \theta \cos \theta - 2RS_x S_y \cos^2 \theta + 2RS_x S_y \sin^2 \theta + 2S_x^2 \sin \theta \cos \theta = 0.$$

Using the identities $\sin 2\theta = 2\sin \theta \cos \theta$ and $\cos 2\theta = \cos^2 \theta - \sin^2 \theta$ one obtains

$$(S_x^2 - S_y^2)\sin 2\theta - 2RS_x S_y \cos 2\theta = 0.$$

Making use of the identity $\sin 2\theta = \tan 2\theta \cos 2\theta$ this becomes

$$\cos 2\theta [\tan 2\theta (S_x^2 - S_y^2) - 2RS_x S_y] = 0.$$

If the product of two numbers is zero, one or the other, or both are zero.
In case

$$\tan 2\theta = 2RS_x S_y / (S_x^2 - S_y^2)$$

then

$$\theta = \frac{1}{2} \tan^{-1} [2RS_x S_y / (S_x^2 - S_y^2)].$$

We use this value for θ and eliminate the $x''y''$ term. Thus leaving terms containing x''^2 and y''^2 with coefficients we indicate by $c_{x''}$ and $c_{y''}$, where

$$c_{x''} = S_y^2 \cos^2 \theta - 2RS_x S_y \sin \theta \cos \theta + S_x^2 \sin^2 \theta$$

$$c_{y''} = S_y^2 \sin^2 \theta + 2RS_x S_y \sin \theta \cos \theta + S_x^2 \cos^2 \theta$$

and

$$K = -2(1 - R^2) \ln(P[z > z]) S_x^2 S_y^2.$$

With the elimination of the $x''y''$ term we get the equation of the ellipse in the translated rotated axes:

$$K = x''^2 c_{x''} + y''^2 c_{y''}$$

or

$$1 = \frac{x''^2}{K/c_{x''}} + \frac{y''^2}{K/c_{y''}}.$$

We see from this that the semimajor axis A and the semiminor B have the values

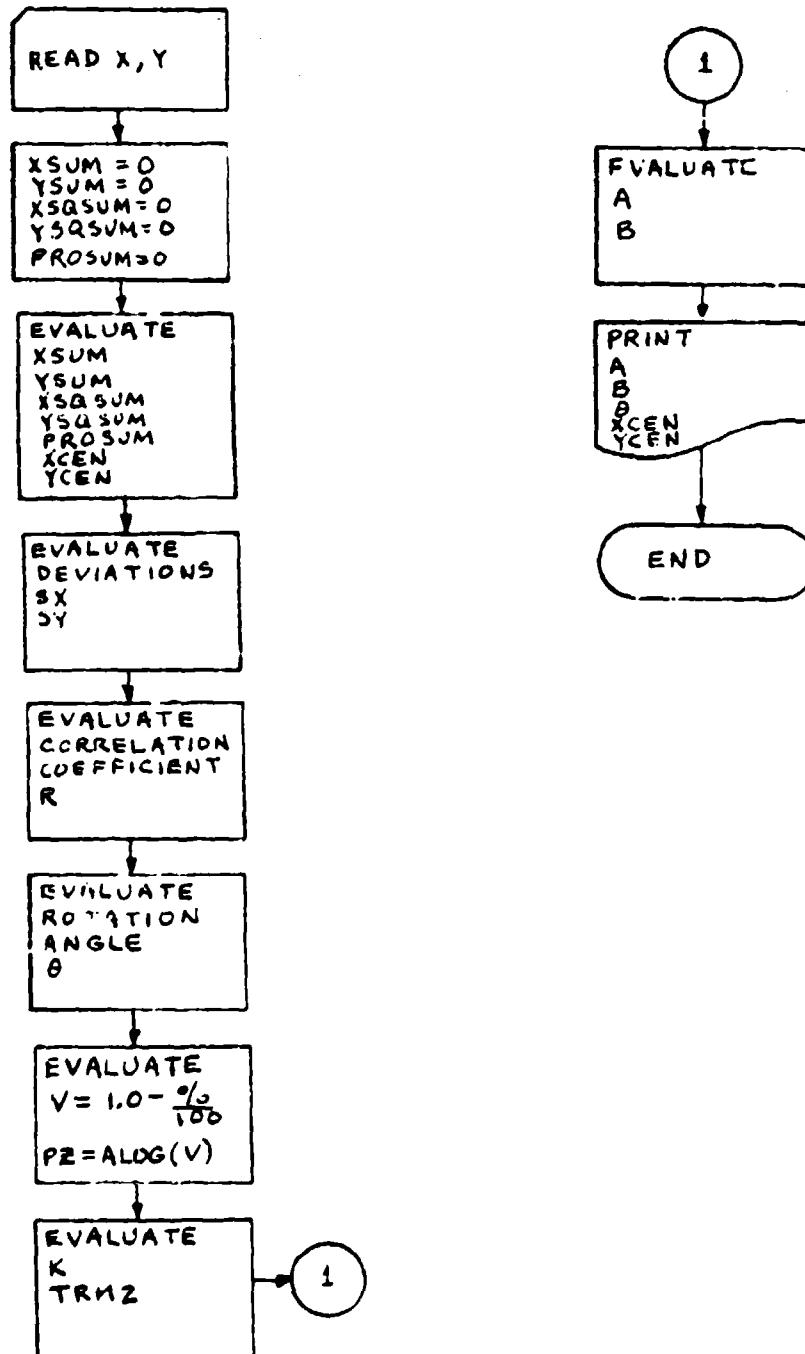
$$A = \sqrt{K/c_{x''}}, \quad B = \sqrt{K/c_{y''}}$$

when the rotated angle θ is

$$\theta = \frac{1}{2} \tan^{-1} [2S_x S_y R / (S_x^2 - S_y^2)].$$

ELLP3 PROGRAM FLOWCHART

ELLP3 IS A PROBABILITY APPROACH TO FIT AN ELLIPSE OF MINIMUM AREA TO PLOTTED IMPACT PATTERN



THREE PROGRAM COMPARISON FOR ELLIPSE PARAMETERS @95%

TAPE	PROGRAM	ROTATION(RAD,DEG)		PLOTTED	IN ELLIPSE	95 % IN ?
		ROTATION	ANGLE			
1821	PARAM SELLIPSE ELLP3	1.58825 0.972417 0.03391	,91.000 ,55.715 , 1.943	614 614 614	584 593 577	95.110 96.580 93.974
1823	PARAM SELLIPSE ELLP3	0.10472 0.788211 -0.25284	, 6.000 ,45.161 ,-14.487	623 623 623	592 583 618	95.024 95.579 99.197
1828	PARAM SELLIPSE ELLP3	1.55334 0.467355 0.27014	,89.000 ,26.548 ,15.478	642 642 642	610 615 612	95.016 95.794 95.327
1841	PARAM SELLIPSE ELLP3	2.38450 0.792313 -0.27040	,170.999 ,45.396 ,-15.493	503 503 503	478 483 493	95.030 96.073 98.011
1843	PARAM SELLIPSE ELLP3	0.15708 0.4633849 0.16115	, 9.000 ,26.550 , 9.233	506 506 506	482 486 496	95.257 96.047 98.023
1848	PARAM SELLIPSE ELLP3	0.48869 0.478499 0.49286	,28.000 ,27.415 ,28.239	638 638 638	607 618 623	95.14 96.865 97.619

<u>TAPE</u>	<u>PROGRAM</u>	<u>A-AXIS</u>	<u>B-AXIS</u>	<u>AREA</u>	<u>XCEN</u>	<u>YCEN</u>
1821	PARAM SELLIPSE ELLP3	232.060 255.666 180.074	174.067 228.750 218.572	126901.222 183731.64 123650.37	424446.375 424446.875 424447.00	622575.937 622575.687 622588.875
1823	PARAM SELLIPSE ELLP3	459.723 380.246 507.229	372.158 394.185 461.352	537494.050 470946.66 735167.6	487020.625 487020.187 487020.375	489769.562 489768.812 489762.625
1828	PARAM SELLIPSE ELLP3	192.969 165.868 166.390	153.713 200.397 191.482	93185.561 104424.81 100093.31	483793.937 483794.625 483794.062	241720.687 241720.500 241716.250
1841	PARAM SELLIPSE ELLP3	464.394 367.528 602.328	301.784 486.431 333.525	440283.78 561644.55 631119.09	486297.625 486298.625 486297.187	493071.687 493071.687 493080.187
1843	PARAM SELLIPSE ELLP3	428.650 453.3 556.280	283.253 300.6 303.841	381440.684 428079.64 530994.1	483421.875 483422.375 483420.125	241589.062 241589.562 241584.000
1848	PARAM SELLIPSE ELLP3	546.773 605.190 647.773	261.299 279.539 271.756	448843.934 531476.45 553034.03	482347.125 482347.500 482344.062	250196.250 250196.250 250188.812

METHODS FOR APPROXIMATING MATHEMATICAL FUNCTIONS

Donald W. Rankin
Army Materiel Test and Evaluation Directorate
US Army White Sands Missile Range
White Sands Missile Range, NM 88002

ABSTRACT. The increasing use of small computers to control complex systems and equipment often gives rise to the need for approximating mathematical functions under restrictive conditions. These conditions may be so varied that no single solution can be called optimum.

Several methods are discussed in some detail, with a view toward simplification. In fact, some of the procedures easily can be committed to memory. An unexpected dividend allows the analyst to employ the powerful features of some programming language even though it may lack a needed mathematical function.

A partial list of subjects addressed includes:

- (1) Power series. A method for developing another power series which converges more rapidly.
- (2) Padé approximations (rational functions). Developing a Padé expression from a truncated power series. Reducing the Padé coefficients to integers (although the function is unique, its coefficients are not --- a powerful advantage).
- (3) Operations which increase accuracy.
 - (a) A linear combination of two approximations can be much more accurate than either alone.
 - (b) Properly restricting the variable range can markedly improve the rate of convergence.
 - (c) A Padé expression can be "optimized" for a stated variable range, in effect embracing both of the above advantages in a single expression.
- (4) Tschebychev polynomials. Tschebychev series. The necessity of employing a transformation of variables. Choosing a transformation which simplifies the Tschebychev expression and reduces the labor of computing the coefficients.
- (5) Maehly's method of developing a rational function from a Tschebychev series.
- (6) An efficient square root algorithm which does not require access to assembly language. Extension to higher roots.

I. INTRODUCTION. There is increasingly widespread use of embedded small computers to perform monitor and control functions in all manner of complex systems. The mushrooming demand has triggered a virtual explosion in the computer industry. But equipments are so different, and conditions can be so varied that no library of software routines can hope to avoid either paucity or obsolescence.

This paper, then, is not a catalogue of mathematical approximations. Rather, it is a discussion of several methods which can be used to produce required approximations.

The methods can be applied to any mathematical function which can be developed in a power series. Analogously, a digital filter can be synthesized to approximate any physical process capable of being expressed as a linear difference equation of any order.

An unexpected dividend accrued recently. An analysis program was written in COBOL to take advantage of the powerful "bookkeeping" features of that language, even though it contains no subroutine for computing required logarithms. Employing the methods herein, a suitable subroutine was easily devised.

II. RATIONAL APPROXIMATIONS TO CERTAIN MATHEMATICAL CONSTANTS. Sometimes it is useful to be able to express a mathematical constant as a ratio. This is particularly the case when a computer (or calculator) will compute with greater precision than it will store or accept inputs. The author owns a calculator which computes to eleven digits but accepts, at most, eight-digit entries.

A carefully chosen ratio will deliver as many significant digits as there are total digits in the fraction (reduced, of course, to lowest terms).

There are several methods for searching for these approximations. We illustrate a method of continued fractions.

Suppose we need an approximation for $\pi = 3.141592\ 653589\ 793238\ 46\dots$. Taking the reciprocal of the fractional part, we express it as

$$\pi = 3 + \frac{1}{7.062513\ 305931\ 045769\ 8}$$

Repeating the procedure,

$$\pi = 3 + \frac{1}{7 + \frac{1}{15.996594\ 406685\ 7199}}$$

It is obvious that

$$\pi \approx 3 + \frac{1}{7 + \frac{1}{16}}$$

will be an excellent approximation. Unscrambling the continued fraction yields

$$\pi \approx 3 + \frac{16}{113} = \frac{355}{113}, \text{ in error by only } 2.67 \times 10^{-7}.$$

If every numerator is unity and every convergent begins with an integer, convergence is assured, but may be slow (e.g., $\ln_6 10$). A convenient way to hasten the process is to begin with almost any recognizable approximation, then apply the continued fraction technique to the residual. Thus

$$\begin{aligned}\ln_6 10 &= 2.302585\ 092994\ 045684\ 018 \dots \\ &= \frac{23}{10} + \frac{1}{386.833279\ 229539\ 353860 \dots}\end{aligned}$$

It is immediately apparent that

$$\ln_6 10 \approx \frac{23}{10} + \frac{1}{386 + \frac{5}{6}}$$

will be an excellent approximation. In vulgar fraction form

$$\ln_6 10 \approx \frac{53443}{23210}, \text{ which errs by } 3.6 \times 10^{-10}.$$

Similarly, $e = 2.718281\ 828459\ 045235\ 36 \dots$ can be expressed as

$$\frac{19}{7} + 0.003996\ 114173\ 330949\ 646 \dots$$

whence

$$e = \frac{19}{7} + \frac{1}{250.243100\ 328250\ 339641 \dots}$$

from which

$$e \approx \frac{19}{7} + \frac{4}{1001} = \frac{2721}{1001} \quad (\text{error} \approx 1.1 \times 10^{-7}).$$

One more step will produce

$$e = \frac{19}{7} + \frac{1}{250 + \frac{1}{4.113527\ 970929\ 849525 \dots}}$$

which leads to

$$e = \frac{19}{7} + \frac{1}{250 + \frac{1}{4 + \frac{1}{9}}}$$

Unscrambling,

$$e = \frac{176180}{64813} \quad (\text{error} \approx 2.3 \times 10^{-9}).$$

It should be apparent that any mathematical constant can be approximated with arbitrary accuracy by an easily-found rational fraction.

Some of the common ones are listed:

$$\pi = 3.141592 653589 793238 46 \dots$$

$$e = 2.718281 828459 045235 36 \dots$$

$$\pi = \frac{355}{113} - 2.67 \times 10^{-7}$$

$$e = \frac{193}{71} - 2.80 \times 10^{-5}$$

$$\pi = \frac{312689}{99532} - 2.90 \times 10^{-11}$$

$$e = \frac{49171}{18089} - 2.77 \times 10^{-10}$$

$$\pi^2 = 9.869604 401089 358618 83 \dots$$

$$\sqrt{e} = 1.648721 270700 128146 85 \dots$$

$$\pi^2 = \frac{227}{23} + 3.92 \times 10^{-5}$$

$$\sqrt{e} = \frac{61}{37} + 7.26 \times 10^{-5}$$

$$\pi^2 = \frac{98548}{9985} - 5.52 \times 10^{-9}$$

$$\sqrt{e} = \frac{34361}{20841} + 1.28 \times 10^{-10}$$

$$\pi/180 = 0.017453 292519 943295 77 \dots$$

$$e^2 = 7.389056 098930 650227 23 \dots$$

$$180/\pi = 57.295779 513082 320876 80 \dots$$

$$e^2 = \frac{2431}{329} - 1.65 \times 10^{-6}$$

$$180/\pi = \frac{4068}{71} + 4.87 \times 10^{-6}$$

$$e^2 = \frac{176761}{23922} - 5.77 \times 10^{-11}$$

$$180/\pi = \frac{829471}{14477} + 7.52 \times 10^{-10}$$

$$\ln_e 10 = 2.302585 092994 045684 018 \dots$$

$$\sqrt{\pi} = 1.772453 850905 516027 30 \dots$$

$$\ln_e 10 = \frac{175}{76} - 4.65 \times 10^{-5}$$

$$\sqrt{\pi} = \frac{296}{167} - 1.24 \times 10^{-6}$$

$$\ln_e 10 = \frac{53443}{23210} + 3.62 \times 10^{-10}$$

$$\sqrt{\pi} = \frac{8545}{4821} + 3.16 \times 10^{-9}$$

Euler's constant

$$\sqrt{2\pi} = 2.506628 274631 000502 42 \dots$$

$$\gamma = 0.577215 664901 532860 6065 \dots$$

$$\sqrt{2\pi} = \frac{945}{377} - 3.025 \times 10^{-6}$$

$$\gamma = \frac{228}{395} + 4.75 \times 10^{-7}$$

$$\sqrt{2\pi} = \frac{221987}{88560} + 1.492 \times 10^{-11}$$

$$\gamma = \frac{33841}{58628} + 3.15 \times 10^{-11}$$

$$\ln_e 2 = 0.693147 \ 180559 \ 945309 \ 4172\ldots \quad \ln_e 3 = 1.098612 \ 288668 \ 109691 \ 395\ldots$$

$$\ln_e 2 = \frac{61}{88} - 3.46 \times 10^{-5}$$

$$\ln_e 3 = \frac{78}{71} + 2.07 \times 10^{-5}$$

$$\ln_e 2 = \frac{25469}{36744} + 6.79 \times 10^{-11}$$

$$\ln_e 3 = \frac{24621}{22411} + 5.98 \times 10^{-11}$$

III. POWER SERIES. Power series are and will remain most useful tools. Whenever great accuracy is required, a properly chosen power series will deliver all the precision of which the computer is capable. Let us look at a familiar MacLaurin expansion:*

$$\frac{\sin \theta}{\theta} = 1 - \frac{\theta^2}{13} + \frac{\theta^4}{15} - \frac{\theta^6}{17} + \frac{\theta^8}{19} - \dots \quad (1)$$

This series converges quite rapidly, providing the value of θ is not too large. But we note that if θ exceeds $\frac{\pi}{4}$, we need merely compute $\cos(\frac{\pi}{2} - \theta)$.

Hence $\frac{\pi}{4}$ will be the largest value of the argument employed. The general term can be written

$$\psi_{2n} = (-1)^n \frac{\theta^{2n}}{12n + 1} \quad (2)$$

Any term can be computed from its predecessor by means of a term-to-term recurrence ratio. Thus

$$\psi_{2n} = \frac{-\theta^2}{2n(2n+1)} \psi_{2n-2} \quad (3)$$

There remains only to compare the size of the latest computed term with some pre-established criterion. If the term is small enough, the computation is finished. Since it is not known in advance how many terms will be required, running time is unpredictable, making the method unsuitable for real-time situations.

Another method employs the "nested" polynomial technique. The series is truncated at an arbitrary point (e.g., after the term $\frac{\theta^8}{19}$) and the arithmetic begun at the other end. We can write

$$((((.. - \frac{1}{19})\theta^2 - \frac{1}{17})\theta^2 + \frac{1}{15})\theta^2 - \frac{1}{13})\theta^2 + 1 = \frac{\sin \theta}{\theta}$$

*For the sake of uniformity, throughout this paper we shall employ series wherever possible whose leading term is unity. The advantage when employing a recurrence ratio is obvious.

Since most computers multiply faster than they divide, it may be more efficient to multiply both sides by $19 = 362880$, yielding

$$(((\theta^2 - 72)\theta^2 + 3024)\theta^2 - 60480)\theta^2 + 362880$$

$$= 362880 \frac{\sin \theta}{\theta} \quad (4)$$

In this form, the algorithm will have a fixed running time (which is very fast), but the error will be a function of the argument. The maximum error, however, can be closely estimated. In this case (for $\sin \theta$) it is given by

$$|\epsilon| < \frac{1}{111} (\frac{\pi}{4})^{11} = 1.757 \times 10^{-9}$$

IV. RESTRICTING THE RANGE OF THE ARGUMENT. Supposing that, in the previous example, we had computed $\sin \frac{\theta}{3}$, then recovered the wanted value by means of the identity $\sin 3\phi = 3 \sin \phi - 4 \sin^3 \phi$, or, in handier form:

$$\frac{\sin 3\phi}{\sin \phi} = 3 - 4 \sin^2 \phi \quad (5)$$

The errors will be as -9:1, hence (for $\sin 3\phi$)

$$|\epsilon| < \frac{9}{111} (\frac{\pi}{12})^{11} = 0.89 \times 10^{-13}$$

This is a dramatic reduction in the maximum error. If preferred, the series can be truncated one more term, simplifying the nested polynomial to

$$((42 - \theta^2)\theta^2 - 840)\theta^2 + 5040 \approx 5040 \frac{\sin \theta}{\theta} \quad (6)$$

Maximum error is now $|\epsilon| < \frac{9}{19} (\frac{\pi}{12})^9 = 1.433 \times 10^{-10}$

We can carry the process another step, using the identity $\sin 5\phi = 5 \sin \phi - 20 \sin^3 \phi + 16 \sin^5 \phi$ or

$$\frac{\sin 5\phi}{\sin \phi} = 5 - 4 \sin^2 \phi (5 - 4 \sin^2 \phi) \quad (7)$$

The errors will be as 25:1, hence

$$|\epsilon| < \frac{25}{17} (\frac{\pi}{20})^7 = 1.17 \times 10^{-8}$$

for the extremely simple expression

$$(\theta^2 - 20)\theta^2 + 120 \approx 120 \frac{\sin \theta}{\theta} \quad (8)$$

When faced with a very slowly converging series, some such technique is virtually mandatory. The series for logarithms and for the inverse trigonometric functions offer typical examples. Let us look at

$$\frac{\arctan z}{z} = 1 - \frac{z^2}{3} + \frac{z^4}{5} - \frac{z^6}{7} + \frac{z^8}{9} - \dots \quad (9)$$

For values of $z > 1$, we employ the identity

$$\arctan z = \frac{\pi}{2} - \arctan \frac{1}{z}$$

but this alone is insufficient. Clearly, if we summed a billion terms, the maximum error would still be

$$|\epsilon| < (2,000,000,001)^{-1} = 5 \times 10^{-10}$$

An additional step is required. We offer two choices.

The first method does not involve square roots. (Some computers perform square root awkwardly or slowly.) The identity

$$\arctan \frac{1}{a} = \arctan \frac{1}{1+a} + \arctan \frac{1}{1+a+a^2} \quad (10)$$

is employed. This requires summing two series, but the worst-case argument cannot exceed $\frac{1}{2}$. For $\arctan z$

$$|\epsilon| < \frac{1}{29} (\frac{1}{2})^{29} = 0.64 \times 10^{-10}$$

after only fourteen terms. If desired, the identity can be employed a second time, resulting in

$$\arctan \frac{1}{a} = \arctan \frac{1}{2+a} + \arctan \frac{1}{1+a+a^2} + \arctan \frac{1}{3+3a+a^2} \quad (11)$$

Three series must now be summed, but after only nine terms each, the maximum error is given by

$$|\epsilon| < \frac{2}{19} (\frac{1}{3})^{19} = 0.9 \times 10^{-10}$$

In the second method, the identity

$$\arctan x = 2 \arctan \frac{x}{1+\sqrt{1+x^2}} \quad (12)$$

is employed. The argument does not now exceed $\frac{1}{1 + \sqrt{2}} = 0.4142\dots$, and

after 11 terms, maximum error is $|\epsilon| < 1.36 \times 10^{-10}$. Obviously, the identity can be applied a second time, yielding

$$\arctan x = 4 \arctan \frac{x}{1 + u + \sqrt{2}(u^2 + u)} \quad (13)$$

where $u^2 = 1 + x^2$. The argument now is no greater than 0.1989..., and after only six terms, $|\epsilon| < 2.35 \times 10^{-10}$.

The identity can be applied, of course, as many times as desired. One more application yields, after only four terms,

$$|\epsilon| < 7.75 \times 10^{-10}$$

Of all the tools that can be used to increase the accuracy of an approximation, restricting the argument range is perhaps the most important, and should be given the highest priority.

V. PADE APPROXIMATIONS. Just as the ratio of two integers can be used to approximate an irrational number, so can the ratio of two polynomials be used to approximate a transcendental function. If the function can be expanded in a power series, there is available a particularly easy method (called a Padé approximation) for obtaining such an approximation. The expression will look like

$$\frac{1 + a_1 x + a_2 x^2 + a_3 x^3}{1 + b_1 x + b_2 x^2 + b_3 x^3} \approx 1 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + \dots \quad (14)$$

It is easy to see that we can multiply the coefficients of the rational function by any arbitrary constant without changing its value. Thus although the function may be unique, its coefficients never are. Herein lies a second advantage of the Padé. In many cases the coefficients can be reduced to small integers. In this form, the Padé is very economical of both running time and storage space, qualities which cannot lightly be ignored.

There is no restriction on the degree of polynomial used in either numerator or denominator. However, the better approximations occur when the degrees of numerator and denominator do not greatly differ, and are significantly more accurate than the truncated power series from which they have been derived. The Padé is much better behaved in the region of maximum error (the slope of the Padé error curve is less steep).

A word of caution is necessary. All polynomials have zeroes. All polynomials of odd degree have at least one real zero. Zeroes in the denominator usually produce poles in the rational function. If such a pole is fictitious (i.e., there is no corresponding pole in the function being approximated), the approximation will "blow up" as the argument approaches the pole. For real values of the argument, complex poles usually cause no difficulty, nor do real

poles which lie well outside the argument range being used. (We have found an additional reason for restricting the argument range.)

It will be found that if a power series is truncated after that term whose exponent is equal to the sum of the degrees of the corresponding Padé numerator and denominator, it provides just enough known coefficients to enable us to compute the unknown ones (in the Padé expression).

The basic form for computation is written

$$\frac{1 + a_1 x + \cdots + a_n x^n}{1 + b_1 x + \cdots + b_m x^m} = 1 + c_1 x + c_2 x^2 + \cdots + c_{n+m} x^{n+m} \quad (15)$$

in which the c_i 's are known, the a_i 's and b_i 's unknown. Both sides are multiplied by the Padé denominator, yielding

$$1 + a_1 x + \cdots + a_n x^n = \\ (1 + b_1 x + \cdots + b_m x^m)(1 + c_1 x + c_2 x^2 + \cdots + c_{n+m} x^{n+m}) \quad (16)$$

The indicated multiplication is performed on the right. Like terms are collected, dropping all terms of degree greater than $m + n$. These terms will be small, but dropping them explains why the Padé and power series approximations yield different results.

The coefficients of terms of like degree are now equated (after supplying the left side with m terms equal to zero). This gives rise to a system of $m + n$ simultaneous linear equations in $m + n$ unknowns. Now it is seen why the higher degree terms are dropped. Retaining them would lead to additional equations which might be (indeed usually are) inconsistent.

The required set of simultaneous linear equations can be written directly by means of the following algorithm:

Step 1. Imagining a checkerboard enter the known coefficients --- the coefficients of the given power series --- along the principal diagonal.
 $c_0 = 1$ is a known coefficient.

Step 2. Below c_{n+m} draw a horizontal line. No terms will be entered below this line.

1					
b_1	c_1				
	b_1	c_2			
		b_1	c_3		
			b_1		
				b_1	c_{m+n}

Step 3. Enter the first unknown coefficient (b_1) everywhere along the next diagonal.

Step 4. Repeat step 3 for each successive coefficient of the Padé denominator.

Step 5. When all the b_1 's have been entered, there will remain at the lower left corner an empty $n \times n$ triangular pattern. Fill it up with zeroes.

Step 6. Multiply each b_1 by the known coefficient at the head of the column.

1					
b_1	c_1				
	$c_1 b_1$	c_2			
b_m		$c_2 b_1$	c_3		
0	$c_1 b_m$		$c_3 b_1$		
0	0	$c_2 b_m$			c_{m+n}

Step 7. Each horizontal row represents the right side of one of the required equations. Add + signs between appropriate terms. Place an = sign to the left of each expression.

Step 8. Arrange the coefficients of the Padé numerator (the a 's) in a vertical column, supplying m zeroes at the bottom. Each element becomes the left side of the appropriate equation. The complete set of equations can now be written.

$$1 = 1$$

$$a_1 = b_1 + c_1$$

$$a_2 = b_2 + c_1 b_1 + c_2$$

$$a_3 = b_3 + c_1 b_2 + c_2 b_1 + c_3$$

 \cdot
 \cdot

$$0 = 0 + \dots + c_n b_m + \dots + c_{m+n}$$

(17)

Step 9. Those equations whose left member is zero form a set of m equations in m unknowns, and can be solved for the b_i 's.

Step 10. The a_i 's are found by direct substitution.

Step 11. Multiplying all the Padé coefficients by their L.C.D. converts them to integers.

Let us illustrate by example. Suppose it is required to write a routine for $\arctan z$. Let us arbitrarily decide to write a Padé of the form

$$\arctan z = \frac{kx(1+ax^2)}{1+b_2x^2+b_4x^4}$$

using the transformation $x = \frac{z}{1+\sqrt{1+z^2}}$ to restrict the range of the argument.

Under this transformation

$$\arctan z = 2 \arctan x$$

so we already know that $k = 2$.

Note that for an even function, we can mentally make the substitution $w = x^2$ and derive the Padé expression as though it were a function of w .

The appropriate power series is

$$\frac{\arctan x}{x} = 1 - \frac{x^2}{3} + \frac{x^4}{5} - \frac{x^6}{7} + \dots \quad (9)$$

We require a rational function such that

$$\frac{1+ax^2}{1+b_2x^2+b_4x^4} \approx \frac{\arctan x}{x}$$

Immediately we write the known coefficients as a diagonal, then add the b 's.

1

$$b_2 \quad -\frac{1}{3}$$

$$b_4 \quad b_2 \quad \frac{1}{5}$$

$$0 \quad b_4 \quad b_2 \quad -\frac{1}{7}$$

Next we multiply by the known coefficient at the head of each column:

$$1$$

$$b_2 - \frac{1}{3}$$

$$b_4 - \frac{1}{3} b_2 \quad \frac{1}{5}$$

$$0 - \frac{1}{3} b_4 \quad \frac{1}{5} b_2 - \frac{1}{7}$$

After adding the column of a's (and zeroes) to the left and supplying the needed = and + signs, we have the required set of equations:

$$1 = 1$$

$$a_2 = b_2 - \frac{1}{3}$$

$$0 = b_4 - \frac{1}{3} b_2 + \frac{1}{5}$$

$$0 = 0 - \frac{1}{3} b_4 + \frac{1}{5} b_2 - \frac{1}{7}$$

Those equations whose left member is zero form a set which can be solved for the b's. Thus, after multiplying the last equation by 3, we find

$$0 = b_4 - \frac{1}{3} b_2 + \frac{1}{5}$$

$$0 = -b_4 + \frac{3}{5} b_2 - \frac{3}{7}$$

Adding,

$$0 = \frac{4}{15} b_2 - \frac{8}{35}$$

from which

$$b_2 = \frac{8}{35} \cdot \frac{15}{4} = \frac{6}{7}$$

$$0 = b_4 - \frac{2}{7} + \frac{1}{5} \quad \text{or} \quad b_4 = \frac{3}{35}$$

and

$$a_2 = \frac{6}{7} - \frac{1}{3} = \frac{11}{21}$$

The required rational function is thus

$$\frac{1 + \frac{11}{21}x^2}{1 + \frac{6}{7}x^2 + \frac{1}{35}x^4}$$

Multiplying numerator and denominator by 105 (the L.C.D.) converts all the coefficients to integers:

$$\frac{105 + 55x^2}{105 + 90x^2 + 9x^4}$$

The complete approximation is

$$\arctan z = \frac{x(210 + 110x^2)}{105 + 90x^2 + 9x^4}$$

where x is given by

$$x = \frac{z}{1 + \sqrt{1 + z^2}}$$

The maximum error occurs when $x = \sqrt{2} - 1$ where, of course $z = 1$ and
 $\arctan z = \frac{\pi}{4}$

$$\left(\frac{\pi}{4}\right) = 0.78539\ 81634)$$

The Padé approximation yields (at $z = 1$)

$$\arctan z \approx 0.78539\ 52528\ 427$$

in error by -2.91×10^{-6}

The truncated power series from which the Padé was derived yields

$$\arctan z \approx 0.78532\ 81810\ 156$$

in error by -7.00×10^{-5}

The Padé is noticeably better, as it often is when the power series converges slowly --- another plus for the Padé. It works best when needed most.

The algorithm for computing Padé coefficients also can be used to develop a reciprocal power series whenever one is required. Since all the a_i 's are zero (except $a_0 = 1$, of course), the left sides of all pertinent equations become zero, eliminating the need for a simultaneous set. The b_i 's can be determined in order, for any arbitrarily large number of them.

After truncation, both the original power series and the reciprocal series can be thought of as limiting-case Padé approximations.

Suppose that a function is expressible in power series form and that the Padé algorithm has been used to develop the reciprocal series. We can write

$$f(x) = 1 + g_1x + g_2x^2 + \dots = \frac{1}{1 + h_1x + h_2x^2 + \dots}$$

The latter can be written

$$f(x) = \frac{1}{1 + h_1(1 + k_1x + k_2x^2 + \dots)}$$

and the process repeated upon the interior series, yielding

$$f(x) = \frac{1}{h_1x} \cdot \frac{1}{1 + \frac{1}{1 + l_1x + l_2x^2 + \dots}}$$

Continuing in this manner develops the function in continued fraction form, and emphasizes the close relationship between Padé approximations and continued fractions.

If a continued fraction is terminated at some n th convergent, it becomes an approximation to the function. It is easy to evaluate by a process of "nested division" not unlike the evaluation of polynomials by "nested multiplication."

Use of the continued fraction technique often results in a program of significantly fewer instructions. However, so many divisions must be performed that running time may be quite slow. There are some models of computers which divide quite rapidly (usually in single precision only). For them, the use of continued fraction approximations seems attractive.

If a terminated continued fraction is "unscrambled," it will be found that the resulting rational function is a member of the set of Padé approximations.

VI. DEVELOPING A POWER SERIES WHICH CONVERGES MORE RAPIDLY. Let us continue to use as an example the series for $\arctan x$. Repeating for convenience

$$\frac{\arctan x}{x} = 1 - \frac{1}{3}x^2 + \frac{1}{5}x^4 - \frac{1}{7}x^6 + \frac{1}{9}x^8 - \dots \quad (9)$$

a series which converges very slowly. Since this series (as the right side is stated) is an even function of x , we mentally make the substitution $w = x^2$ and treat it as an expansion in w .

Now if the ratio of two successive coefficients approaches some definite limit l ($\neq 0$) as the exponent of w increases without bound, multiplying by $(1 - w/l)^n$ will produce a new series which converges more rapidly (n is a positive integer). It will be necessary to include enough terms to accommodate the multiplier. Thus, since in the present case $l = -1$,

$$\frac{1+x^2}{x} \arctan x = 1 + \frac{2}{3}x^2 - \frac{2}{3 \cdot 5}x^4 + \frac{2}{5 \cdot 7}x^6 - \frac{2}{7 \cdot 9}x^8 + \dots \quad (18)$$

and

$$\frac{(1+x^2)^2}{x} \arctan x = 1 + \frac{5}{3}x^2 + \frac{8x^4}{1 \cdot 3 \cdot 5} - \frac{8x^6}{3 \cdot 5 \cdot 7} + \frac{8x^8}{5 \cdot 7 \cdot 9} - \dots \quad (19)$$

Using the series for $\frac{1+x^2}{x} \arctan x$, let us develop a Padé expression which has one less term in the denominator than did our previous effort. The work layout is so simple, we could almost do it in the head.

$$l = 1$$

$$a = b + \frac{2}{3}$$

$$0 = 0 + \frac{2}{3}b - \frac{2}{15}$$

from which $b = \frac{1}{5}$ and $a = \frac{13}{15}$. The required expression is

$$(1+x^2) \frac{\arctan x}{x} = \frac{1 + \frac{13}{15}x^2}{1 + \frac{1}{5}x^2}$$

Dividing both sides by $1+x^2$,

$$\frac{\arctan x}{x} = \frac{1 + \frac{13}{15}x^2}{1 + \frac{6}{5}x^2 + \frac{1}{5}x^4}$$

For purposes of comparison, let us multiply by the same integer as before. Thus

$$\frac{\arctan x}{x} = \frac{105 + 91x^2}{105 + 126x^2 + 21x^4}$$

Since we used fewer terms of the power series to develop this expression, we expect less accuracy, and indeed at $z = 1$,

$$\arctan z \approx 0.78530\ 37156\ 499$$

in error by -9.445×10^{-5}

VII. OPTIMIZING A PADÉ APPROXIMATION FOR A STATED VARIABLE RANGE. Once the decision has been made which sets the range over which the argument will be allowed to vary, the Padé approximation can be "optimized" for that specific range. If the range restrictions change, so must the optimizing coefficients.

Let us continue with the example, $\arctan z$.

Arranging the Padé coefficients in matrix form, and labeling them primary and secondary according to the order in which they were developed, we find

$$P \begin{pmatrix} 105 & 55 & 0 \\ 105 & 90 & 9 \end{pmatrix}$$

and

$$S \begin{pmatrix} 105 & 91 & 0 \\ 105 & 126 & 21 \end{pmatrix}$$

Let us produce what we shall call the "delta matrix" by performing the subtraction

$$S - P = \Delta$$

$$\Delta \begin{pmatrix} 0 & 36 & 0 \\ 0 & 36 & 12 \end{pmatrix}$$

If the work has been performed correctly, the first column will be zeroes and the second constant. If desired, the delta matrix can be reduced to lowest terms.

It turns out that any multiple of Δ , when added to P , produces a linear combination of P and S , which will define another Padé-like approximation. Its error curve will be a linear combination of two monotonic error curves which are not congruent, and hence the combination error curve will not be monotonic, but will have two zeroes. One will be at the origin. The other can be positioned arbitrarily. By choosing that point near the maximum allowable value of the argument, the approximation can be markedly improved.

To return to the example, the maximum errors of P and S suggest the form

$$\begin{aligned} P - \frac{\Delta}{36} &= P' \begin{pmatrix} 105 & 54 & 0 \\ 105 & 89 & 8\frac{2}{3} \end{pmatrix} \\ &= P' \begin{pmatrix} 315 & 162 & 0 \\ 315 & 267 & 26 \end{pmatrix} \end{aligned}$$

The full expression is

$$\arctan z = \frac{x(630 + 324x^2)}{315 + 267x^2 + 26x^4}$$

where

$$x = \frac{z}{1 + \sqrt{1 + z^2}}$$

When $z = 1$, $x = \sqrt{2} - 1$, $\arctan z = \pi/4$. The approximation yields 0.78539 79371 287, in error by -2.2627×10^{-7} . This is an order of magnitude improvement.

Let us define relative error as follows:

$$\text{relative error} = \frac{\text{approximation}}{\text{true value}} - 1 \quad (20)$$

Graphing the relative error of the approximation P' , we see that it varies from $+2.814 \times 10^{-7}$ at $z = 0.77$ to -2.881×10^{-7} at $z = 1.00$, leaving virtually no room for improvement. For all practical purposes, then, P' optimizes P . (See Fig. 1)

Supposing that we apply the transformation a second time, add a term to the numerator, and optimize. How much accuracy would be gained?

$$\arctan z \approx 4x \left(\frac{1 + a_2 x^2 + a_4 x^4}{1 + b_2 x^2 + b_4 x^4} \right)$$

$$\text{where } v^2 = 1 + z^2 \text{ and } x = \frac{z}{1 + v + \sqrt{2(v^2 + v)}}$$

$$l = 1$$

$$a_2 = b_2 - \frac{1}{3}$$

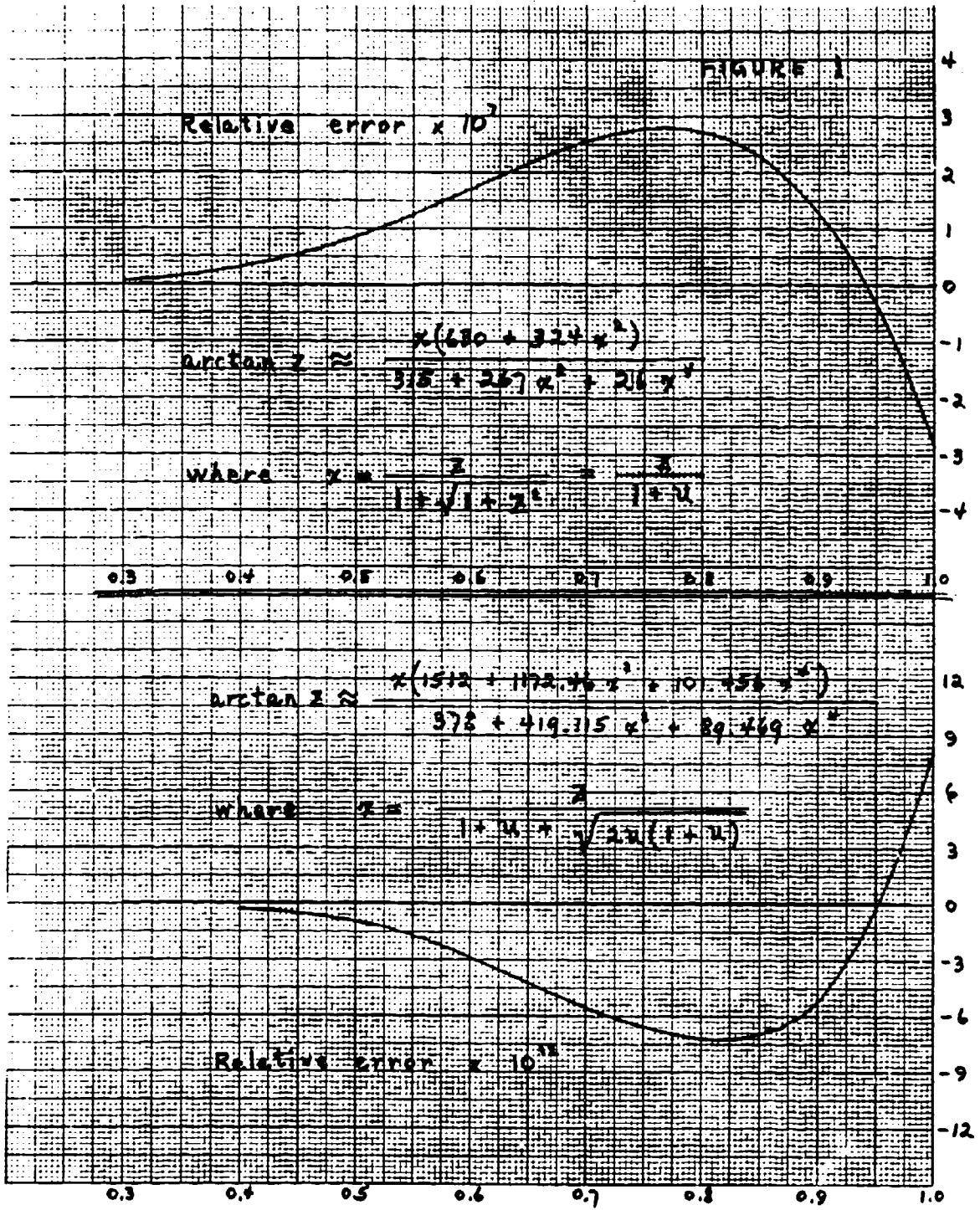
$$a_4 = b_4 - \frac{1}{3} b_2 + \frac{1}{5}$$

$$0 = 0 - \frac{1}{3} b_4 + \frac{1}{5} b_2 - \frac{1}{7}$$

$$0 = 0 + 0 + \frac{1}{5} b_4 - \frac{1}{7} b_2 + \frac{1}{9}$$

The basic approximation then computes to be

$$\arctan z \approx 4x \left(\frac{945 + 735x^2 + 64x^4}{945 + 1050x^2 + 225x^4} \right)$$



When $z = 1$, $x = 0.1989 \dots$, r.e. $\approx 1.3 \times 10^{-10}$

The second form is derived from

$$1 = 1$$

$$a_2 = b + \frac{2}{3}$$

$$a_4 = 0 + \frac{2}{3}b - \frac{2}{15}$$

$$0 = 0 + 0 - \frac{2}{15}b + \frac{2}{35}$$

from which

$$b = \frac{3}{7}$$

$$a_2 = \frac{23}{21}$$

$$a_4 = \frac{16}{105}$$

$$\frac{1 + \frac{23}{21}x^2 + \frac{16}{105}x^4}{1 + \frac{10}{7}x^2 + \frac{3}{7}x^4} \text{ reduces to } \arctan z = 4x \left(\frac{945 + 1035x^2 + 144x^4}{945 + 1350x^2 + 405x^4} \right)$$

yielding

$$\Delta \begin{pmatrix} 0 & 300 & 80 \\ 0 & 300 & 180 \end{pmatrix}$$

$P = 0.007375\Delta$ is virtually optimum, giving

$$\arctan z = 4x \left(\frac{945 + 732.7875x^2 + 63.41x^4}{945 + 1047.7875x^2 + 223.6725x^4} \right)$$

r.e. is within $\pm 8.1 \times 10^{-12}$

Multiplying through by 400 obviously recovers integer coefficients.

The relative error curves for these two "optimized" approximations are plotted in Figure 1. The form of the error curve near the maximum argument is typical.

The approximation $P = 0.0074\Delta$ is close to optimum and yields smaller coefficients, viz.,

$$\arctan z = 4x \left(\frac{1575 + 1221.3x^2 + 105.68x^4}{1575 + 1746.3x^2 + 372.78x^4} \right)$$

Multiplying by 50 recovers integers.

Additional Padé approximations are given in the appendices.

VIII. THE USE OF TSCHEBYCHEV POLYNOMIALS. Let us introduce the following shorthand notation:

$$T_0 = \cos(0) = 1$$

$$T_1 = \cos \theta$$

$$T_2 = \cos 2\theta = 2 \cos^2 \theta - 1$$

$$T_3 = \cos 3\theta = 4 \cos^3 \theta - 3 \cos \theta$$

.

.

.

$$T_n = \cos n\theta$$

(21)

Recalling that

$$\cos(n\theta + \theta) = \cos \theta \cos n\theta - \sin \theta \sin n\theta$$

and

$$\cos(n\theta - \theta) = \cos \theta \cos n\theta + \sin \theta \sin n\theta$$

we obtain by simple addition

$$\cos(n\theta - \theta) + \cos(n\theta + \theta) = 2 \cos \theta \cos n\theta \quad (22)$$

which, stated in the shorthand, becomes

$$T_{n-1} + T_{n+1} = 2T_1 T_n \quad (23)$$

This trigonometric identity --- sometimes called a three-term recurrence relation --- enables us to compute any T_{n+1} from the previous two. The result will be an expression in the various powers of $\cos \theta$.

Further to simplify, let us make the parametric substitution $x = \cos \theta$. This results in

$$T_0 = 1$$

$$T_1 = x$$

$$T_2 = 2x^2 - 1$$

$$T_3 = 4x^3 - 3x$$

$$T_4 = 8x^4 - 8x^2 + 1$$

.

.

.

etc.

(24)

We have just generated the set of Tschebychev polynomials.

Since any T_n can be expressed as the cosine of some angle, it can assume only values between the limits -1 and 1. Thus each T_n can be said to be an equal ripple function, with known extrema.

It should be apparent that any analytic function can be expressed as a series in the T_n 's, this being merely a special case of a Fourier series.

Turning our attention to the parameter $x = \cos \theta$, it is seen that it, too, is subject to the constraint $-1 < x < 1$. For values of the argument outside these limits a Tschebychev series will not converge. Sometimes this difficulty can be overcome by a suitable transformation of variables. For example, suppose it is desired to expand $\ln z$ in a Tschebychev series. The argument is very badly behaved ($0 < z < \infty$), but the transformation

$$x = \frac{z-1}{z+1}$$

suggests itself. Solving for z ,

$$z = \frac{1+x}{1-x}$$

Hence $\ln_e(\frac{1+x}{1-x})$ can be expanded in a Tschebychev series in the variable x .

Theoretically, the number of these transformations is endless, hence there is no unique Tschebychev series for any mathematical function. Thus when stating a Tschebychev expansion, it is necessary to state also the transformation used.

It is useful to be able to express the various powers of x in terms of the Tschebychev polynomials. This is easily done by what I call "half" a binomial expansion. When there is an even number of terms in the quasi-binomial expansion, no manipulation is necessary. But when there is an odd number, the middle term must be halved. This always happens to the term involving T_0 , so that it is convenient to use $T_0/2$ and then employ the binomial coefficients. To illustrate:

$$1 = T_0$$

$$x = T_1$$

$$x^2 = \frac{1}{2}(T_2 + 2\frac{T_0}{2}) = \frac{1}{2}(T_2 + T_0)$$

$$x^3 = \frac{1}{4}(T_3 + 3T_1)$$

$$x^4 = \frac{1}{8}(T_4 + 4T_2 + 6\frac{T_0}{2}) = \frac{1}{8}(T_4 + 4T_2 + 3T_0)$$

$$x^5 = \frac{1}{16}(T_5 + 5T_3 + 10T_1)$$

.

.

$$x^n = \frac{1}{2^{n-1}}(T_n + nT_{n-2} + \frac{n(n-1)}{2}T_{n-4} + \dots) \quad (25)$$

As a check, the sum of the interior coefficients must equal the exterior denominator.

If a given function can be expressed as a power series, and if the argument behaves properly (i.e., its absolute value does not exceed unity), direct substitution of the various expressions for x^n , followed by a collection of terms, results in a Tschebychev series, in which the function is expanded in terms of the successive Tschebychev polynomials, rather than in ascending powers of the argument.

Tschebychev series have two extremely useful properties, which doubtless are responsible for the considerable popularity of the method. They are:

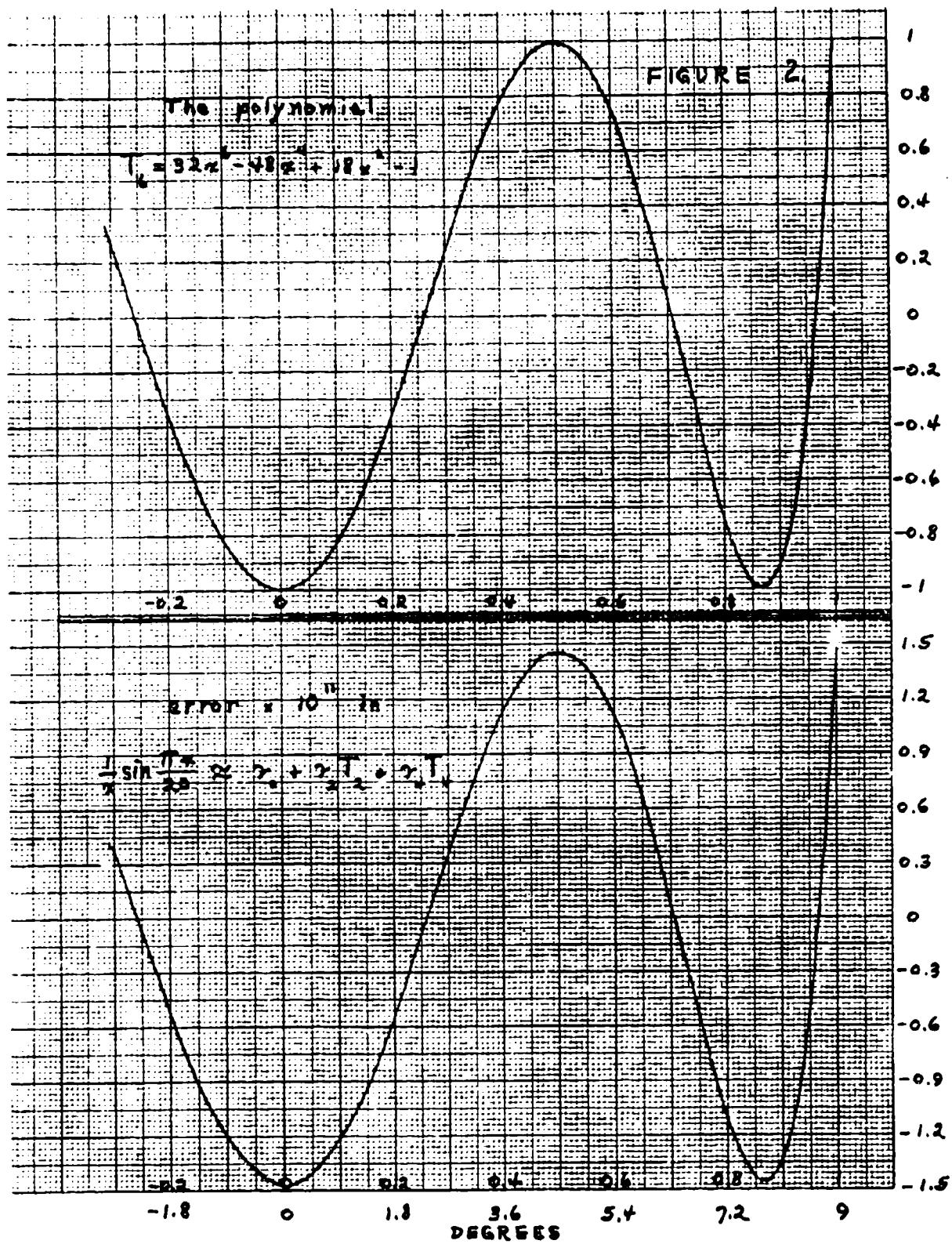
1. A Tschebychev series will converge more rapidly than any other series (given the same function and argument). Thus, the desired accuracy often can be attained with fewer terms.

2. Each Tschebychev polynomial is an equal ripple function (between the argument limits -1 and +1). There will be $n + 1$ of these extrema, interlaced by n zeroes. The value of these extrema, of course, is ± 1 , since $T_n(x) = \cos nx$. It has been proven that, of all polynomials of like degree whose highest-degree terms have the same coefficient, the Tschebychev polynomial has the smallest extrema. In other words the maximum (extremum) is minimized. Hence the term "minimax." Thus, when an approximation is computed from a truncated Tschebychev series, the error is closely given by the coefficient of the first neglected Tschebychev term (see Fig. 2).

A Tschebychev series is essentially a Fourier cosine series, and the coefficients of course can be computed by evaluating the pertinent definite integrals. This can be laborious. Since usually we will be dealing with well-behaved analytic functions, it normally will be better to expand the function in a power series and then develop the Tschebychev series by direct substitution and collection of terms, in the manner already seen. It will be observed that each Tschebychev coefficient is itself the sum of an infinite series. These latter series are easily summed by a programmable calculator, since they involve nothing more complicated than powers of a constant, factorials, and binomial coefficients.

When the function being approximated has a zero in the range (-1, 1), a Tschebychev approximation may behave very badly near this zero. Simply stated, the error in the approximation may exceed the value of the function. Often, this is less a fault than a result of clumsy handling. In most cases, the zero can be removed by altering the function slightly. We shall illustrate by example.

FIGURE 2



We wish to develop a Tschebychev series for $\sin \theta$. Now

$$\sin \theta = \theta - \frac{\theta^3}{15} + \frac{\theta^5}{15} - \frac{\theta^7}{15} + \dots$$

Letting the argument vary from $-\frac{\pi}{4}$ to $\frac{\pi}{4}$, we see that the proper transformation is $\theta = \frac{\pi}{4}x$. Direct substitution yields a series in those Tschebychev polynomials of odd subscript. Truncation, however, produces an approximation which is of little use when $\sin \theta$ is very small. The problem is very easily corrected. Noticing that θ factors the right side,

$$\frac{\sin \theta}{\theta} = 1 - \frac{\theta^2}{15} + \frac{\theta^4}{15} - \frac{\theta^6}{15} + \dots \quad (1)$$

The function $\frac{\sin \theta}{\theta}$ contains no zero within the chosen argument range. In fact, $\frac{\sin \theta}{\theta} > 0.9$.

$$\frac{\sin \theta}{x} = \frac{\pi}{4} \left(1 - \frac{\theta^2}{15} + \frac{\theta^4}{15} - \frac{\theta^6}{15} + \dots \right)$$

does virtually as well and leads to some simplification. It is obvious that, in the last two cases, the two Tschebychev series will involve only those polynomials of even subscript.

It is instructive to show the complete derivation of a Tschebychev series for $\sin \theta$. From the last equation, we have

$$\frac{\sin \frac{\pi}{4} x}{x} = \frac{\pi}{4} - \left(\frac{\pi}{4}\right)^3 \frac{x^2}{15} + \left(\frac{\pi}{4}\right)^5 \frac{x^4}{15} - \left(\frac{\pi}{4}\right)^7 \frac{x^6}{15} + \dots$$

Repeating for convenience

$$x^2 = \frac{1}{2}(T_2 + T_0)$$

$$x^4 = \frac{1}{8}(T_4 + 4T_2 + 3T_0)$$

$$x^6 = \frac{1}{32}(T_6 + 6T_4 + 15T_2 + 10T_0)$$

$$x^8 = \frac{1}{128}(T_8 + 8T_6 + 28T_4 + 56T_2 + 35T_0)$$

$$x^{10} = \frac{1}{512}(T_{10} + 10T_8 + 45T_6 + 120T_4 + 210T_2 + 126T_0)$$

$$x^{12} = \frac{1}{2048}(T_{12} + 12T_{10} + 66T_8 + 220T_6 + 495T_4 + 792T_2 + 462T_0)$$

$$x^{14} = \frac{1}{8192}(T_{14} + 14T_{12} + 91T_{10} + 364T_8 + 1001T_6 + 2002T_4 + 3003T_2 + 1716T_0)$$

$$x^{16} = \frac{1}{32768} (T_{16} + 16T_{14} + 120T_{12} + 560T_{10} + 1820T_8 + 4368T_6 + 8008T_4 + 11440T_2 + 6435T_0) \quad (25)$$

etc.

To shorten the notation, set $p = \frac{\pi}{4}$. Then

$$\frac{\sin px}{x} = p - p^3 \frac{x^2}{13} + p^5 \frac{x^4}{15} - p^7 \frac{x^6}{17} + \dots \quad (26)$$

The desired Tschebychev series is

$$\frac{\sin px}{x} = a_0 T_0 + a_2 T_2 + a_4 T_4 + \dots \quad (27)$$

Substituting and collecting similar terms,

$$a_0 = p - \frac{1}{12} p^3 + \frac{1}{320} p^5 - \frac{1}{16128} p^7 + \dots$$

The general term ψ_n is given by

$$\psi_{2n+1} = \frac{1}{2} \frac{(-1)^n}{2^{2n-1}} \frac{p^{2n+1}}{12n+1} C(2n, n) \quad (28)$$

The term-to-term recurrence ratio is

$$\psi_{2n+1} = -\frac{1}{4} \frac{2n-1}{2n+1} \frac{p^2}{n^2} \psi_{2n-1} \quad (29)$$

$$\begin{aligned}
 & 0.785398 \ 163397 \ 448309 \ 62 \\
 & -0.040372 \ 756094 \ 140390 \ 85 \\
 & 0.000933 \ 897963 \ 822270 \ 06 \\
 & -0.000011 \ 430063 \ 806930 \ 39 \\
 & 0.000000 \ 085684 \ 836846 \ 28 \\
 & -0.000000 \ 000432 \ 447669 \ 64 \\
 & 0.000000 \ 000001 \ 567473 \ 35 \\
 & -0.000000 \ 000000 \ 004275 \ 39 \\
 & 0.000000 \ 000000 \ 000009 \ 09 \\
 & \underline{-0.000000 \ 000000 \ 000000 \ 02} \\
 a_0 & \underline{0.745947 \ 960457 \ 275642 \ 10}
 \end{aligned}$$

In a like fashion,

$$a_2 = -\frac{1}{12} p^3 + \frac{1}{240} p^5 - \frac{1}{10732} p^7 + \frac{1}{829440} p^9 - \dots$$

The general term ψ_n is given by

$$\psi_{2n+1} = \frac{(-1)^n}{2^{2n-1}} \frac{p^{2n+1}}{12n+1} C(2n, n-1)$$

Note the absence of the leading factor $\frac{1}{2}$. That factor applies only to t_0 .

The term-to-term recurrence ratio is

$$\psi_{2n+1} = -\frac{1}{4} \frac{2n-1}{2n+1} \frac{p^2}{n^2-1} \psi_{2n-1}$$

-0.040372 756094 140390 85
0.001245 197285 096360 08
-0.000017 145095 710395 59
0.000000 137095 738954 04
-0.000000 000720 746116 06
0.000000 000002 687097 16
-0.000000 000000 007481 94
0.000000 000000 000016 16
-0.000000 000000 000000 03
-0.039144 567527 081957 02

s_2

Continuing

$$s_4 = \frac{1}{960} p^5 - \frac{1}{26880} p^7 + \frac{1}{1658880} p^9 - \dots$$

The general term ψ_n is given by

$$\psi_{2n+1} = \frac{(-1)^n}{2^{2n-1}} \frac{p^{2n+1}}{12n+1} C(2n, n-2)$$

The term-to-term recurrence ratio is

$$\psi_{2n+1} = -\frac{1}{4} \frac{2n-1}{2n+1} \frac{p^2}{n^2-4} \psi_{2n-1}$$

0.000311 299321 274090 02
-0.000006 858038 284158 23
0.000000 068547 869477 02
-0.000000 000411 854923 46
0.000000 000001 679435 73
-0.000000 000000 004987 96
0.000000 000000 000011 31
-0.000000 000000 000000 02
0.000304 509420 678944 41

s_4

For any coefficient a_{2k} ($k = 1, 2, 3, \dots$)

$$\psi_{2n+1} = \frac{(-1)^n}{2^{2n-1}} \frac{p^{2n+1}}{12n+1} C(2n, n-k) \quad (30)$$

The first term is obtained by setting $n = k$. The term-to-term recurrence ratio is

$$\psi_{2n+1} = -\frac{1}{4} \cdot \frac{2n-1}{2n+1} \cdot \frac{p^2}{n^2 - k^2} \psi_{2n-1} \quad (31)$$

To complete the example, the rest of the calculations follow:

	-0.000001 143006 380693 04
	0.000000 019585 105564 86
	-0.000000 000154 445596 30
	0.000000 000000 746415 88
	-0.000000 000000 002493 98
	0.000000 000000 000006 17
	<u>-0.000000 000000 000000 01</u>
a_6	-0.000001 123574 976796 42
	0.000000 002448 138195 61
	-0.000000 000034 321243 62
	0.000000 000000 223924 76
	-0.000000 000000 000906 90
	0.000000 000000 000002 57
	<u>-0.000000 000000 000000 01</u>
a_8	0.000000 002414 039972 41
	-0.000000 000003 432124 36
	0.000000 000000 040713 59
	-0.000000 000000 000226 73
	<u>0.000000 000000 000000 79</u>
a_{10}	-0.000000 000003 391636 71
	0.000000 000000 003392 80
	-0.000000 000000 000034 88
	<u>0.000000 000000 000000 17</u>
a_{12}	0.000000 000000 003358 09
	-0.000000 000000 000002 49
	<u>0.000000 000000 000000 02</u>
a_{14}	-0.000000 000000 000002 47
	$a_{16} \approx 1.4 \times 10^{-21}$

As a simple yet powerful check on the calculations, the sum of the coefficients should equal the value of the function at $x = 1$. (When $x = 1$, every Tschebychev polynomial also equals unity.) In the present instance

$$\frac{\sin \pi x}{x} = \frac{1}{2} \sqrt{2} \quad (x = 1)$$

When employing a Tschebychev series in the basic form, the values of the successive Tschebychev polynomials corresponding to the stated value of the argument are easily obtained by repeatedly applying the proper trigonometric identity (three-term recurrence relationship).

When a Tschebychev series has been truncated for use, the maximum error in the approximation is given (nearly) by the first neglected coefficient. The equal ripple feature distributes this maximum error (small though it be) throughout the argument range. It is obvious, then, that restricting the range of the argument will not reduce the maximum error. Again, this may be less a fault than a faux pas.

Consider what happens when the principles of range restriction are applied before selecting a transformation.

Continuing to use the sine function as an example, suppose we let $\theta = 5\phi$. Computing $\sin \phi$, we recover the wanted function by means of the identity

$$\frac{\sin 5\phi}{\sin \phi} = 5 - 4 \sin^2 \phi (5 - 4 \sin^2 \phi) \quad (7)$$

But ϕ need not exceed 9 degrees. Hence we can develop a Tschebychev series for the transformation

$$p = \frac{\pi}{20}$$

All formulae remain the same. It is only necessary to substitute the new value of p .

	0.157079 632679 489661 92
	-0.000322 982048 753123 13
	0.000000 298847 348423 13
	-0.000000 000146 304816 73
	0.000000 000000 043870 64
	<u>-0.000000 000000 000008 86</u>
a_0	0.156756 949331 824006 98
	-0.000322 982048 753123 13
	0.000000 398463 131230 84
	-0.000000 000219 457225 09
	0.000000 000000 070193 02
	<u>-0.000000 000000 000014 76</u>
a_2	-0.000322 583805 008939 13

	0.000000	099615	782807	71
	-0.000000	000087	782890	04
	0.000000	000000	035096	51
	<u>-0.000000</u>	<u>000000</u>	<u>000008</u>	<u>43</u>
a_4	0.000000	099528	035005	75
	-0.000000	000014	630481	67
	0.000000	000000	010027	57
	<u>-0.000000</u>	<u>000000</u>	<u>000003</u>	<u>16</u>
a_6	-0.000000	000014	620457	26
	0.000000	000000	001253	45
	<u>-0.000000</u>	<u>000000</u>	<u>000000</u>	<u>70</u>
a_8	0.000000	000000	001252	74

$$a_{10} = -7.0 \times 10^{-20}$$

The resulting economy is obvious.

A Tschebychev series is rarely encountered in its basic form. The computational power and efficiency of a simple polynomial in nested form is so great that one cannot gainsay its use. Hence most Tschebychev series, after truncation, are converted to this form by substituting for the T_n 's and collecting terms.

We illustrate, using

$$\frac{1}{x} \sin \frac{\pi x}{20} = a_0 T_0 + a_2 T_2 + a_4 T_4$$

Thus

$$\begin{aligned} \frac{\sin \frac{\pi x}{20}}{x} &= a_0 + a_2(2x^2 - 1) + a_4(8x^4 - 8x^2 + 1) \\ &= (a_0 - a_2 + a_4) + (2a_2 - 8a_4)x^2 + 8a_4x^4 \end{aligned}$$

Simplifying,

$$\sin \frac{\pi x}{20} = x(0.157079 632665 - 0.000645 963834 x^2 + 0.000000 796224 x^4)$$

Maximum error occurs at $x = 1$ and is less than 1.5×10^{-11} , exactly as predicted by a_6 . The error of this approximation to $\frac{1}{x} \sin \frac{\pi x}{20}$ is shown in Figure 2.

Note that when a Tschebychev approximation has been reconverted to a simple polynomial form, there is no practical way to estimate the maximum error by inspecting the resulting coefficients.

IX. MAEHLY'S METHOD. A method attributed to Maehtly develops a rational function approximation from a Tachebychev series in much the same way that a Padé approximation is developed from a simple power series. The basic form is like

$$\frac{a_0 + a_1 T_1 + a_2 T_2 + a_3 T_3}{1 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3} = Y_0 + Y_1 T_1 + Y_2 T_2 + \dots \quad (32)$$

Since the coefficients of a rational function are not unique, we can arbitrarily choose any one of them, and so set $\beta_0 = 1$.

As before, both sides are multiplied by the Maehtly denominator and like terms collected on the right. The multiplication requires use of the trigonometric identity

$$T_n T_m = \frac{1}{2}(T_{n-m} + T_{n+m}) \quad (m < n) \quad (33)$$

It will be noticed that this uses terms of the basic Tachebychev series out to a degree equal to the sum of the degree of the Maehtly numerator and twice the degree of the denominator.

The computations, if systematized, are less complicated than at first appears. If we designate the product-series as

$$\zeta_0 T_0 + \zeta_1 T_1 + \zeta_2 T_2 + \zeta_3 T_3 + \dots \quad (34)$$

then

$$\begin{aligned} \zeta_0 &= Y_0 + \frac{1}{2}(\beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \dots) \\ \zeta_1 &= Y_1 + \frac{1}{2}\beta_1(2Y_0 + Y_2) + \frac{1}{2}\beta_2(Y_1 + Y_3) + \frac{1}{2}\beta_3(Y_2 + Y_4) \\ &\quad + \frac{1}{2}\beta_4(Y_3 + Y_5) + \frac{1}{2}\beta_5(Y_4 + Y_6) + \dots \\ \zeta_2 &= Y_2 + \frac{1}{2}\beta_1(Y_1 + Y_3) + \frac{1}{2}\beta_2(2Y_0 + Y_4) + \frac{1}{2}\beta_3(Y_1 + Y_5) \\ &\quad + \frac{1}{2}\beta_4(Y_2 + Y_6) + \frac{1}{2}\beta_5(Y_3 + Y_7) + \dots \\ \zeta_3 &= Y_3 + \frac{1}{2}\beta_1(Y_2 + Y_4) + \frac{1}{2}\beta_2(Y_1 + Y_5) + \frac{1}{2}\beta_3(2Y_0 + Y_6) \\ &\quad + \frac{1}{2}\beta_4(Y_1 + Y_7) + \frac{1}{2}\beta_5(Y_2 + Y_8) + \dots \\ \zeta_4 &= Y_4 + \frac{1}{2}\beta_1(Y_3 + Y_5) + \frac{1}{2}\beta_2(Y_2 + Y_6) + \frac{1}{2}\beta_3(Y_1 + Y_7) \\ &\quad + \frac{1}{2}\beta_4(2Y_0 + Y_8) + \frac{1}{2}\beta_5(Y_1 + Y_9) + \dots \end{aligned} \quad (35)$$

etc.

Within each set of parentheses, the subscripts of the γ_i are given by the sum and difference of the ζ_i and β_i subscripts. (If a negative value occurs, simply use the absolute value.) When the subscripts of ζ_i and β_i are equal, the coefficient γ_0 will appear and must be doubled.

Since any reasonable approximation will contain a finite number of β_i 's, each of the expressions for ζ_i will terminate. Notice that within each set of parentheses the subscripts are either odd or even in pairs. This means that for odd or even functions, half the parenthetical terms will vanish.

It is now possible to develop a set of $m + n + 1$ simultaneous linear equations in the same number of unknowns (the α_i 's and β_i 's). That the technique is adapted from the Padé method is obvious. The wanted equations are:

$$\begin{aligned}\alpha_0 &= \zeta_0 \\ \alpha_1 &= \zeta_1 \\ \alpha_2 &= \zeta_2 \\ \alpha_3 &= \zeta_3 \\ 0 &= \zeta_4 \\ 0 &= \zeta_5 \\ 0 &= \zeta_6\end{aligned}$$

for the example form given.

Exactly as with the Padé, those equations whose left member is zero are solved for the β_i 's, after which the α_i 's are found by direct substitution.

The error curve of a Maehly strongly resembles that of a Tschebychev approximation. In fact, it is usually possible to select a Maehly and a Tschebychev so that the error curves have the same number of ripples, similarly spaced. Under these conditions, a linear combination of the two approximations can achieve fantastic accuracy.

Let us illustrate the method, using the Tschebychev series for $\frac{\sin px}{x}$; $p = \frac{\pi}{20}$. The Tschebychev coefficients, previously computed, are repeated for convenience:

$$\begin{aligned}\gamma_0 &= 0.156756 949331 824007 \\ \gamma_2 &= -0.000322 583805 008939 \\ \gamma_4 &= 0.000000 099528 035006 \\ \gamma_6 &= -0.000000 000014 620457 \\ \gamma_8 &= 0.000000 000000 001253\end{aligned}$$

We choose to develop a Maehly of the form

$$\frac{\sin px}{x} = \frac{a_0 + a_2 T_2}{1 + \beta_2 T_2} = M$$

The required equations are:

$$a_0 = \zeta_0 = \gamma_0 + \frac{1}{2} \beta_2 \gamma_2$$

$$a_2 = \zeta_2 = \gamma_2 + \frac{1}{2} \beta_2 (2\gamma_0 + \gamma_4)$$

$$0 = \zeta_4 = \gamma_4 + \frac{1}{2} \beta_2 (\gamma_2 + \gamma_6)$$

The last equation yields β_2 immediately.

$$\beta_2 = 0.000617 067744 563840$$

$$a_0 = 0.156756 849803 793512$$

$$a_2 = -0.000225 854117 132272$$

Now a_2 and β_2 are coefficients of $T_2 = 2x^2 - 1$.

Hence

$$M = \frac{(a_0 - a_2) + 2a_2 x^2}{(1 - \beta_2) + 2\beta_2 x^2}$$

The coefficients of a rational function are never unique. This allows us arbitrarily to set any one of them. Let us choose unity as the coefficient of x^2 in the denominator. Thus, dividing through by $2\beta_2$,

$$M = \frac{127.200542 6502 - 0.366011 8668 x^2}{809.783804 9871 + x^2}$$

If desired, a partial division will produce a form more suitable for subsequent linear combination. Thus

$$M = 0.157079 632695 59 - \frac{0.523091 4995 x^2}{809.783804 9871 + x^2}$$

or, better still,

$$M = -0.366011 866803 88 + \frac{423.591024 8211}{809.783804 9871 + x^2}$$

We wish to compare this approximation to the Tschebychev approximation

$$T = \gamma_0 + \gamma_2 T_2 + \gamma_4 T_4$$

which, expressed in terms of x^2 , is

$$\frac{\sin px}{x} = T = (Y_0 - Y_2 + Y_4) + (2Y_2 - 8Y_4)x^2 + 8Y_4x^4$$

$$Y_0 - Y_2 + Y_4 = 0.157079\ 632664\ 87$$

$$2Y_2 - 8Y_4 = -0.000645\ 963834\ 298$$

$$8Y_4 = 0.000000\ 796224\ 2800$$

An inspection of the error plots (Figure 3) suggests the linear combination $0.524 T + 0.476 M$. Performing the multiplication and combining the resulting constant terms, we get

$$\begin{aligned}\frac{\sin px}{x} = & -0.091911\ 921082\ 257 \\& -0.000338\ 485049\ 17211 x^2 \\& +0.000000\ 417221\ 522744\ 10 x^4 \\& + \frac{201.629327\ 81484}{809.783804\ 98711 + x^2}\end{aligned}$$

an approximation so accurate that a thirteen-digit calculator could detect no error.

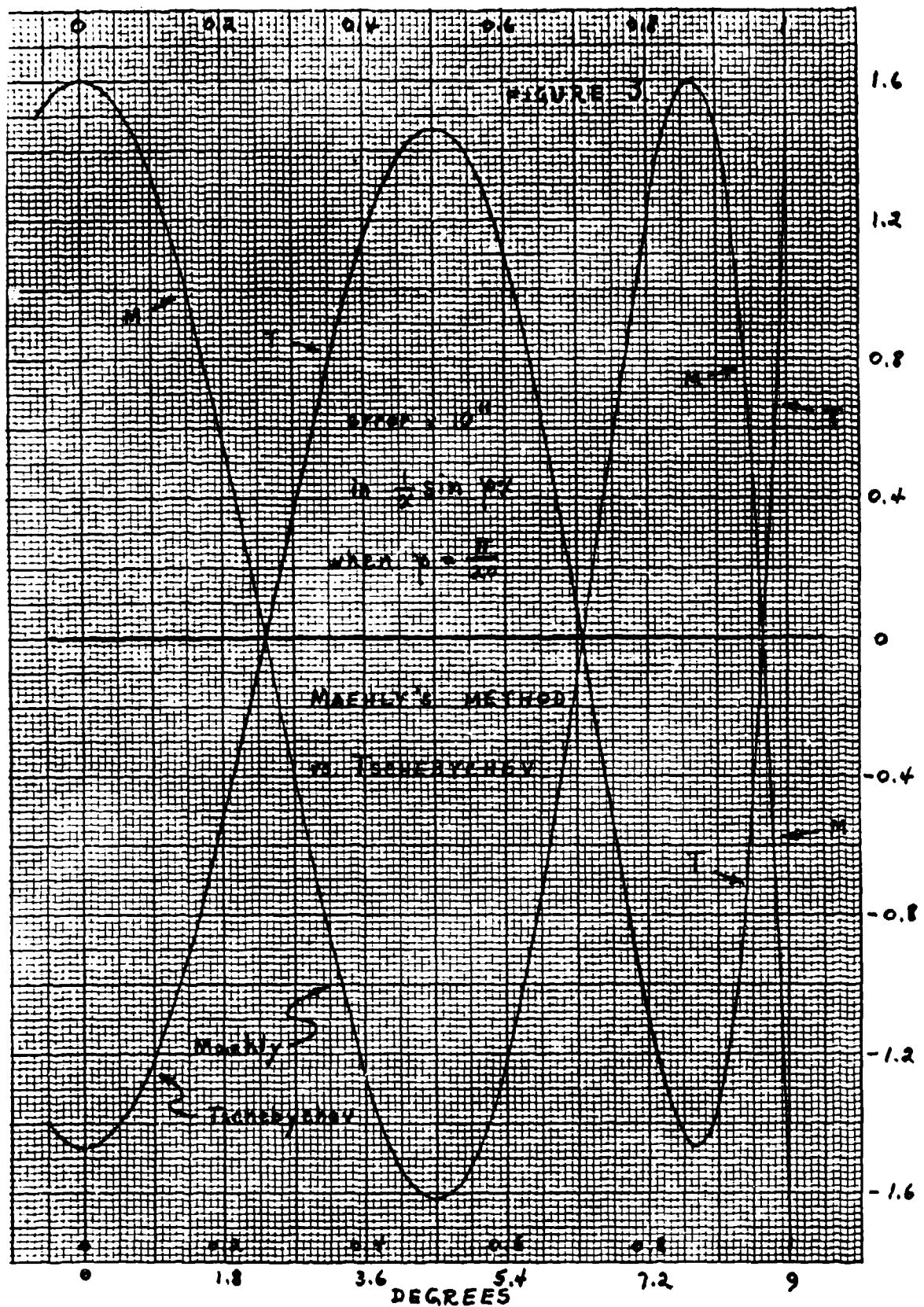
Unfortunately, the functions which are encountered in real life rarely are as well-behaved as is the common sine. But even though at first glance they may look solution-proof, sometimes a little guile will go a long way. For an example, let us return to that recalcitrant function, the inverse tangent. Previously, we have seen that if $z = \tan \theta$,

$$\frac{\arctan z}{z} = 1 - \frac{1}{3}z^2 + \frac{1}{5}z^4 - \frac{1}{7}z^6 + \dots \quad (9)$$

We can restrict our interest to $|z| < \frac{\pi}{4}$, whence $|z| < 1$, but z is not a suitable Tschebychev variable. It is easy to see that the ratio of successive coefficients of the Maclaurin series tends toward the limit -1 , making it useless for computing Tschebychev coefficients. Moreover, the Tschebychev series (for the transformation $x = z$) converges rather slowly, requiring a large number of terms for any stated accuracy. Altogether, a seemingly formidable task, with a disappointingly cumbersome solution.

Let us employ some of the tools we have acquired. First, let

$$y = \frac{z}{1 + \sqrt{1 + z^2}} \quad (12)$$



Then $\arctan z = 2 \arctan y$.

$$\text{But if } |z| < 1, \text{ then } |y| < \frac{1}{1 + \sqrt{2}} = \sqrt{2} - 1$$

We make the parametric substitution $y = px$ and set $x = 1$ when $y = \sqrt{2} - 1$.

Thus, $p = \sqrt{2} - 1$. This is a particularly handy transformation, since $p^{-1} = 2 + p$ and $-p^2 = 2p - 1$. The solution then, is

$$\arctan z = 2 \arctan px$$

We will, however, compute the Tschebychev coefficients for

$$\frac{\arctan px}{x} = \gamma_0 + \gamma_2 T_2 + \gamma_4 T_4 + \dots \quad (36)$$

in order to avoid a zero within the stated argument range.

Now p is a constant ($-p^2 = -0.17157 \dots$), and hence the various powers of p in

$$\frac{\arctan px}{x} = p - \frac{p^3}{3} x^2 + \frac{p^5}{5} x^4 - \frac{p^7}{7} x^6 + \dots \quad (37)$$

become part of the coefficients. Convergence is quite rapid, and the computations are easily made on a programmable calculator. The resulting Tschebychev series is:

$$\frac{\arctan px}{x} = \gamma_0 + \gamma_2 T_2 + \gamma_4 T_4 + \dots \quad (36)$$

$$x = \frac{z(2+p)}{1 + \sqrt{1+z^2}} \quad p = \sqrt{2} - 1$$

$$\begin{aligned}\gamma_0 &= 0.403199\ 719161\ 511495\ 80 \\ \gamma_2 &= -0.010749\ 968804\ 390963\ 96 \\ \gamma_4 &= 0.000256\ 378716\ 684566\ 71 \\ \gamma_6 &= -0.000007\ 264267\ 589573\ 12 \\ \gamma_8 &= 0.000000\ 223914\ 266710\ 62 \\ \gamma_{10} &= -0.000000\ 007256\ 851307\ 14 \\ \gamma_{12} &= 0.000000\ 000243\ 155037\ 30 \\ \gamma_{14} &= -0.000000\ 000008\ 343268\ 50 \\ \gamma_{16} &= 0.000000\ 000000\ 291421\ 29 \\ \gamma_{18} &= -0.000000\ 000000\ 010320\ 90\end{aligned}$$

$$\begin{aligned}
 Y_{20} &= 0.000000 000000 000369 59 \\
 Y_{22} &= -0.000000 000000 000013 36 \\
 Y_{24} &= 4.86 \times 10^{-19} \quad Y_{26} = -1.78 \times 10^{-20}
 \end{aligned}$$

From these coefficients, let us compute the simplest possible Maehly:

$$M = \frac{\alpha_0}{1 + \beta_2 T_2}$$

$$\alpha_0 = \zeta_0 = Y_0 + \frac{1}{2}\beta_2 Y_2$$

$$0 = \zeta_2 = Y_2 + \frac{1}{2}\beta_2 (2Y_0 + Y_4)$$

$$\beta_2 = \frac{-Y_2}{Y_0 + \frac{1}{2}Y_4}$$

$$\beta_2 = 0.026653 173701 372558 84$$

$$\alpha_0 = 0.403056 458768 597611 48$$

Thus,

$$M = \frac{\alpha_0}{1 + \beta_2(2x^2 - 1)} = \frac{\alpha_0}{(1 - \beta_2) + 2\beta_2 x^2}$$

Dividing through by $2\beta_2$,

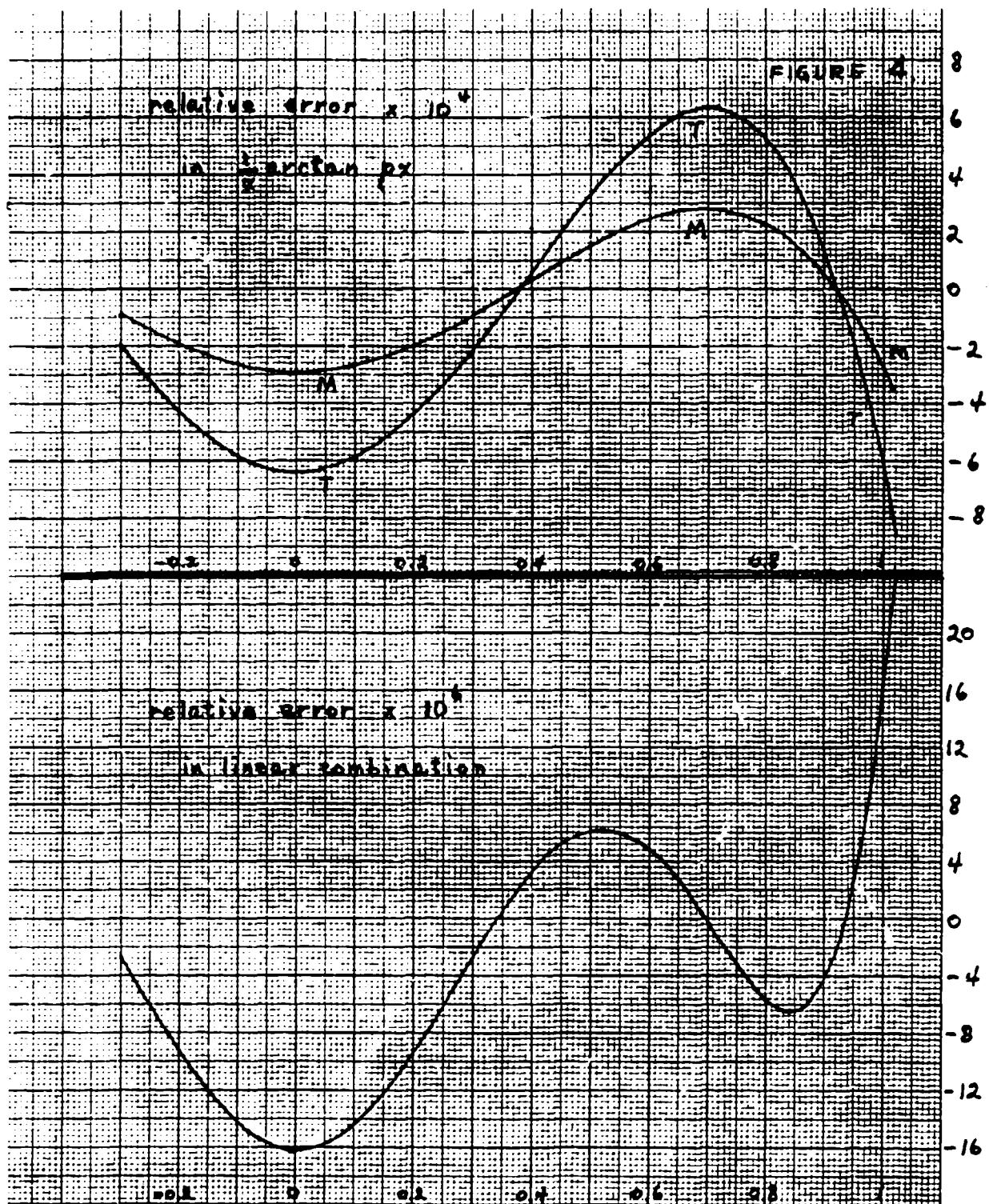
$$\frac{\arctan px}{x} = \frac{7.561134 431579}{18.259492 044066 + x^2}$$

The relative error in this approximation does not exceed 2.9×10^{-4} . Figure 4 compares this error with that of the simple Tschebychev approximation $T = Y_0 + Y_2 T_2$. Also shown is the error curve for the linear combination

$$1.78981 M - 0.78981 T.$$

This error curve (for the linear combination) is shown on an expanded scale in the lower graph, where it clearly reveals the shape of the SECOND missing Tschebychev polynomial. It can be shown that the linear combination method produces an approximation virtually as accurate as the Tschebychev (or Maehly) of next higher order.

FIGURE 2



X. ASYMPTOTIC SERIES --- A LOOK AHEAD. The author presently is investigating the theory of asymptotic series. First results are very promising and may warrant a future paper. Some tentative findings are:

1. To be useful, an asymptotic series must be an alternating series.

2. Provided the value of the argument is not too small, there will be a smallest term which is not the first term. "Smallest" is taken in the sense of the absolute value. There may be two consecutive smallest terms (equal in absolute value but opposite in sign, of course).

3. After the smallest term, the series diverges. None of these divergent terms should be included in any approximation.

4. Truncation immediately after the smallest term produces an approximation which errs less than that produced by truncation at any other point.

There is a widespread --- but mistaken --- belief that this represents the best approximation of which the asymptotic series is capable. It is simply not true, as we shall see immediately when we pursue the New Look.

5. If, in place of the first divergent term, there is substituted a term of like sign whose absolute value is exactly half the absolute value of the smallest term, there is a definite reduction in the error of the approximation, often by more than an order of magnitude.

6. If the remaining error is plotted as a function of the argument, it is seen to be a "sawtooth." See [9] for an example. This function appears to possess continuous derivatives. If so, it is a sufficient condition for the existence of an exact analytical expression (e.g., a Fourier or Tschebychev series).

7. Discovery of such an expression is the next logical step. (Only an approximation was developed in [9].)

8. The composite expression should converge for all values of the argument down to include the point where the sum of the first two terms of the asymptotic series is zero, thereby reducing the total expression at that point to something like $f(z) = \frac{1}{2} + \epsilon(z)$. For two commonly used asymptotic series, this minimum value of the argument may turn out to be:

$$\text{for } \ln_e \Gamma(z), z = \frac{1}{\sqrt{3}} \approx 0.289 \text{ ([2] and [8])}.$$

$$\text{for erfc } z, z = \frac{1}{\sqrt{2}} \approx 0.707 \text{ ([9])}.$$

APPENDIX A
AN ALGORITHM FOR SQUARE ROOT

Before actually developing an approximation, let us address a few ancillary matters.

First. The Newton-Raphson technique. If r_1 is any reasonable estimate of \sqrt{z} , then

$$r_{i+1} = \frac{1}{2} [r_i + \frac{z}{r_i}]$$

is a better one.

The process can be repeated endlessly and is found to converge quadratically upon the true value. Examining the form of the iterative equation suggests that the first estimate be in the form of a rational function approximation, since if $r_1 = N/D$, then

$$r_2 = \frac{1}{2} \left[\frac{N}{D} + \frac{zD}{N} \right] = \frac{N^2 + zD^2}{2DN}, \quad (38)$$

thereby saving a division. In fact

$$\begin{aligned} r_3 &= \frac{1}{2} \left[\frac{N^2 + zD^2}{2DN} + \frac{2zDN}{N^2 + zD^2} \right] \\ &= \frac{1}{2} \left[\frac{(N^2 + zD^2)^2 + 4zD^2N^2}{2DN(N^2 + zD^2)} \right] \\ &= \frac{N^4 + 6zD^2N^2 + z^2D^4}{4DN(N^2 + zD^2)} \end{aligned} \quad (39)$$

and still another division is saved.

Second. The bilinear transformation. Simply stated,

$$wz + aw + bz + c = 0$$

w and z are variables, a, b and c are coefficients, any of which might be complex. Should the term in wz appear with a coefficient other than unity, division by that coefficient will produce the above form without loss of generality. The expression can be regarded as linear in w or as linear in z, but is NOT linear in both together; hence, the term "bilinear." It is easy to solve for either variable in terms of the other, viz.,

$$wz + aw = -bz - c$$

$$w = \frac{-bz - c}{z + a} \quad (41)$$

Two (or more) successive bilinear transformations can always be replaced by a SINGLE bilinear transformation (with, of course, different coefficients). Let us examine the bilinear transformation

$$w = \frac{z - 1}{z + 1}$$

Solving for z ,

$$z = \frac{1 + w}{1 - w} \quad (42)$$

As z varies through the argument range 0 to ∞ , w varies from -1 to +1, suggesting that a transformation of this or a similar form ($b = -1$, $c = a$) will be useful for square roots, logarithms, or any function whose argument must be non-negative.

Third. Catering to very large argument ranges. Superficially, it would seem that the bilinear transformation would be enough. But we find severe warping near the band edges (as $w \rightarrow \pm 1$ in the above example). This introduces an acute scaling problem similar to that of the tangent function near 90 degrees. Clearly, an additional device is needed.

A possible approach is to develop several approximations, each for use with a different stated argument range. Since this method would seem to require more perseverance than ingenuity, it will not be further pursued in this paper.

Instead, we shall define a process and call it "normalization." Basically, it amounts to a transformation which separates a floating point number into exponent and mantissa. The floating point number can be expressed to any convenient base. The approximation is applied to the mantissa, after which a suitable inverse transformation recovers the desired result. For the square root, this amounts to dividing by any arbitrarily chosen perfect square, computing the approximation, then multiplying by the perfect root.

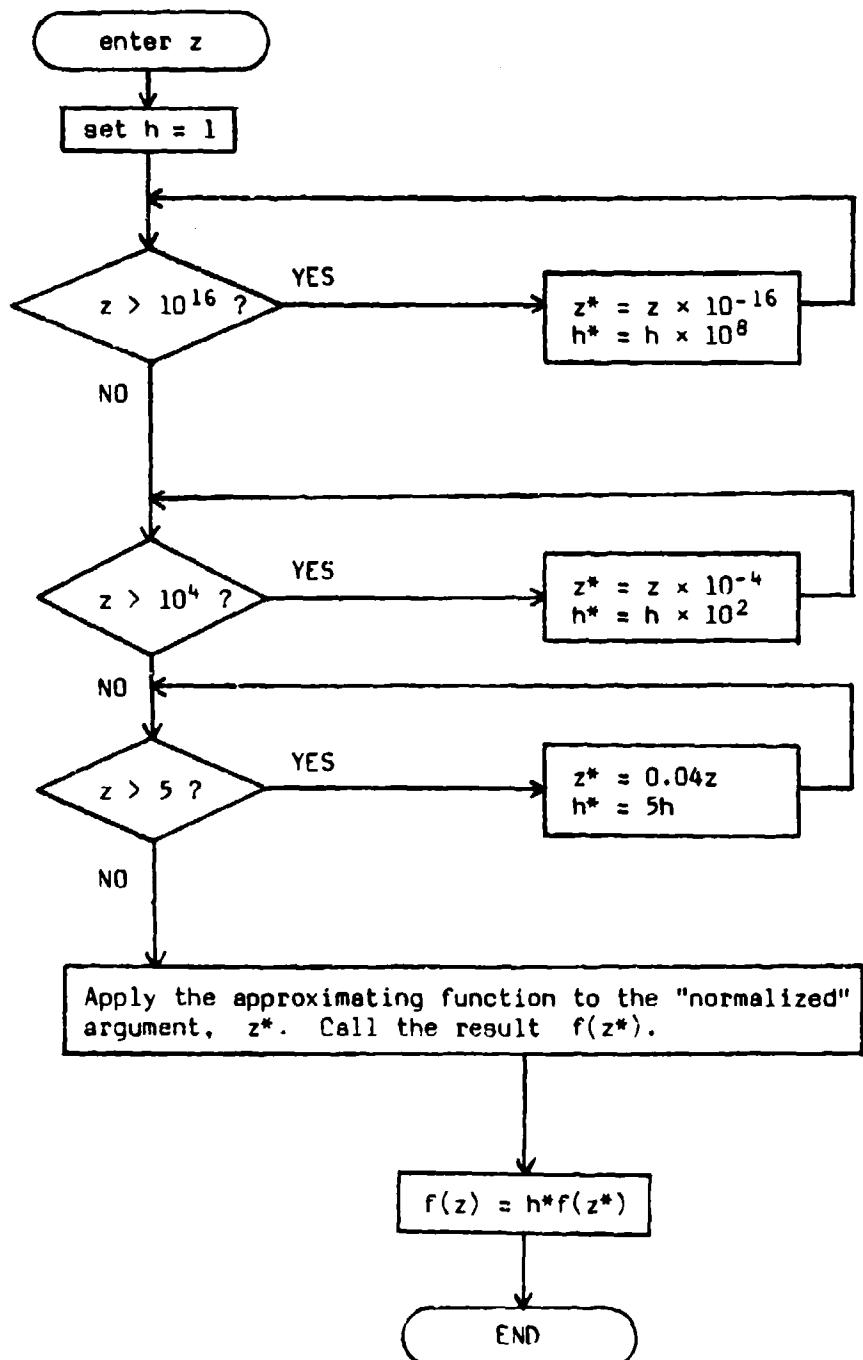
$$\sqrt{z} = k \sqrt{\frac{z}{k^2}}$$

A suggested algorithm follows.

Consider separately $0 < z < 1$ and $1 < z < \infty$.

Denote the latest transform by $*$.

Case 1. $1 \leq z < \infty$



Case 2. $0 < z < 1$.

Either apply the "normalization" algorithm to the reciprocal, $\frac{1}{z}$, or replace all the constants within the algorithm by their reciprocals (and $>$ by $<$).

The resulting normalized argument will always lie in the range $0.2 < z^* < 5$. This suggests the bilinear transformation

$$px = \frac{z^* - 1}{z^* + 1}$$

Setting $x = 1$ when $z^* = 5$, we find that $p = \frac{2}{3}$. If we can find a power series for the square root function, it should be easy to compute the Tschebychev coefficients for this transformation, and hence the Mashly rational function.

Now

$$z^* = \frac{1 + px}{1 - px} = \frac{(1 + px)(1 - px)}{(1 - px)^2} \quad (44)$$

Thus

$$(1 - px) \sqrt{z^*} = (1 - p^2x^2)^{1/2} \quad (45)$$

and the right side can be expanded by means of the binomial theorem, yielding

$$\begin{aligned} \sqrt{1 - p^2x^2} &= 1 - \frac{1}{2} p^2x^2 - \frac{1}{8} p^4x^4 - \frac{1}{16} p^6x^6 - \frac{5}{128} p^8x^8 \\ &\quad - \frac{7}{256} p^{10}x^{10} - \frac{21}{1024} p^{12}x^{12} - \frac{33}{2048} p^{14}x^{14} - \dots \end{aligned}$$

The general term (after the first) is

$$-\frac{1}{2^{2n-1}} \frac{p^{2n}x^{2n}}{\underline{1}_n \underline{n-1}} \quad (47)$$

and the recurrence ratio is $\frac{2n-3}{2n} p^2x^2$. Inserting the value $p = \frac{2}{3}$, the general term becomes $-\frac{2}{3^{2n}} \frac{\underline{1}_{2n-2} x^{2n}}{\underline{1}_n \underline{n-1}}$ and the recurrence ratio $\frac{4n-6}{9n} x^2$.

The series for computing the coefficients of the Tschebychev series require the additional multipliers (for the general term) of

$$\frac{\underline{1}_{2n}}{2^{2n-1} \underline{1}_n \underline{n-1} \underline{n+a}}$$

where γ_{2n} is the Tschebychev coefficient being computed and ψ_{2n} is a term

of the series used to compute it. Temporarily skipping over the problem of computing γ_0 , we find that the first non-zero term in the series for γ_2 ($a = 1, 2, 3, \dots$) occurs when $n = a$. The expression for it simplifies to

$$\frac{-4 \ln 2n-2}{8^{2n} \ln \ln 1}$$

The recurrence ratio is $\frac{(2n-3)(2n-1)}{9(n^2 - a^2)}$.

It is found that the second term of the series for γ_0 is identical with the first term of that for γ_2 . The recurrence ratio is the same. Only the value of a is different. Wherefore we compute this term, reset a to zero, and sum the resulting series --- yielding, of course, $\gamma_0 = 1$.

Thus we have the Tschebychev series

$$(1 - px) \sqrt{z^*} = \gamma_0 + \gamma_2 T_2 + \gamma_4 T_4 + \dots \quad (48)$$

where

$$px = \frac{z^* - 1}{z^* + 1} \quad p = \frac{2}{3}$$

$$\begin{aligned}
 \gamma_0 &= 0.877328 215244 7546 \\
 \gamma_2 &= -0.126982 320508 2891 \\
 \gamma_4 &= -0.004619 211325 3075 \\
 \gamma_6 &= -0.000336 513800 0460 \\
 \gamma_8 &= -0.000030 660625 1123 \\
 \gamma_{10} &= -0.000003 129636 9144 \\
 \gamma_{12} &= -0.000000 342323 7095 \\
 \gamma_{14} &= -0.000000 039230 6243 \\
 \gamma_{16} &= -0.000000 004649 4341 \\
 \gamma_{18} &= -0.000000 000565 1841 \\
 \gamma_{20} &= -0.000000 000070 0800 \\
 \gamma_{22} &= -0.000000 000008 8292 \\
 \gamma_{24} &= -0.000000 000001 1270 \\
 \gamma_{26} &= -0.000000 000000 1454 \\
 \gamma_{28} &= -0.000000 000000 0189 \\
 \gamma_{30} &\approx -0.000000 000000 0025 \\
 \gamma_{32} &\approx -0.000000 000000 0003
 \end{aligned}$$

It seems possible that a Maehly as simple as

$$(1 - px) \sqrt{z^2} = \frac{\alpha_0 + \alpha_2 T_2}{1 + \beta_2 T_2} = \frac{N}{D}$$

may be sufficient for our needs. Before computing the Maehly coefficients, however, let us take notice of how the bilinear transformation we have used simplifies the application of the Newton-Raphson technique. Of course

$$r_1 = \frac{N}{D(1 - px)}$$

Then

$$r_2 = \frac{1}{2} \left[\frac{N}{D(1 - px)} + \frac{(1 + px)}{(1 - px)} \cdot \frac{D(1 - px)}{N} \right]$$

$$r_2 = \frac{N}{2D(1 - px)} + \frac{D(1 + px)}{2N}$$

which combines to

$$r_2 = \frac{N^2 + D^2(1 - p^2x^2)}{2DN(1 - px)}$$

In similar fashion,

$$r_3 = \frac{N^2 + D^2(1 - p^2x^2)}{4DN(1 - px)} + \frac{DN(1 + px)}{N^2 + D^2(1 - p^2x^2)} \quad (49)$$

which states the final solution in terms of the original Maehly approximation. The Maehly coefficients are:

$$\alpha_0 = 0.881935 217627$$

$$\alpha_2 = -0.190474 825654$$

$$\beta_2 = -0.072561 319783$$

Since the coefficients of a rational function are not unique, neither are N and D --- only their ratio is. Thus there are a limitless number of forms in which they can be expressed. Perhaps the simplest is

$$(\frac{3}{2} - x) \sqrt{z^2} = \frac{11.08452 - 3.93753 x^2}{7.39072 - x^2}$$

In final form

$$\sqrt{z^*} = \frac{N^2 + D^2 (\frac{9}{4} - x^2)}{4DN(\frac{3}{2} - x)} + \frac{DN(\frac{3}{2} + x)}{N^2 + D^2 (\frac{9}{4} - x^2)}$$

An alternate form:

$$\text{Let } u = \frac{D}{N}(\frac{3}{2} - x) \text{ and } v = \frac{D}{N}(\frac{3}{2} + x)$$

$$\text{Then } \sqrt{z^*} = \frac{1 + uv}{4u} + \frac{v}{1 + uv}$$

The relative error of this approximation at $x = 1$ ($z^* = 5$) is less than 10^{-15} . We use this relative error to estimate the precision with which we must state our Maehly coefficients. Twice taking the square root (to remove the Newton-Raphson effect), we multiply the result by the least coefficient.

$$3.93753 \times 10^{-15/4} = 7 \times 10^{-4}$$

This result suggests inclusion of the fourth digit after the decimal point. The fifth is shown, but is superfluous. In fact, expressing the approximation as

$$(\frac{3}{2} - x) \sqrt{z^*} = \frac{11.08 - 3.94x^2}{7.39 - x^2} = \frac{N}{D}$$

results in a relative error ($0.2 < z^* < 5$) of no worse than 3×10^{-12} .

It is interesting to note what happens when the Maehly rational function is extended to

$$(1 - px) \sqrt{z^*} = \frac{\alpha_0 + \alpha_2 T_2 + \alpha_4 T_4}{1 + \beta_2 T_2 + \beta_4 T_4} = \frac{N}{D}$$

after which the Newton-Raphson technique is applied only once

$$\alpha_0 = \xi_0 = \gamma_0 + \frac{1}{2}\beta_2\gamma_2 + \frac{1}{2}\beta_4\gamma_4$$

$$\alpha_2 = \xi_2 = \gamma_2 + \frac{1}{2}\beta_2(2\gamma_0 + \gamma_4) + \frac{1}{2}\beta_4(\gamma_2 + \gamma_6)$$

$$\alpha_4 = \xi_4 = \gamma_4 + \frac{1}{2}\beta_2(\gamma_2 + \gamma_6) + \frac{1}{2}\beta_4(2\gamma_0 + \gamma_8)$$

$$0 = \zeta_6 = Y_6 + \frac{1}{2}\beta_2(Y_4 + Y_8) + \frac{1}{2}\beta_4(Y_2 + Y_{10})$$

$$0 = \zeta_8 = Y_8 + \frac{1}{2}\beta_2(Y_6 + Y_{10}) + \frac{1}{2}\beta_4(Y_4 + Y_{12})$$

Computing,

$$\alpha_0 = 0.891040 789794$$

$$\alpha_2 = -0.316214 857627$$

$$\alpha_4 = 0.011427 219988$$

$$\beta_2 = -0.216071 134840$$

$$\beta_4 = 0.002611 917418$$

Simplification results in

$$\left(\frac{3}{2} - x\right) \sqrt{z^*} = \frac{87.4847865 - 51.9623632x^2 + 6.5625467x^4}{58.3231998 - 21.6812755x^2 + x^4}$$

$$\sqrt{z^*} = \frac{N}{2D\left(\frac{3}{2} - x\right)} + \frac{D\left(\frac{3}{2} + x\right)}{2N} = \frac{1 + uv}{2u}$$

At $x = 1$ ($z^* = 5$), this approximation errs by 10^{-13} . To obtain that accuracy, the coefficients must be stated to five or six decimal places (seven are shown).

APPENDIX B

CUBE ROOT

Many of the problems involved in computing the higher roots are similar to those of the square root and submit to similar solutions. Additionally, the odd-numbered roots admit negative values of the argument. So saying, the matters of "normalization" and accounting for sign are left to the reader.

To develop a power series for cube root, we employ the transformation

$$px = \frac{z^* - 1}{z^*}$$

Solving for z^* , $z^* = \frac{1}{1 - px} = \frac{(1 - px)^2}{(1 - px)^3}$ so that

$$(1 - px) \sqrt[3]{z^*} = (1 - px)^{2/3}$$

$$= 1 - \frac{2}{3}px - \frac{1}{9}p^2x^2 - \frac{4}{81}p^2x^3$$

$$- \frac{7}{243}p^4x^4 - \frac{14}{729}p^5x^5 - \dots$$
(50)

Let us develop a Padé approximation of the form

$$(1 - px) \sqrt[3]{z^*} = \frac{1 + a_1 px + a_2 p^2 x^2 + a_3 p^3 x^3}{1 + b_1 px + b_2 p^2 x^2} \quad (51)$$

With a Padé, it is a matter of indifference whether we restrict the argument range before or after developing the rational function. Therefore, in the interest of convenience, set $p = 1$. The simultaneous system of equations then is

$$1 = 1$$

$$a_1 = b_1 - \frac{2}{3}$$

$$a_2 = b_2 - \frac{2}{3}b_1 - \frac{1}{9}$$

$$a_3 = 0 - \frac{2}{3}b_2 - \frac{1}{9}b_1 - \frac{4}{81}$$

$$0 = 0 - 0 - \frac{1}{9}b_2 - \frac{4}{81}b_1 - \frac{7}{243}$$

$$0 = 0 - 0 - 0 - \frac{4}{81}b_2 - \frac{7}{243}b_1 - \frac{14}{729}$$

Solving,

$$b_1 = -\frac{14}{15} \quad b_2 = \frac{7}{45}$$

$$a_1 = -\frac{8}{5} \quad a_2 = \frac{2}{3} \quad a_3 = -\frac{4}{81}$$

Dividing both sides of the expression by $1 - x$ and multiplying all coefficients by their L.C.D. yields

$$\sqrt[3]{z^*} \sim \frac{405 - 648x + 270x^2 - 20x^3}{405 - 783x + 441x^2 - 63x^3}$$

It is desirable to choose an argument range (for the "normalized" variable) the ratio of whose end points is a perfect cube (e.g., 8). Noticing that x passes through zero as z^* passes through unity, we can see at once that

$\frac{1}{8} < z^* < 1$ and $1 < z^* < 8$ will be a good choice for these end points. It turns out that in the range $0.35 < z^* < 2.8$, the relative error is less than 5×10^{-4} . However, we can do nearly an order of magnitude better than that.

The two term divisor enables us to use the original series as a Padé (of which it is a special case) and write, after multiplying by 405:

$$S \begin{pmatrix} 405 & -270 & -45 & -20 \\ 405 & -405 & 0 & 0 \end{pmatrix}$$

Subtracting and reducing to lowest terms, we find

$$4 \begin{pmatrix} 0 & 6 & -5 & 0 \\ 0 & 6 & -7 & 1 \end{pmatrix}$$

The Padé is "optimized" at $P = 0.8A$ from which

$$\sqrt[3]{z^*} = \frac{405 - 652.8x + 274x^2 - 20x^3}{405 - 787.8x + 446.6x^2 - 63.8x^3}$$

Within the argument range $0.38296 < z^* < 3.06853$ the relative error does not exceed 7.167×10^{-5} . The ratio of these end points slightly exceeds 8 --- a perfect cube --- so that the upper limit of z^* (for which the "normalization" routine searches) arbitrarily can be set anywhere between 3.0637 and 3.0685 --- say 3.066.

For any odd-numbered root, there is a specialized adaptation of the Newton-Raphson technique which converges more rapidly than any other. Before developing it, however, let us review Newton-Raphson in simple form:

$$\text{Let } f(y) = z - y^n \quad (n = 3, 5, 7, 9, \dots)$$

$$f'(y) = -ny^{n-1}$$

Suppose we have an estimate of the root, y_1 . We also know that at the true root, \hat{y} , $f(\hat{y}) = 0$. This enables us to write the approximation

$$\frac{f(y_1) - 0}{y_1 - \hat{y}} = f'(y_1) \quad (52)$$

Now \hat{y} is the only unknown, but this is only an approximation, so we will not recover \hat{y} , but y_2 , another (and hopefully better) estimate. Substituting and rearranging,

$$y_2 = y_1 - \frac{f(y_1)}{f'(y_1)}$$

$$y_2 = y_1 + \frac{z - y_1^n}{ny_1^{n-1}}$$

$$y_2 = \frac{1}{n} [(n-1)y_1 + \frac{z}{y_1^{n-1}}] \quad (53)$$

Examining the second derivative, it is found that

$$f''(y) = -n(n-1)y^{n-2}. \quad (54)$$

The curvature is not negligible, and increases with higher roots --- introducing error into the approximation and thereby slowing the rate of convergence.

We could just as well have written the original expression in the form

$$y^{\frac{n-1}{2}} f(y) = z - y^n \quad (n = 3, 5, 7, 9, \dots) \quad (55)$$

The left side is still a function of y which drives to zero at the desired root. But $y^{\frac{n-1}{2}} = 0$ only for the trivial case $y = z = 0$. Thus we can apply the Newton-Raphson technique to

$$f(y) = \frac{z}{y^{\frac{n-1}{2}}} - y^{\frac{n+1}{2}} \quad (56)$$

But now look at the derivatives!!

$$f'(y) = -\left(\frac{n-1}{2}\right) \frac{z}{y^{\frac{n+3}{2}}} - \left(\frac{n+1}{2}\right) y^{\frac{n-1}{2}} \quad (57)$$

and

$$f''(y) = \left(\frac{n^2-1}{4}\right) \frac{z}{y^{\frac{n+5}{2}}} - \left(\frac{n^2-1}{4}\right) y^{\frac{n-3}{2}} \quad (58)$$

It is seen that there is a point of inflection exactly at the desired root. This means that as the estimate approaches the true value, the slope becomes virtually constant, thereby hastening convergence. The rate of convergence is never worse than quadratic, and ultimately tends toward order of magnitude n .

Perhaps this process is best expressed in digital filter form; i.e., as an output/input ratio. Thus

$$\frac{y_2}{y_1} = \frac{(n+1)z + (n-1)y_1^n}{(n-1)z + (n+1)y_1^n} \quad (59)$$

For the cube root, this reduces to

$$\frac{y_2}{y_1} = \frac{2z + y_1^3}{z + 2y_1^3} \quad (60)$$

To cite an example, we find the following errors in the approximation to

$\sqrt[3]{2.5}$:

after Padé ("optimized"), -0.000092

after modified Newton-Raphson, -0.28×10^{-12}

APPENDIX C

LOGARITHMS TO ANY BASE

The argument is "normalized" to a value in the range $3.165 > z^* > (3.165)^{-1}$. The ratio of these two limits slightly exceeds 10. The method is strikingly similar to that used for square root. The parameter h is set to zero.

If $z^* = z \times 10^{-n}$,

then $h^* = h + n$.

(61)

After the approximation has been applied,

$$\log_b z = \log_b z^* + h^* \log_b 10. \quad (62)$$

The most commonly used bases are 2, e, 10, and 16. The necessary constants are found in the first part of this paper.

We develop a Maehly rational function for $\ln_e z^*$ using the transformation

$$px = \frac{z^* - 1}{z^* + 1} ; \quad p = 0.52$$

Thus $z^* = \frac{1 + px}{1 - px}$, and

$$\frac{1}{2x} \ln_e z^* = p + \frac{p^3}{3} x^2 + \frac{p^5}{5} x^4 + \frac{p^7}{7} x^6 + \dots \quad (63)$$

The coefficients of the Tschebychev series are now computed. They are:

$$Y_0 = 0.546850\ 950695\ 9441$$

$$Y_2 = 0.028096\ 097358\ 0741$$

$$Y_4 = 0.001314\ 425494\ 3168$$

$$Y_6 = 0.000073\ 490993\ 0960$$

$$Y_8 = 0.000004\ 482077\ 7161$$

$$Y_{10} = 0.000000\ 287824\ 8525$$

$$Y_{12} = 0.000000\ 019125\ 8782$$

$$Y_{14} = 0.000000\ 001302\ 1921$$

$$Y_{16} = 0.000000\ 000090\ 2873$$

$$Y_{18} = 0.000000\ 000006\ 3490$$

$$Y_{20} = 0.000000 000000 4515$$

$$Y_{22} = 0.000000 000000 0324$$

$$Y_{24} = 0.000000 000000 0023$$

$$Y_{26} = 0.000000 000000 0002$$

A Maehly rational function of the form

$$\frac{1}{2x} \ln_e z^* \approx \frac{a_0 + a_2 T_2 + a_4 T_4 + a_6 T_6}{1 + B_2 T_2 + B_4 T_4 + B_6 T_6}$$

is now developed.

$$a_0 = 0.543339 498483 1167$$

$$a_2 = -0.108861 411680 5337$$

$$a_4 = 0.002641 510414 0964$$

$$a_6 = -0.000008 962142 2536$$

$$B_2 = -0.250375 253205 3509$$

$$B_4 = 0.008877 426771 5024$$

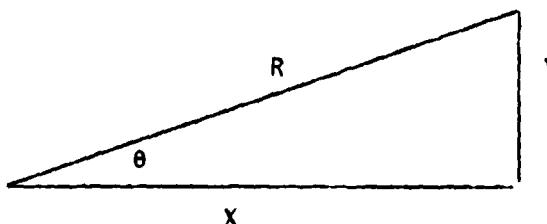
$$B_6 = -0.000076 902174 0061$$

At $x = 1$, ($z^* = 19/6$), the relative error is -4.957×10^{-12} .

APPENDIX D

RECOVERING AN ANGLE FROM RECTANGULAR CO-ORDINATES

Let us label the sides of
a right triangle X, Y and R.



Given any two, it is required to
find the angle θ .

We have immediately that $R^2 = X^2 + Y^2$.

(64)

Most programmers unhesitatingly (and unthinkingly!) select one of the common ratios

$$\frac{Y}{X} = \tan \theta, \frac{Y}{R} = \sin \theta, \frac{X}{R} = \cos \theta$$

and compute the inverse. There are serious objections to this cavalier approach:

(1) Near $\theta = 90^\circ$, $\tan \theta$ presents scaling difficulties as the slope increases without bound.

(2) Also near $\theta = 90^\circ$, $\sin \theta$ becomes a most imprecise measure of angle, since the slope approaches zero, rendering the function insensitive to changes in the argument.

(3) Near $\theta = 0$, $\cos \theta$ exhibits the same disadvantages, plus the additional one of failing to change sign as θ passes through zero.

The answer usually is taught during the first week of most college trigonometry courses, then promptly forgotten:

$$\tan \frac{\theta}{2} = \frac{Y}{X + R} \quad (65)$$

The function $\tan \frac{\theta}{2}$ behaves very well indeed. Provided $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, the absolute value of $\tan \frac{\theta}{2}$ varies between 0 and 1, while its slope varies between 1 and 2. This fundamental identity appears in many forms, two of which are

$$\tan \frac{\theta}{2} = \frac{\sin \theta}{1 + \cos \theta} = \frac{\tan \theta}{1 + \sqrt{1 + \tan^2 \theta}}$$

Applying the identity a second time yields (since $Y^2 + (X + R)^2 = 2R(X + R)$):

$$\tan \frac{\theta}{4} = \frac{Y}{(X + R) + \sqrt{2R(X + R)}} \quad (66)$$

For simplicity in notation, let $Q = \sqrt{2R(X + R)}$. Then

$$\tan \frac{\theta}{8} = \frac{Y}{(X + R = Q) + \sqrt{2(2R + Q)(X + R)}} = t$$

or

$$t = \tan \frac{\theta}{8} = \frac{Y}{(X + R + Q) + \sqrt{2Q(X + R + Q)}} \quad (67)$$

The Padé approximation now is used to recover the angle θ . It is

$$\theta = 8 \arctan t = 8t \left(\frac{189 + 147t^2 + 12.8t^4}{189 + 210t^2 + 45t^4} \right)$$

The error in θ is less than 2.2×10^{-10} radians.

Combining factors and employing mixed polynomial form yields

$$\theta_{\text{radians}} = + \left\{ \frac{(147t^2 + 1176)t^2 + 1512}{(3t^2 + 210)t^2 + 189} \right\}$$

or

$$\theta_{\text{degrees}} = \frac{t}{\pi} \left\{ \frac{(1228.8t^2 + 14112)t^2 + 18144}{(3t^2 + 14)t^2 + 12.6} \right\}$$

It is to be remarked that in order to achieve this accuracy, a Maclaurin series would have to employ the t^{13} term, while a Tschebychev series requires six coefficients expressed to 10-digit accuracy.

BIBLIOGRAPHY

- [1] Abramowitz, M. and Stegun, I. A., eds., HANDBOOK OF MATHEMATICAL FUNCTIONS; Dover Publications, Inc., New York, 1972.
- [2] Boole, G., THE CALCULUS OF FINITE DIFFERENCES; 5th Ed., Chelsea Publishing Co., New York, 1970.
- [3] Churchill, R. V., INTRODUCTION TO COMPLEX VARIABLES AND APPLICATIONS; McGraw-Hill Book Co., Inc., New York, 1948.
- [4] Fike, C. T., COMPUTER EVALUATION OF MATHEMATICAL FUNCTIONS; Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968.
- [5] Hamming, R. W., NUMERICAL METHODS FOR SCIENTISTS AND ENGINEERS; McGraw-Hill Book Co., Inc., New York, 1962.
- [6] Hastings, C., APPROXIMATIONS FOR DIGITAL COMPUTERS; Princeton University Press, Princeton, N. J., 1955.
- [7] Peirce, B. O. and Foster, R. M., A SHORT TABLE OF INTEGRALS; 4th Ed., Ginn and Co., New York, 1957.
- [8] Rankin, D. W., ESTIMATING RELIABILITY FROM SMALL SAMPLES; in Proc. 22nd Conf. on DOE in Army RD&T, Dept. of Defense, 1977.
- [9] _____, COMPUTING THE DEFINITE INTEGRAL $\int_0^\infty e^{-(px^2 + qx + r)} dx$ ON A PROGRAMMABLE CALCULATOR; in Proc. 23rd Conf. on DOE in Army RD&T, Dept. of Defense, 1978.
- [10] Saff, E. B. and Varga, R. S., eds., PADÉ AND RATIONAL APPROXIMATION THEORY AND APPLICATIONS; Academic Press, Inc., New York, 1977.
- [11] Smith, J. M., SCIENTIFIC ANALYSIS ON THE POCKET CALCULATOR; John Wiley and Sons, New York, 1975.

MOS TRAINING COURSE SELECTION
CRITERIA: AN APPLICATION OF DISCRIMINANT ANALYSIS

Pat Cassidy and Lounell Snodgrass

Analysis Branch II
Training Effectiveness Analysis Division
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico

I. INTRODUCTION

This is a study of criteria by which soldiers are selected for Military Occupational Speciality (MOS) training schools. Three distinct MOS's and their associated training courses are considered. For simplicity, they will be referred to as MOS A, MOS B, and MOS C.

Intelligence screening of new recruits is accomplished with the Armed Forces Qualification Test (AFQT). Job or occupation qualifications are determined with the Armed Forces Vocational Aptitude Battery (ASVAB). These tests are described in Tables 1 and 2. In the development of ASVAB, training course performance was taken as the measure of soldier performance. Aptitude composites were developed to maximize validity coefficients. Consequently, the composites are composed of several tests and are highly intercorrelated. For a description of the development of the aptitude composites see Fuchs and Maier (1973, 78). Composite scores normally range from 40 to 160, with an average score near 100 and standard deviation near 20.

Typically the selection criteria for a specific training school (course) will consist of a minimum score on a single ASVAB composite. Unlike raw test scores, aptitude composites are maintained in a soldier's personnel file where they can be easily obtained by a particular school. Rarely, minimum scores on two composites may be required. As weapon systems, the Army population, and training courses have evolved; some schools have experienced high attrition rates. TRADOC Systems Analysis Activity (TRASANA) was asked to study samples from three MOS school (courses) and recommend improved selection criteria.

TABLE 1
TESTS IN THE ARMED SERVICES VOCATIONAL APTITUDE BATTERY (ASVAB)

CATEGORY	TEST TITLE	TEST SYMBOL
General Ability Tests	Arithmetic Reasoning General Information Mathematics Knowledge Science Knowledge Word Knowledge	AR GI MK SK WK
Mechanical Ability Tests	Automotive Information Electronics Information Mechanical Comprehension Trade Information	AI EI MC TI
Perceptual Ability Tests	Attention to Detail Pattern Analysis	AD PA
Classification Inventory	Attentiveness Scale Combat Scale Electronics Scale Maintenance Scale	CA CC CE CM

TABLE 2
APTITUDE AREAS AND RELATED ARMY JOBS

<u>Aptitude Area Symbol</u>	<u>Title</u>	<u>Composite ACB Tests</u>	<u>Major Related Jobs</u>
CO	Combat	AR+TI+PA+AD+CC	infantry, Armor, Combat Engineer
FA	Field Artillery	AR+GI+MK+EI+CA	Field Cannon and Rocket Artillery
EL	Electronics Repair	AR+EI+MC+TI+CE	Missiles Repair, Air Defense Repair, Tactical Electronics Repair, Fixed Plant Communication Repair
OF	Operators and Food	GI+AI+CA	Missiles Crewman, Air Defense Crewman, Driver, Food Services
SC	Surveillance and Communications	AR+WK+MC+PA	Target Acquisition and Combat Surveillance, Communication Operations
MM	Mechanical Maintenance	MK+AI+EI+TI+CM	Mechanical & Air Maintenance, Rails
GM	General Maintenance	AR+SK+AI+MC	Construction and Utilities, Chemical, Marine, Petroleum
CL	Clerical	AR+WK+AD+CA	Administrative, Finance Supply
ST	Skilled Technical	AR+MK+SK	Medical, Military Police, Intelligence, Data Processing, Air Control, Topography and Printing Information, and Audio Visual
GT	General	AR+WK	Used only to qualify for special tests, as Officer Candidate Test

Of course, high attrition rates might be remedied by improving the courses.

This remedy has been done; however, this was not part of the TRASANA study.

II. DATA

One sample was provided for such MOS. The current selection criteria, the number passing and failing, course scores, ASVAB composite and AFQT scores were available for each MOS sample. The type of failure -academic, non-academic was known for MOS C. These data are summarized in Table 3.

TABLE 3

MOS	CURRENT SELECTION CRITERIA	NUMBER PASSING COURSE	NUMBER FAILURES	NUMBER NON-ACADEMIC FAILURES
A	EL > 90 CL Σ 90	114	69	N/A
B	EL \geq 90	227	78	N/A
C	EL \geq 120	109	73	23

III. ANALYSIS

Stepwise discriminant analysis was the technique chosen to determine improved selection criteria. This method produces a linear combination of ASVAB composites which best discriminates between the pass and fail groups. This linear discriminant function allows the incorporation of posteriori probabilities and the costs of misclassification. The resulting classification procedure minimizes the expected cost of misclassification under certain conditions. For a description of discriminant analysis see A.A. Afifi and S. P. Azen or T. W. Anderson.

A. MOS A ANALYSIS

This analysis produced MM and ST as the variables which best discriminate between the two groups. For simplicity of application, selection criteria are traditionally given as minimum scores on one or two (rarely) composites. Consequently, the linear discriminant function is not a practical classification tool. To determine a more practical classification procedure, the 114 sample cases were ranked first by MM, then by ST. The following classification procedure was then determined: if $MM \geq 100$ and $ST \geq 100$ classify as pass; otherwise classify as fail. A graphical comparison of the two procedures is given in Figure I. The attrition rate using the proposed MM/ST criteria would be 15.5%, while reducing the number of soldiers chosen for the course by 56 (see TABLE 4).

An alternate criterion, proposed by Army School A, using $EL \geq 105$ was also considered. Course attrition and the course attendees available for this sample are shown in TABLE 4.

TABLE 4
RELATIVE EFFECTIVENESS OF THREE ALTERNATIVE COURSE
SELECTION CRITERIA FOR MOS A

SELECTION CRITERIA	ATTENDEES SELECTED	GRADUATES	NON-GRADUATES	ATTRITION RATE (%)
ACTUAL: $EL \& CL \geq 90$	114	69	45	39.5
ALTERNATIVES: $MM > 100$, $ST \geq 100$	58	49	9	15.5
$EL \geq 105$	56	42	14	25.0

LINEAR CLASSIFICATION AND PROPOSED MOS A SELECTION CRITERIA

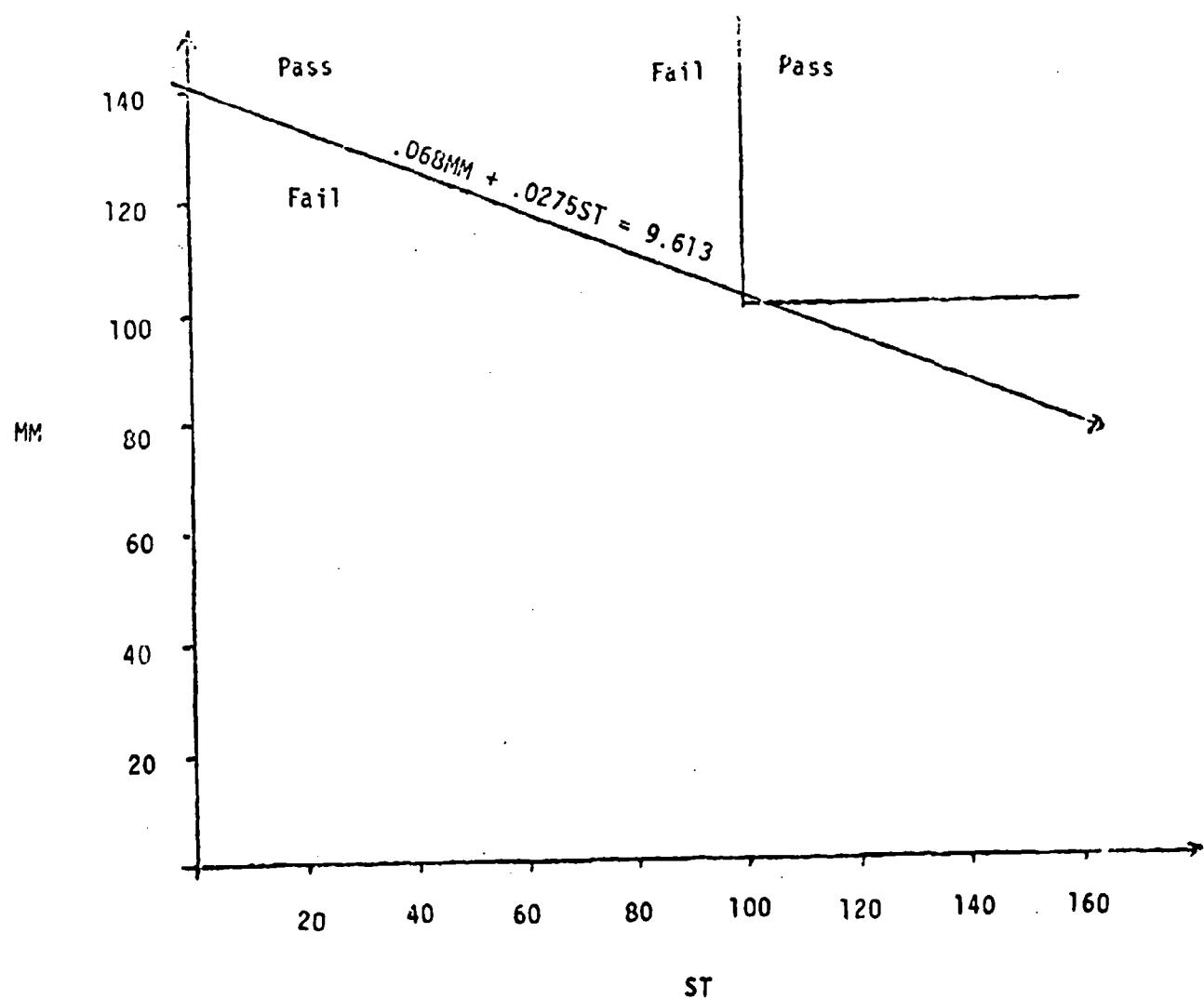


Figure 1

B. MOS B ANALYSIS

This analysis was different from the MOS A analysis since all criteria considered included the original EL criteria. The discriminant analysis chose SC as the best predictor of passing or failing the MOS B course. Chi-square tests were performed to test for the independence of a selection criterion from the pass or fail classification. Table 5 summarizes the proposed criteria, their attrition rates, and the number of soldiers selected from the sample.

TABLE 5
RELATIVE EFFECTIVENESS OF ALTERNATIVE COURSE SELECTION
CRITERIA FOR MOS B

SELECTION CRITERIA	ATTENDEES SELECTED	GRADUATES	NON-GRADUATES	ATTRITION RATE (%)
ACTUAL:				
<u>EL > 90</u>	305	227	78	25.6
ALTERNATIVES:				
<u>EL > 95</u>	235	184	51	21.7
<u>EL > 100</u>	157	132	25	15.9
<u>EL > 90</u> <u>SC \sum 100</u>	132	116	16	12.1
<u>EL > 90</u> <u>SC \sum 95</u>	182	154	28	15.4

C. MOS C ANALYSIS:

This analysis also considered the original criteria ($EL \geq 120$) as a necessary condition for any new criteria since the sample used for the discriminant analysis was chosen by this criteria. The discriminant analysis chose GM as the score that best discriminates between pass or fail groups (academic failures). The non-academic failures were not included in the analysis. Chi-square tests were performed to test for the independence of a selection criteria from the pass or fail classification. Table 6 summarizes the proposed criteria, their attrition rates, and the number of soldiers selected from the sample.

TABLE 6
RELATIVE EFFECTIVENESS OF ALTERNATIVE COURSE SELECTION CRITERIA FOR MOS C

SELECTION CRITERIA	ATTENDEES SELECTED	GRADUATES	NON-GRADUATES	ATTRITION RATE (%)
ACTUAL:				
$EL \geq 120$	182	109	73	40.1
ALTERNATIVES:				
$EL \geq 125$	103	69	34	33.0
$EL \geq 130$	56	46	10	17.9
$EL \geq 120 \text{ & } GM \geq 125$	78	60	18	23.1
$EL \geq 120 \text{ & } GM \leq 120$	102	75	27	26.5

IV. SUMMARY

The proposed criteria that best discriminates between graduating or non-graduating from the MOS A course were an MM \geq 100 and ST \geq 100. The best selection criteria for MOS B were EL \geq 90 and SC \geq 95. Finally the most promising selection criteria for MOS C were EL \geq 120 and GM \geq 120. These conclusions are based on the alternate criteria that lower course attrition while do not substantially reduce the attendees selected.

V. POINTS OF DISCUSSION:

A. The data for the MOS A and MOS B analysis contained no distinction between academic and non-academic failures. Since ASVAB composites are intended to indicate subject aptitudes, their use to predict non-academic failures might be questioned. Depending on the sample their inclusion or exclusion could significantly alter the conclusions.

B. All samples in the study were selected from current courses. Therefore, all cases met the current selection criteria for each sample. For MOS B and MOS C the current criteria were included as part of the new criteria. For MOS A the current criterion was omitted. The justification for this is that current criterion is not very restrictive. The statistical justification for such a generalization is lacking.

C. In the MOS A analysis a ranking procedure was used to develop absolute cut-off scores for a classification procedure. The relation of such a procedure to that of the discriminant function and its "optimal" properties, if any, were ignored.

Bibliography

A. A. Afifi & S. P. Azen, Statistical Analysis: A Computer Oriented Approach,
2nd Ed., Academic Press, 1979.

T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley,
1958.

M. H. Maier & E. F. Fuchs, "Differential Validity of the Army Aptitude Areas
for Predicting Army Job Training Performance of Blacks and Whites," US Army
Research Institute, Technical Paper 312, 1978.

M. H. Maier & E. F. Fuchs, "Effectiveness of Selection and Classification
Testing," US Army Research Institute, Research Report 1179, AD-768 168, 1973.

THE ARMOR COMBAT FOR MODEL SUPPORT (ARCOMS) FIELD EXPERIMENT

Roger F. Willis
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico 88002

ABSTRACT.

The Armor Combat Operations Model Support (ARCOMS) Test, Phase II, is a force-on-force field experiment aimed at collecting target acquisition and engagement data for use in the design and running of combined arms simulations and war games. TRASANA is the proponent and the test will be conducted by TCATA at Fort Hood in January-February 1981. Tactical realism would be unacceptably sacrificed if certain key variables were controlled. Due to resource constraints, very few replications can be run under given conditions. This presentation will pose the question (specifically for ARCOMS) of how to extract the maximum amount of valid information from relatively uncontrolled field experiments (operational tests) carried out with very small sample sizes.

1. Background.

a. ARCOMS will be the first in a series of field experiments carried out to provide better input data for combined arms models and war games.

b. It has been recognized for years by the modeling community that we do not have adequate data on attacker detection rates, in realistic force-on-force conditions, and on attacker engagement dynamics and fire distribution. The ARCOMS test also presents the opportunity to gain valuable insights concerning alternative attacker tactics and defender detection rates and defender fire distribution.

c. For the first time intervisibility between combat vehicles will be measured dynamically and recorded automatically.

2. Purpose.

The purpose of ARCOMS is to examine the combat processes in a force-on-force environment and to provide input data for TRADOC combined arms models, simulations and games. Emphasis is to be placed on identifying the process by which the attacker acquires and uses information during the attack. Data on detection probabilities will be keyed to the times at which intervisibility starts. TRASANA will use the data to develop algorithms and to provide input parameters for the revision of combat models. In addition to serving as an empirical source for probability distributions and other data for models, the ARCOMS test outputs will be used for testing a number of hypotheses about the basic nature of combat processes.

3. Scope. ARCOMS will consist of a series of force-on-force experiments of a platoon-sized unit defending against a company force, with deliberate variations in terrain and attacker tactics. This phase will examine intervisibility, the fire and maneuver interactions. To the extent possible, low visibility conditions will be considered.

4. Gross Design of Test.

a. The conditions for the twenty-four test runs (individual battles) are defined in Table 1. Four major factors are varied: attacker tactics, type of defense, light level, and terrain type (A or B). Note that we will not have enough runs to investigate some of the interactions that one suspects might be important. For example, the rapid approach tactic will not be run in Type B terrain; there will be no night trials in Type B terrain; and the rapid approach tactic will not be run at night.

b. It will be possible to develop estimates of the impact of some of these major factors on key measures (e.g., on average attacker detection rate). For the impact of "light level" we will compare run set A with run set B, or set E with set F. For the impact of "attacker tactics" we will compare set A with set D (or set E with set H), etc. In order to increase sample sizes in some cases we will lump sets together. For example, we get a sample of six battles by lumping set A with set E, assuming that the differentiation between hasty defense and deliberate defense might not be significant (for some measures).

5. Quantities to be Measured.

a. Before listing the outputs ultimately needed from the experiment we will discuss the quantities that will actually be measured. Briefly we need to measure things like who could have detected whom and when, who actually detected whom and when and why, who "killed" (laser hits) whom and when, and how was information transmitted and used.

b. For each combat vehicle (attacker as vehicles and defender vehicles) the test time-tagged data of the following types will be collected:

- (1) position location
- (2) line-of-sight (laser A)
- (3) detection
- (4) firing (laser B)
- (5) hit and/or kill
- (6) video through the gunner sight
- (7) audio

6. Uncontrolled Factors.

In order to approach tactical realism many important factors (that could influence the quantities measured) will not be controlled. However, to the extent possible, the values assumed by these uncontrolled factors will be measured or estimated during the trials or recaptured after the trials. Some of these uncontrolled factors are listed in Table 2. The attacker task force commander and the defender platoon commander, who will be varied extensively during the course of the test, will each be given a broad mission. The details of how they carry out their missions will be up to them. Actually data on how much variation we observe between individual commanders presumably carrying out the same mission will also be important information.

7. Major Test Outputs.

We list here only the most important measures (dependent variables) expected to be produced by analysis of the data collected during the trials. The next step would be to correlate each of these measures with the controlled variables (Table 1) and also with the uncontrolled variables (Table 2). (An example is: correlation of attacker detection rate with the force ratio - obtained by dividing the number of attacker weapons ready by the number of defender weapons ready. This initial force ratio will usually vary from battle to battle, depending on the states of readiness of the individual weapons plus their instrumentation added on for the test.) The major output measures are as follows:

- a. attacker detection time - conditional (given a detection)
- b. attacker detection time - unconditional
- c. defender detection time - conditional (given a detection)
- d. defender detection time - unconditional
- e. attacker engagement time
- f. defender engagement time
- g. time at least 3 attackers in LOS, etc.
- h. time at least 3 defenders in LOS, etc.
- i. attacker fire distribution patterns:
 - (1) defender sites intervisible with attackers but not engaged by attackers
 - (2) defenders engaged "simultaneously" by 2 attackers, etc.
 - (3) number of rounds fired per engagement

j. defender fire distribution patterns:

- (1) attackers intervisible but not engaged by defenders
- (2) attackers engaged "simultaneously" by 2 defenders, etc.
- (3) number of rounds fired per engagement

k. frequency of attackers engaging false targets

l. frequency of defenders engaging false targets

3. Additional Hypotheses.

Although the primary purpose of ARCOMS is to collect data on detection rates, engagement rates, etc. to provide inputs for combined arms models and war games, the same data set will be used by TRASANA to investigate a number of tactical hypotheses in the areas of:

- a. detections by attacker
- b. attacker communications
- c. attacker control of movements
- d. degradation or enhancement
- e. defender allocation of fire
- f. defender disengagements

These hypotheses, after field testing, will be either rejected, accepted, or qualified and the accompanying analyses will provide insights that will be even more valuable than model inputs. These insights might contribute to improvement in the structures of the combat models and to more credible theories of combat.

9. Analysis Procedures.

The following types of analyses will be carried out with the ARCOMS data:

- a. Plotting and graphics of battles
- b. Serial correlations
- c. Descriptive statistics
- d. Analysis of covariance
- e. Theoretical distribution fitting
- f. Hypotheses testing

- g. Evaluation of tactics
- h. Comparison of LOS data with digitized terrain
- i. Analysis of detection data and model improvement (by NVEOL)

TABLE 1 - ARCOMS TEST RUNS
(COMPANY TASK FORCE VERSUS PLATOON SLICE)

<u>RUN SET</u>	<u>TEST RUNS</u>	<u>ATTACKER TACTICS</u>	<u>DEFENSE</u>	<u>LIGHT LEVEL</u>	<u>TERRAIN</u>
A	1 - 3	Maneuver	Hasty	Day	A
B	4 - 6	Special	Hasty	Day	B
C	7 - 9	Maneuver	Hasty	Day	B
D	10 - 12	Rapid approach	Hasty	Day	A
E	13 - 15	Maneuver	Deliberate	Day	A
F	16 - 18	Maneuver	Deliberate	Night	A
G	19 - 21	Maneuver	Deliberate	Day	B
H	22 - 24	Rapid approach	Deliberate	Day	A

TABLE 2 - ARCOMS - UNCONTROLLED FACTORS

A. DEFENDER

1. Decision to move
2. Number of weapons ready
3. Communications (target handoff)
4. Frequency of firing
5. Distribution of fire
6. Amount of concealment
7. Open-fire ranges

B. ATTACKER

1. Velocity
 - a. individual weapons
 - b. platoons
2. Specific movement patterns (use of terrain, trees, etc.)
3. Number of weapons ready
4. Use of overwatchers
5. Familiarity with terrain
6. Communications (target handoff)
7. Distribution of fire
8. Frequency of firing

C. ENVIRONMENT

1. Visibility
2. Weather
3. Other obscuration
4. Range (distance)
5. Vegetation
6. Angle of sun
7. Target background

EXTREME VALUE QUANTILE RESPONSE EXPERIMENTAL DESIGN

Jill H. Smith
Jerry Thomas
Probability and Statistics Branch
Ballistic Modeling Division
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

ABSTRACT. An experimental design has been developed to be used to determine the shielding thickness required between rounds stored in a storage area to prevent round-to-round propagation from an initial explosion. Extreme value quantile response techniques were used with shielding thickness as the stimulus variable. The developed design drastically reduces the sample size required for a given quantile and confidence when compared with known distribution-free extreme value designs.

1. INTRODUCTION. The Terminal Ballistics Division of the Ballistic Research Laboratory encountered the problem of determining how thick the shielding should be between rounds of ammunition stored in a storage area to prevent round-to-round propagation from an initial explosion. Vulnerability analysis indicated that the probability of survival of the storage area would drastically decrease with an increase in the number of rounds exploding. Prior testing has shown that shielding material placed between rounds could prevent neighboring rounds from exploding. Due to space limitations in the storage area, it was desired to keep the shielding thickness to a minimum and simultaneously minimize the probability of round-to-round propagation.

It was decided that the specific objective of the test would be to find the shielding thickness needed to be 90% confident that the probability of a neighboring round exploding is less than 0.1.

The problem appeared to fit into the category of extreme value quantile response problems. Defining X as the stimulus variable, in this case the thickness of the shielding which effects the stimulus, and the probability of a response associated with a given X , x , is described by a nonresponse function $M(x)$. (Usual notation has $M(x)$ as the probability of response. However, defining $M(x)$ as a nonresponse is more natural for this problem.) This function is assumed to be monotonically nondecreasing with increasing stimulus levels.

A discussion of available designs and the modified design chosen for the experiment is contained in the following chapters.

2. AVAILABLE DESIGNS. A nonparametric approach was taken because of the lack of information about the response function. As stated, the quantile in which we are interested is $\alpha = .10$, and therefore is in the tail of the response distribution. From a review of the available designs in

in the literature the only nonparametric test designs available for testing in the tail regions are the Alexander Extreme Value Design and the Rothman Design. Of these, the Alexander Extreme Value Design is preferred since it:

- 1) is "generally more efficient than other available nonparametric designs, and is asymptotically as efficient as the best parametric stochastic approximation when distributional assumptions are valid,"
- 2) has significantly simpler design rules and analysis procedures than the Rothman Design, and
- 3) does not differ in median required sample size.

3. ALEXANDER EXTREME VALUE DESIGN. The Alexander Extreme Value Design assumes only a monotone nondecreasing response function as the stimulus increases.

A. Design Rules

- 1) The first test is at level (shielding thickness) X_1 , the a priori best guess of X_α .
- 2) Testing is performed by alternately increasing and decreasing sequences of test levels. The test levels are increased or decreased by a step size δ , where δ is a fraction of an estimate of the standard deviation. Terms such as "higher" and "level above" refer to thicker shielding levels, and "below" and "lowest" refer respectively to thinner and thinnest shielding thickness levels.
- 3) The first sequence decreases the levels until a response (explosion) is observed.
- 4) The first test of an increasing sequence is at the level above the highest level at which a response has been observed. The increasing sequence ends at level X_i such that in the corresponding zero region² less than or equal to X_i at least N nonresponses have been observed. Values for N can be found from

$$(1 - \alpha)^N = 1 - P \quad (2.1)$$

where $N = [n] + 1$ and P is some specified probability.

¹D. Rothman, M. J. Alexander and J. M. Zimmerman, The Design and Analysis of Sensitivity Experiments, NASA CR-62026, Vol. I, p. 74.

²Zero region - stimulus region above the highest level at which a response has been observed.

5) The first test of a decreasing sequence is at the level above the highest level at which a response has been observed. If the result is a response, the sequence ends; otherwise, one more test at the next lower level is performed.

6) Testing terminates when there are three adjacent levels, X_r , $X_r + \delta$ and $X_r + 2\delta$ such that at least one response has been observed at X_r and none at a higher level, and a total of N nonresponses have been observed at $X_r + \delta$ and $X_r + 2\delta$. (δ is the step size between levels.)

7) The maximum likelihood estimate of X_α , \hat{X}_α , is found by the method of reversals and linear interpolation (see Appendix).

B. Analysis

We are interested in the $\alpha = .1$ quantile of the response distribution, that is, the value, $X_{.1}$, at which the probability of a response is $.1$. Therefore, the probability of a nonresponse at the $X_{.1}$ quantile is $(1 - .1)$. The probability of n nonresponses, assuming the n tests are independent, is $(1 - .1)^n$. The probability of at least one response out of n tests is $1 - (1 - .1)^n$. Specifying the probability of at least one response out of n tests at the $X_{.1}$ quantile to be $P = .9$, we have

$$.9 = 1 - (1 - .1)^n.$$

This, with a slight algebraic manipulation, is Equation 2.1 with $\alpha = .1$ and $P = .9$. Solving, $N = [n] + 1 = 22$. Hence, we would expect with probability $.9$ at least one response out of 22 tests at the $X_{.1}$ quantile.

If we observe 22 nonresponses at some level X_* , we can assume we are not at the $X_{.1}$ quantile and, in fact, the

$$\text{Prob}\{X_{.1} < X_*\} > .9.$$

Using the above argument, we can conclude from the Alexander Extreme Value Design that the level at which the true probability of response is $.1$ is less than $X_r + 2\delta$ with ninety percent confidence. The point estimate of the $X_{.1}$ quantile can be found using the method of reversals outlined in the Appendix.

C. Simulation

Based on "guesimates" for $X_{.5}$ and $X_{.75}$, a response distribution was hypothesized with which to Monte Carlo the Alexander Extreme Value Design for $\alpha = .1$ and $P = .9$. The response distribution assumed was the cumulative

normal distribution with mean = .5 and variance = .14. (Note, however, that the test design and analysis procedures are distribution-free.) The smallest practical step size of shielding thickness was 1/8 inch.

Figures 1 and 2 are examples of the Alexander Extreme Value Design Monte Carloed to illustrate the design rules. Responses are denoted by "X" 's and nonresponses by "0" 's. I_i denotes the i-th increasing sequence and D_j the j-th decreasing sequence. The number of rounds required (NR), the maximum likelihood estimate of the .1 quantile ($\hat{X}_{.1}$), and the ($X_r + 26$) level are given for each simulation.

Figure 3 shows the distribution of the number of rounds required for 500 simulations of the above design. The number of rounds required is twice the number of responses and nonresponses shown for each simulation since a donor round must be detonated for each test round. The average number of rounds required to complete the test was 166, the median was 164 and ten percent of the tests required 184 rounds or more.

The distribution of the maximum likelihood estimates of $\hat{X}_{.1}$ for the 500 simulations is given by the histogram in Figure 4. The distribution of $\hat{X}_{.1}$ is asymptotically normal about the true $X_{.1}$ quantile = 7.83. The distribution generated by the test data shown in Figure 4 has a mean of 7.77, which is in good agreement for 500 simulations, and is approximately normally distributed as shown by the overlying normal curve.

Figure 5 shows the distribution of level $X_r + 26$ for 500 simulations. This is the level about which we can conclude that the

$$\text{Prob} \{X_{.1} < X_r + 26\} > .9.$$

4. MODIFICATION OF THE ALEXANDER EXTREME VALUE DESIGN. The median number of rounds required for the Alexander Extreme Value Design (EVD), as described in the previous section, was 164 as determined by the 500 simulations. Since the number of rounds available for testing was considerably smaller, the major objective in modifying the Alexander EVD was to reduce the number of rounds required, while maintaining the confidence level and the ability to compute the point estimate of the $X_{.1}$ quantile.

The Alexander EVD requires that a donor round be detonated for each test. The number of donors needed can be reduced by using one donor to detonate up to four test rounds (acceptors). Figure 6 shows the configuration of four acceptors per donor. Steel shielding will be placed between acceptors, as shown by the dotted lines, if interaction between acceptors is observed. Optimizing the number of acceptors per donor in the Alexander EVD reduces the number of rounds required by approximately 33 percent.

It was noticed that the rounds above level $X_r + 26$ were neither used to establish the confidence statement, nor to terminate the test design, nor to compute the point estimate of the $X_{.1}$ quantile. By limiting each increasing sequence above the highest stimulus level at which a response has been observed, the rounds "wasted" above level $X_r + 26$ can be eliminated. There is a trade-off in eliminating these rounds since the level $X_r + 26$ can change if a response is observed at a higher level. Therefore, some testing should be above $X_r + 26$ until more than half the number of rounds required to demonstrate the chosen probability are at levels $X_r + \delta$ and $X_r + 26$. Testing at X_r and below is used in the determination of the point estimate of $X_{.1}$.

The following test design is the result of many Monte-Carlo simulations in which different starting levels, number of acceptors per donor and sequences of testing have been tried in order to minimize the required number of rounds, yet retain the confidence level and point estimate of the quantile $X_{.1}$.

A. Modified Design Rules

- 1) The first test level is X_1 , the best a priori guess of X_a . δ is the step size between levels.
- 2) One acceptor per donor is used, in a decreasing sequence, until a response is observed. Let X_r be the highest level at which a response is observed.
- 3) After the first response, the number of acceptors per donor in each test is increased to alternately three and then four. After the first response, three acceptors per donor are tested at the next three levels above X_r . Then four acceptors per donor having shielding at levels X_r and the next three higher levels are tested.
- 4) If another response is observed at a higher level, it becomes X_r , and testing continues alternating three and then four acceptors per donor until at least 12 (more than half the required 22) nonresponses have been observed at the two levels immediately above X_r .
- 5) When at least 12 nonresponses have occurred at $X_r + \delta$ and $X_r + 26$, the number of acceptors per donor is reduced to alternately two above X_r and then three, starting at X_r , for the remainder of the test.
- 6) Testing terminates when at least N (22) nonresponses have been observed at the two levels immediately above the highest level at which a response has been observed.

B. Analysis of the Modified Design

As in the Alexander Extreme Value Design, we have $N = 22$ nonresponses at $X_{\bar{X}} + \delta$ and $X_{\bar{X}} + 2\delta$ and can conclude that we are not at level the $X_{.1}$ quantile and in fact,

$$\text{Prob } \{X_{.1} < X_{\bar{X}} + 2\delta\} > .9.$$

The point estimate can again be found using the method of reversals. Therefore, the changes in the test design have not affected the confidence statement or the point estimate.

C. Simulations Using Modified Design

Using the same response function that was used when simulating the Alexander Extreme Value Design, 500 simulations of the modified design were also Monte-Carloed.

Figures 7 and 8 are examples of the modified test design illustrating the modified design rules. Again, responses are denoted by "X" 's and nonresponses by "O" 's. The abscissa represents individual tests rather than sequences of tests as shown in the Alexander Extreme Value Design.

Figure 9 shows the distribution of the required number of rounds for the 500 simulations. The median number of rounds required was 67 and the mean number of rounds, 70. Only ten percent of the simulations required 93 or more rounds.

The histogram in Figure 10 is the distribution of the maximum likelihood estimates of $X_{.1}$ for the 500 simulations. Again, the distribution of the maximum likelihood estimates are asymptotically normal about the true $X_{.1}$ quantile = 7.83. The distribution shown has a mean of 8.00, and is approximately normally distributed as shown by the overlying normal curve. Figure 12 shows the distribution of number of rounds required for both the Alexander EVD and the Modified Alexander EVD. The Modified Alexander EVD is on the left and the Alexander EVD is on the right.

5. SUMMARY. The Alexander EVD was modified, mainly, by using multiple rounds per test and by limiting the number of rounds above the highest response. These changes resulted in a design that required less than half the rounds of the Alexander EVD in the simulations performed. The range of the required number of rounds (NR) for the Alexander EVD was from 134 to 256 and for the modified Alexander EVD was 46 to 140. The Modified Alexander EVD has simple design rules that permit the estimation of an extreme value of a quantile response function and the associated confidence interval.

This report used only the normal distribution as the assumed underlying distribution for the Monte Carlo simulations. Other distributions are currently being used for this purpose. A reduction in the number of rounds required for these distributions is also expected. Recall, however, that neither the experimental design nor the analysis methods require the assumption of a response distribution. The design is distribution-free.

APPENDIX

METHOD OF REVERSALS FOR SENSITIVITY DATA

1. METHOD. The method of reversals is a maximum-likelihood procedure for obtaining distribution-free estimates of a monotone nondecreasing response function. The test stimulus levels are X_i ($i = 1, 2, \dots, k$) and are ordered from thickest to thinnest shielding thickness,

$$X_1 > X_2 > \dots > X_k \quad (\text{A.1})$$

If \hat{p}_i is the estimate of the probability of response at X_i , and if we assume that the response function is monotone nondecreasing, then necessarily

$$\hat{p}_1 \leq \hat{p}_2 \leq \dots \leq \hat{p}_k. \quad (\text{A.2})$$

The algorithm below can be used to find the estimates of the response distribution and their associated stimulus levels.

- 1) Let X_i ($i = 1, 2, \dots, k$) be the k stimulus levels at which data have been collected, where $X_1 > X_2 > \dots > X_k$. We wish to find the estimates, \hat{p}_i , of the values $p_i = M(X_i)$, the response probabilities at the levels X_i , which satisfy Equation A.2.
- 2) Let n_i ($i = 1, 2, \dots, k$) be the number of tests performed at level X_i and f_i ($i = 1, 2, \dots, k$) be the number of responses observed in the n_i tests.

Consider the sequence

$$\frac{f_1}{n_1}, \frac{f_2}{n_2}, \dots, \frac{f_k}{n_k}.$$

If this sequence is nondecreasing, then the estimates \hat{p}_i are simply given by

$$\hat{p}_i = \frac{f_i}{n_i}.$$

- 3) If for some i , $\frac{f_i}{n_i} > \frac{f_{i+1}}{n_{i+1}}$, replace both by

$$\frac{f_{i,i+1}}{N_{i,i+1}} = \frac{f_i + f_{i+1}}{n_i + n_{i+1}}.$$

The new sequence is then

$$\frac{f_1}{n_1}, \frac{f_2}{n_2}, \dots, \frac{f_{i-1}}{n_{i-1}}, \frac{F_{i,i+1}}{N_{i,i+1}}, \frac{f_{i+2}}{n_{i+2}}, \dots, \frac{f_k}{n_k} .$$

If this sequence still contains a reversal, a pair of consecutive fractions for which the first is greater than the second, replace the pair with a single term as above. This process is continued until one obtains a non-decreasing sequence:

$$\frac{\phi_1}{n_1}, \frac{\phi_2}{n_2}, \frac{\phi_3}{n_3}, \dots,$$

where $\frac{\phi_j}{n_j} = \frac{f_i + \dots + f_{i+s}}{n_i + \dots + n_{i+s}}$ for appropriate i and s .

4. The final estimates are given by

$$\hat{p}_i = \dots = \hat{p}_{i+s} = \frac{\phi_j}{n_j} .$$

5. Linear interpolation is used to compute the values of the response function between stimulus levels tested.

2. EXAMPLE. If the results of the experiment were as shown in Figure A1, the maximum likelihood estimate found by the method of reversals is as follows:

Shielding Thickness (Inches)	f_i/n_i	$\frac{F_{i,i+1}}{N_{i,i+1}}$	$\frac{\phi_j}{n_j}$
4/8	1/1		1.0
5/8	0/2		.3
6/8	1/2	1/4	.3
7/8	2/6		.3
8/8	1/9		.11
9/8	0/12		0
10/8	0/10		0

The shielding thickness corresponding to the .1 quantile is found by linear interpolation.

$$.13 \left[\begin{array}{ccc} \frac{8/8 - 1}{.01} & .11 & .01 \\ & .1 & \\ 9/8 - 1.13 & & .0 \end{array} \right] .11$$

The shielding thickness associated with the .1 quantile is 1.01 inches.

FIGURE 1

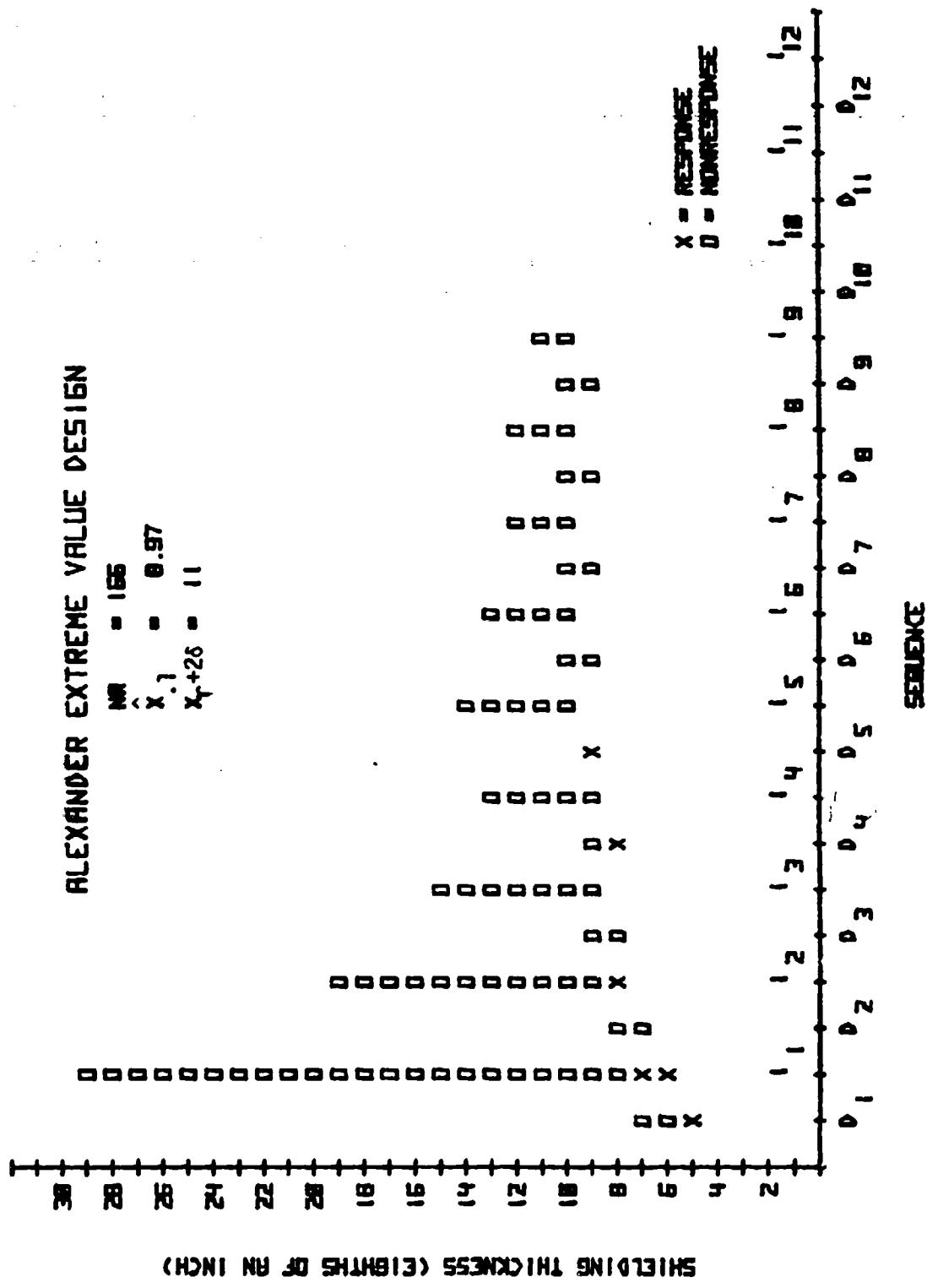


FIGURE 2

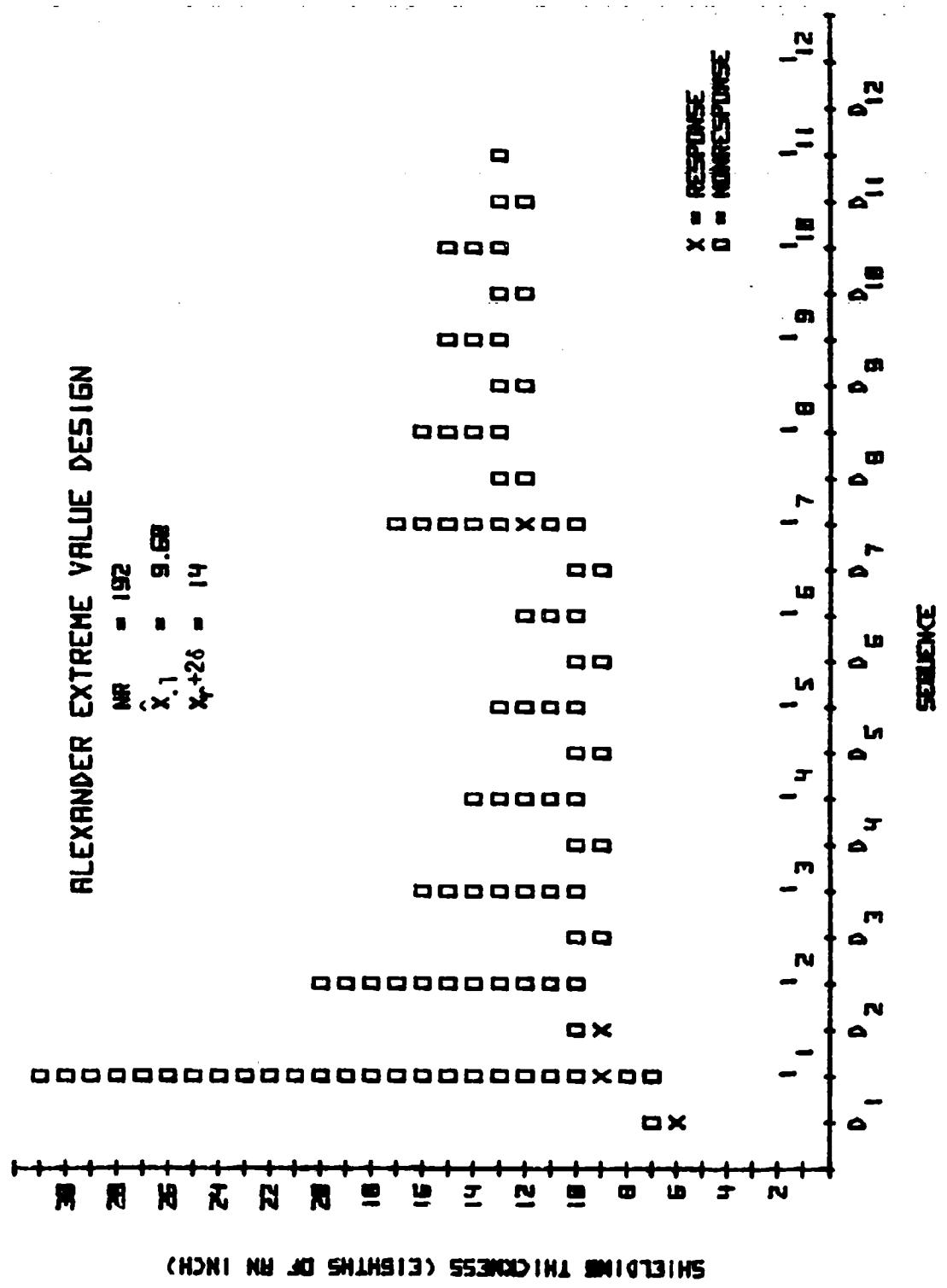


FIGURE 3

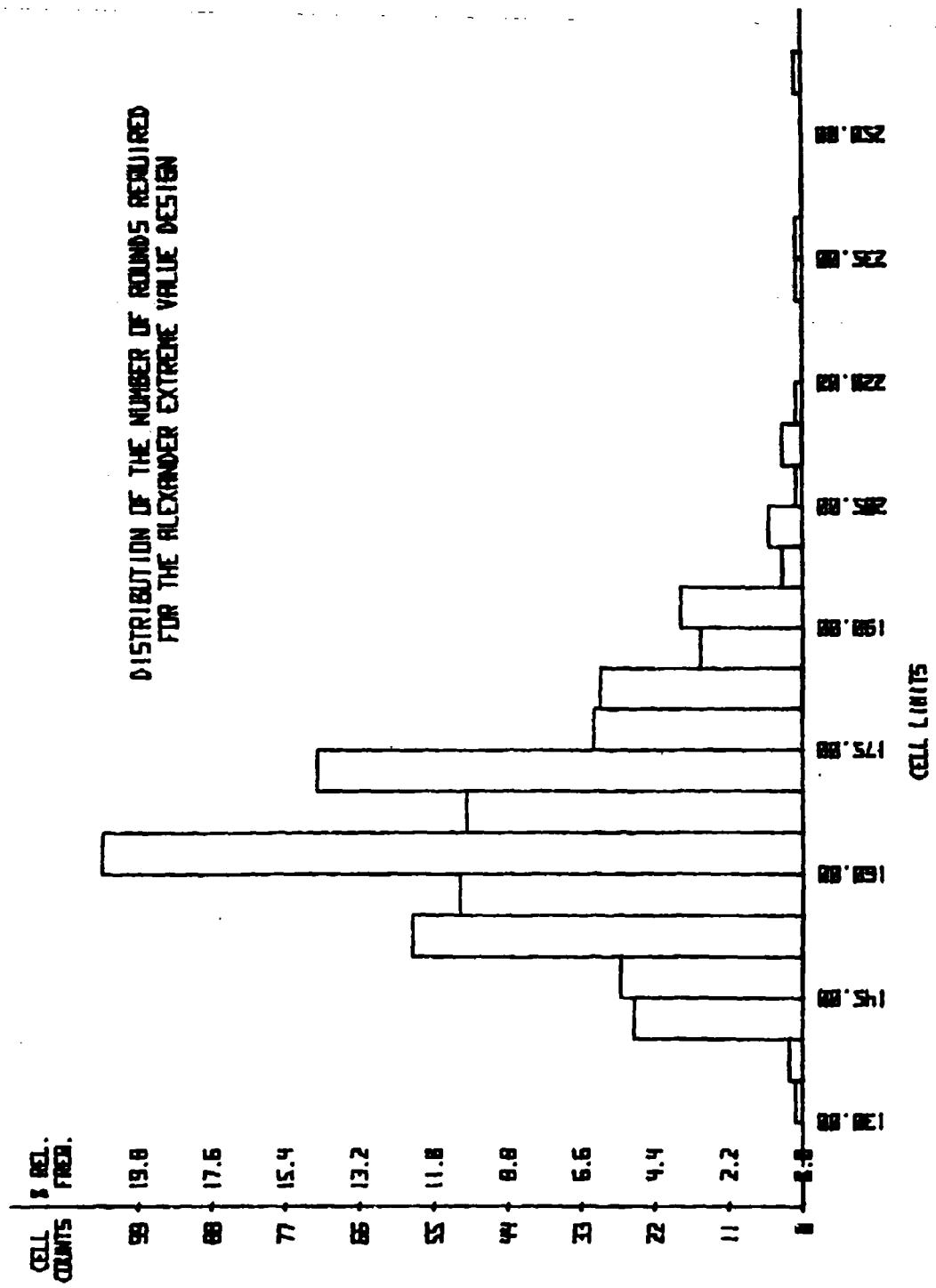


FIGURE 4

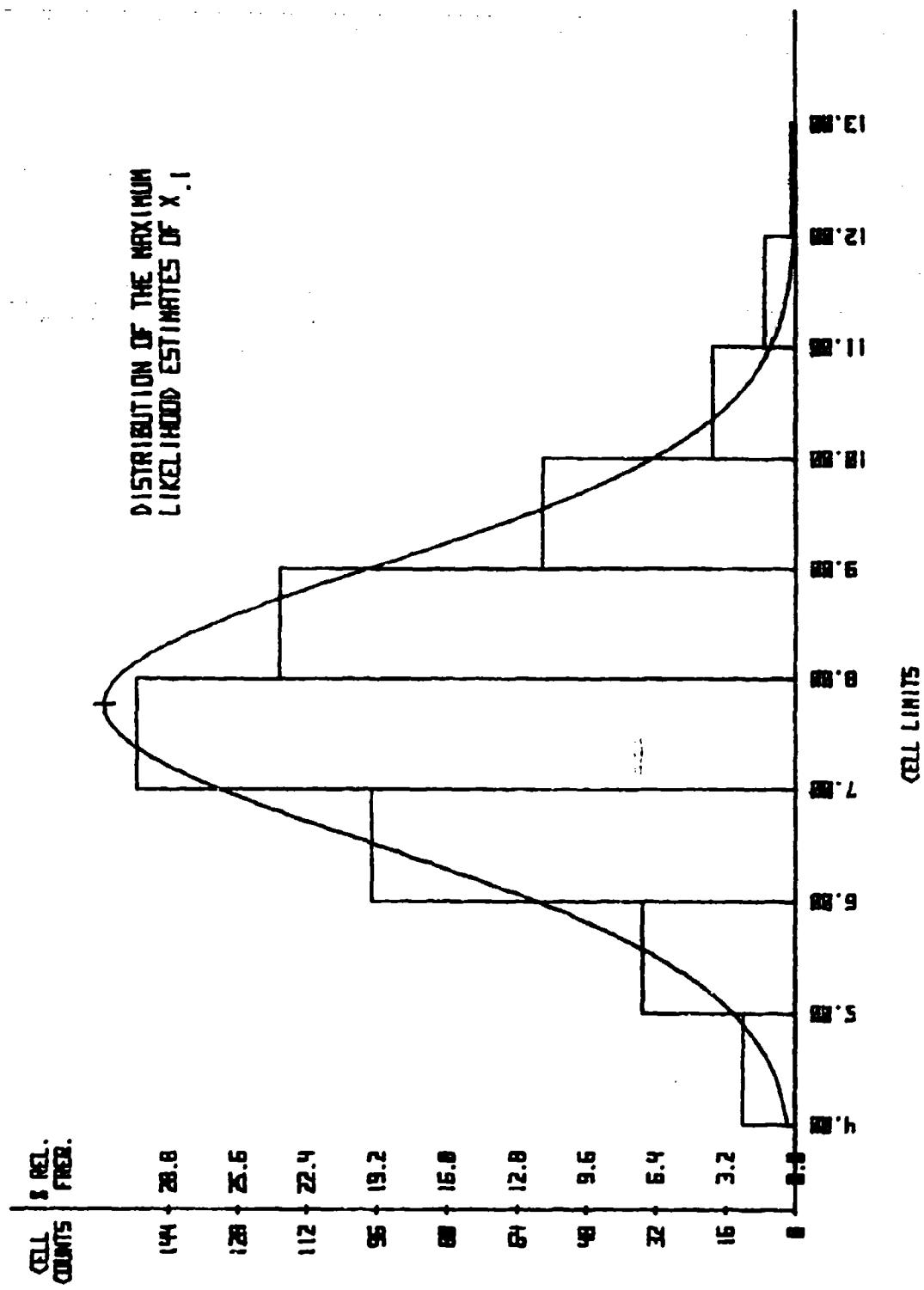
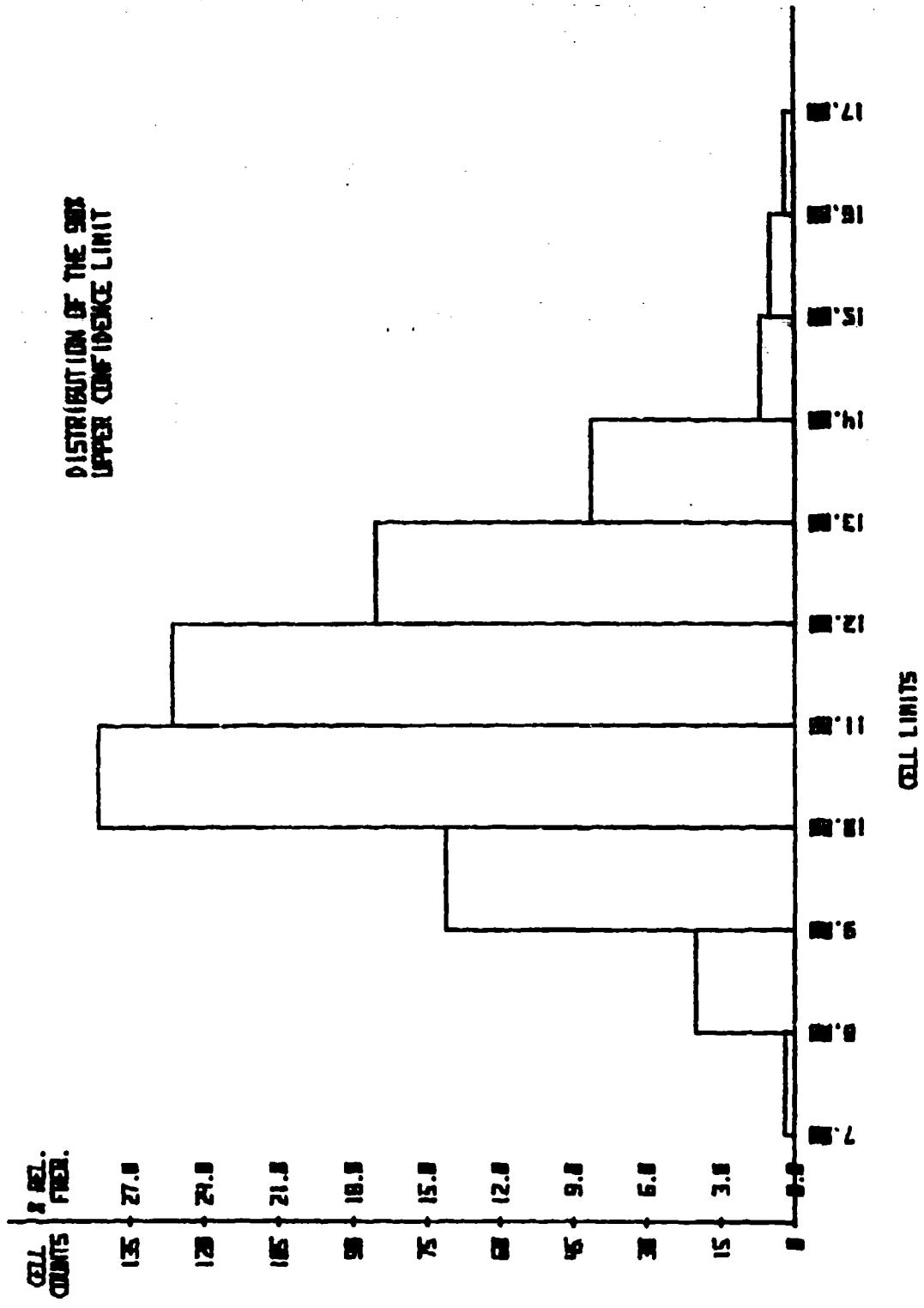


FIGURE 5



MODIFIED TEST CONFIGURATION

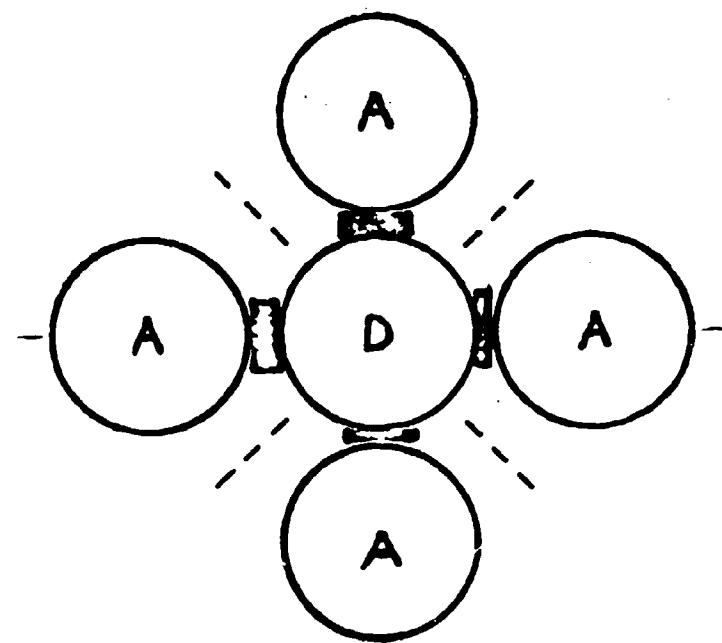
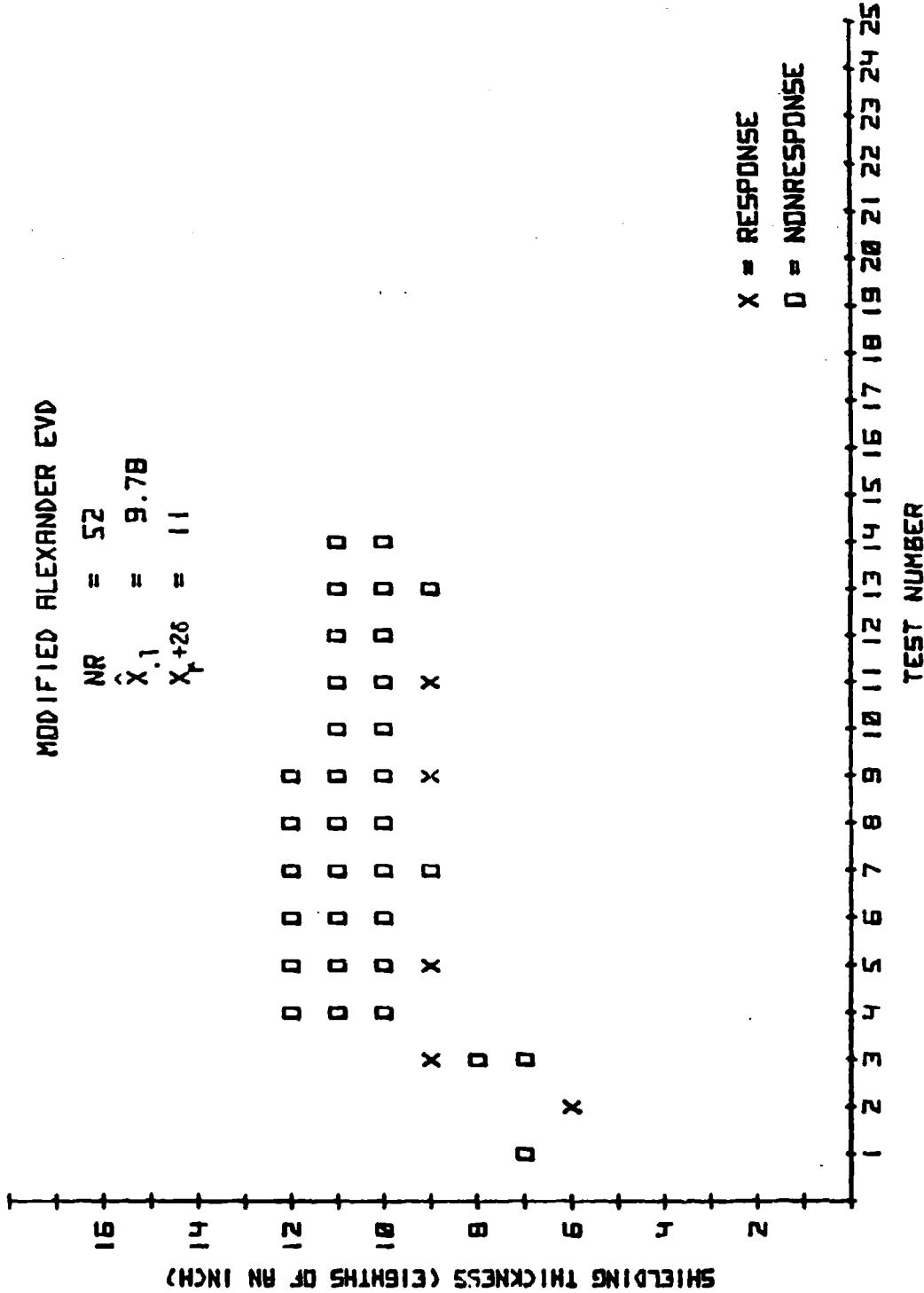


FIGURE 6

MODIFIED ALEXANDER EVD

$$\begin{aligned} NR &= 52 \\ \bar{X}_1 &= 9.78 \\ \bar{X}_{+26} &= 11 \end{aligned}$$



X = RESPONSE

O = NONRESPONSE

TEST NUMBER

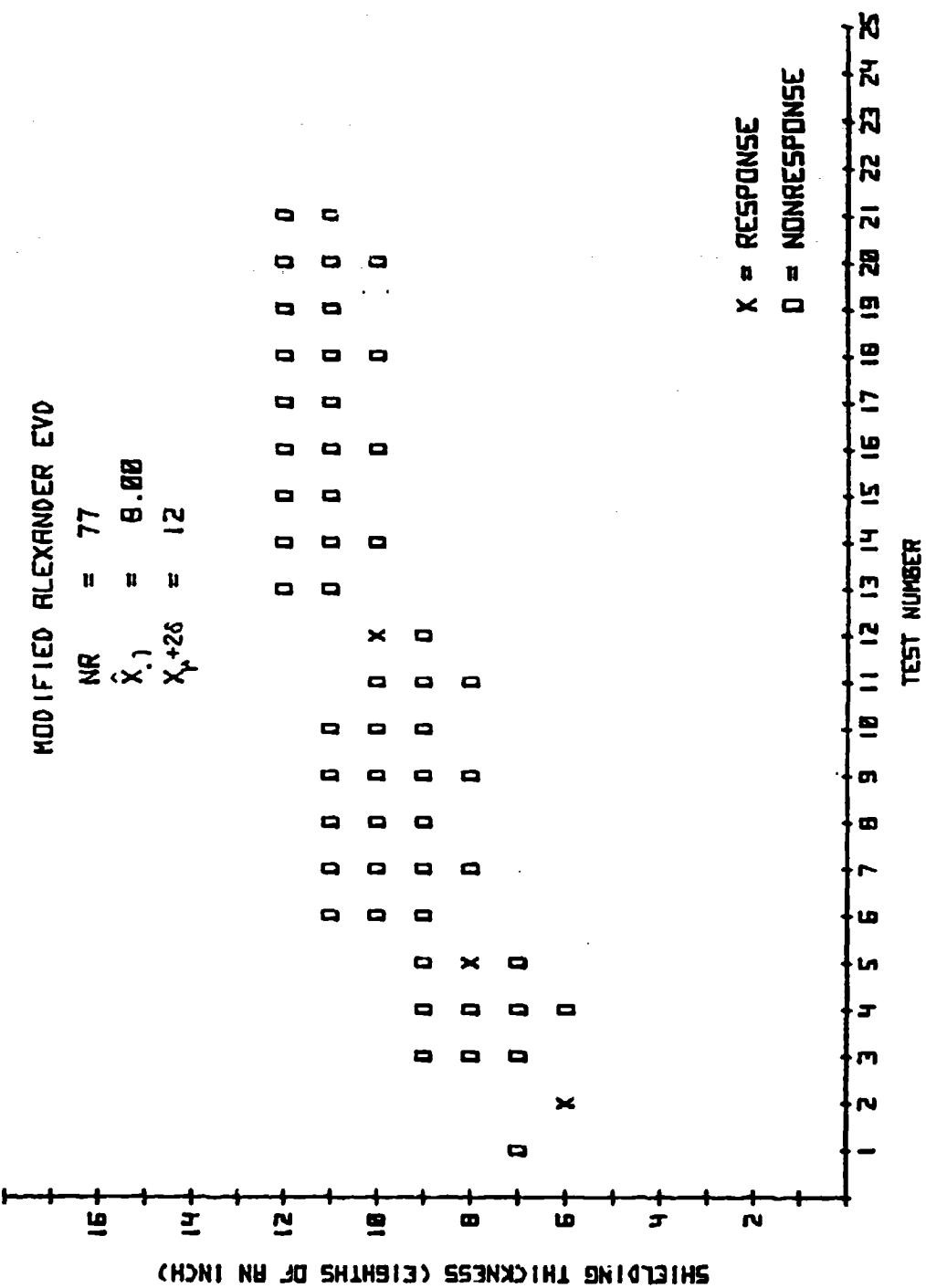


FIGURE 8

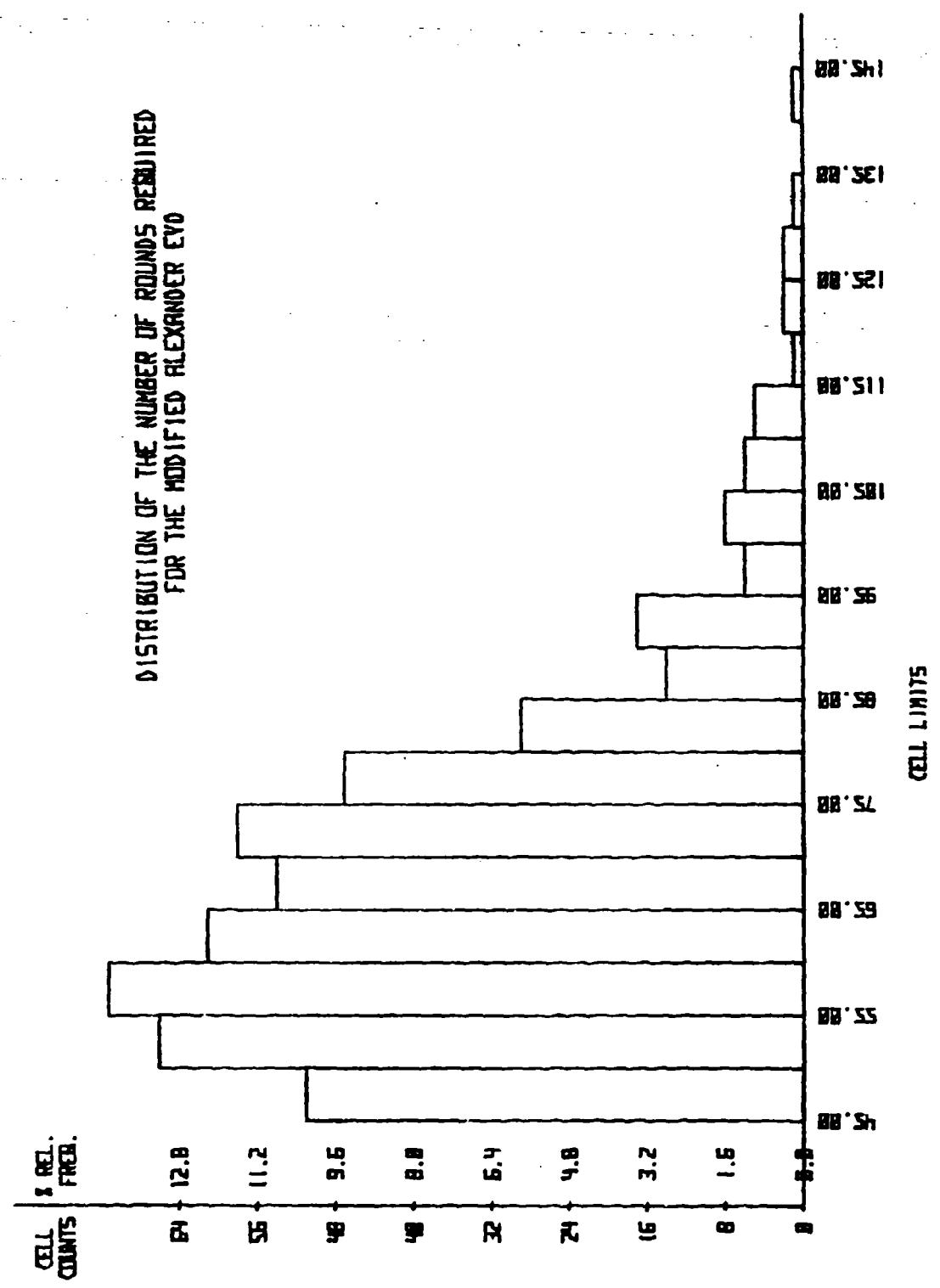
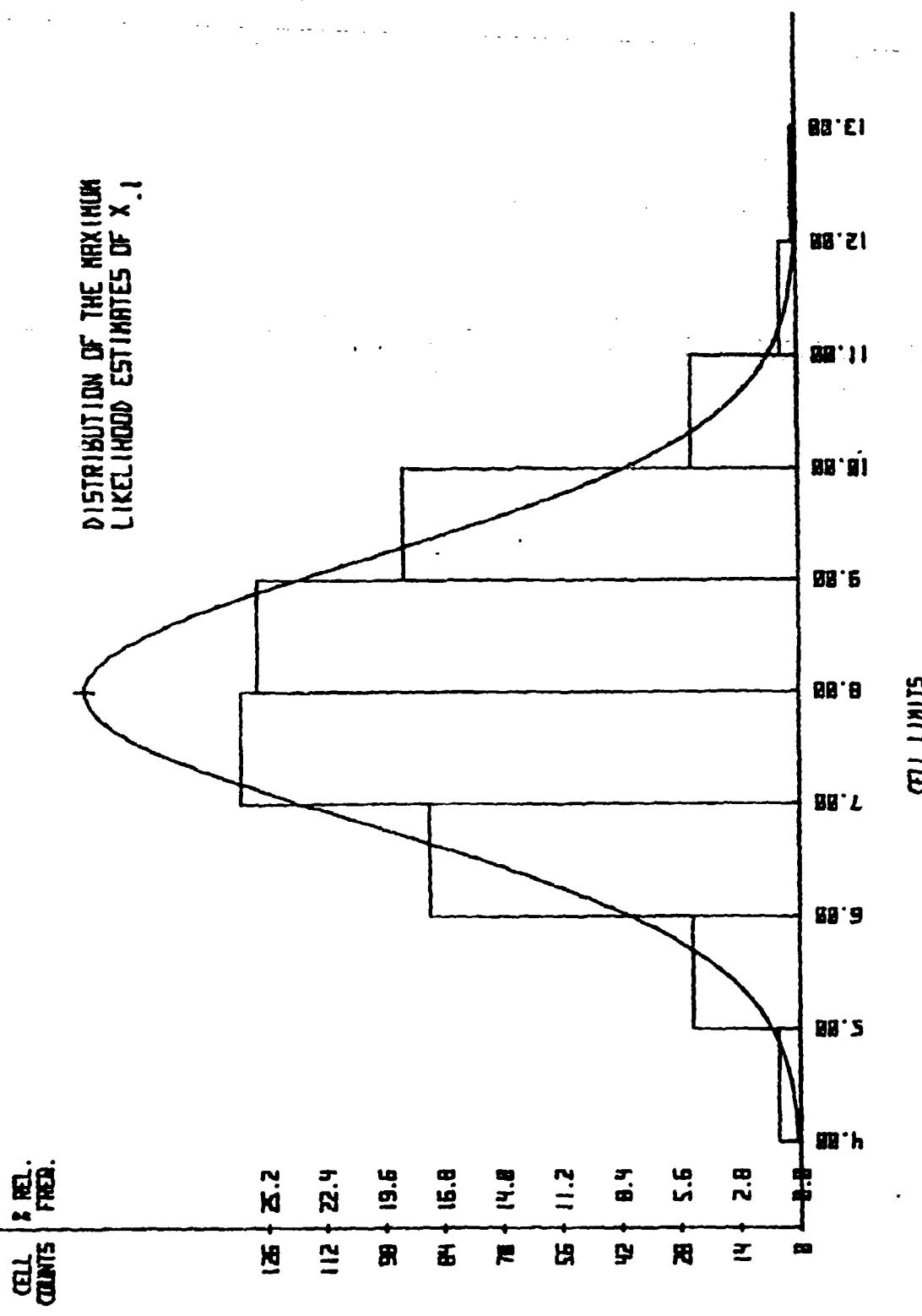


FIGURE 10



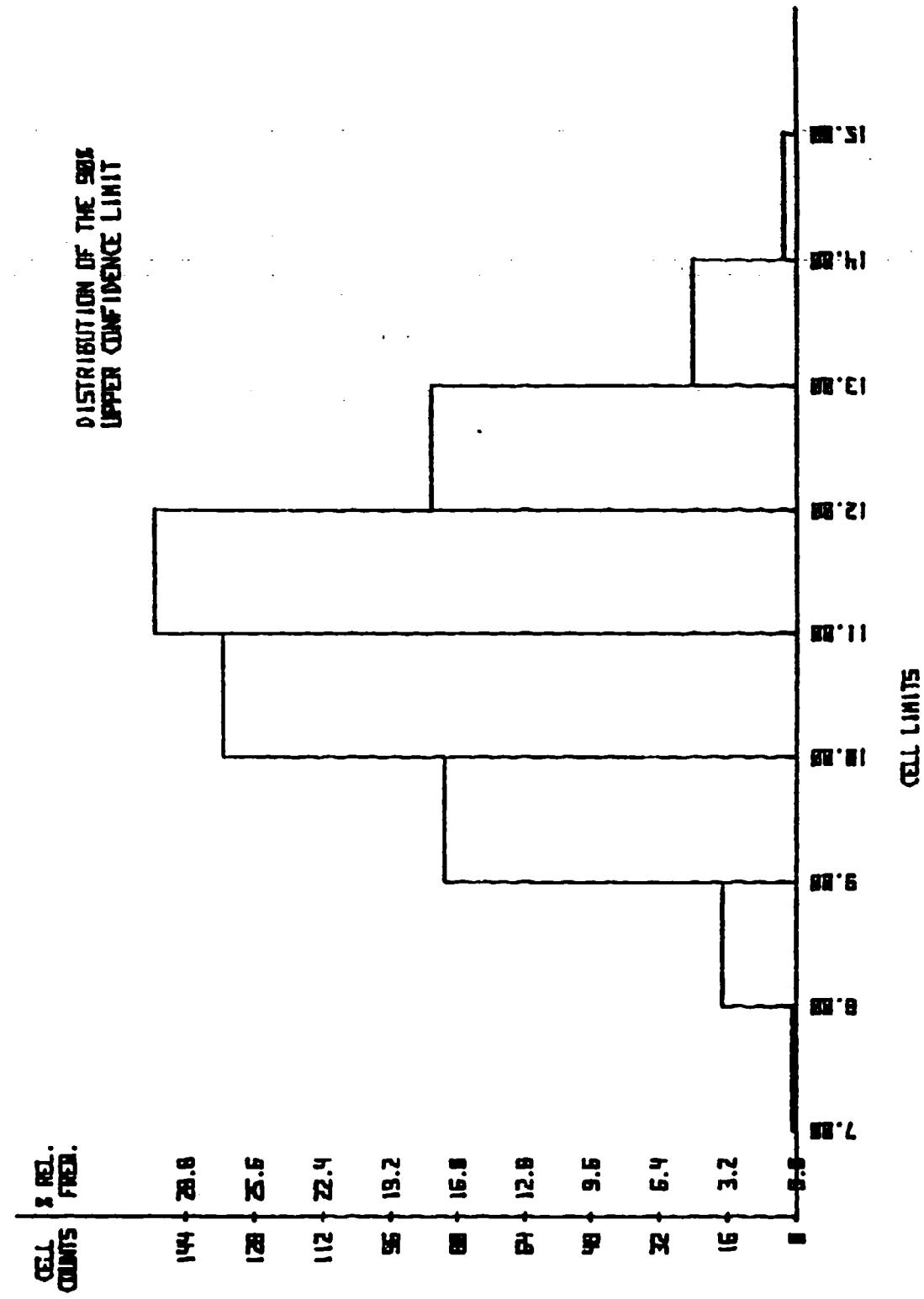
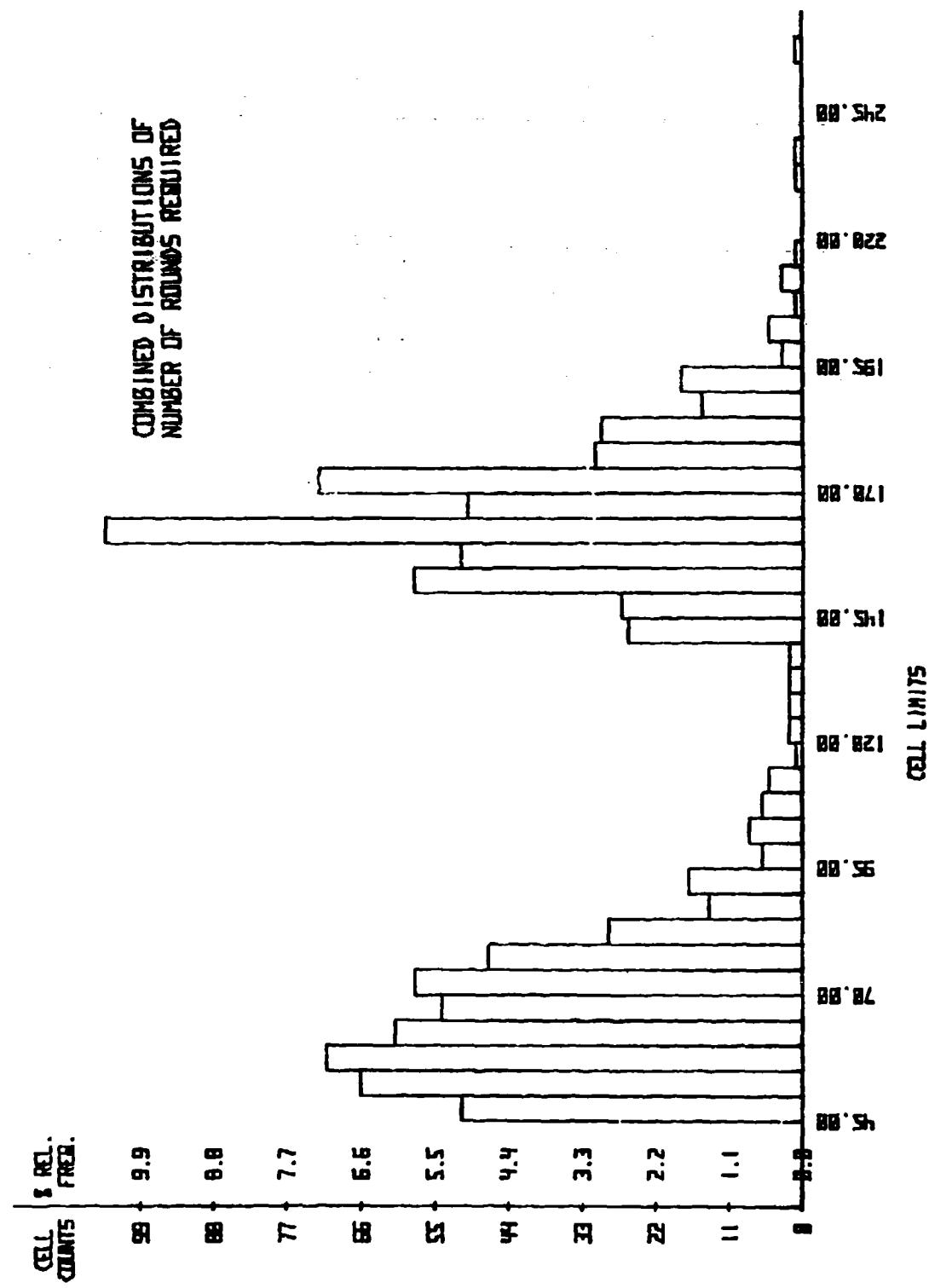


FIGURE 12



The Rank Transformation as a Robust
and Powerful Tool for the Analysis of Experimental Data

W. J. Conover, Texas Tech University

Abstract

Rank Transformation procedures are ones in which the usual parametric procedure is applied to the ranks of the data instead of to the data themselves. In the one way layout the rank transformation procedure is equivalent to the Kruskal-Wallis test. Simulation results using various distributions show that this procedure tends to have more power than either the F test or Fisher's randomization test, a well known nonparametric procedure.

The rank transformation procedure for the two way layout is compared with the F test and Fisher's randomization test under normality and several types of nonnormality. Overall the rank transformation procedure seems to be the best.

The Fisher's LSD multiple comparisons procedure in the one way and two way layouts is compared with a randomization procedure and with the same procedure computed on ranks. In nonnormal situations the rank transformation procedure appears to maintain power better than Fisher's LSD or the randomization procedures. The conclusion of this study is that the rank transformation provides a reasonable alternative to the usual analysis of experimental designs.

PRECEDING PAGE BLANK-NOT FILMED

1. INTRODUCTION

Three methods for analyzing experimental data are compared in this study. The first is the standard analysis of variance procedure based on the assumption of normality and assumed to be robust in most situations encountered in practical applications. The second is a randomization procedure attributed to R.A. Fisher (1935), which is known to be "most efficient" in some sense, and is assumed by many practitioners to be the best test one could possibly use, although it is difficult to use even with a computer. The third procedure involves a rank transformation of the data prior to the application of the first procedure, that is it is an analysis of variance on the ranks.

These three procedures are compared in a completely randomized design (a one-way layout) and in a randomized block design. Other designs could just as easily have been selected for comparison, but the randomization test involves such extensive computer time that only a limited study is possible. The robustness of all three procedures is estimated under the null hypothesis by computer simulation, and the power is estimated under the assumed existence of treatment effects, also by computer simulation. A multiple comparisons procedure is used whenever the null hypothesis is rejected, and comparisons of the three multiple comparisons procedures are made also.

These results were obtained by Ronald L. Iman of Sandia National Laboratories in some joint research work with the author. More extensive results appear in the unpublished manuscripts by Iman and Conover (1980a and 1980b) and by Conover and Iman (1980).

The Rank Transformation as a Robust
and Powerful Tool for the Analysis of Experimental Data

W. J. Conover, Texas Tech University

Abstract

Rank Transformation procedures are ones in which the usual parametric procedure is applied to the ranks of the data instead of to the data themselves. In the one way layout the rank transformation procedure is equivalent to the Kruskal-Wallis test. Simulation results using various distributions show that this procedure tends to have more power than either the F test or Fisher's randomization test, a well known nonparametric procedure.

The rank transformation procedure for the two way layout is compared with the F test and Fisher's randomization test under normality and several types of nonnormality. Overall the rank transformation procedure seems to be the best.

The Fisher's LSD multiple comparisons procedure in the one way and two way layouts is compared with a randomization procedure and with the same procedure computed on ranks. In nonnormal situations the rank transformation procedure appears to maintain power better than Fisher's LSD or the randomization procedures. The conclusion of this study is that the rank transformation provides a reasonable alternative to the usual analysis of experimental designs.

2. THE COMPLETELY RANDOMIZED DESIGN

Let X_{ij} , $1 \leq i \leq n_j$, $1 \leq j \leq k$ be random variables representing the i th observation in treatment j in a completely randomized design. Let $\bar{X}_{.j}$ and $\bar{X}_{..}$ represent the sample treatment mean and the overall mean respectively. The F statistic is given by

$$F = \frac{(N-K) \sum_j n_j (\bar{X}_{.j} - \bar{X}_{..})^2}{(k-1) \sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2} \quad (2.1)$$

Where $N = \sum n_j$ is the total sample size. The F test compares the F statistic with the F distribution, $k-1$ and $n-k$ degrees of freedom, and rejects the null hypothesis of equal treatment means if F is in the upper α tail of the F distribution. Such a test is exact under assumptions of identical normal distributions, but is robust even for some nonnormal distributions. If the null hypothesis is rejected, Fisher's LSD procedure is used to declare treatments j_1 and j_2 significantly different when the inequality

$$|\bar{X}_{.j_1} - \bar{X}_{.j_2}| > t_{\alpha/2, N-k} \sqrt{\text{MSE}} \sqrt{\frac{1}{n_{j_1}} + \frac{1}{n_{j_2}}} \quad (2.2)$$

is satisfied, where

$$\text{MSE} = \frac{1}{N-K} \sum_i \sum_j (X_{ij} - \bar{X}_{.j})^2 \quad (2.3)$$

and where $t_{p,m}$ is the $(1-p)$ quantile from a student's t distribution with m degrees of freedom.

For Fisher's randomization test the F statistic from Equation (2.1) is compared with the distribution of all possible F statistics arising from the $N!/\prod_j (n_j)!$ ways the same N observations can be partitioned into k groups of size n_j each, $j=1, \dots, k$. In practice, even with high speed computers and moderate

sample sizes the total number of combinations is too large to handle, so the suggestion of Dwass (1957) is followed. That is, a random subset of the total number possible is used to obtain an unbiased and consistent estimate of the distribution function of the randomization statistic. In this paper $k=4$ and (n_1, n_2, n_3, n_4) is $(7, 8, 9, 10)$. A subset of 1000 partitions, out of the more than 10^{18} partitions possible, was used to estimate α_{hat} .

Whenever the α_{hat} was 5% or less multiple comparisons were made using a procedure similar to that described above, only restricting the permutations to the ways the observations in the two samples being compared can be partitioned. Here again, only 1000 of the possible permutations were used for each comparison. The treatments were considered significantly different if the observed value of $|\bar{X}_{j_1} - \bar{X}_{j_2}|$ was among the largest 5% obtained.

The third test consists of replacing data by the ranks from 1 to N , and performing an F test on the ranks. This is equivalent to the Kruskal-Wallis test. Multiple comparisons were made by computing (2.2), as in the Fisher LSD procedure, but using the same ranks used above instead of the data. These three procedures are called the F, R and RT methods respectively.

Comparisons of these three tests were made for three population distributions, the normal, lognormal and exponential distributions. The null case was examined, along with three non-null settings corresponding to slight, medium, and strong differences in treatment effects. The parameters used are summarized in Table 1.

In each of these 12 combinations of distributions with treatment effects, 500 replications were made to compare the robustness and power of the three tests. These results are given in Table 2. They show that the Fisher randomization test and the rank transform test are robust for all three distributions, as expected because they are both nonparametric procedures. The F test

TABLE 1. The population effects used in the completely randomized design simulation study: Means of the normal ($\sigma^2 = 4$), means of the log of the lognormal (σ^2 of logs = 4), means of the exponential.

<u>EFFECTS</u>	<u>NORMAL</u>	<u>LOGNORMAL</u>	<u>EXPONENTIAL</u>
Null	(0,0,0,0)	(0,0,0,0)	(1,1,1,1)
Slight	(0,0,0,1)	(0,0,0,1)	(2,2,2,3)
Medium	(0,0,1,2)	(0,0,1,2)	(1,1,2,3)
Strong	(0,1,2,3)	(0,1,2,3)	(1,2,3,4)

TABLE 2. The percent of time the null hypothesis was rejected in the completely randomized design, four treatments, $n_1=7$, $n_2=8$, $n_3=9$, $n_4=10$.

<u>EFFECTS</u>	<u>NORMAL</u>			<u>LOGNORMAL</u>			<u>EXPONENTIAL</u>		
	<u>R</u>	<u>F</u>	<u>RT</u>	<u>R</u>	<u>F</u>	<u>RT</u>	<u>R</u>	<u>F</u>	<u>RT</u>
Error Rate in Null Case:	5%	5%	5%	6%	2%	7%	4%	4%	5%
Power Under									
Slight Effects:	19%	19%	19%	9%	4%	17%	12%	10%	12%
Medium Effects:	52%	52%	49%	18%	12%	43%	47%	40%	46%
Strong Effects:	72%	72%	70%	22%	13%	69%	43%	37%	53%

on the other hand is robust for the normal and exponential distributions, but quite conservative for the lognormal distribution. The conservative nature of the F test carries over to inhibit its power for detecting differences in lognormal distributions. The rank transform procedure shows the most power in the lognormal and exponential cases, and about the same power as the other two procedures when the distributions are normal.

When the null hypothesis was rejected using the previous procedures, the corresponding multiple comparisons tests were made as previously described. The results, summarized in Table 3, show the same types of results as in Table 2.

TABLE 3. The number of times treatment pairs were declared significantly different in 500 simulations, using CR design with 4 treatments, $n_1 = 7$, $n_2 = 8$, $n_3 = 9$, $n_4 = 10$.

EFFECTS	TREATMENT PAIR	NORMAL			LOGNORMAL			EXPONENTIAL		
		R	F	RT	R	F	RT	R	F	RT
<u>Null</u>	1,2*	6	6	7	10	9	8	6	8	9
	1,3*	5	6	5	8	7	7	5	9	9
	2,3*	6	4	5	7	7	12	10	10	10
	1,4*	4	4	3	16	5	9	4	4	8
	2,4*	5	5	4	10	6	9	3	5	7
	2,4*	6	6	5	4	3	6	6	5	11
<u>Slight</u>	1,2*	17	21	17	3	5	17	7	7	13
	1,3*	14	19	12	4	4	13	10	7	12
	2,3*	21	23	20	4	2	22	7	4	15
	1,4	54	53	51	11	20	57	20	24	23
	2,4	56	53	52	11	18	57	20	28	15
	3,4	48	48	54	16	18	51	30	34	24
<u>Medium</u>	1,2*	16	24	19	10	0	20	9	0	13
	1,3	63	73	77	8	4	71	56	37	93
	2,3	81	78	83	15	5	61	75	42	89
	1,4	189	200	198	26	38	184	143	174	186
	2,4	206	210	203	45	36	180	172	180	185
	3,4	81	79	75	36	36	86	54	86	63
<u>Strong</u>	1,2	69	77	64	9	0	60	40	8	64
	1,3	207	220	209	22	4	210	86	56	146
	2,3	64	80	88	14	4	90	36	27	49
	1,4	332	338	339	49	59	328	148	137	208
	2,4	220	236	235	51	58	233	84	108	95
	3,4	79	81	76	39	49	87	36	76	38
<u>Simple totals:</u>										
Identical populations		100	118	97	76	48	123	67	59	107
Some effects present		1749	1826	1804	362	349	1775	1000	1017	1278

*These populations are identical.

That is, the LSD procedure on the ranks has more overall power to detect differences where they exist than the other two types of procedures do. In summary, for the CR design the transformation to ranks prior to the usual analysis improves the robustness and power of the usual analysis in nonnormal situations without losing much of the fine qualities of the usual analysis in the normal situation.

3. THE RANDOMIZED COMPLETE BLOCK DESIGN

Let X_{ij} , for $1 \leq i \leq b$ and $1 \leq j \leq k$, be random variables associated with the i th block and the j th treatment, and let $\bar{X}_{i.}$, $\bar{X}_{.j}$ and $\bar{X}_{..}$ be the sample block, treatment and grand means respectively. The F statistic is given by

$$F = \frac{\frac{b(b-1)}{k} \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2}{\sum_i \sum_j (X_{ij} - \bar{X}_{.j} - \bar{X}_{i.} + \bar{X}_{..})^2} \quad (3.1)$$

The parametric F test compares the F statistic with quantiles of the F distribution with $k-1$ and $(b-1)(k-1)$ degrees of freedom. These quantiles are exact under normality, additivity, and equal variances, and are reasonable approximations under mild violations of the normality assumption. If the F statistic is in the upper α tail of the F distribution, the null hypothesis of equal treatment means is rejected, and multiple comparisons are made. Treatments j_1 and j_2 are declared significantly different if the inequality

$$|\bar{X}_{.j_1} - \bar{X}_{.j_2}| > t_{\alpha/2, (b-1)(k-1)} \sqrt{\frac{2(SSE)}{b(b-1)(k-1)}} \quad (3.2)$$

is satisfied, where SSE is the denominator of Equation (3.1), and where $t_{p,m}$ is the p th quantile from a t distribution with m degrees of freedom. This is the well known Fisher's LSD procedure.

For Fisher's randomization test, as presented by Welch (1937) and Pitman (1938), the F statistic is used, but not the F distribution. The F statistic is computed for each of the $(k!)^b$ configurations of the observations, obtained by permuting the observations within blocks. If the observed F statistic is one of the $(k!)^b \cdot \alpha$ largest of these, the null hypothesis is rejected. In this study $k=3$ and $b=5$, so the $(3!)^5(.05) = 384$ largest values of F constitute the critical region. (The actual value 388.8 is rounded down to the first multiple of 6,

because the configurations appear in a multiplicity of 6 and the alpha level should be $\leq .05$.) The actual number of possible F values greater than or equal to the observed F value is divided by $(k!)^b$ to obtain "alphahat," sometimes known as the p value or the critical level.

If alphahat is less than or equal to .05, multiple comparisons are made by permuting only those observations in the treatment pair being considered. Because there are effectively only $(2!)^5 = 32$ different permutations, the treatment pair is declared significantly different if and only if all pairwise differences have the same sign, at a level of significance $2/32 = .0625$. For comparison purposes these same values of .05 and .0625 were used in the F test described previously and in the following test.

The third test is a rank transform procedure found by Iman (1974) and Conover and Iman (1976) to have good properties of power and robustness in randomized block designs. First all $b \cdot k$ observations are replaced by their ranks from 1 to $b \cdot k$. The F statistic of Equation (3.1) is computed on these ranks and compared with the F distribution, $k-1$ and $(b-1)(k-1)$ degrees of freedom as an approximation procedure, just as in the first method described. Multiple comparisons are made using Equation (3.2) just as in the parametric case, but using the same ranks used above rather than reranking each pair of samples in a Mann-Whitney fashion.

This study examines normal, lognormal and exponential distributions, under the null case and with slight, medium and strong treatment effects. Under each of these 12 population-treatment combinations 500 replications were made, and the three tests conducted. Thus 6000 computations of the F test (F) and the rank transform procedure (RT) were made, and 7,776,000 F statistics were computed for the randomization test as a different null distribution must be found

TABLE 4. The population effects present in the simulation study: means of the normal ($\sigma^2 = 4$), means of the log of the lognormal (σ^2 of logs = 4), means of the exponential. Add block effects (1, 2, 3, 4, 5) to the means in the five blocks.

<u>EFFECTS</u>	<u>NORMAL</u>	<u>LOGNORMAL</u>	<u>EXPONENTIAL</u>
Null	(0,0,0)	(0,0,0)	(0,0,0)
Slight	(0,0,1)	(0,0,1)	(0,0,1)
Medium	(0,1,2)	(0,1,2)	(0,4,6)
Strong	(0,1,3)	(0,1,3)	(0,7,9)

in each case. Specific values of the parameters used are listed in Table 4.

The results of the three tests are summarized in Table 5 for the twelve situations described in Table 4. The results are similar to the results for CR designs presented in the previous chapter. That is, the usual F test on the ranks has better robustness and power in the nonnormal cases examined than the F test on the data, and essentially the same robustness and power in the normal situation. The randomization procedure has power somewhere between the power of the other two tests.

TABLE 5. The percent of time the null hypothesis was rejected in the randomized complete blocks design, five blocks, three treatments.

<u>EFFECTS</u>	<u>NORMAL</u>			<u>LOGNORMAL</u>			<u>EXPONENTIAL</u>		
	<u>R</u>	<u>F</u>	<u>RT</u>	<u>R</u>	<u>F</u>	<u>RT</u>	<u>R</u>	<u>F</u>	<u>RT</u>
Error Rate									
in Null Case:	5%	5%	5%	5%	1%	5%	5%	3%	4%
Power Under									
Slight Effects:	10%	10%	10%	8%	1%	8%	10%	7%	10%
Medium Effects:	22%	22%	21%	15%	2%	23%	18%	12%	23%
Strong Effects:	42%	42%	42%	34%	4%	41%	20%	17%	27%

TABLE 6. The number of times treatment pairs were declared significantly different in 500 simulations, using an RCB design with three treatments, five blocks, one observation per cell.

EFFECTS	TREATMENT PAIR	NORMAL			DISTRIBUTION			EXPONENTIAL		
		R	F	RT	R	F	RT	R	F	RT
<u>NULL</u>	1,2*	11	14	17	9	2	13	9	8	13
	2,3*	12	12	18	9	4	15	12	9	13
	1,3*	7	11	13	12	4	18	12	10	10
SLIGHT	1,2*	15	16	21	3	0	18	4	4	10
	2,3	34	41	42	17	4	30	18	19	23
	1,3	29	36	33	18	4	27	11	17	23
MEDIUM	1,2	41	51	52	16	2	50	37	17	81
	2,3	37	48	50	37	6	53	26	42	37
	1,3	74	97	96	61	5	110	60	53	109
STRONG	1,2	51	64	65	15	0	56	51	45	106
	2,3	96	136	135	105	20	117	22	40	27
	1,3	163	204	204	117	19	198	64	57	118
SIMPLE TOTALS:										
IDENTICAL POPULATIONS		45	53	69	33	10	64	37	31	46
SOME EFFECTS PRESENT		525	677	677	386	60	641	289	290	524

*These populations are identical.

Multiple comparisons were made when the null hypothesis was rejected using the previous tests. The multiple comparisons results given in Table 6 are similar to the results obtained for the CR design in the previous section.

Overall the rank transformation allows more real differences to be detected than when either of the other two procedures is used.

4. CONCLUSIONS

The usual F test, followed by Fisher's LSD procedure for multiple comparisons, shows approximately the same robustness and power as Fisher's randomization test and the rank transform procedure when the populations are normal, slightly less power than the other two procedures with exponential distributions, and considerably less power than the other two procedures when the distributions are lognormal. This latter result may be due in part to the extreme conservative nature of the parametric procedure under the lognormal distribution, or it may be due in part to the nonhomogeneity of variances in the models considered. Nonhomogeneity of variances is a natural consequence of positive valued data when the means are different. It occurs often in actual data analysis, so no attempt was made to alter the situation in this study either.

Fisher's randomization test is a difficult and time consuming procedure to use in experimental designs. This study indicates that the extra work required is probably not justified, because while Fisher's randomization test shows better power and robustness overall than the F test on the untransformed data, it compares unfavorably with the F test on the ranks of the data.

The F test on the ranks of the data, with the subsequent LSD procedure on the ranks, is an easy procedure to use. It has essentially the same power as the F test in normal situations, and more power than either the F test or Fisher's randomization test when populations are lognormal or exponential, at least in the cases studied.

REFERENCES

- Conover, W. J. and Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs, Comm. in Statist. - Theo. and Meth., A5, 1349-1368.
- Conover, W. J. and Iman, Ronald L. (1980). Small sample efficiency of Fisher's randomization test when applied to experimental designs. Presented at the National Meeting of the American Statistical Association, Houston, August.
- Dwass, Meyer (1957). Modified randomization tests for nonparametric hypotheses. The Annals of Mathematical Statistics, 28, pp. 181-187.
- Fisher, R. A. (1935). The Design of Experiments. New York: Hafner.
- Iman, R. L. (1974). A power study of a rank transform for the two-way classification model when interaction may be present. Canad. J. of Statist. Sect. C: Appl., 2, pp. 227-239.
- Iman, Ronald L. and Conover, W. J. (1980a). Multiple comparisons procedures based on the rank transformation. Presented at the National Meeting of the American Statistical Association, Houston, August.
- Iman, Ronald L. and Conover, W. J. (1980b). A comparison of distribution free procedures for the analysis of complete blocks. Presented at the National Meeting of the American Institute of Decision Sciences, Las Vegas, November.
- Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. Biometrika, 29, pp. 322-335.
- Welch, B. L. (1937). On the z-test in randomized blocks and latin squares. Biometrika, 29, pp. 21-52.

ALTERNATIVE QUANTILE ESTIMATION

W. D. Kaigh
The University of Texas at El Paso

Abstract. An alternative to the conventional sample quantile is proposed as a nonparametric estimator of a continuous population quantile. The alternative estimator is a "generalized sample quantile" obtained by averaging an appropriate subsample quantile over all subsamples of a fixed size. Since the resulting statistic is a U-statistic with representation also as a linear combination of order statistics, known results are employed then to establish asymptotic normality. The alternative estimator is shown to be asymptotically efficient in the class of nonparametric models specified by Pfanzagl (1975). Analytic results and Monte Carlo studies with moderate sample sizes indicate that the proposed estimator usually produces mean square error of estimation less than that of the conventional sample quantile and also jackknifes to provide approximate confidence intervals.

1. Introduction. Suppose that F is an absolutely continuous c.d.f. with corresponding p.d.f. f . For $0 < u < 1$ let $G(u) = \inf \{x: F(x) = u\}$ be an inverse of F and denote the derivative $G'(u) = 1/f[G(u)]$ when it exists. For $0 < p < 1$ define ξ_p to be the pth. quantile of F which satisfies $\xi_p = G(p)$.

We assume throughout that $f(\xi_p) > 0$.

Suppose that X_1, \dots, X_n are i.i.d. r.v.'s with c.d.f. F and denote the corresponding order statistics by $X_{1:n}, \dots, X_{n:n}$. Assuming no further information regarding F , the conventional estimator of the population quantile ξ_p is the pth. sample quantile $X_{[(n+1)p]:n}$, where $[x]$ denotes the integral part of x . The asymptotic distribution of the sample quantile is given by the following well known result (e.g. Wilks (1962), page 273):

$$(1.1) \quad n^{\frac{1}{2}}(X_{[(n+1)p]:n} - \xi_p) \xrightarrow{D} N(0, \sigma_p^2(F)) \text{ as } n \rightarrow \infty$$

where

$$\sigma_p^2(F) = p(1-p)/f^2(\xi_p) = p(1-p)[G'(p)]^2.$$

In a nonparametric context assuming a positive differentiable p.d.f., Pfanzagl (1975) has shown that the sample quantile is efficient among the class of all translation-equivariant and asymptotically median unbiased estimators. However, Reiss (1980) has demonstrated that quasiquantiles may perform considerably better than sample quantiles when comparisons are based on the notion of deficiency as introduced by Hodges and Lehmann (1970).

Kaigh and Lachenbruch (1981) propose and study another alternative to the sample quantile in an attempt to improve the precision of the estimation of population quantiles. The alternative estimator is a U-statistic with representation as a linear combination of order statistics and may be viewed as a "generalized sample quantile" obtained by averaging a sample quantile estimate over subsamples of the complete sample. Although subsampling schemes are common and, in fact, our generalized sample median was obtained first by Yanagawa (1969) as a robust estimator of location for symmetric distributions, the procedure provides a natural local "smoothing" of the entire sample quantile function. In a related study Harrell and Davis (1981) consider a similar quantile estimator obtained through application of the bootstrap.

In Section 2 we provide the introduction of the alternative estimator and a discussion of its elementary properties; in Section 3 we determine the asymptotic distribution of the alternative estimator as an application of known results concerning linear combinations of order statistics and U-statistics; in Section 4 we employ both analytic methods and Monte Carlo results to compare the alternative estimator with the conventional sample quantile estimator; finally, in Section 5 we jackknife the alternative estimator to obtain interval estimates for population quantiles.

2. The Alternative Estimator $K_{[(k+1)p]:k;n}$. For a fixed integer k satisfying $1 \leq k \leq n$, consider the selection of a simple random sample (without replacement) from the complete sample X_1, \dots, X_n and denote the ordered observations in the subsample by $Y_{1:k;n}, \dots, Y_{k:k;n}$. An elementary combinatorial argument shows that for each integer r satisfying $1 \leq r \leq k$

$$\Pr(Y_{r:k;n} = X_{j:n}) = \frac{\binom{j-1}{r-1} \binom{n-j}{k-r}}{\binom{n}{k}}, \quad r \leq j \leq r + n - k.$$

For $0 < p < 1$ a sample quantile estimator of ξ_p based on the observations in a single subsample would be $Y_{[(k+1)p]:k;n}$. We define the alternative quantile estimator $K_{[(k+1)p]:k;n}$ to be the subsample quantile averaged over all $\binom{n}{k}$ subsamples of size k so that

$$(2.1) \quad K_{[(k+1)p]:k;n} = \sum_{j=r}^{r+n-k} [(\binom{j-1}{r-1} \binom{n-j}{k-r}) / \binom{n}{k}] X_{j:n}, \quad r = [(k+1)p].$$

The estimator of (2.1) is obviously translation-equivariant (i.e., $K_{[(k+1)p]:k;n}(X_1+c, \dots, X_n+c) = K_{[(k+1)p]:k;n}(X_1, \dots, X_n) + c$) and satisfies $K_{[(k+1)p]:k;n} = E(Y_{[(k+1)p]:k;n}|X_1, \dots, X_n)$ with expectation $\mu_{r:k}(F)$, the mean of the $r = [(k+1)p]$ th. order statistic in a random sample of size k from F .

From the development through averaging the symmetric kernel $f^*(x_1, \dots, x_k) = x_{r:k}$ over all subsamples, it follows that $K_{[(k+1)p]:k;n}$ is a U-statistic with representation also as the linear combination of order statistics given by (2.1). In a specialized application to reliability theory Takahasi (1970) also considered the U-statistic above as an estimator of its mean $\mu_{r:k}(F)$. The weights which appear in the summation of (2.1) correspond to the probability distribution of a negative hypergeometric random variable representing the number of individual selections (without replacement) required to obtain a total of r "special

"items" from a dichotomous population consisting of exactly k "special items" and $n-k$ "ordinary items". The mean and mode of the negative hypergeometric distribution appearing in (2.1) are $r(n+1)/(k+1)$ and $[(r-1)n/(k-1)] + 1$, respectively, which indicate a weight function centered appropriately about $[np]$.

A sample quantile is not in general an unbiased estimator of the corresponding population quantile, although (1.1) shows that any bias becomes negligible with increasing sample size. Appeal to a monotonicity principle would suggest that the subsampling scheme provides an estimator $K_{[(k+1)p]:k;n}$ of ξ_p with bias magnitude exceeding that of the conventional estimator $X_{[(n+1)p]:n}$. However, it would seem plausible also that the averaging procedure might result in a reduction of sampling variability adequate to decrease the overall mean square error of estimation.

Subject to the obvious constraint $1 \leq k \leq n$, the assumed subsample size is arbitrary and the choice $k = n$ in (2.1) gives $K_{[(n+1)p]:n;n} = X_{[(n+1)p]:n}$ so the statistics defined by (2.1) form a collection of "generalized quantile estimators" which includes the usual sample quantile. As an illustration consider a complete sample size $n = 99$ and the estimation of $\xi_{0.05}$. Permissible subsample sizes are then $k = 19, 39, 59, 79, 99$ with corresponding $[(k+1)p] = 1, 2, 3, 4, 5$, where for convenience we have chosen to avoid the use of fractional order statistics (see Stigler (1977)) and adopted a convention that a quantile ξ_p is estimable from a sample of size k only if $(k+1)p$ is an integer. The estimation problem becomes that of choosing the subsample size appropriate to the minimization of $E(K_{[(k+1)p]:k;n} - \xi_p)^2$. Although the theory of U-statistics as developed by Hoeffding (1948) would suggest a choice of the minimal permissible subsample size to provide a ker-

nel estimator of minimum variance, the substitution $k=1$ in (2.1) provides the sample mean as an estimator of a possibly asymmetric population median. Obviously the minimization of mean square error of estimation requires consideration of bias magnitude as well as sampling variance.

Finally, since the exact distributional properties of U-statistics and of linear combinations of order statistics typically are intractable, our subsequent analyses are concerned with asymptotic development and simulation. Although the robustness properties of an averaging process are suspect, in practice the estimator $K_{[(k+1)p]:k;n}$ is computed from a trimmed sample and is quite robust provided that care is exercised to avoid the sample extremes.

3. Asymptotic Distribution of $K_{[(k+1)p]:k;n}$. Our immediate objective is to obtain asymptotic distribution results for the statistic $K_{[(k+1)p]:k;n}$ to facilitate comparison of the estimators introduced in Section 2. Theorem 3.1 requires a fixed subsample size k whereas theorem 3.2 considers a subsample size increasing in proportion with the total sample size n .

For a fixed subsample size we first formulate and then apply the results of Hoeffding (1948). Suppose X_1, \dots, X_n are i.i.d. r.v.'s and let $f^*(X_1, \dots, X_m)$ be a real-valued symmetric statistic with mean η and second moment $E[f^*(X_1, \dots, X_m)]^2 < \infty$. The corresponding U-statistic for η is then

$$U_n(X_1, \dots, X_n) = \binom{n}{m}^{-1} \sum_{C_n} f^*(X_{\alpha_1}, \dots, X_{\alpha_m})$$

where C_n indicates that the summation is over all combinations $\{\alpha_1, \dots, \alpha_m\}$ of m integers selected from $\{1, \dots, n\}$. Then U_n has expectation η for all $n \geq m$ and

$$(3.1) \quad n^{\frac{1}{2}}(U_n - \eta) \xrightarrow{D} N(0, m^2 \zeta_1)$$

where

$$\zeta_1 = \text{Var}[E(f^*(X_1, \dots, X_m) | X_1)].$$

Moreover, $n \text{Var } U_n$ is a decreasing function of n with limit $m^2 \zeta_1$.

Let $r = [(k+1)p]$ and recall now from Section 2 that $K_{[(k+1)p]:k;n}$ is the U-statistic for $\mu_{r:k}(F)$ corresponding to the kernel $f^*(x_1, \dots, x_k) = x_{r:k}$. Denoting the beta p.d.f. $m_{r:k}(x) = [1/B(r, k-r+1)]x^{r-1}(1-x)^{k-r}$, $0 < x < 1$, we write the expectation as

$$(3.2) \quad \mu_{r:k}(F) = \int_0^1 G(u) m_{r:k}(u) du.$$

Employing the formula for the variance of the projection of an order statistic given in lemma 2 of Stigler (1969) yields a convenient representation of the asymptotic variance $k^2 \text{Var } E(X_{r:k} | X_1)$ as

$$(3.3) \quad \sigma_{r:k}^2(F) = \int_0^1 \int_0^1 (u \wedge v - uv) G'(u) G'(v) m_{r:k}(u) m_{r:k}(v) du dv.$$

Application of (3.1) provides

THEOREM 3.1. For $0 < p < 1$ and k fixed,

$$n^{\frac{k}{2}}(K_{[(k+1)p]:k;n} - \mu_{r:k}(F)) \xrightarrow{D} N(0, \sigma_{r:k}^2(F)) \text{ as } n \rightarrow \infty, \text{ where } r = [(k+1)p].$$

Although not presented here, a multivariate extension of theorem 3.1 follows easily from further results in Hoeffding (1948). The univariate development given here appears also in Takahasi (1970) and it should be noted that the conclusion requires only the existence of the variance of the r th. order statistic in a random sample of size k from F . Also, $n \text{Var } K_{[(k+1)p]:k;n}$ decreases with limit $\sigma_{r:k}^2(F)$ of (3.3).

For the case of a subsample size increasing in proportion with the total sample size we formulate the results of Bickel (1967). Let $\{m_{j,n}\}$, $1 \leq j \leq n$, $n \geq 1$ be a double sequence of constants such that $m_{j,n} = 0$ for $j \leq \delta n$, $j \geq (1-\delta)n$ for some $\delta > 0$ and consider the statistic $T_n = \sum_{j=1}^n m_{j,n} X_{j:n}$. If there exists $M(u)$ of bounded variation on $I = [\delta, 1 - \delta]$ such that $M_n(u) = \sum_{j=n\delta}^{(1-\delta)n} m_{j,n} + M(u)$

Definition: $u \wedge v = \min(u, v)$.

on a dense set of I and that $\sup_n V_0^1(M_n) < \infty$ (V_0^1 denotes total variation), then

$$(3.4) \quad n^{1/2}(T_n - E(T_n)) \xrightarrow{D} N(0, \sigma^2(M, F))$$

where

$$\sigma^2(M, F) = \int_0^1 \int_0^1 (u \wedge v - uv) G'(u) G'(v) dM(u) dM(v).$$

Next we apply the above version of theorem 4.1 of Bickel to the statistic

$K_{[(k+1)p]:k;n}$ when $k/n \rightarrow \lambda$, $0 < \lambda < 1$, as $n \rightarrow \infty$.

The negative hypergeometric probabilities given in (2.1) specify the probability distribution of a random variable $U_{r:k}(n)$ corresponding to the r th order statistic in a simple random sample (without replacement) from the finite population $\{j/(n+1): 1 \leq j \leq n\}$. The mean and variance of $U_{r:k}(n)$ are respectively $r/(k+1) = p$ and $r(k-r+1)(n-k)/(k+1)^2(k+2)(n+1) \rightarrow 0$ as $k, n \rightarrow \infty$. It follows by Chebyshev's inequality that $U_{r:k}(n)$ converges in distribution to the unit mass assigned to the point p . Application of (3.4) gives

THEOREM 3.2. For $0 < p < 1$ and $k/n \rightarrow \lambda$, $0 < \lambda < 1$, as $n \rightarrow \infty$

$$n^{1/2}(K_{[(k+1)p]:k;n} - \xi_p) \xrightarrow{D} N(0, \frac{\sigma^2(F)}{p})$$

where

$$\frac{\sigma^2(F)}{p} = p(1-p)/f^2(\xi_p) = p(1-p)[G'(p)]^2.$$

Although not developed here, it follows from theorem 4.3 of Bickel (1967) that the conclusion above holds whenever k tends to infinity with n , provided that F has finite second moment and that ξ_p is replaced with $\mu_{[(k+1)p]:k}$. Since under the conditions of theorem 3.2 it even can be shown that $n^{1/2}(K_{[(k+1)p]:k;n} - X_{[(n+1)p]:n}) \rightarrow 0$ with probability one, the rationale for inclusion of Bickel's results and theorem 3.2 is the demonstration that, in a certain sense, the generalized sample quantile is efficient in the nonparametric models discussed in Section 1. In addition, the development of theorem 3.2 illustrates the applicability of re-

sults concerning linear combinations of order statistic to the alternative estimator. Under more restrictive hypotheses, Bickel's theorem 4.3, in fact, will provide our theorem 3.1 since $n \frac{(j-1)(n-j)}{r-1} / \binom{n}{k} \approx m_{r:k}(j/n)$ and $U_{r:k}(n)$ has asymptotic beta distribution for fixed k as $n \rightarrow \infty$.

4. Comparisons of the Quantile Estimators $K_{[(k+1)p]:k;n}$. Although preferences among competing estimators often are established through extensive Monte Carlo studies, our treatment here is more in the spirit of the "small sample asymptotics" of Stigler (1977). However, some simulation results are included primarily to evaluate the adequacy of certain analytic approximations. It is of interest that the limited small sample numerical comparisons in Yanagawa (1969), (1970) suggest merit of the generalized sample median in the specialized application as a location estimator for symmetric distributions.

As an initial step in the comparisons we consider the asymptotic variances of the estimators $X_{[(n+1)p]:n}$ and $K_{[(k+1)p]:k;n}$. The equality of the variances given in (1.1) and theorem 3.2 indicate asymptotic equivalence so we consider $\sigma_{r:k}^2(F)$ of (3.3). First we investigate some specific distributions which permit explicit calculation and provide some insight regarding the behavior of the alternative estimator. In addition, the examples supply motivation for a subsequent approximation and its limitations. Although the result probably is available elsewhere, we include details of the calculation of $\sigma_{r:k}^2(F)$ for the uniform distribution since the derivation is probabilistic and possibly new. In the other example we simply list $\sigma_p^2(F)$ of (1.1) and $\sigma_{r:k}^2(F)$ of (3.3), the derivations being quite similar. We assume throughout that $(k+1)p$ and $(n+1)p$ are integers.

EXAMPLE 4.1. Standard uniform distribution. Let $F(x) = x$, $0 < x < 1$, $G(u) = u$, $0 < u < 1$, $\xi_p = p$. From (1.1) we obtain

$$(4.1) \quad \sigma_p^2(F) = p(1-p) = r(k-r+1)/(k+1)^2.$$

From (3.3) we have

$$(4.2) \quad \sigma_{r:k}^2(F) = 2 \int_0^1 \int_0^v u(1-v) m_{r:k}(u) m_{r:k}(v) du dv.$$

An easy manipulation of the integrand provides

$$\sigma_{r:k}^2(F) = [r(k-r+1)/(k+1)^2].$$

$$2 \int_0^1 \int_0^v [1/B(r+1, k-r+1)B(r, k-r+2)] u^r (1-u)^{k-r} v^{r-1} (1-v)^{k-r+1} du dv.$$

The integral above admits an interpretation as the probability that a random variable $V_{r+1:k+1}$ distributed as beta with parameters $r+1, k-r+1$ is less than a random variable $W_{r:k+1}$ distributed independently as beta with parameters $r, k-r+2$. Consider two independent random samples, each consisting of $k+1$ observations from the continuous uniform distribution on $(0,1)$. Then it follows that $\Pr(V_{r+1:k+1} < W_{r:k+1})$ may be computed as the probability that the $(r+1)$ st. order statistic in the first sample is less than the r th. order statistic in the second independent sample. A combining of the two independent samples and an elementary combinatorial argument regarding the sample origin of the smallest $2r$ observations shows that

$$\Pr(V_{r+1:k+1} < W_{r:k+1}) = \sum_{x=0}^{r-1} \binom{k+1}{x} \binom{k+1}{2r-x} / \binom{2k+2}{2r}.$$

Symmetry of the hypergeometric distribution indicated above provides

$$\Pr(V_{r+1:k+1} < W_{r:k+1}) = \frac{1}{2} [1 - \binom{k+1}{r} \binom{k+1}{r} / \binom{2k+2}{2r}].$$

It follows that

$$(4.3) \quad \sigma_{r:k}^2(F) = [r(k-r+1)/(k+1)^2] [1 - \binom{k+1}{r} \binom{k+1}{r} / \binom{2k+2}{2r}], \quad 1 \leq r \leq k.$$

From (4.1) and (4.3) we obtain

$$(4.4) \quad \sigma_p^2(F) / \sigma_{r:k}^2(F) = [1 - \binom{k+1}{r} \binom{k+1}{r} / \binom{2k+2}{2r}]^{-1} > 1.$$

Here both estimators are unbiased for ξ_p , so it follows from (4.4) that

$K_{[(k+1)p]:k;n}$ has (asymptotic) mean square error less than that of

$X_{[(n+1)p]:n}$ for any allowable subsample size if the population c.d.f. is that of the uniform distribution on (0,1).

EXAMPLE 4.2. Standard logistic distribution. Let $F(x) = (1+e^{-x})^{-1}$,

$-\infty < x < \infty$, $G(u) = \log[u/(1-u)]$, $0 < u < 1$, $\xi_p = \log[p/(1-p)]$.

Then

$$\sigma_p^2(F) = 1/p(1-p) = (k+1)^2/r(k-r+1).$$

$$\sigma_{r:k}^2(F) = [k^2/(r-1)(k-r)][1 - \frac{(k-1)}{r-1} \frac{(k-1)}{r-1} / \binom{2k-2}{2r-2}], \quad 1 < r < k.$$

EXAMPLE 4.3. Standard exponential distribution. Let $F(x) = 1-e^{-x}$,

$0 < x < \infty$, $G(u) = -\log(1-u)$, $0 < u < 1$, $\xi_p = -\log(1-p)$.

Then

$$\sigma_p^2(F) = p/(1-p) = r/(k-r+1)$$

$$\sigma_{r:k}^2(F) = [r/(k-r)][1 - \frac{k}{r} \frac{k}{r} / \binom{2k}{2r}], \quad 1 \leq r < k.$$

EXAMPLE 4.4. Standard power function distribution (ζ). Let $F(x) = x^{\frac{1}{k}}$,

$0 < x < 1$, $G(u) = u^2$, $0 < u < 1$, $\xi_p = p^2$.

Then

$$\sigma_p^2(F) = 4p^3(1-p) = 4r^3(k-r+1)/(k+1)^4$$

$$\sigma_{r:k}^2(F) = [4r^2(r+1)(k-r+1)/(k+1)^2(k+2)^2]$$

$$\cdot [1 - \frac{(k+2)}{r+1} \frac{(k+2)}{r+1} / \binom{2k+4}{2r+2}], \quad 1 \leq r \leq k.$$

Although many standard distributions such as the normal do not possess

inverse cumulatives which permit calculation of (3.3), the use of Tukey's lambda distributions (see Joiner and Rosenblatt (1971) and Ramberg and Schmeiser (1974)) can provide suitable approximations. However, as an alternative approach we observe instead that the examples presented above suggest an approximation of $\sigma_{r:k}^2(F)$ adequate at least for qualitative purposes. As our primary objective is to ascertain the behavior of the alternative estimator over a large class of "well-behaved" distributions, we implicitly assume throughout the necessary smoothness conditions on the inverse c.d.f. G.

Computation of the variance ratios ... the preceding examples suggests that for $r, r-k$, and k of moderate size

$$(4.5) \quad \sigma_p^2(F)/\sigma_{r:k}^2(F) \approx [1 - (\frac{k+1}{r})(\frac{k+1}{r})/(\frac{2k+2}{2r})]^{-1} > 1.$$

NOTE. Examination of $\sigma_{r:k}^2(F)$ in examples 4.2 - 4.4 indicates the importance of the qualifying statement "of moderate size".

Now the respective mean and variance of the beta density $m_{r:k}$ are $r/(k+1) = p$ and $r(k-r+1)/(k+1)^2(k+2) = p(1-p)/(k+2)$. Assuming that the continuous p.d.f. f is relatively constant near ξ_p , the approximation $G'(x) \approx G'(p)$ in (3.3) provides

$$\sigma_{r:k}^2(F) \approx [G'(p)]^2 \int_0^1 \int_0^v u(1-v)m_{r:k}(u)m_{r:k}(v) du dv.$$

The integral above is precisely that of (4.2) computed in the uniform case of example 4.1 so we obtain

$$(4.6) \quad \begin{aligned} \sigma_{r:k}^2(F) &\approx p(1-p)[G'(p)]^2 [1 - (\frac{k+1}{r})(\frac{k+1}{r})/(\frac{2k+2}{2r})] \\ &= \sigma_p^2(F) [1 - (\frac{k+1}{r})(\frac{k+1}{r})/(\frac{2k+2}{2r})], \quad r = [(k+1)p]. \end{aligned}$$

The preceding presents some justification of (4.5) and the resultant implication

that the ratio of the asymptotic variances $\sigma_p^2(F)/\sigma_{r:k}^2(F)$ is "approximately independent" of F for r , $r-k$, and k of moderate size. Another consequence of (4.5) is the suggestion that the ratio of the asymptotic variances decreases to 1 with increasing subsample size.

To investigate the adequacy of the approximations of theorem 3.1 and (4.5) we performed 10,000 simulations of median estimation for seven symmetric distributions based on a complete sample size $n = 99$ and subsample sizes $k = 9, 19, 29, 39, 79$. The results appear in Table 4.1 and are qualitatively as predicted and quantitatively in quite reasonable agreement with (4.5) (and the results of examples 4.1 and 4.2).

Recall now that the formulation of theorem 3.1 provides an estimator of ξ_p which is asymptotically biased in most cases. The simulations of Table 4.1 were selected to avoid confounding bias considerations (both estimators are unbiased for the medians of these symmetric distributions) so that the theoretical variance ratios equal the theoretical mean square error ratios. In the more general case, a decrease in variance by the subsampling scheme can be insufficient to achieve the desired reduction in mean square error of estimation. It is clear then that asymptotic bias magnitude should be considered in the evaluation of the alternative estimator.

For this objective, ignoring all but the first three terms of a Taylor's expansion of G in (3.2) gives

$$\mu_{r:k}(F) - \xi_p \approx p(1-p)G''(p)/2(k+2)$$

which in conjunction with (4.6) provides the approximate mean square error

$$\begin{aligned}
 \text{MSE}_{r:k}(F) &\approx \sigma_{r:k}^2(F)/n + [\mu_{r:k}(F) - \xi_p]^2 \\
 &\approx p(1-p)[G'(p)]^2[1 - (\frac{k+1}{r})(\frac{k+1}{r})(\frac{2k+2}{2r})]/n \\
 &\quad + [p(1-p)G''(p)]^2/4(k+2)^2, \quad r = [(k+1)p].
 \end{aligned}$$

Similarly, $X_{[(n+1)p]:n}$ has approximate mean square error

$$\begin{aligned}
 \text{MSE}_p(F) &\approx \sigma_p^2(F)/n + [\mu_{[(n+1)p]:n}(F) - \xi_p]^2 \\
 &\approx p(1-p)[G'(p)]^2/n + [p(1-p)G''(p)]^2/4(n+2)^2.
 \end{aligned}$$

It follows that for moderate $r, r=k$, and k and large n

$$\begin{aligned}
 (4.7) \quad \text{MSE}_p(F)/\text{MSE}_{r:k}(F) &\approx \{1 + p(1-p)[G''(p)/G'(p)]^2/4n\} \\
 &\cdot \{[1 - (\frac{k+1}{r})(\frac{k+1}{r})/(\frac{2k+2}{2r})] + n p(1-p)[G''(p)/G'(p)]^2/4k^2\}^{-1}, \quad r = [(k+1)p].
 \end{aligned}$$

If k is not too small, both bias terms in (4.7) are negligible and the remarks immediately following (4.6) apply to mean square error as well.

Finally in Table 4.2 we present further simulation results for both symmetric and asymmetric distributions for $n = 99$, $k = 39$, $k = 79$, and p varying from 0.05 to 0.95. The results are in quite reasonable agreement with (4.5) (the bias terms in (4.7) are indeed negligible) for $0.2 < p < 0.8$. The alternative estimator performed better than the conventional estimator for all deciles of all distributions except the heavy-tailed double exponential and Cauchy for $k = 39, 79$. Problems were encountered for extreme quantiles of the power function, exponential, logistic, and normal distributions, also. The corresponding values of r were 2 and 38 (not moderate) indicating difficulty with the accuracy of (4.5) and/or bias magnitude. However, it should be noted that the intent of Table 4.2 is to suggest the existence of a single subsample size k providing simultaneously better estimates for a spectrum of quantiles over a class of different distributions, and a larger subsample size would eliminate the aberrant cases. In practice different subsample sizes probably should be employed for estimation of different quantiles.

TABLE 4.1 Ratio of Mean Square Error $X_{[(n+1)p]:n}$ to Mean Square Error $K_{[(k+1)p]:k;n}$ Based on 10,000 Simulations.

<u>k</u>	<u>E</u>	<u>Uniform</u>	<u>Arcsine</u>	<u>Triangular</u>	<u>Logistic</u>	<u>Normal</u>	<u>Double Exponential</u>	<u>Cauchy</u>
9	5	1.48	4.93	1.5	1.22	1.57	0.5	0
19	10	1.28	1.73	1.21	1.26	1.30	0.96	0.94
29	15	1.20	1.39	1.15	1.19	1.21	1.01	1.06
39	20	1.16	1.26	1.12	1.15	1.16	1.03	1.08
79	40	1.06	1.07	1.05	1.05	1.05	1.04	1.06

n = 99, p = 0.5

*Theoretical ratio of asymptotic variances.

TABLE 4.2 Ratio of Mean Square Error $\chi_{[(n+1)p]:n}^2$ to Mean Square Error $\chi_{[(k+1)p]:k;n}^2$ Based on 10,000 Simulations.

<u>p</u>	<u>Uniform</u>	<u>Arcsine</u>	<u>Triangular</u>	<u>Power Function</u>	<u>Exponential</u>	<u>Logistic</u>	<u>Normal</u>	<u>Double Exponential</u>	<u>Cauchy</u>
0.05	1.41	1.01	1.23	0.99	1.38	0.89	1.00	0.83	0.13
0.10	1.28	1.09	1.20	1.07	1.26	1.02	1.10	0.98	0.51
0.20	1.21	1.14	1.17	1.10	1.18	1.10	1.14	1.06	0.80
0.30	1.18	1.18	1.15	1.12	1.14	1.12	1.14	1.08	0.96
0.40	1.16	1.19	1.13	1.13	1.13	1.13	1.14	1.08	1.05
0.50	1.16	1.20	1.10	1.14	1.11	1.13	1.13	1.04	1.08
0.60	1.17	1.20	1.13	1.16	1.10	1.13	1.13	1.08	1.05
0.70	1.18	1.19	1.16	1.19	1.09	1.13	1.12	1.09	0.97
0.80	1.20	1.14	1.16	1.21	1.05	1.09	1.12	1.05	0.79
0.90	1.28	1.07	1.19	1.29	0.97	1.01	1.09	0.97	0.49
0.95	1.42	1.03	1.24	1.43	0.83	0.86	0.97	0.83	0.14
 <u>n = 99, k = 39</u>									
0.05	1.13	1.08	1.10	1.08	1.13	1.05	1.08	1.04	0.82
0.10	1.10	1.07	1.08	1.07	1.09	1.06	1.07	1.06	0.96
0.20	1.08	1.07	1.07	1.06	1.07	1.06	1.07	1.06	1.02
0.30	1.06	1.07	1.06	1.06	1.06	1.06	1.06	1.05	1.03
0.40	1.06	1.07	1.06	1.06	1.06	1.06	1.06	1.05	1.05
0.50	1.06	1.07	1.05	1.06	1.05	1.05	1.05	1.04	1.05
0.60	1.07	1.07	1.06	1.06	1.06	1.06	1.06	1.05	1.05
0.70	1.07	1.07	1.07	1.07	1.06	1.05	1.06	1.04	1.04
0.80	1.07	1.07	1.07	-	1.07	1.05	1.06	1.05	1.01
0.90	1.10	1.08	1.09	1.10	1.06	1.07	1.06	1.06	0.95
0.95	1.13	1.09	1.10	1.13	1.04	1.04	1.06	1.04	0.85
 <u>n = 99, k = 79</u>									

5. Quantile Interval Estimation. The asymptotic results for the point estimators $X_{[(n+1)p]:n}$ and $K_{[(k+1)p]:k;n}$ may be employed to obtain large-sample confidence intervals for population quantiles provided that sample estimates may be obtained for the asymptotic variances of (1.1) and (3.3). Here we consider application of the jackknife procedure to obtain sample variance estimates of the generalized sample quantile, although an alternative method using the sample quantile function and tables of incomplete beta functions is described by Maritz and Jarrett (1978).

First we develop briefly the jackknife estimator of an unknown population parameter θ based on a random sample X_1, \dots, X_n . Let $\hat{\theta}_n^0$ be the estimate of θ based on all n observations and let $\hat{\theta}_{n-1}^i$, $i=1, \dots, n$, be the estimate obtained by deletion of the i th observation. The pseudo-values are defined by

$$\hat{\theta}_i^* = n \hat{\theta}_n^0 - (n-1) \hat{\theta}_{n-1}^i, \quad i=1, \dots, n$$

and the jackknife estimator of θ is then

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i^*/n.$$

A sample estimate of the variance of the jackknife estimator is given by

$$S_{\hat{\theta}}^2 = S^2/n \text{ where}$$

$$S^2 = \frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i^* - \hat{\theta})^2 / (n-1).$$

Under certain conditions (e.g. Miller (1964)) $S_{\hat{\theta}}^2$ is consistent and the standardized statistic $(\hat{\theta} - \theta)/S_{\hat{\theta}}$ is asymptotically standard normal. The jackknife estimator may be employed then as a pivotal statistic for robust interval estimation of θ .

Although the sample quantile represents a classic failure of the jackknife procedure (see Efron (1979)), we show that the generalized sample quantile estimator "jackknifes well" and the asymptotic behavior of the statistic

$K_{r:k;n}$ under the jackknife follows easily by application of results of Arvesen (1969) concerning U-statistics.

Deletion of the i th observation provides pseudo-value $nK_{r:k;n} - (n-1) K_{r:k;n-1}^1$, where the weights required for the computation of $K_{r:k;n-1}^1$ are $\binom{j-1}{r-1} \binom{n-j-1}{k-r} / \binom{n-1}{k}$, $j=r, \dots, r+n-k-1$. Since $K_{r:k;n}$ is a U-statistic the pseudo-values provide average $K_{r:k;n}$ and sample variance $S_{r:k;n}^2$ given by

$$S_{r:k;n}^2 = (n-1) \sum_{i=1}^n (K_{r:k;n-1}^1 - K_{r:k;n})^2.$$

Application of Arvesen's theorem 6 in conjunction with our theorem 3.1 provides

THEOREM 5.1. For r and k fixed, as $n \rightarrow \infty$

- i) $S_{r:k;n}^2 \xrightarrow{P} \sigma_{r:k}^2(F)$
- ii) $n^{\frac{1}{2}} (K_{r:k;n} - \mu_{r:k}(F)) / S_{r:k;n} \xrightarrow{D} N(0,1).$

We remark that the results of Parr and Schucany (1981) concerning jackknifed linear combinations of order statistics will produce theorem 5.1 under more restrictive conditions on the c.d.f. F .

The preceding yields an approximate $1-\alpha$ confidence interval for the population quantile ξ_p given by

$$(5.1) \quad K_{r:k;n} \pm \phi^{-1}(1-\alpha/2) S_{r:k;n} / n^{\frac{1}{2}}, \quad r = [(k+1)p]$$

where ϕ is the standard normal c.d.f.

To investigate the validity of (5.1) for small and moderate sample sizes, additional simulations were performed for the uniform and exponential distributions. Since the number of distinct pseudo-values obtained is $n-k+1$, the standard normal percentage points in (5.1) were replaced by those of the t distribution with $n-k$ degrees of freedom. Results of median interval estimates for the uniform distribution based on various sample sizes appear

in Table 5.1 while results of other quantile interval estimates for the uniform and exponential distributions based on a single moderate sample size appear in Table 5.2. Taking into account Monte Carlo variability, there is only small deviation of the empirical confidence levels from the nominal levels even for small sample sizes. Although further simulation studies should be performed, the results of Section 4 suggest adequate validity of (5.1) for other distributions as well.

BIBLIOGRAPHY

- Arvesen, J. N. (1969). Jackknifing U-statistics. Ann. Math. Statist. 40 2076-2100.
- Bickel, P. J. (1967). Some contributions to the theory of order statistics. Proc. Fifth Berkeley Symp. Math. Statist. Prob. I 575-591.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. Ann. Statist. 7 1-26.
- Harrell, F. E. and Davis, C. E. (1981). Improved distribution-free quantile estimation. In preparation.
- Hodges, J. L. and Lehmann, E. L. (1970). Deficiency. Ann. Math. Statist. 41 783-801.
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Ann. Math. Statist. 19 293-325.
- Joiner, B. L. and Rosenblatt, J. R. (1971). Some properties of the range in samples from Tukey's symmetric lambda distributions. J. Amer. Statist. Assoc. 66 394-399.
- Kaigh, W. D. and Lachenbruch, P. A. (1981). A generalized quantile estimator. To appear.
- Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median. J. Amer. Statist. Assoc. 73 194-196.
- Miller, R. G., Jr. (1964). A trustworthy jackknife. Ann. Math. Statist. 35 1594-1605.
- Parr, W. C. and Schucany, W. R. (1981). L-statistics with smooth weight functions jackknife well. To appear.

- Pfanzagl, J. (1975). Investigating the quantile of an unknown distribution.
In: Statistical Methods in Biometry (dedicated to A. Linder) (ed.
W. J. Ziegler) 111-126, Birkhäuser Verlag, Basel.
- Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for
generating asymmetric random variables. CACM 17, 2 78-82.
- Reiss, R. -D. (1980). Estimation of quantiles in certain non-parametric
models. Ann. Math. Statist. 8 87-105.
- Stigler, S. M. (1969). Linear functions of order statistics. Ann. Math.
Statist. 40, 770-788.
- Stigler, S. M. (1977). Fractional order statistics, with applications.
J. Amer. Statist. Assoc. 72 544-550.
- Takahasi, K. (1970). Estimation of several characteristics of distributions
of order statistics. Ann. Inst. Statist. Math. 22 403-412.
- Yanagawa, T. (1969). A small sample robust competitor of Hodges-Lehmann
estimate. Bull. Math. Statist. 13 1-14.
- Yanagawa, T. (1970). On some robust estimate of a location parameter.
Mem. Fac. Gen. Educ., Kumamoto Univ., Ser. Nat. Sci. 5 7-16.
- Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York.

TABLE 5.1 Quantile Interval Escimation $K_{[(k+1)p]:k;n} \pm t_{n-k,\alpha/2} s_{[(k+1)p]:k;n} / n^{1/2}$
 Based on 1,000 Simulations.

p = 0.5

Uniform Distribution

<u>Sample Size</u>	<u>Subsample Size</u>	<u>Observed Confidence Level</u>	<u>Ratio of Observed Variance to Mean Jackknife Variance Estimate</u>
		<u>0.68*</u>	<u>0.95*</u>
n=99	k=79	0.66	0.92
	k=39	0.67	0.93
n=49	k=39	0.66	0.91
	k=19	0.68	0.92
n=39	k=19	0.68	0.94
	k= 9	0.70	0.95
n=29	k=19	0.65	0.92
	k= 9	0.66	0.93
n=19	k= 9	0.68	0.93

*nominal confidence level

TABLE 5.2 Quantile Interval Estimation $K_{[(k+1)p]:k;n} \pm t_{n-k,\alpha/2} s_{[(k+1)p]:k;n}/n^{1/2}$

Based on 1,000 Simulations.

<u>P</u>	Uniform Distribution		Exponential Distribution	
	Observed Confidence Level	Confidence Level	Observed Confidence Level	Confidence Level
0.10	0.65	0.91	0.66	0.92
0.20	0.65	0.92	0.66	0.93
0.30	0.66	0.94	0.67	0.94
0.40	0.66	0.93	0.67	0.93
0.50	0.67	0.93	0.67	0.93
0.60	0.66	0.94	0.66	0.95
0.70	0.67	0.93	0.67	0.94
0.80	0.68	0.92	0.67	0.93
0.90	0.67	0.93	0.65	0.94

*nominal confidence level

n=99, k=39

The Nonparametric Estimation
of Probability Densities in Ballistics Research*

Chih-chy Fwu
Richard A. Tapia
James R. Thompson

Department of Mathematical Sciences
Rice University

Abstract. The problem of nonparametric probability density estimation is considered for higher dimensions. An "onion peel" algorithm is suggested for 3-dimensions. For dimensions of 4 or more, a decomposition procedure is proposed, which first finds the centers of mass using nearest neighbor techniques, then estimates the density around these centers using fixed mesh procedures.

Acknowledgement. The authors wish to thank Dr. Malcolm Taylor and Mr. Jerry Thomas of Aberdeen Proving Ground for bringing the data set used in this paper to their attention and for their insightful comments on its analysis.

*This research was supported in part by the Army Research Office (Durham) under DAAG29-78-G-0187 .

Introduction

There are many reasons for the possible failure of standard parametric statistical procedures. Among these, the problem of tailiness beyond that in the model assumed has attracted the most interest. As one example, for some years now, rank tests have been used as an alternative to likelihood ratio tests [7]. More recently, notions of robustness as delineated in the Princeton Robustness Study have moved to center stage in statistical investigation [1]. Both these sets of techniques tend to assume symmetry and unimodality of the underlying distributions. Both are somewhat tied to one dimensional probability densities.

A second type of pathology, and the one to which we shall address ourselves in this paper, is departures of the underlying distributions from unimodality and symmetry. In this case protection against tailiness will be of little avail. Procedures are required which will be robust against the unexpected "in the center."

Of such techniques, the oldest is the histogram, which existed in crude form as long ago as 1662 [4]. The "shifted histogram" of Rosenblatt [12] gave greater efficiency and flexibility than those of the histogram. The still more general kernel estimates of Parzen [11] have found wide applicability

Another approach has been that of series estimates [6, 15]. These have a loyal group of users but do not presently enjoy the popularity of kernel estimates.

A suggestion of Good and Gaskins [3] to pose Bayesian estimation in a function space setting for density estimation (with a prior measure on the space of densities) was successfully pursued by de Montricher [10]. However,

the practical difficulties of algorithmic implementation have given preference to the related concept (also suggested by Good and Gaskins) of maximum penalized likelihood density estimation [14, 15]. This algorithm has been included as a standard routine in the widely disseminated IMSL package [5].

The three categories of density estimation--histogram (including the shifted), series, and maximum penalized likelihood-- are by no means exhaustive of the techniques robust against the possibility of multimodality, but are the most commonly used. Each of these can be generalized to several dimensions. The technique used in this paper, however, is based on the shifted histogram.

Discussion

The nonparametric estimation of densities in higher dimensions presents the investigator with difficulties not encountered in the well explored one dimensional case. If we use evaluation on a standard fixed mesh grid, we have the problem of exponentially exploding cost of computation with increasing dimension. Moreover, with kernel (shifted histogram) techniques we face the empty space problem-- namely the vast majority of grids will contain no data points. So a great deal of our computation will be effectively wasted. A preferred procedure, then, would be to use a variable grid which increases in size in regions of low density but decreases in regions with many data points.

This leads us to the k-nearest neighbor algorithm [2,8]. To delineate it, we let

$$(1) \quad \hat{f}(x) = \frac{k}{N} \frac{1}{V_m(x, d(x, k))}$$

where $d(x, k)$ = Euclidean distance to the k th nearest data point from x
 $V_m(x, d(x, k))$ = the volume of the m -dimensional sphere centered at x
with radius $d(x, k)$
 N = the sample size.

We note that as k increases, the variability of our estimate for f decreases, but at the expense of increased bias. Sufficient conditions for consistency of the estimate in (1) are (p.84, [15])

$$(2) \quad \lim_{N \rightarrow \infty} k = \infty$$

$$\lim_{N \rightarrow \infty} k/N = 0.$$

For density estimation problems in 1, 2, or 3 dimensions, it is an easy matter to choose the appropriate k interactively.

As yet, no completely automated rule for the selection of k is available, although an iterative procedure developed for fixed kernel width selection [13] appears well suited to this task using the formula for the mean square error of nearest neighbor estimates given in [9]. For low dimensional densities (1, 2, or 3), it is not difficult to choose k interactively. We simply start with k large-- say $N/2$ -- and sequentially reduce it by powers of 2 until the graphs of the estimated density begin to display high frequency wiggles. Then we return to the preceding value of k .

It is interesting to note that it is the graphing of \hat{f} (or some machine alternative to graphing) which is the greatest problem in density estimation in higher dimensions. The use of a data based "grid" does not liberate us from the curse of dimensionality. As an example of this point, suppose we have a sample of 300 from a 5 dimensional density. An investigator who estimated the density using fixed mesh (20/dim.) would be required to evaluate \hat{f} at 3.2×10^6 points. The nearest neighbor advocate might argue with some validity that we could make do with evaluating \hat{f} only at the 300 data points. The argument for this attitude might be that he is interested at points where f is large-- and these are most likely to be near the data points. But what sense can he make of $\{\hat{f}(x_i); i = 1, 2, \dots, 300\}$? He must somehow exploit the assumed continuity of f to "get a picture" of it. In one dimension, the eye itself would perform this task from a simple plotting of $\{\hat{f}(x_i); i = 1, 2, \dots, 300\}$. In two dimensions, one would need to use some care in selecting the appropriate graphical technique (2 dimensional contour plots, 3-d Calcomp plots, etc.).

In 3-dimensions (which, due to the added dimension from f , is really a 4-d plotting problem), one will have to be somewhat clever. And in higher dimensions, where our essentially 3-d perceptions fail us, what to do is unclear. There is, unfortunately, a vast difference in knowing the functional form of f and simply knowing it on a regular (let alone an irregular) mesh. A knowledge of f on the continuum would return us to the happy world of parametric probability densities (a low dimensional problem). A knowledge of f at a discrete number of points leaves us with a problem of high dimensionality. Of course, we shall not even know f at a finite number of points-- only an estimate \hat{f} . But from a practical point of view, inferential difficulties would remain-- even if we knew f exactly at a discrete number of points.

Let us consider the following question: would we rather have a random sample of 300 from our unknown 5 dimensional density f or would we rather know f precisely at 300 points selected from a uniform distribution over the 5-dimensional hypercube in which we know a-priori the bulk of the mass of f is imbedded? A little thought reveals that the first of the two cases is the more informative (though we would surely pick the second for the 1-dimensional problem) on those regions having the greatest density. This again argues against fixed mesh width shifted histogram estimation and in favor of nearest neighbor techniques in higher dimensions.

But it also points us toward the desirability of focusing on local centers of high density. Let us consider a three dimensional ballistics data set. As a first step we translate the data to the sample mean and rescale it so that the marginal sample variances are equal. We now consider the estimation of f in the three planes $MV = 0$, $\phi = 0$, $\theta = 0$. We show, in Figure 5, the procedure whereby this estimation is carried out. Some

important but mathematically trivial computational savings can be made. For example, suppose we have determined the distance $d(Q, P_1)$ of a point P_1 , with coordinates (x_1, y_1, z_1) from a sample point Q with coordinates (x, y, z) . Then going to the next point $P_2(x_1 + \delta, y_1, z_1)$ in the grid, we have the simple update formula:

$$(3) \quad d^2(Q, P_2) = d^2(Q, P_1) + \delta^2 + 2\delta(x - x_1).$$

The gain in computational efficiency using this simple update formula is of the order of 2 to 3 (note that if we had not used a regular grid, this saving would not have been available).

Next, we note that if we use (1) for estimating $\hat{f}(x_1)$ for a predetermined $k = pN$ it is not essential that we use precisely this value of k in the formula, as long as we know what k is. Consequently, we select randomly a subset of the N data points of size $M = 2^r$ (with, typically, $r = 6$ or 7). Then we find the distance to the 2^s th nearest neighbor to the grid point x_1 where $2^s/2^r \sim p$. Call this distance d . Returning to the full data set, count the number of sample points at least as close to x_1 as d - let the number of such points be called $k'(\approx k)$.

Then we use the formula

$$(3) \quad \hat{f}(x_1) = \frac{k'}{N} \cdot \frac{1}{V_m(x_1, d)}.$$

The information loss caused by this latter "pilot study" algorithm is negligible, while the improvement in computational efficiency is of the order $\{\log_2 N / \log_2 M\}$.

For each of the three planes, we now interpolate to obtain the (conditional) iso- \hat{f} level curves. (Such curves for $MV = 0$ are given in

Figure 6.) We note that such curves will enclose less and less area as \hat{f} increases. Connecting points on the level curves for a fixed value of \hat{f} gives us the level surface (with MV coming out of the page) in Figure 7. We next let \hat{f} increase to give us the level surface in Figure 8. We continue to increase \hat{f} until first one bump, then the other disappears. In using the "onion peel" procedure for the present problem, it was noted that at the \hat{f} levels of disappearance of the two bumps, the MV values were identical - thus indicating only that only one modal MV value is appropriate. Naturally, we might find it desirable to make one or two additional sets of onion peel plots in determining the coordinates of the modes (each corresponding to one of the two angular coordinates being used as the coordinate coming out of the page), since the "out of the page" coordinate is not as easily dealt with as the two on the page. In the example at hand, we found two modes with coordinates: MV = 722.51 gm/s, $\phi = -8.15^\circ$ and $\theta = 24.18^\circ, 45.50^\circ$.

It is interesting to note that for the present example, although we have used the nearest neighbor variant of the shifted histogram procedure, we have used a fixed mesh grid to determine where the density should be estimated. One would be justified in asking the question: would we have not have done as well to stay with fixed mesh estimation as well? The answer is, "yes, for the present well behaved data set." In general, if we estimate f at a point in its support, we are implicitly assuming it to be significantly greater than zero at that point. And, if such be the case, the many practical advantages of a fixed mesh may be decisive.

In general, the greatest value of a variable mesh should be in pointing to those regions of relatively high density. Once we have determined the rough boundaries of these regions, we might do well to use a tuned fixed mesh estimation on each of the regions. Thus we would be using

$$(4) \quad \hat{f} = \sum_{j=1}^k \hat{f}_j$$

where $\hat{f}_j(x) = 0$ for $x \in R_j^c$.

So we are advocating a kind of decomposition into regions of high density using k-nearest neighbor techniques followed by a fixed (for each region) mesh estimation of the density in each region.

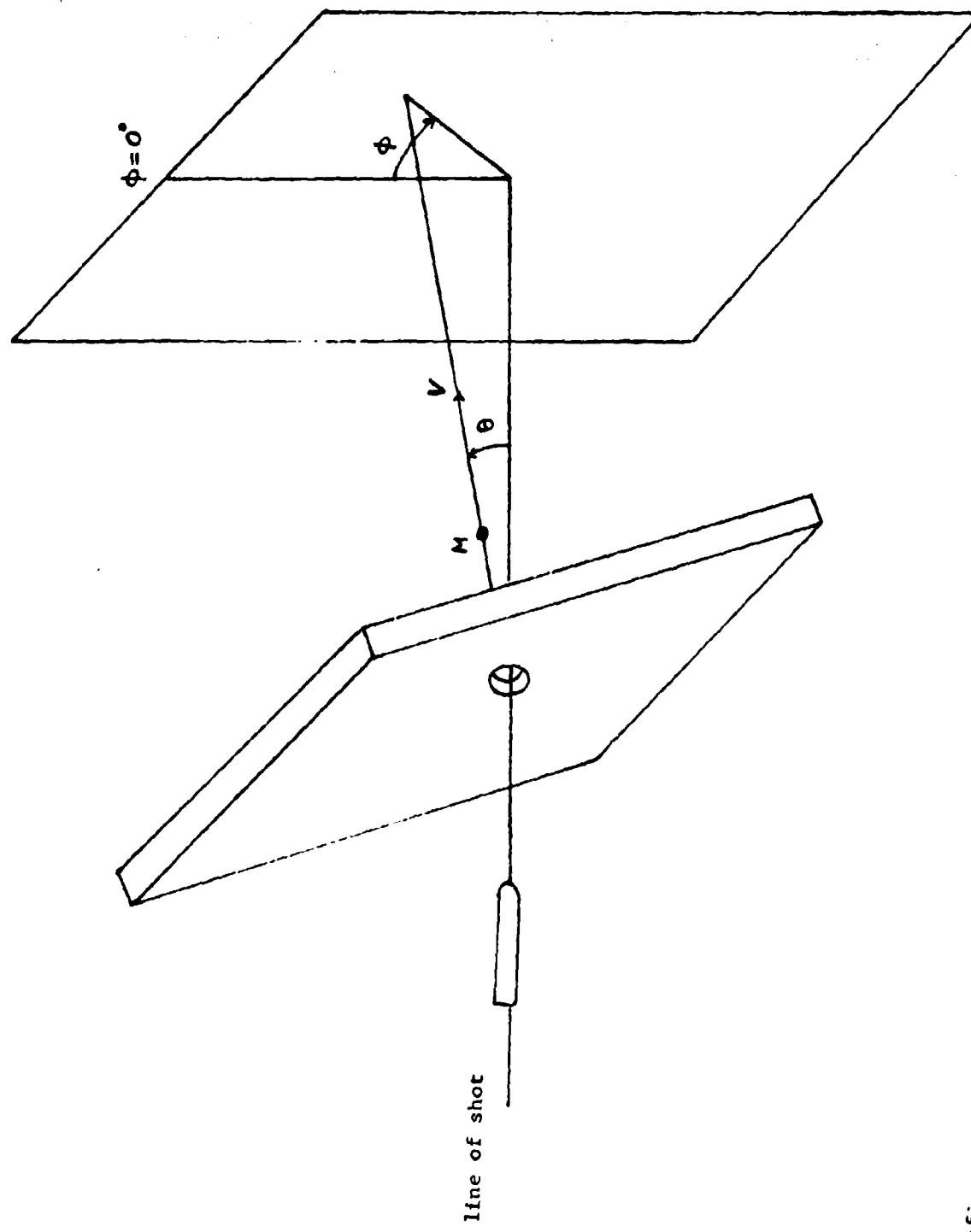
Although we are still working on this two step algorithm, we can already indicate some preliminary results. First, we start on the fringes of the data set. Then taking a data point as the first iterate x_1 , we let

$$(5) \quad x_n = \text{Ave}(k \text{ nearest neighbors of } x_{n-1}).$$

Experience shows that, at least in dimension of 3 or less, the algorithm in (5) will stop (or cycle) prematurely - i.e., before a bona fide local maximum of f has been essentially reached. However, it generally brings us into the domain of attraction (for Newton's method) of a local maximum. So then, a two stage averaging and Newton's method algorithm appears to work well for finding the local maxima of f .

Following the location of centers of high density, we can investigate estimation around each locally. This might involve, for example, a preliminary investigation using nearest neighbor techniques to determine the contours of \hat{f} values $1/4, 1/8$ and $1/32$ of that at the local maximum. In many cases, it may be possible to use parametric techniques for some of the local densities. In others, a fixed mesh technique - e.g., shifted histogram or maximum penalized likelihood - might prove useful.

FIGURE 1 *



*Figure courtesy of Dr. Malcolm Taylor and Mr. Jerry Thomas

FIGURE 2

$N = 1380$

$k = 300$ (too large)

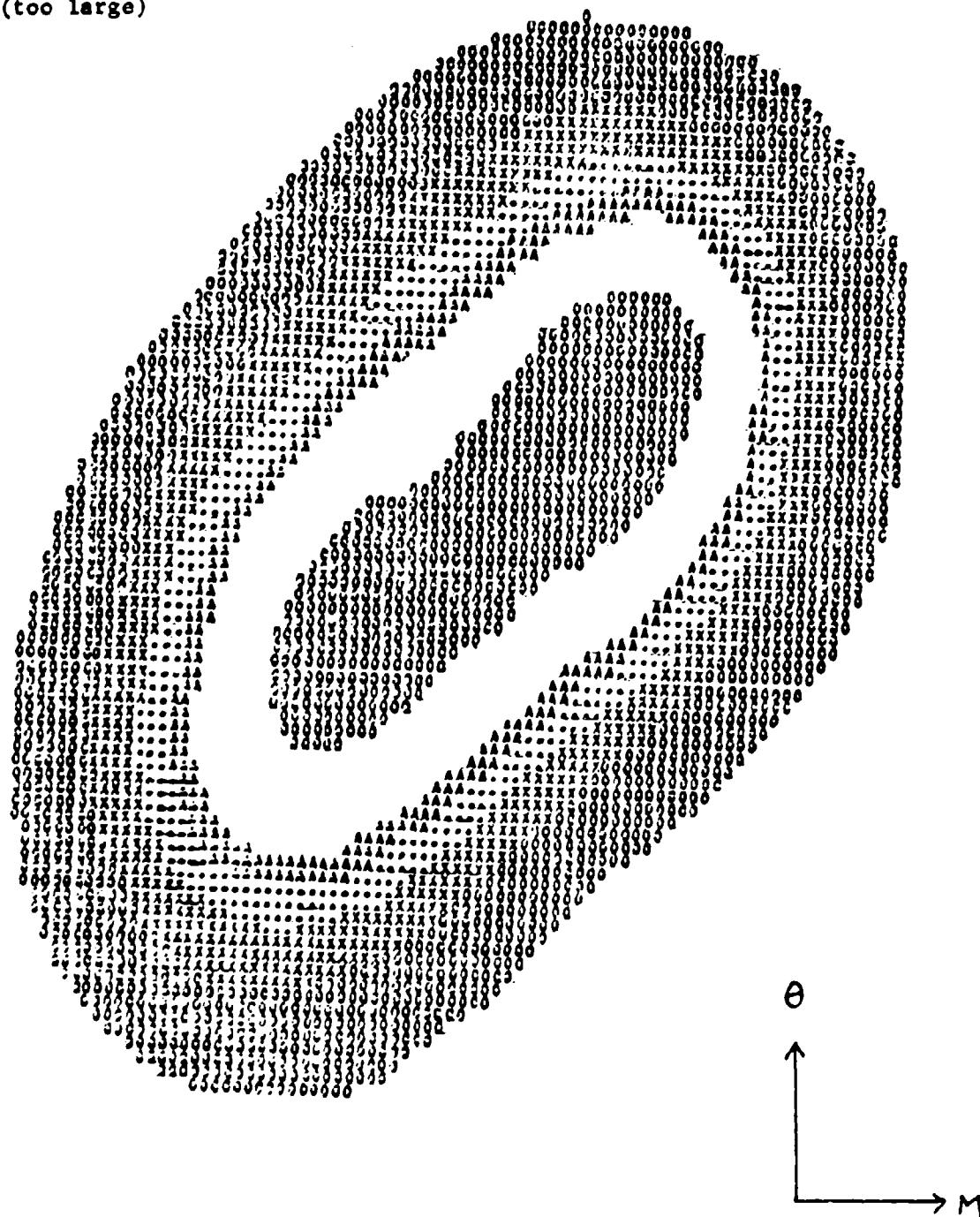


FIGURE 3

$N = 1380$

$k = 150$ (about right)

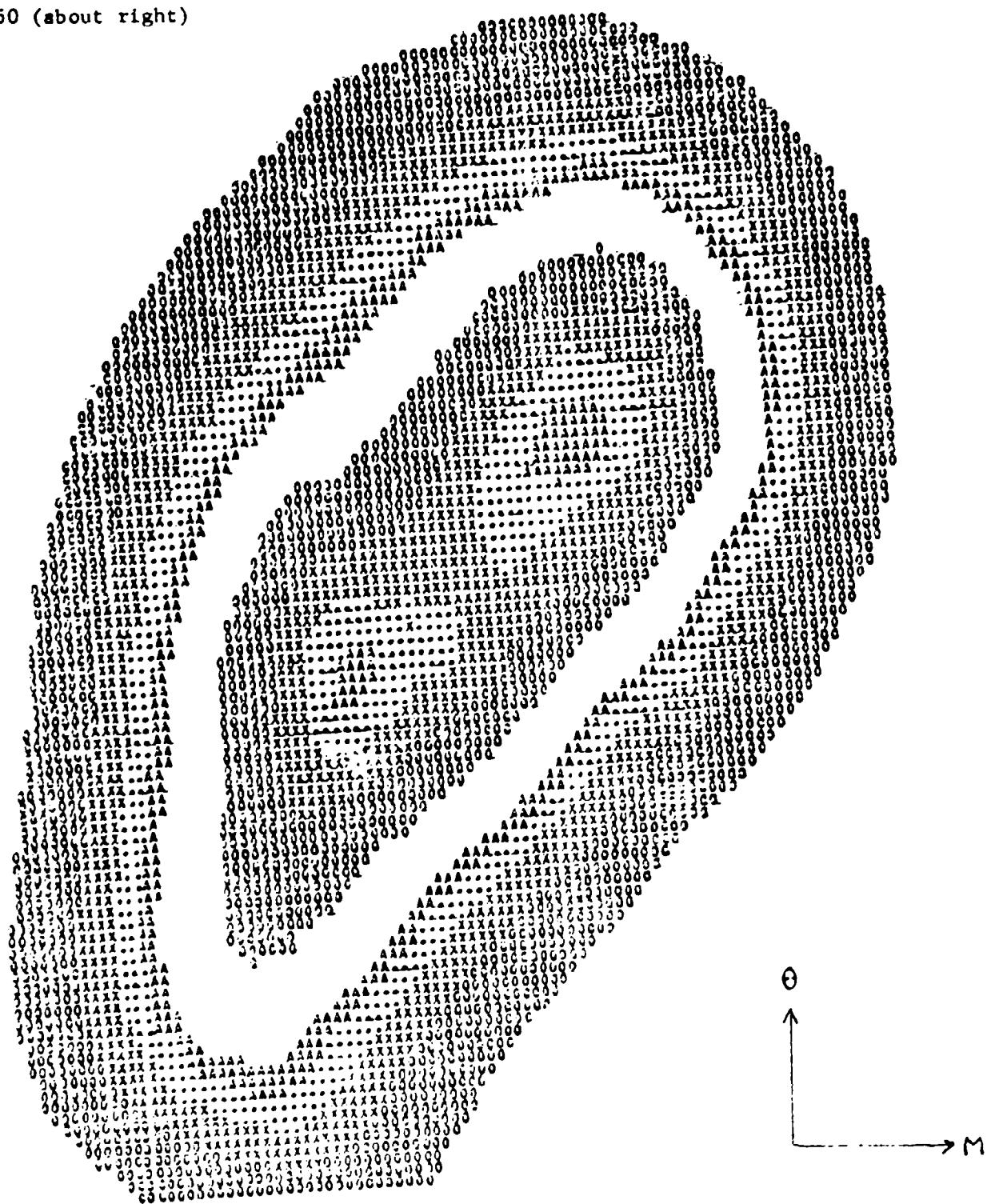


FIGURE 4

$N = 1380$

$k = 30$ (too small)

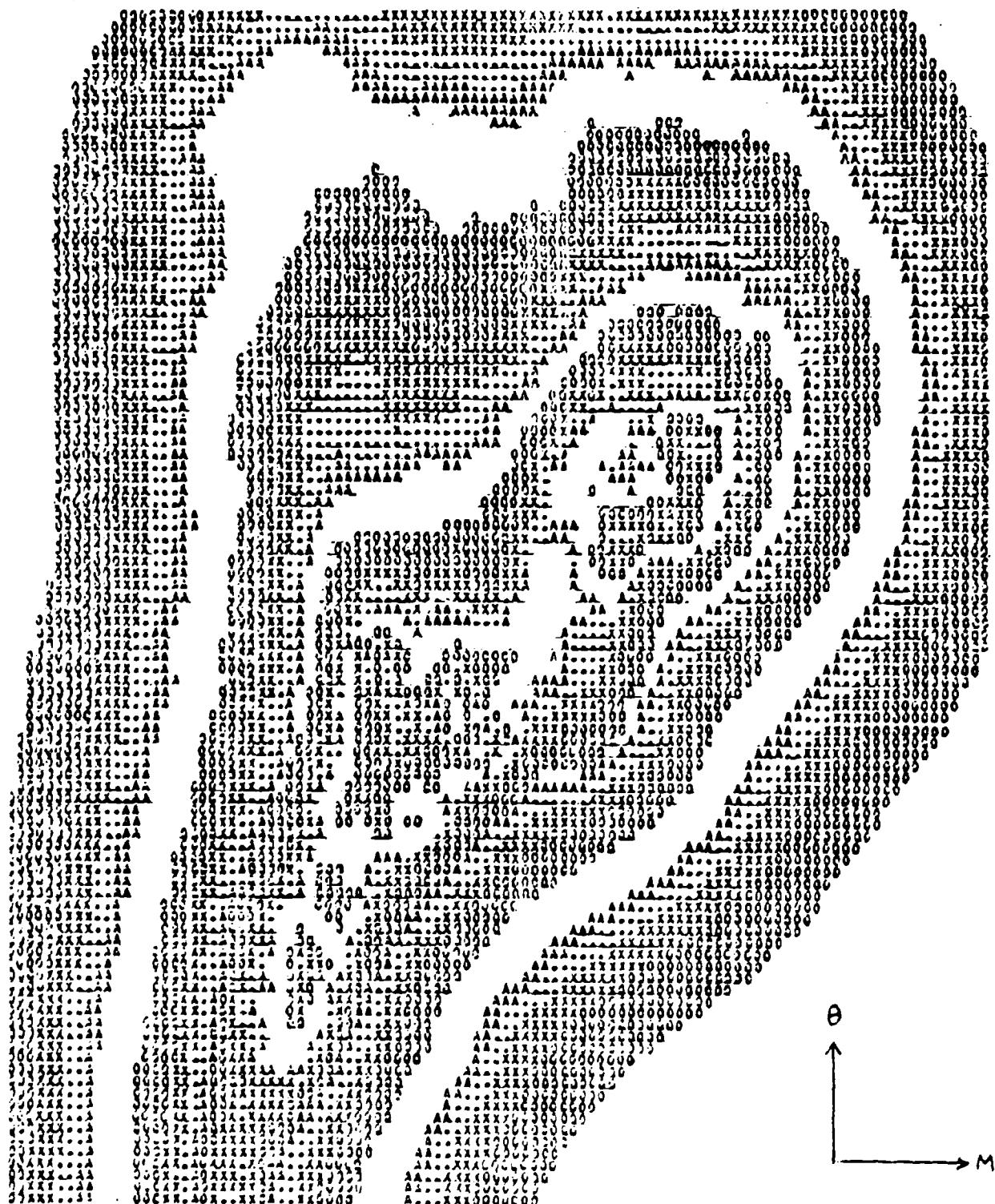
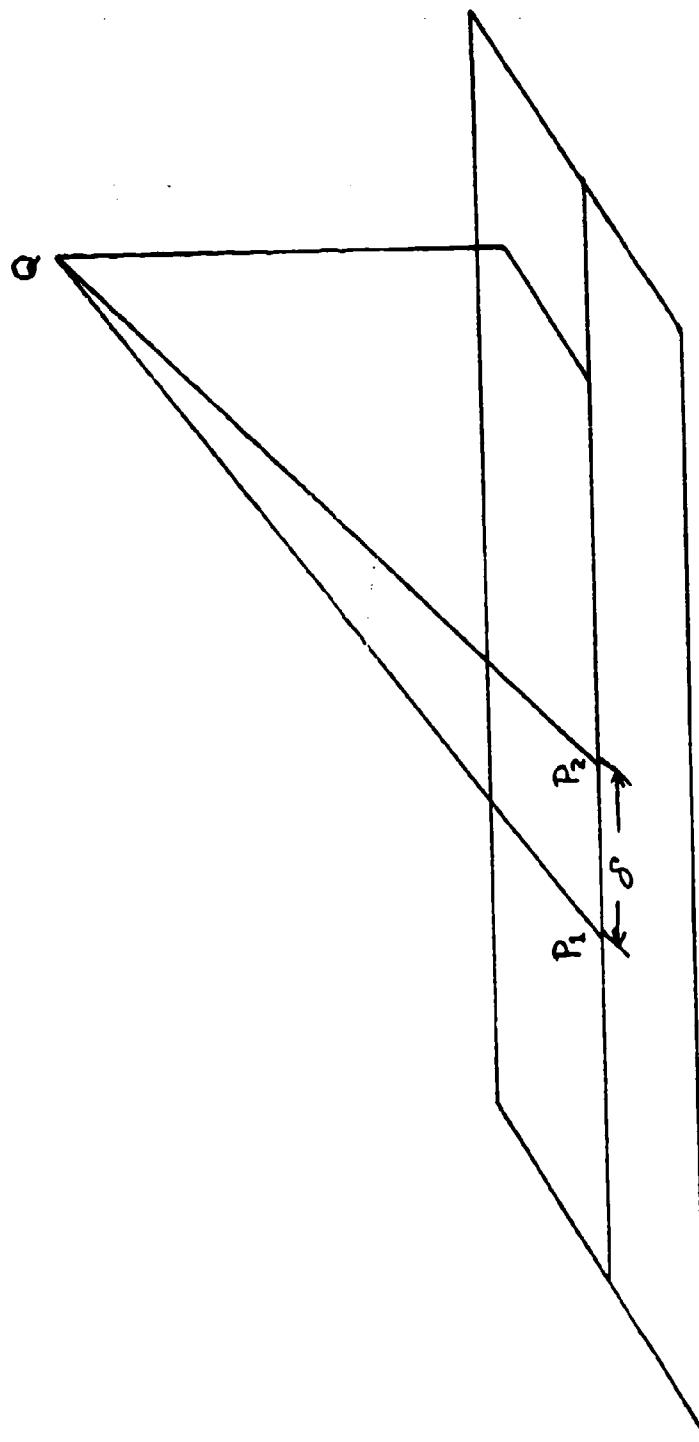


FIGURE 5



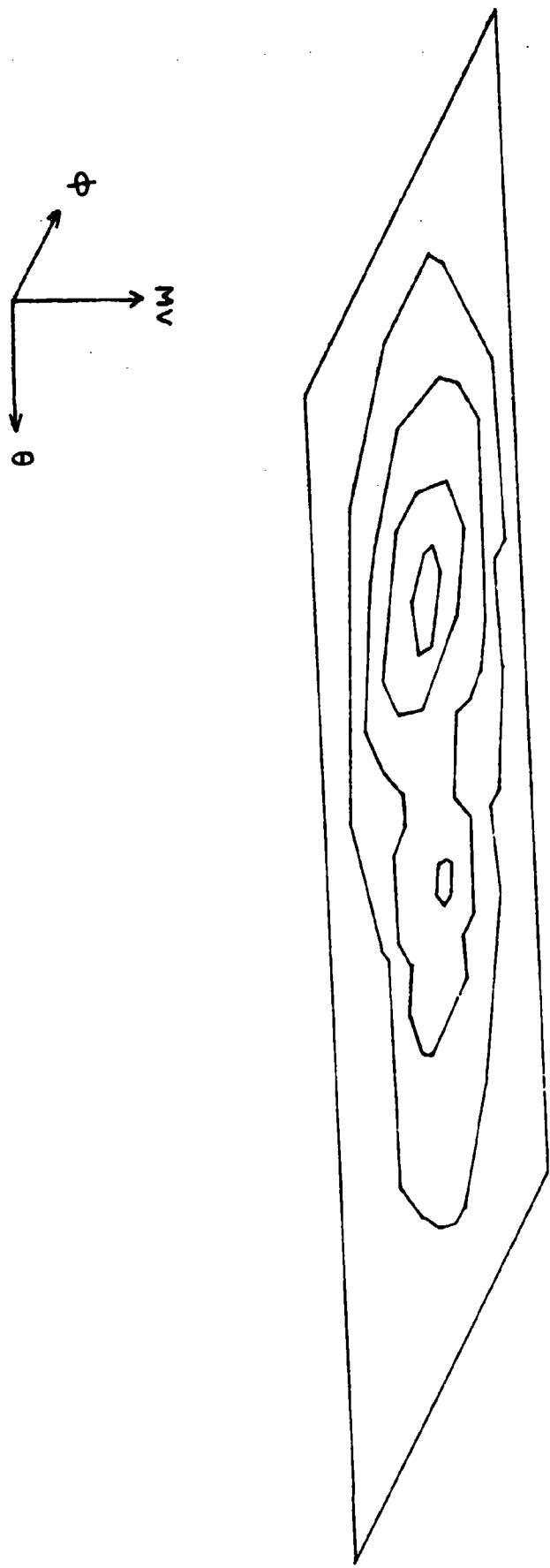
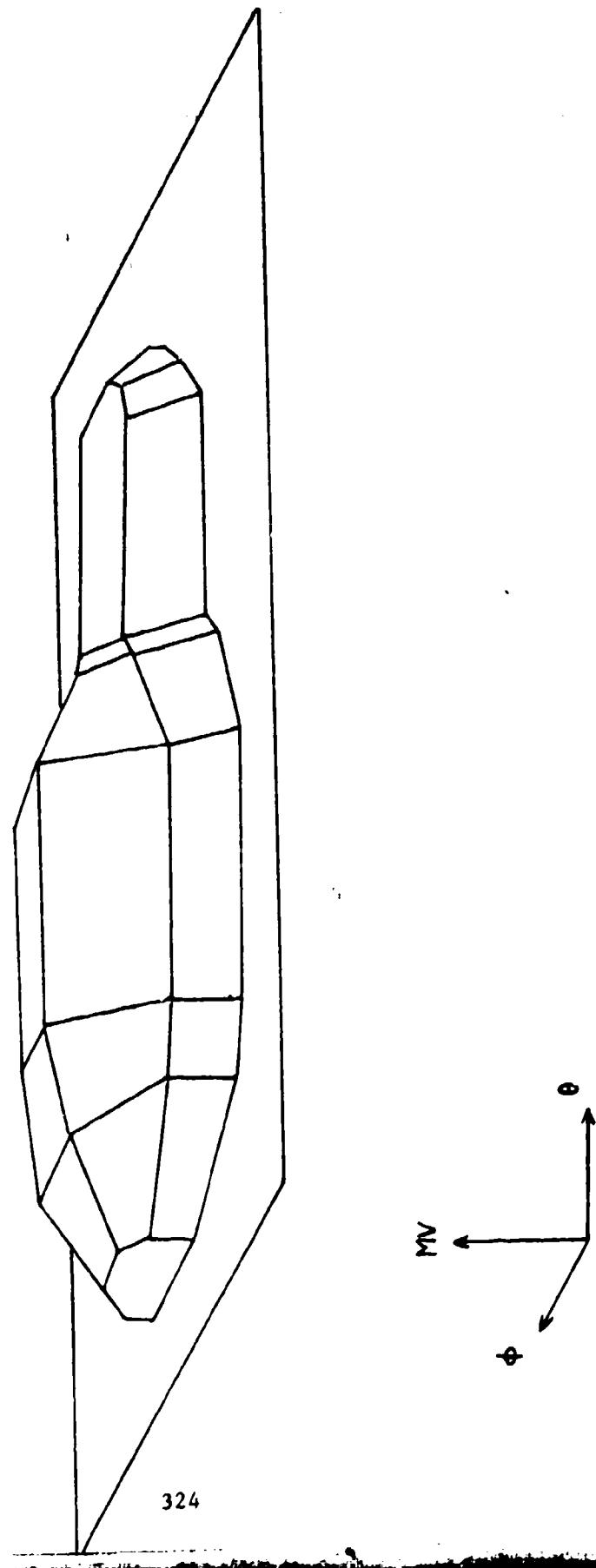
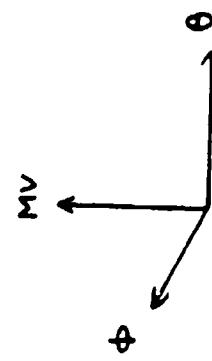
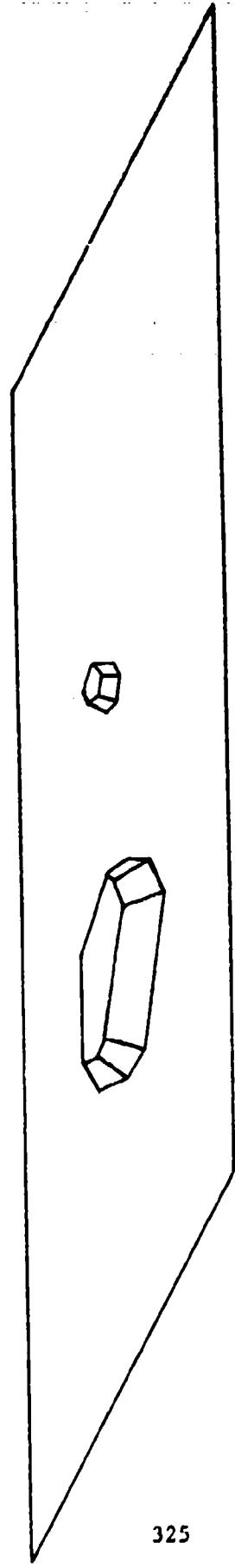


FIGURE 6
($MV = 0$)

FIGURE 7





325

REFERENCES

1. Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., Tukey, J. (1972). Robust Estimates of Location, Princeton University Press: Princeton.
2. Fukunaga, K. and Hostetler, L.D. (1973). Optimization of k-nearest neighbor density estimates. IEEE Trans. Inform. Theory, V. 19, pp. 320-326.
3. Good, I. and Gaskins, R. (1971). Nonparametric roughness penalties for probability densities. Biometrika, V. 36, pp. 149-176.
4. Graunt, J. (1662) Natural and Political Observations on the Bills of Mortality.
5. Subroutine NDMPLE, International Mathematical and Statistical Libraries, Houston, Texas.
6. Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. J. Am. Stat. Assn., V. 63, pp. 925-952.
7. Lehmann, E. (1975). Nonparametrics. Holden-Day: San Francisco.
8. Loftsgaarden, D.O. and Quesenberry, C.P. (1965). A nonparametric estimate of a multivariate density function. Ann. Math. Statist., V. 36, pp. 1049-1051.
9. Mack, Y.P. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. J. Multivariate Analysis, V. 9, pp. 1-15.
10. de Montricher, G. (1973). Nonparametric Bayesian Estimation of Probability Densities by Function Space Techniques. Doctoral dissertation. Rice University.
11. Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Stat., V. 33, pp. 1065-1076.
12. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Stat., V. 27, pp. 832-835.
13. Scott, D., Tapia, R. and Thompson, J. (1977). Kernel density estimation revisited. J. Nonlinear Analysis, V. 1, pp. 339-372.
14. Scott, D., Tapia, R. and Thompson, Jr. (1980). Nonparametric probability density function estimation by discrete maximum penalized-likelihood criteria. Ann. Stat., V. 8, pp. 820-832.
15. Tapia, R. and Thompson, J. (1978). Nonparametric Probability Density Estimation. Johns Hopkins University Press: Baltimore.

TESTABILITY OF LINEAR HYPOTHESES IN NORMAL LINEAR MODELS

Gerald S. Rogers

New Mexico State University

Las Cruces, New Mexico

ABSTRACT. Let a normal linear model be represented by $Y = X\theta + e$. It is shown that the usual F statistic derived from the likelihood ratio can be used to test the hypothesis $H\theta = 0$ independently of any conditions of estimability provided that $\rho(X') + \rho(H') - \rho(X' , H')$ is positive. (ρ denotes the rank of a matrix.) The inherent non-uniqueness leads to the definition of an effective hypothesis: $X\theta$ in the range space of $X(I - H^+H)$; ($+$ denotes the Moore-Penrose generalized inverse.) It is shown that this hypothesis has an estimable form $TX\theta = 0$ and that the procedure is equivalent to a previous definition of "effective".

I. THE LIKELIHOOD RATIO TEST. A basic linear model is representable by $Y = X\theta + e$ where Y is n by 1, X is a given n by p matrix with rank $r \leq p < n$, θ is p by 1, e is an n by 1 normal random variable with mean 0 and covariance matrix $\sigma^2 I_n$.

If the vector θ is an arbitrary element of the p -fold cartesian product with real components, say $\theta \in \Phi$, the parameter space is $\Delta = \{(\theta, \sigma^2) : \theta \in \Phi, \sigma^2 > 0\}$. The hypothesis that θ is in a subspace \emptyset of Φ is represented by $\delta = \{(\theta, \sigma^2) : \theta \in \emptyset, \sigma^2 > 0\}$. Denote a likelihood function by $\text{lik}(\Delta, Y)$ and the ordinary Euclidean norm by $\|\cdot\|$. A generalized likelihood ratio test of the hypothesis is based on $\sup \text{lik}(\Delta, Y) / \sup \text{lik}(\delta, Y)$ which reduces to $\min_{\theta \in \emptyset} \|Y - X\theta\|^2 / \min_{\theta \in \Phi} \|Y - X\theta\|^2$.

Notation: for a matrix W , $R(W)$ is the column range space; $N(W)$ is the column null space; $\rho(W)$ is the rank; W^+ is the Moore-Penrose generalized inverse; $\text{tr}(W)$ is the trace when W is square. $\dim V$ is the dimension of a vector space V .

Let H be a given h by p matrix of rank $h < p$. Suppose that the hypothesis is $H\theta = 0$; that is $\theta \in \emptyset = N(H) = R(Q)$ where $Q =$

$I - H^+H$. Note that Φ is the direct sum of $N(H)$ and $R(H')$; also, H^+H and Q are symmetric idempotent matrices. Then,

$$SSE = \min_{\theta \in \Phi} \|Y - X\theta\|^2 = Y'(I - XX^+)Y$$

$$SSH = \min_{\theta \in \emptyset} \|Y - X\theta\|^2 = \min_{\theta \in \emptyset} \|Y - XQ\theta\|^2 = Y'(I - XQ(XQ)^+)Y$$

(These minima are done neatly in Albert (1972), pages 30-36.)

Now $A = XX^+$, $B = XQ(XQ)^+$, $I - A$, $I - B$ are also symmetric idempotent matrices. It is easily seen that $AB = B$ from which it follows that $BA = B$, $A - B$ is symmetric idempotent and $(A - B)(I - A) = 0$. "Large values" of SSH/SSE correspond to "large values" of $(SSH - SSE)/SSE = Y'(A - B)Y/Y'(I - A)Y$. The quadratic forms in this numerator and denominator are σ^2 multiples of independent chi-square random variables with $t = \rho(A - B)$ and $n - r$ degrees of freedom respectively. The non-centrality parameters are $\theta'X'(A - B)X\theta = \theta'X'(I - B)X\theta$ and $\theta'X'(I - A)X\theta = C$ respectively. Thus $F = Y'(A - B)Y/t + Y'(I - A)Y/(n-r)$ will be a proper F random variable when t is positive.

The import of the non-centrality parameter is discussed in section III; the rank condition is examined first.

II. RANK IN THE FOUR CASES. The basic result is the THEOREM: Given X n by p, H h by p, let $Q = I - H^+H$, $A = XX^+$, $B = XQ(XQ)^+$. Then $t = \rho(A - B) = \dim R(H') \cap R(X') = \rho(X') + \rho(H') - \rho(X', H')$.

Proof: Both A and B are symmetric idempotent matrices and $AB = B$ so also $BA = B$. It follows that $A - B$ is symmetric idempotent so that its rank is equal to its trace: $\rho(A - B) = \text{tr}(A - B) = \text{tr}(A) - \text{tr}(B) = \rho(A) - \rho(B) = \rho(X) - \rho(XQ)$. As part of their Theorem 1, Baksalary and Kala (1976) show that $\rho(X) - \rho(XQ) = \dim R(H') \cap R(X')$. As part of their Theorem 2, they show that $\rho(XQ) = \rho(X', H') - \rho(H')$. The conclusion follows by substitution. #

Of course, this is merely a restatement of the standard result on dimensions of subspaces V_1, V_2 : $\dim(V_1 + V_2) = \dim V_1 +$

$\dim V_2 - \dim V_1 \cap V_2$. The following four cases can occur; only the first three cases have been recognized previously.

1. $R(H') \cap R(X') = \{0\}$ iff $t = 0$ iff $A = B$. This is the case in which the error sums of squares are equal and there is no F test.

2. $R(H') \cap R(X') = R(X')$ iff $R(X') \subset R(H')$ iff $t = \rho(X)$ iff $\rho(XQ) = 0$ iff $B = 0$ iff $X = XH^+H$. Here $SSH = Y'Y$ and the F test is equivalent to that for testing $X\theta = 0$ versus $X\theta \neq 0$. Note that $\rho(X) \leq \rho(H)$.

3. $R(H') \cap R(X') = R(H')$ iff $R(H')$ is a proper subset of $R(X')$ (the equality to be considered in 2) iff $t = \rho(H)$ only if $H = HX^+X$. This is the usual condition that $H\theta$ be estimable and the F test is the common one. Note that $\rho(H) < \rho(X)$. In particular, this is the case when X has full rank p as in ordinary regression.

4. $R(H') \cap R(X')$ is a subspace with dimension $0 < t < \min(\rho(X), \rho(H))$ as in the example below. Within the context of section III, the F test is valid.

EXAMPLE Factor one has 2 levels and factor two has 3 levels; there are k observations in each cell and only main effects are to be considered. Let $\theta = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3)'$ and

$$M = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \quad H = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Then $X = (I_6 \otimes l_k)M = M \otimes l_k$ where \otimes denotes the Kronecker product and l_k is an k by 1 vector of all ones.

$$X'X = k \begin{pmatrix} 6 & 3 & 3 & 2 & 2 & 2 \\ 3 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{pmatrix} \quad (X'X)^+ = \frac{1}{k} \begin{pmatrix} 2/3 & -1/3 & 0 & -1/2 & -1/2 & 0 \\ -1/3 & 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & 0 & 0 & 1 & 1/2 & 0 \\ -1/2 & 0 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

It is easily seen that $H \neq HX^+X$. But columns 1 and 2 of H' sum to column 2 of M' and columns 1 and 3 of H' sum to column 3 of M' . There are no other such linear dependencies so in this case, $\dim R(H') \cap R(X') = 2 < 3 = \rho(H') < 4 = \rho(X')$. #

On page 194, Searle (1971) says that when $H \neq HX^+X$, the rows of H and X^+X are linearly independent. The example shows clearly that this is not true in general: the difference of the last two rows of X^+X is equal to $2k$ times the difference of the last two rows of H . In a subsequent publication, Searle (1973) has discussed other errors and their corrections.

The usual manipulations (as appear for example in Rao and Mitra (1971) chapter 7) may be used to find the rank conditions when $H\theta = h \neq 0$, when there are also constraints $G\theta = g$ consistent with $H\theta = h$ and when the covariance matrix is $\sigma^2 V$ with V known. Such details are given in Rogers and Urquhart (1980).

III. THE EFFECTIVE HYPOTHESIS. Note that $\theta'X'(I - B)\theta = 0$ iff $(I - B)\theta = 0$ iff $X\theta \in N(I - B) = R(XQ)$. It is obvious from the geometry that the generalized likelihood ratio procedure will lead to this same F whenever a hypothesis projects $X\theta$ into $R(XQ)$ whether or not $H\theta$ is estimable. Therefore, an anomaly of F is that it is a basis for a test of not one but many different hypotheses all of which lead to the same effective hypothesis: $X\theta \in R(XQ)$. (Actually, it is not necessary that H have full row rank but only that $\rho(H) < p$.)

There is in fact an estimable one of these hypotheses as is implicit in results of Scheffé (1959, page 34). Let T' be an n by t matrix whose columns generate the orthogonal complement of $R(XQ)$ in $R(X)$. Since $TXQ = 0$, $TX\theta = 0$ when $H\theta = 0$. But $TX\theta = 0$ implies $X\theta \in N(T)$ which is the orthogonal complement of $R(T')$; thus $X\theta \in R(XQ)$. Therefore, SSE is the minimum when $TX\theta = 0$ of $\|Y - X\theta\|^2$ and the same F is obtained. Obviously, $TX\theta$ is an estimable function. Since $d'TX = 0$ iff $T'd$ is orthogonal to $R(X)$, $d'TX = 0$ implies $T'd = 0$ which in turn implies $d = 0$; thus $\rho(TX) = \rho(T) = t$.

EXAMPLE For the model of section II with $k = 1$, $X = M$.

$$XQ = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ .5 & -.5 & 1 & 1 & 0 & 0 \\ .5 & -.5 & 1 & 0 & 0 & 0 \\ .5 & -.5 & 1 & 0 & 0 & 0 \end{pmatrix} \quad T' = \begin{pmatrix} .4 & 0 \\ 1 & .5 \\ 1 & -.5 \\ -.4 & 0 \\ .2 & .5 \\ .2 & .5 \end{pmatrix}$$

$$TX = \begin{pmatrix} 2.4 & 2.4 & 0 & 0 & 1.2 & 1.2 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} #$$

HSC (Hocking, Speed and Coleman (1980)) used "effective model", "effective constraint" and "effective hypothesis" in a discussion of the cell means model. In order to see the equivalence with the present definition of effective, it is convenient to use a factorization of the X matrix as is illustrated in the following

EXAMPLE Suppose that in the model of section II, observations are available only for cells 12 13 21 23. The corresponding incidence matrix is

$$U = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{Let } S \text{ be a block}$$

diagonal matrix with "diagonal" $l_{k1}, l_{k2}, l_{k3}, l_{k4}$ representing (possibly) different numbers of replications in the "observed" cells. Then the new $X = SUM$. For the cell means model, $v = M\theta$ and $X\theta = SUv = S(v_{12}, v_{13}, v_{21}, v_{23})'$. #

In general for a cell means model, $E[Y] = SUv = Sv_0$ where v_0 is the vector of means of cells for which there are observations; say v_0 is q by 1. Index the cells so that $v = (v_m', v_o')$ where v_m is the vector of means of cells which are "missing"; say v_m is m by 1 so $p = m + q$. Now U takes the form $(0, I_q)$. For a hypothesis $Gv = 0$ (G being g by p of rank g), without loss of generality, one may use row reductions to write this as

$$\begin{pmatrix} G_{mn} & G_{mo} \\ 0 & G_o \end{pmatrix} \begin{pmatrix} v_m \\ v_o \end{pmatrix} = 0$$

where G_o has maximal rank. The effective hypothesis of HSC is $G_o v_o = 0$ in the effective model $Y = S v_o + e$. By the techniques of the present paper, this leads to $E[Y] \in R(SQ_o)$ with $Q_o = I - G_o^T G_o$. For $Gv = 0$ in the model $Y = SUv + e$, the relevant subspace is $R(SUQ)$ with $Q = I - G^T G$.

Since G has full row rank, $G^+ = G'(GG')^{-1}$. Let $(GG')^{-1}$ be partitioned appropriately with "rows" $C D / J L$. It turns out that UQ has "columns"

$$-(G'_{mo} C + G' J) G_{mm} / I_q = (G'_{mo} C + G' J) G_{mo} = (G'_{mo} D + G' L) G_o$$

Then from $GG^+ G = G$, one gets $G_o UQ = 0$ and hence $Q_o UQ = UQ$.

Now $R(SUQ) \subset R(SQ_o)$ iff $SQ_o (SQ_o)^{-1} SUQ = SUQ$. (See Rao and Mitra (1971); $"-"$ denotes any generalized inverse.) Since $S^+ = (S'S)^{-1} S'$ and Q_o is symmetric idempotent, one form of $(SQ_o)^{-1}$ is $Q_o S^+$: $SQ_o Q_o S^+ S Q_o = S Q_o Q_o Q_o = S Q_o$. Hence, $R(SUQ) \subset R(SQ_o)$ iff $SQ_o UQ = SUQ$ iff $Q_o UQ = UQ$ which has just been demonstrated. But $Q_o UQ = UQ$ iff $R(UQ) \subset R(Q_o)$ since $Q_o^+ = Q_o$. Now $\rho(UQ) = \rho(U', G')$ $= \rho(G') = \rho(G_{mm}) + q = \rho(G) = q = (\rho(G) - \rho(G_{mm})) = q - \rho(G_o)$ $= \rho(Q_o)$ so that these subspaces are equal. Thus $Y = SUv + e$ with $Gv = 0$ and $Y = S v_o$ with $G_o v_o = 0$ lead to projection of $E[Y]$ into the same subspace and consequently have the same SSH with degrees of freedom $n - q + \rho(G_o)$.

When $Gv = 0$ represents a constraint rather than a hypothesis, $Y'(I - SUQ(SUQ)^+)Y = Y'(I - SQ_o (SQ_o)^+)Y = SSE$ rather than SSH. Then in the method of HSC, $G_{mm} v_m + G_{mo} v_o = 0$ is used to eliminate all or part of v_m from the hypothesis $Hv = 0$. Suppose that this produces, $H_{ml} v_{ml} + H_{mo} v_o = 0$; then row reductions are used to write (H_{ml}, H_{mo}) as $\begin{pmatrix} H_{ml}^* & H_{mo}^* \\ 0 & H_o \end{pmatrix}$

with H_o of maximal rank; their effective hypothesis is $H_o v_o = 0$. As above (with messier details), $SSH = Y'(I - SUQ(SUQ)^+)Y = Y'(I - SQ_o (SQ_o)^+)Y$ for $Q = Q' = I - (G', H')(G', H')^+$ and $Q_o = Q' = I - (G_o', H_o')(G_o', H_o')^+$. Here the degrees of freedom is $n - q + \rho(G_o', H_o')$ and $t = \rho(G_o', H_o') - \rho(G')$.

IV. REMARKS. For the balanced model considered previously

with no missing cells, let $H\theta = (0, I_5)\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3)'$. Then, $Q = I - H^+H$ has rows $1_0 / 0_0$ and $XQ = (1_n, 0)$ where the zeroes denote zero matrices of appropriate dimensions. Now $I - XQ(XQ)^+ = I - 1_n 1_n' / n$ has rank $n - 1$, $n = 6k$, and $\rho(X) = 4$. The numerator in the F test for $H\theta = 0$ has degrees of freedom $(n - 1) - (n - 4) = 3 = t$.

In fact, for any X n by p with first column 1_n , and $H = (0, I_{p-1})$, $H\theta = 0$ is testable in the same way though $H\theta$ need not be estimable; $R(XQ) = R(1_n, 0) = R(1_n)$.

In analysis of variance, the columns of X' represent cell means: t will be positive iff the hypothesis $H\theta = 0$ contains implicitly the hypothesis that (at least) one linear combination of means of cells for which there are observations is zero. An example is the TX in section III which yields

$$\nu_{12} + \nu_{13} = 0 \quad \text{and} \quad \nu_{22} - \nu_{23} = 0$$

Rao and Mitra (1971) show that varying W arbitrarily in $H^+ + W - H^+HWHH^+$ generates all generalized inverses H^- . Then, $Q_a = I - H^-H = (I - H^+H)(I - WH)$ say QW . The results in Baksalary and Kala (1976) are proved actually for H^- so $\rho(XQ_a) = \rho(X', H')$ = $\rho(H') = \rho(XQ)$ and $R(XQ_a) = R(XQ)$. The use of "−" usually simplifies the calculations.

Note that the ranks may be calculated at any convenient stage in $\rho(X) = \rho(X'X) = \rho(XX^+)$ = $\text{tr}(XX^+)$ and similarly for $\rho(H)$, $\rho(X', H')$. Hemmerle (1979) discusses computer aspects. This chore can be further simplified by factorization of X as indicated above. The new $X = \text{SUM}$ and $\rho(X) = \rho(X'X) = \rho(M'U'S'\text{SUM}) = \rho(M'U'UM) = \rho(UM)$ since $S'S$ is positive definite. Similarly, for any Q , $\rho(XQ) = \rho(UMQ)$, etc. When there are no missing cells, U is an identity matrix.

The author extends his thanks to Mary Ann Maher, N. Scott Urquhart and David V. Hinkley for stimulating discussions of this topic which consequently appears in a different form than that presented at the Conference.

REFERENCES

- [1] Albert, A. (1972) Regression and the Moore-Penrose Pseudoinverse, Academic Press, New York.
- [2] Baksalary, J.K. and Kala, R. (1976) Extensions of Milliken's estimability criterion, Annals of Statistics 4 639-641.
- [3] Hemmerle, W.J. (1979) Balanced hypotheses and unbalanced data, J. Amer. Statist. Assoc. 74 794-798.
- [4] Hocking, R.R., Speed, F.M. and Coleman, A.T. (1980) Hypotheses to be tested with unbalanced data, Commun. Statist. Theor-Meth A9(2) 117-129.
- [5] Rao, C.R. and Mitra, S.K. (1971) Generalized Inverse of Matrices and its Applications, Wiley and Sons, New York.
- [6] Rogers, G.S. and Urquhart, N.S. (1980) Testability of linear hypotheses in normal linear models, Tech. Rpt. no. 36 NMSU Statist. Lab., Las Cruces, N.M.
- [7] Scheffe, H. (1959) The Analysis of Variance, Wiley and Sons, New York.
- [8] Searle, S.R. (1971) Linear Models, Wiley and Sons, New York.
- [9] Searle, S.R. (1973) Testing non-testable hypotheses in linear models: corrections, Biometrics Unit, Cornell University Ithaca, N.Y.

THE POTENTIAL UTILITY OF CROSSING A FRACTIONAL FACTORIAL WITH A FULL FACTORIAL IN THE DESIGN OF FIELD TESTS

Carl T. Russell¹
US Army Cold Regions Test Center
Fort Greely, Alaska

ABSTRACT. Typically in the design of a field test, there are more factors than can be completely tested (in a full factorial) within resource constraints. Usually one or two "primary" factors and a few "secondary" factors are selected as design variables and all other factors of possible interest are either fixed or ignored. Traditionally, "replicates" of several overlapping full factorials in the primary factor and one or two secondary factors at a time are conducted, yielding a design made up of many side tests and sensitivity tests. This paper presents an alternative to the traditional design approach in terms of an example which calls for conducting full factorials in the primary factor with combinations of levels of secondary factors specified by a fractional factorial in the secondary factors. Potential pragmatic and analytical advantages and disadvantages of such designs are discussed with emphasis on design flexibility and inherent blocking schemes.

I. INTRODUCTION. Within the Army testing community, a test design for a field test is usually regarded as synonymous with a test matrix, that is, a matrix identifying combinations of controlled test conditions and the number of times each combination is to be tested. As a result, both resource managers and test operators tend to become infatuated with sample size: the resource manager often bases judgements primarily on the "number of replications" while the test operator concentrates primarily on obtaining the "requisite sample size" in each cell as efficiently as possible. I think that as statisticians, we should find ways to deemphasize such summary matrices and emphasize test structure in a manner marketable to the testing community.

The structure I refer to is blocking. In my experience, all field tests are conducted in blocks of time or space. Block sizes are usually dictated by resource constraints, but block contents are most often dictated by expedient completion of the "requisite sample size," leading to extensive confounding of possible block effects with factor effects of interest. The standard request that order of trials be randomized (as much as possible within test constraints) over the entire test matrix does not give the test operator the usable systematic statistical advice he should have for test conduct. Consequently, extensive randomization seldom occurs in test conduct, a fact usually ignored during data analysis.

¹ Much of the work underlying this paper was done while the author was a Research Staff Member, Systems Evaluation Division, Institute for Defense Analyses, Arlington, Virginia.

I believe that for most field tests, the resource structure suggests a natural structure which can be exploited by the test designer to produce a fairly small basic test matrix addressing test factors of interest. This matrix could then be conducted in blocks with both block order and order of trials in blocks randomized as much as possible within test constraints. (Notice that this requested randomization is fairly small-scale and should fit naturally into test conduct, hopefully yielding substantial actual randomization and hopefully precluding gross damage in case of non-randomization.) The basic matrix would then be repeated, possibly with some factors or factor levels deliberately changed in sensitivity tests. This intuitively simple "basic matrix approach" can yield highly structured designs which not only are executable but also permit refined analysis.

After discussing the traditional approach to field test design in the special case where the goal is to compare levels of a primary factor in the presence of several secondary factors, this paper applies the basic matrix approach to a particular test resource structure involving primary and secondary factors. The resulting example illustrates the flexibility and potential analytic richness inherent in the basic matrix approach.

II. THE TRADITIONAL DESIGN APPROACH. Often in field testing there is a controllable factor of primary interest² (say, G, at four levels) and there are several controllable factors of secondary interest (say, A, B, C, and D, each at two levels). The traditional approach then evolves a design along the following lines. Initially it is determined that a sample size of 20 trials per cell is desirable based on the "rule of fingers and toes."³ A $2^4 \times 4$ full factorial with 20 "replications per cell, however, requires 1280 trials, so the designer shrinks the number of cells by selectively deleting cells to obtain a test design with 20 "replications" per cell made up of three overlapping $2^2 \times 4$ full factorials, $A \times B \times G$, $A \times C \times G$, and $B \times D \times G$ (the suppressed factors are fixed at their low levels). This yields a matrix of 640 trials which appears in the approved Test Design Plan. By test date, however, Higher Headquarters has determined that resource constraints will permit only 320 trials and that two other factors (previously considered to be fixed) should be varied in sensitivity tests. Adding two new factors E and F and redistributing the shrunken sample size yields the test matrix of Table 1, which is fairly typical of a design for a large field test.

² The illustrations in this paper deal with just one primary factor at several levels. An obvious generalization regards these levels as combinations of levels of two or more factors making up a full factorial in those factors.

³ This rule states that anything likely to be operationally significant is likely to show up in samples countable on fingers and toes and vice-versa.

TABLE 1. A Typical Test Matrix Arising From
the Traditional Design Approach

Secondary Factors				Sensitivity Tests		Primary Factor(G)			
A	B	C	D	E	F	g_1	g_2	g_3	g_4
a_1	b_1	c_1	d_1	e_1	f_1	9	9	9	9
					f_2	4	4	4	4
		c_2	d_2	e_2	f_1	4	0	0	4
				e_2	f_1	4	0	0	4
		c_2	d_1	e_1	f_1	9	9	9	9
	b_2	c_1	d_1	e_1	f_1	9	9	9	9
		c_2	d_2	e_1	f_1	9	9	9	9
	b_1	c_1	d_1	e_1	f_1	9	9	9	9
	b_2	c_1	d_1	e_1	f_1	9	9	9	9
a_2									

Table entries are the number of trials to be conducted under each test condition. The order in which trials are conducted is to be randomized as much as possible within test constraints.

This design still contains the three overlapping $2^2 \times 4$ full factorials referred to earlier (now with only 9 "replications" per cell); these subdesigns examine the effects of A, B, C, and D and their possible interactions with G in "side tests," each involving 45 percent of all test trials. In addition, the design contains two "sensitivity tests," each involving 10 percent of all test trials: the 2×4 design F \times G (with 4 "replications" per cell⁴ and with all other factors fixed at low levels) and the 2^3 design D \times E \times G* (with 4 "replications" per cell⁴, with factor G restricted to two levels, and with all other factors fixed at low levels). Analysis would typically consist of analyzing data from each subdesign separately, assuming randomization was complete.

* Actually these subdesigns call for either 4 or 9 trials per cell and would probably be analyzed as unbalanced designs; in spirit, however, they have 4 "replications" per cell.

The traditional design approach has two main practical advantages. First, it is accepted: it is at heart a simple "vary-one-factor-at-a-time" approach to experimentation, and it is time-proven. Second, it is flexible: pre-test reduction of resources or addition of factors can easily be accommodated by re-distributing the available sample size, and the lack of detailed structure hides damage due to data loss during test execution.

However, the traditional design approach has three great statistical disadvantages. The first two disadvantages are technical: the resulting designs are inefficient and give main effects biased by (non-estimable) 2-factor interactions. The third and most important disadvantage is pragmatic: by emphasizing sample size rather than design structure the traditional approach permits sample size to be whittled down cell-by-cell, and it fails to give systematic statistical guidance for detailed test conduct which might preclude day-to-day test scheduling from totally dominating randomization and which could provide a solid base for accurate statistical inference.

III. REQUIREMENTS FOR AN ALTERNATIVE DESIGN APPROACH. To be marketable to the testing community (and executable in the field) any alternative to the traditional design approach

- must be intuitively (if not analytically) simple
- must conscientiously consider problems of test scheduling and control
- must permit insertion of meaningful sensitivity tests
- must degrade gracefully in the face of resource reductions or substantial data loss

In actuality, the traditional design approach meets only the first of these requirements well. The complete randomization prescribed by the traditional approach does not consider problems of test scheduling and control, and it is seldom well-followed during test execution. Sensitivity tests frequently involve too few trials for conclusive results, and many of those trials are not usable for any other purpose. Furthermore, reductions in sample size, through either resource reductions or data loss, are accommodated easily only because traditional designs lack detailed structure which could illuminate the consequences of sample size reductions in terms other than square root of sample size ratio.

The example given now shows that a design generated by the basic matrix approach can meet these requirements.

VI. AN EXAMPLE BASED ON CROSSING A FRACTIONAL FACTORIAL WITH A FULL FACTORIAL. The basic matrix approach was implemented, as summarized in Figure 1, in a design proposed for a test with the goal of comparing 5 levels of a primary factor under a variety of operational conditions, some of which could be specified as combinations of secondary factor levels.

Identify Basic Test Matrix

Secondary Factors	Primary Factor q ₁	q ₂	•	•	•	q _n
List of combinations	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•

Run Basic Test Matrix in Blocks

Secondary Factors	Primary Factor q ₁	q ₂	•	•	•	q _m
Comb. 1	✓	✓	✓	•	•	✓
Comb. 2	✓	✓	✓	✓	•	✓

•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•

Comb. n	✓	✓	✓	•	•	✓
•	•	•	•	•	•	•

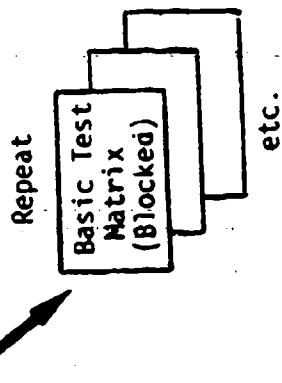


FIGURE 1. Basic Matrix Approach

The resource structure for the test was as follows. There were three six-week test periods, between which substantial variations in equipment, personnel, and SOP's were expected. Each test period consisted of three two-week test segments, between which personnel were expected to change but equipment and SOP's were not. It was suspected that crew learning might be substantial within test segment, and it was anticipated that trials could be conducted at the rate of 4 trials per day for 10 days per test segment, yielding 40 trials per test segment. Since there were 5 levels of the primary factor and 40 trials per test segment appeared to be feasible, it was natural to define a basic matrix calling for all five levels of the primary factor to be tested under 8 combinations of secondary factor levels during each test segment. Since $8 = 2^3 = \frac{1}{2} \times 2^4$, either 3 or possibly 4 secondary factors, each at 2 levels, could be accommodated in the basic matrix. Conveniently, there are half replicates of the 2^4 design which are of resolution IV (that is, all four main effects are estimable free from 2-factor interactions provided 3-factor and higher-order interactions are suppressed) furthermore, these half-replicates are fold-over designs (they can be run in blocks of size 2 without losing the resolution IV property). One of those half replicates was therefore chosen to define the 8 combinations of secondary factor levels for the basic test matrix which appears as Table 2.

TABLE 2. The Basic Test Matrix

Secondary Factors				Primary Factor(G)				
A	B	C	D	g ₁	g ₂	g ₃	g ₄	g ₅
a_1	b_1	c_1	d_1	α	α	α	α	α
		c_2	d_2	β	β	β	β	β
	b_2	c_1	d_2	γ	γ	γ	γ	γ
		c_2	d_1	δ	δ	δ	δ	δ
a_2	b_1	c_1	d_2	δ	δ	δ	δ	δ
		c_2	d_1	γ	γ	γ	γ	γ
	b_2	c_1	d_1	β	β	β	β	β
		c_2	d_2	α	α	α	α	α

The basic matrix is to be run in four blocks ($\alpha, \beta, \gamma, \delta$) of ten trials each. The order in which the four blocks are run should be chosen at random each time the matrix is run, and the order in which trials are conducted within blocks should be randomized as much as possible within test constraints.

The blocking scheme in Table 2, together with its associated randomization scheme, was chosen to exploit the fold-over property of the fractional factorial while providing an executable design and a randomized block structure on the primary factor. Within the basic matrix, main effects of both primary and secondary factors, as well as primary-secondary 2-factor interactions, are estimable free from blocks and secondary-secondary 2-factor interactions. Secondary-secondary two factor interactions are confounded with each other in pairs and also confounded with blocks. The randomization scheme does not severely constrain the test operator; instead the randomization scheme together with the blocking scheme provide the test operator with the sort of systematic statistical guidance he should have to obtain a statistically defensible data set. Because block order is randomized (this should be entirely executable because of similar structure within different blocks), some inference regarding order effects (possible "crew learning") might be possible even though such order effects would be confounded in a complicated random way with suppressed 2-factor interactions between secondary factors. The requested randomization of trials within blocks could probably not be complete because it is unlikely that all of the changes in factor levels between an arbitrary pair of trials would be practical. All trials within a block, however, could reasonably be conducted within a few days--which would give each block reasonable analytic interpretation--and even partial randomization should preclude systematic within block bias. Moreover, each original block could be regarded as two blocks in the levels of the primary factor, each defined by one combination of secondary factor levels. Thus the blocking and randomization schemes make the core of the design a very simple randomized complete blocks design in the levels of the factor of primary interest.

Several assumptions are made in what follows. First, it will be assumed that 3-factor and higher-order interactions among primary and secondary factors are zero. For convenience, possible 2-factor secondary-secondary interactions will also be suppressed (which ignores a possible source of bias in the block effects discussed). It will also be assumed that the error terms, ε , are uncorrelated random variables with mean zero and the same variance. Although least squares estimation is implicit in the discussion, linear models are discussed more to illuminate the potential structure of data sets produced by basic matrix designs than to provide specific analytic methods.

A. The Initial Model. Provided the basic matrix were re-run nine times (once for each test segment in each test period) and there were no interactions of primary or secondary factors with blocks, the following linear model would be appropriate (where β_{pqS} is the [probably random] effect of the q^{th} block in the s^{th} test segment of the p^{th} test period).

Model I:

$$Y_{ijklmnpqs} = \mu + g_i + a_j + (ga)_{ij} + b_k + (gb)_{ik} + c_m + (gc)_{im} + d_n + (gd)_{in} + \beta_{pq} + \epsilon_{ijklmnpqs} \quad (1)$$

Side Conditions

$$\begin{aligned} 0 &= g_i = a_j = b_k = c_m = (ga)_{ij} = (ga)_{ji} = (gb)_{ik} = (gb)_{ki} \\ &= (gc)_{im} = (gc)_{mi} = (gd)_{in} = (gd)_{ni} = \beta_{pq} = 0 \end{aligned} \quad (2)$$

B. Modifications and Extensions of the Initial Model. Although Model I has 60 independent parameters, illustrating the rich analytic structure potentially available in a data set generated from the basic matrix approach, this model can be modified by reinterpreting existing parameters and can be extended by adding new parameters.

The most obvious modification of the initial model is a reinterpretation of the block effects. The q^{th} block in the s^{th} segment of the p^{th} test period contains interblock information on potential effects of period, π_p , segment within period, ψ_{ps} , and order or "crew learning," $\lambda_q + (\pi\lambda)_{pq} + (\lambda\psi)_{qps}$; that is, the 35 independent block effects, β_{pq} , can be reinterpreted as

$$\beta_{pq} = \pi_p + \psi_{ps} + \lambda_q + (\pi\lambda)_{pq} + (\lambda\psi)_{qps}, \quad (3)$$

$$\begin{aligned} 0 &= \pi_p = \psi_{ps} = \lambda_q = (\pi\lambda)_{pq} = (\pi\lambda)_{qps} = (\lambda\psi)_{qps} \\ &= (\lambda\psi)_{ps} = 0. \end{aligned} \quad (4)$$

In addition, the whole basic matrix is crossed with test segments and test periods, so all interactions of test segment and test period with primary and secondary factors are estimable and can be incorporated into the model as desired. In particular, the fact that rather large changes in test conditions were expected from test period to test period suggests that possible interactions of test period with primary and secondary factors be incorporated into the model. This is done in Model II, which also allows, as an example, possible interactions of test segment with factors G and A.

MODEL II

$$\begin{aligned}
 Y_{ijklmnpqrs} = & \mu + \pi_p + \psi_{ps} + g_i + (gn)_{ip} + (g\psi)_{ips} + a_j + (an)_{jp} \\
 & + (a\psi)_{jps} + (ga)_{ij} + (gan)_{ijp} + (ga\psi)_{ijps} + b_k + (bn)_{kp} \\
 & + (gb)_{ik} + (gbn)_{ikp} + c_m + (cn)_{mp} + (gc)_{im} + (gc\pi)_{imp} \\
 & + d_n + (dn)_{np} + (gd)_{in} + (gd\pi)_{inp} + \lambda_q + (\pi\lambda)_{pq} + (\lambda\psi)_{qps} \\
 & + \epsilon_{ijklmnpqrs}
 \end{aligned} \tag{5}$$

Side Conditions in Addition to (2) and (4)

$$\begin{aligned}
 0 &= (gn)_i = (gn)_{\cdot p} = (g\psi)_{ip} = (g\psi)_{\cdot ps} = (an)_j = (an)_{\cdot p} = (a\psi)_{jp} \\
 &= (a\psi)_{\cdot ps} = (gan)_{ij} = (gan)_{\cdot p} = (gan)_{\cdot jp} = (ga\psi)_{ijp} = (ga\psi)_{\cdot ps} \\
 &= (ga\psi)_{\cdot jps} = (bn)_k = (bn)_{\cdot p} = (gbn)_{ik} = (gbn)_{\cdot p} = (gbn)_{\cdot kp} \\
 &= (cn)_m = (cn)_{\cdot p} = (gc\pi)_{im} = (gc\pi)_{\cdot p} = (gc\pi)_{\cdot mp} = (dn)_n \\
 &= (dn)_{\cdot p} = (gd\pi)_{ij} = (gd\pi)_{\cdot p} = (gd\pi)_{\cdot np} = 0
 \end{aligned} \tag{6}$$

C. Inserting Sensitivity Tests. The 102 independent parameters added in passing from Model I to Model II potentially enable stronger inferences by explicitly accounting for interaction effects of nuisance parameters. They also enable the insertion of sensitivity tests by allowing certain factors to change from period-to-period or segment-to-segment.

The simplest way to introduce a sensitivity test is to replace one of the secondary factors in the basic matrix by other factors for one or more test periods. For instance, factor B could be replaced by three factors B_1 , B_2 , and B_3 , each at 2 levels and each varied during one test period (and fixed during the other two periods). Then B_p would be nested in test period p , and its main effect, $(b_p)_k$, as well as its interactions with the primary factor, $(gb_p)_{ik}$, would have meaning only within period p . The 15 independent Model II parameters involving factor B would be replaced by the 15 independent parameters $(b_p)_k$ and $(gb_p)_{ik}$ with $b_p = (gb_p)_k = 0$. Adding a sensitivity test in this manner does not impact the precision or bias of the estimators of any primary or secondary effects except those of B. However, the estimator of each estimable effect involving B_p is based on only one third the observations for corresponding estimators involving other secondary factors, and each is biased by possible 2-factor interactions with test period (which are nonestimable).

Alternatively, a factor could be replaced for one or more test segments in each test period, but such a situation will not be discussed here. Instead, a different method for adding a sensitivity test is considered.

Suppose, for example, that factor A represented two types of jamming and a sensitivity test were desired to investigate whether different methods of employment of each type jamming had any substantial effects. By selecting three different methods of employment and applying each during one randomly selected test segment of each test period, the desired sensitivity test could be accommodated in the design and in the model. This essentially confounds a new factor, E, with test segment. The six independent parameters ψ_{ps} , each reinterpreted as being the effect of the segment of period p for which factor E is at level s, lead to six independent parameters $e_s = \psi_{s.s}$ and $(\pi e)_{ps} = \psi_{ps} - e_s$ with $e_s = (\pi e)_{p.s} = 0$. In addition, the 54 independent interaction parameters $(g\psi)_{ips}$, $(a\psi)_{jps}$, and $(gau)_{ijps}$ lead to 54 independent parameters (not all of which are necessarily worth entertaining in a model) describing possible differing effects of jamming method between jamming types and between levels of the primary factor:

$$\begin{aligned}(ge)_{is} &= (g\psi)_{i.s}, (gne)_{ips} = (g\psi)_{ips} - (ge)_{is}, \\(ae)_{js} &= (a\psi)_{j.s}, (ane)_{jps} = (a\psi)_{jps} - (ae)_{js}, \\(gae)_{ijs} &= (gau)_{ij.s}, (gane)_{ijps} = (gau)_{ijps} - (gae)_{ijs}. \end{aligned}\quad (6)$$

When a sensitivity test is inserted in this manner, none of the estimators of the original primary or secondary factor effects are directly impacted. As when changing a factor from period-to-period, however, estimators of effects involving the sensitivity factor require more judgemental interpretation than those involving the primary and secondary factors because of confounding.

Clearly, several sensitivity tests could be inserted into the same design by proceeding along the lines suggested above. Although results of such sensitivity tests cannot be expected to be as clear as results involving factors examined in each basic matrix, they can involve sufficient numbers of well organized trials to justify their conduct. Moreover, these sensitivity tests follow the spirit of efficient statistical design, interleaving new factors by confounding them with interactions of old factors and reparameterizing rather than stealing a few observations from the existing design and conducting a tiny excursion.

D. Graceful Degradation. Reduction of the number of trials, either before or during testing, could be accommodated by the design in simple ways having clear consequences.

One easy way to accommodate pre-test resource reductions would be to eliminate test periods or test segments. Termination of the test after two test periods, instead of three, would have only a slight impact, primarily that of sample size reduction. Differences between the two test periods could still be estimated, and any period-to-period sensitivity test could still be implemented on a reduced basis. Termination of the test after one

test period, however, would preclude period-to-period sensitivity testing and force the assumption that differences between periods would not have been substantial. Since it was anticipated that differences between periods would be more pronounced than differences between segments, the statistician could state that a test having three test periods of one segment each would be more likely to yield general results than a test having one test period of three test segments. Of course, reducing either the number of test periods or the number of test segments to one limits both the number and the value of sensitivity tests which could be accommodated.

Another way to accommodate resource reductions would be to reduce the number of levels considered for the primary factor. This could be done prior to test or it might be necessary during test execution if (as frequently occurs) trials could not be conducted as quickly as originally anticipated or if a large number of trials resulted in unusable data. Since the levels of a primary factor often represent incremental additions to some baseline conditions, they could often be prioritized for deletion, and the loss of one or two levels might only reduce the scope of potential results.

Unplanned data loss on a large scale (as much as 1/3 to 2/3 loss) is unfortunately not uncommon during a field test. Designs generated by the basic matrix approach degrade gracefully in the presence of such data loss because the same block structure which permits refined data analysis with nearly complete data sets also gives a solid framework for salvaging results from a degraded data set. In particular, the fact that the core of the present design is a randomized complete blocks design in the levels of the primary factor assures that, even with substantial data loss, comparisons between levels of the primary factor can be made in such a way that they are at least free from main effects of test period and test segment. At worst, differences between two primary factor levels could be examined (using either parametric or nonparametric techniques) in all blocks (modified by restriction to one combination of secondary factor levels) having usable trials for each of the two primary factor levels. To a lesser extent, secondary factors could be examined within the remaining block structure too. Each block in the original design could be regarded as up to five blocks, one defined by each level of the primary factor having an observation at both combinations of secondary factor levels. A crude analysis could then be performed using all (modified) blocks by examining differences between levels for each secondary factor separately (ignoring hopefully weak confounding of secondary factors); the apparent effects obtained would at least be free from main effects of the primary factor as well as those of test period and test segment. Data quality, computer facilities, and time permitting, the blocking structure could probably be exploited in more elaborate ways to sort out main effects of secondary factors or even interactions between primary and secondary factors.

In no case does an underlying design structure which permits consideration of models like Model I and Model II preclude accommodation of resource reductions or hinder data analysis in degraded modes. Provided the design structure is based on the resource structure, it can accommodate resource reductions gracefully, if not without consequences, and refined structural

models can serve as guides to the data even when they can no longer be fully exploited.

V. SUMMARY AND CONCLUSION. The design example illustrates the flexibility and analytical richness which can be obtained by applying an intuitively simple structural approach to the design of a field test. Such designs enable formal consideration of a large number of frequently ignored test factors, provide systematic statistical guidance for detailed test conduct, and emphasize test structure rather than sample size.

Although such designs seem to have no statistical disadvantages compared to more traditional designs, they have the pragmatic disadvantages characteristic of all detailed plans and all deviations from traditional methods. Detailed statistical test planning is hard work, forces test conduct issues traditionally decided by convenience during test execution to be deliberately resolved prior to test, and provides guidance to the test operator which not only imposes constraints on his conduct of the test but also makes the inevitable shortcomings in test conduct more obvious. Because the traditional approach is accepted and does not force detailed statistical planning upon the testing community, any highly structured design proposed for a field test is likely to meet stiff resistance, and proposers of such designs must be prepared for extended argument and frequent disappointment.

SOME REMARKS ON CROSSOVER EXPERIMENTS
J. ROBERT BURGE
WALTER REED ARMY INSTITUTE OF RESEARCH

Introduction.

Frequently experimental observations are collected at different times on the same sampling unit. The simplest example of such a repeated-measurements design consists of the administration of a control treatment (a placebo or standard) on one occasion and the treatment of interest on another. The paired observation case, when extended, generates the general setup of k responses collected on the same experimental unit at successive times. Designs in which several treatments are applied in successive periods to each unit in a cyclic sequence are known as change-over designs. These designs are often exercised when units are highly variable, or are expensive or scarce.

Change-over designs are capable of providing treatment comparisons of high precision, because differences between units can be disjoined from experimental error. This advantage is obtained at the risk of possible complications, for performance in a given period might reflect more than the direct effect of the current treatment. These additional effects, known as residual effects, exist if preceding treatment effects crossover to influence responses measured in succeeding periods.

Cox (1958) advises that the crossover design be used only when it can be assumed that residual effects are negligible. However, when the residual effects do not persist for more than one period after application, some provision for the separation of direct and residual effects can easily be made. Still, it is most useful when the residual effects are relatively small.

Case 1: Three Factor Repeated Measurements Experiment

Before discussing the details of a crossover analysis, a repeated measures example is considered. The summary of results presented here resembles the pattern adopted for split-plot procedures. While an experimental unit (whole plot), in a split plot design, is subdivided physically, an experimental unit is split in time in a repeated measures design. That is, each unit receives several treatments in successive time periods. It is assumed in this first example that performance in a given period reflects only the direct effect of the current treatment.

a) Numerical Example.

This repeated measures analysis will be demonstrated using artificial data based on a study now being planned at Walter Reed. The study will be used to compare two positioning techniques in reducing the discomfort from intramuscular injection in the dorsogluteal site. The techniques involve placing patients in the prone position with either hips internally rotated (method A) or with hips externally rotated (method B).

To illustrate the meaning of the data set out in Table 1, ten post-operative patients were randomly assigned to one of two groups. The five patients allocated to group one received their first period injection (a.m.) while placed in position A. In period two (p.m.) they were placed in position B and the second medication was injected into the opposite dorsogluteal site. For the next two days medication was supplied by the oral route. On the final day the period one and two injections were administered in the reverse order (viz., B then A).

TABLE 1
DISCOMFORT SCORES FROM INTRAMUSCULAR INJECTIONS

		Period 1 (a.m.)	Period 2 (p.m.)	Period 1 (a.m.)	Period 2 (p.m.)	
		Patient	Trt. A	Trt. B	Trt. B	Trt. A
Group One	1		1.0	2.8	1.9	3.0
	2		1.0	1.5	1.7	1.2
	3		2.0	4.0	1.9	3.6
	4		1.5	3.0	1.5	2.0
	5		2.5	3.0	3.0	3.0
Group Two	Patient	Trt. B	Trt. A	Trt. A	Trt. B	
	6	3.3	2.0	2.5	3.2	
	7	2.1	1.7	1.5	1.0	
	8	4.5	3.4	4.0	3.8	
	9	4.5	1.5	2.0	4.5	
	10	3.5	1.0	1.5	4.0	

In contrast, the five patients assigned to the second group received the sequence B then A on the first day. On the final day their sequence was A then B. The patients used the discomfort rating scale given in Figure 1 to assign scores.

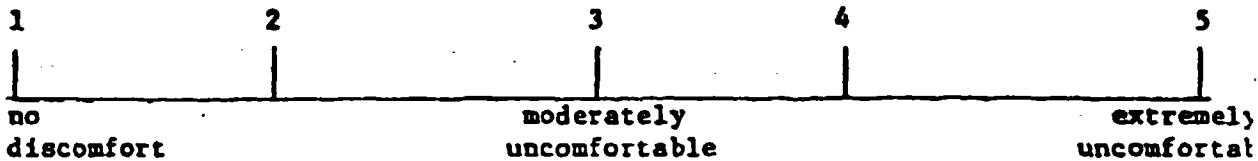


FIGURE 1
DISCOMFORT RATING SCALE

b) Model.

It will be assumed that the linear model upon which the analysis would be based is a function of the main effects of groups (γ_1), periods (ρ_j), treatments or methods (τ_1) and their interactions. Specifically,

$$E[Y_{ijklm}] = \mu + \gamma_1 + \pi_{k(i)} + \rho_j + \gamma\rho_{ij} + \tau_1 + \gamma\tau_{i1} + \rho\tau_{j1} + \gamma\rho\tau_{ij1} .$$

The notation $\pi_{k(i)}$ indicates the effect of patient k nested under group i. The "patient" factor is crossed with periods and treatment, but is nested under groups.

c) Analysis of Variance.

Assuming that all interactions with the "patient" factor are zero, the analysis of variance summary is shown in Table 2. The mean square for patients nested in groups is used to test the group effect. The period and treatment effects are within patient effects.

d) Unbalanced Data.

Suppose the patients were only administered each treatment one time. Members of the first group receiving only the sequence A-B; the second group being handled in the reverse order B-A. Then only four of the eight subclasses in Table 1 would contain data (viz., the non-bracketed data cells). Nevertheless, a feature of the design for the remaining data set, is that it still yields an analysis of variance that is easy to calculate and interpret. Table 3 allows one to examine and compare the analysis of variance procedure for the full ($N=40$) data set with the reduced set ($N=20$). The economy of effort is desirable when the interactions are zero. If not, the adequacy of the reduced design is questionable—estimates of the main effects will be confounded by interaction terms.

TABLE 2
CASE 2
SUMMARY OF AOV THREE-FACTOR EXPERIMENT

Source of Variation	df	MS	F
<u>Between Patients</u>	9		
Groups (G)	1	2.704	1.1704
Error (a) (Patients w. groups)	8	2.3102	
<u>Within Patients</u>	30		
Periods (P)	1	0.841	
GP	1	3.844	
Methods (M)	1	7.056	16.2549
GM	1	2.401	
PM	1	<.001	
GPM	1	0.025	
Error (b) (P x Patients w. groups M x " " " PM x " " ")	24	0.43408	
TOTAL	39		

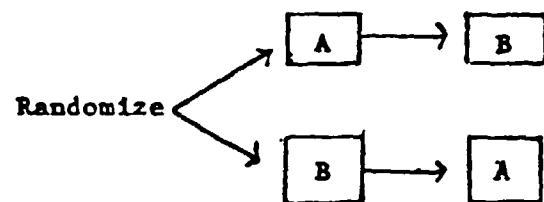
Table 3 AOV Scheme for Balanced Case (N=40) vs Missing Cells (N=20).

Source of Variation	df	df	Source of Variation	Alias
<u>Between Patients</u>	9	9		
Groups	1	1	Groups	PM
Error (a)	8	8	Error (a)	
[Patients W. Groups]				
<u>Within Patients</u>	30	10		
Periods	1	1	Periods	CM
GP	1	-		
Methods	1	1	Methods	GP
GM	1	-		
PM	1	-		
GPM	1	-		
Error (b)	24	8	Error (b)	
[P x Patients w. Groups]				
[M x Patients w. Groups]				
[PM x Patients w. Groups]				
TOTAL	39	19	TOTAL	

Case 2: The Simple Two-Treatment Crossover Experiment

In the example we consider next, each experimental unit (patient) has two periods of treatment supplied. The treatments, say A, B, to be compared are randomly assigned to the two periods. Usually, equal numbers of units are allocated to two groups. Members of the first group receive the treatment sequence AB; the second group is treated in the reverse order BA.

FIGURE 2
NOTATION AND LAYOUT
FOR THE SIMPLE CROSSOVER EXPERIMENT



a) Numerical Example.

The reduced data set ($N=20$) of the previous example fits this layout of a change-over experiment. It will be used to demonstrate an analysis that follows the summarization adopted by Lucas and Patterson (1962).

b) Model.

The model for the observation is $y_{ijk} = \mu + p_k + \phi_m + \lambda_r + \pi_{ij} + e_{ijk}$

where

μ = overall mean

p_k = effect of the k th period

ϕ_m = effect of the m th treatment (method) $m = A, B$

λ_r = effect of the r th treatment in the 1st period on the response in the 2nd period

π_{ij} = the effect of the j th subject in the i th group

$i = 1, 2 \quad j = 1, 2, \dots, N_{ij}$

e_{ijk} = within subject deviation for the k th period

c) Model Equations. Consider the following:

μ	π_{11}	π_{12}	...	π_{23}	π_{24}	ρ	ϕ	λ	δ	
1.0	1	1	0		0	0	1	1	0	1
2.8	1	1	0		0	0	-1	-1	1	1
1.0	1	0	1		0	0	1	1	0	1
1.5	1	0	1		0	0	-1	-1	1	1
2.0	1	0	0		0	0	1	1	0	1
4.0	1	0	0		0	0	-1	-1	1	1
1.5	1	0	0		0	0	1	1	0	1
3.0	1	0	0		0	0	-1	-1	1	1
2.5	1	-1	-1		0	0	1	1	0	1
3.0	1	-1	-1		0	0	-1	-1	1	1
$y =$:								z	
3.3	1	0	0		0	0	1	-1	0	-1
2.0	1	0	0		0	0	-1	1	-1	-1
2.1	1	0	0		0	0	1	-1	0	-1
1.7	1	0	0		0	0	-1	1	-1	-1
4.5	1	0	0		1	0	1	-1	0	-1
3.4	1	0	0		1	0	-1	1	-1	-1
4.5	1	0	0		0	1	1	-1	0	-1
1.5	1	0	0		0	1	-1	1	-1	-1
3.5	1	0	0		-1	-1	1	-1	0	-1
1.0	1	0	0		-1	-1	-1	1	-1	-1

20×13

$$b'_1 = (\mu, \pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{24}, \rho, \phi, \lambda)$$

$b'_2 = (\mu, \rho, \phi, \delta)$ where δ represents the period by treatment interaction.

The components of the vectors b'_1 and b'_2 form subsets of the set of Greek letters used to label the columns of the 20×13 matrix Z .

This association is used to determine the corresponding columns one selects from Z to build the equations $\underline{Y} = X_1 b_1$ and $\underline{Y} = X_2 b_2$ used to perform the analysis presented in Tables 4 and 5.

d) Analysis of Variance.

The regression approach to AOV offers a computationally convenient algorithm for generating, from the $\underline{Y} = X_1 b_1$ equations, the various entries in the AOV table. The R()-notation, employed in the tables, serves to clearly describe the way sums of squares were computed. A complete summary of R()-notation is given in Searle (1971).

TABLE 4
AOV FOR CASE 2
(simple two-treatment change-over design)

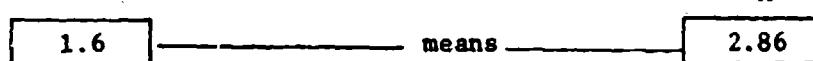
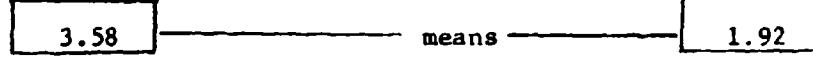
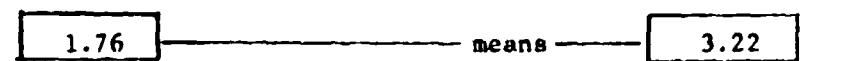
Source of Variation	df	SS	R() - Notation	F	p
Between Units Groups	9 1	10.188 1.352	R ($\delta \mu, \rho, \phi$)	1.224	.3
Units w. Groups	8	8.836	R (π)		
Within Units Periods Methods	10 1 1	0.200 10.658	R (ρ) R (ϕ) = R ($\phi \mu, \pi, \rho$)	25.9	
Error (b)	8	3.292	$Y'Y - R (\mu, \pi, \rho, \phi, \lambda)$		
Total	19	24.338	$Y'Y - R (\mu)$		

Since direct and residual treatment effects are not orthogonal, two partitions of the total sum of squares for treatments are given in Table 5. Only the first of these is required if residual effects are of no interest apart from their use in adjusting direct effects. Both will be required when residual effects are tested. F-tests for direct and residual effects can be carried out by dividing s^2_d and s^2_r , respectively, by s^2_e . A summary of the experiment is presented in Table 6. The first part of this table gives the summary required when residual effects are present. The second part presents the summary required when the analysis shows residual effects are negligible (i.e., estimates of direct effects unadjusted for residual effects, but adjusted for differences between units, are given)

TABLE 5
PARTITION OF TREATMENT SUM OF SQUARES

Source of Variation	df	SS	R() - Notation	
Between Units				
Units w. Groups	8	8.836	$R(\pi)$	
Within Units				
Periods	1	0.200	$R(\rho)$	
Treatments	2	12.010	$R(\phi, \lambda \mu, \pi, \rho)$	
Error	8	3.292	$\mathbf{Y}'\mathbf{Y} - R(\mu, \pi, \rho, \phi, \lambda)$	$s_e^2 = .4115$
Total	19	24.338	$\mathbf{Y}'\mathbf{Y} - R(\mu)$	
Direct Effects (eliminating residual)	1	9.801	$R(\phi \mu, \pi, \rho, \lambda)$	s_d^2
Residual Effects (ignoring direct)	1	2.209	$R(\lambda \mu, \pi, \rho)$	
Treatments (direct and residual)	2	12.010		
AND				
Direct Effects (ignoring residual)	1	10.658	$R(\phi \mu, \pi, \rho)$	
Residual Effects (eliminating direct)	1	1.352	$R(\lambda \mu, \pi, \rho, \phi)$	s_r^2

TABLE 6
SUMMARY OF RESULTS CASE 2 DATA SET

<u>Residual Effects Present</u>	
$(\hat{\mu} + \hat{\rho}_1 + \hat{\phi})$ Group 1 Period 1 Trt. A	$(\hat{\mu} - \hat{\rho}_1 - \hat{\phi} + \hat{\lambda})$ Group 1 Period 2 Trt. B _A
 $n=5$	$n=5$
$Se \sqrt{1/5} = .287$	$.497 = Se \sqrt{3/5}$
 $n=5$	$n=5$
Group 2 Period 1 Trt. B	Group 2 Period 2 Trt. A _B
$(\hat{\mu} + \hat{\rho}_1 - \hat{\phi})$	$(\hat{\mu} - \hat{\rho}_1 + \hat{\phi} - \hat{\lambda})$
<hr/> <hr/> <hr/>	
<u>Residual Effects Negligible</u>	
Trt. A $(\hat{\mu} + \hat{\phi}')$	Trt. B, $(\hat{\mu} - \hat{\phi}')$
 $n=10$	$n=10$
$Se \sqrt{1/10}$	
$= .203$	

Case 3: Alternative Analysis for the Simple Two-Treatment Crossover Experiment

The two period crossover design for clinical trials was compared with other simple designs in terms of statistical precision and cost by Brown (*Biometrics*, March 1980). In Brown's article an alternative analysis of a simple crossover study was illustrated. The data from this example are presented in Table 7. The response of interest is an oral hygiene index used to compare a test compound with a placebo with regard to effect on dental hygiene. Summary statistics and tests of significance for the data are given in Table 8. Results of estimating the treatment effect on the basis of first-period data alone are reproduced in Table 9. For comparative purposes, the summary results from the hygiene data are presented using the "Lucas Format" in Table 10.

Table 7
Improvement in hygiene index for a crossover study

Subject	Group I		Group II	
	Period 1 Placebo	Period 2 Test	Period 1 Test	Period 2 Placebo
1	0.83	1.83	1.67	0.33
2	1.00	2.17	2.50	0.50
3	0.67	1.67	1.00	0.17
4	0.50	1.50	1.67	0.50
5	0.50	2.33	1.83	0.50
6	0.83	1.83	0.50	0.33
7	1.00	0.50	1.33	0.67
8	0.67	0.33	1.33	0.00
9	0.67	0.50	0.50	0.17
10	0.33	0.67	2.17	0.83
11	0.00	0.83	1.67	0.33
12	1.17	1.33	1.50	0.00
13	0.00	0.67	1.33	0.50
14	0.50	1.83	1.50	0.50
15	0.33	1.50	1.33	0.00
16	0.33	1.50	0.67	0.17
17	0.50	1.17	1.67	0.50
18	1.00	1.67	2.50	0.67
19	0.00	1.33	1.83	0.00
20	0.50	1.50	0.83	0.67
21	0.50	2.83	2.33	0.17
22	0.17	2.33	1.17	0.50
23	1.00	1.33	1.33	0.00
24	1.00	1.67	1.33	0.83
25	1.33	0.67	0.33	1.33
26	0.33	0.83	2.17	1.17
27	2.00	1.00	1.00	0.33
28	4.00	0.17	0.33	1.00
29	0.83	1.67	1.17	0.17
30	0.50	1.33	0.50	0.50
31	0.50	1.50		
32	0.50	1.67		
33	2.17	1.33		
34	0.67	1.17		

TABLE 8
SUMMARY STATISTICS AND TESTS OF SIGNIFICANCE
FOR THE HYGIENE DATA

Analysis of Differences

	<u>Group I</u>	<u>Group II</u>
Mean	0.5985	-0.9440
Variance	1.3268	0.5173
Number	34	30
Pooled Var.	0.9482	
Est. Trt. Effect	0.7712	
St. Error	0.1220	
t	6.32	
df = 62		
p	< .001	

Analysis of Sums

	<u>Group I</u>	<u>Group II</u>
Mean	2.1176	1.7883
Variance	0.6059	0.5333
Number	34	30
Pooled Var.	0.5719	
Est. residual effect	-0.8294	
St. Error	.1894	
t	1.73	
df = 62		
p	0.087	

TABLE 9
Summary Statistics and Tests of Significance
of the First-Period Data from the Dental Hygiene Study

	<u>Group I</u>	<u>Group II</u>
Mean	0.7597	1.3663
Variance	0.6	0.3845
Number	<u>34</u>	<u>30</u>
Pooled Var	0.4992	
Trt. Effect	0.6066	
St. Error	0.1770	
t	3.4271	
df	62	
p	0.001	

Basically, Brown's article is critical of crossover studies. The analysis of differences, he points out, relies on the assumption that a therapy has a certain additive effect that does not depend upon the time period the treatments were administered. That is, response to a treatment should not be influenced by whether or not the other treatment was just given. For the hygiene data, the analysis of sums suggests the assumption is questionable; thus, the possibility of a large direct treatment effect bias exists due to residual effects. When the assumptions are contradicted, Brown recommends one only utilize first period data. Although his data set was large enough to evaluate treatment effects in this way, he reminds us that this is the exception, not the rule, for change-over designs.

TABLE 10
PARTITION OF TREATMENT SUM OF SQUARES: HYGIENE DATA SET

Source of Variation	df	SS	R() - Notation	F	t
Between Subjects Groups Subj. w. Groups	63 1 62	18.5993 0.86264 17.7367		3.015	1.736
Within Subjects Periods Treatments Error (b)	1 1 62	0.5 18.95456 29.40679	R($\rho \mu$) R(TD μ, ρ)	39.963	6.32
Total	127	67.46069	$\bar{Y}'\bar{Y} - R(\mu)$		
Direct Effects (eliminating residual)	1	5.86495	R(TD μ, ρ, TR)	12.36	3.52
Residual Effects (ignoring direct effects)	1	13.95225	R(TR μ, ρ)		
Treatments - TD and TR (direct and residual)	2	19.8172			
or					
Direct Effects (ignoring residual)		18.95456	R(TD μ, ρ)		
Residual Effects (eliminating direct effects)	1	0.86264	R(TR μ, ρ, TD)		
Treatments (direct and residual)	2	19.8172			

Summary - Brown vs. Lucas

Brown's format for crossover analysis was applied to the injection data and the results are presented in Tables 11 and 12. Summary results from the injection data are presented using the Lucas format in Table 13. Brown bases his test for a residual effect on the "analysis of sums". Lucas would carry out the F-test for residual effects by comparing s_r^2 with s_e^2 . The tests differ. Brown chooses between group variability for an error term, while Lucas chose within-subject variability. Further differences, as well as similarities, between the other tests of significance utilized by these two approaches are set out in Table 14.

REFERENCES

- Brown, B. W., The Crossover Experiment for Clinical Trials, Biometrics 36, 69-79 (March 1980).
- Lucas, H. L. and Patterson, H. D., Change-Over Designs, Tech. Bul. No. 147, North Carolina Agricultural Experiment Station and U.S. Dept. of Agric., (Sept 1962).
- Searle, S. R., Linear Models, John Wiley & Sons, Inc., New York (1971).
- Speed, F. M. and Hocking, R. R., The Use of the R()-Notation with Unbalanced Data, The American Statistician, February 1976, Vol 30, No 1, p. 30-33.
- Winer, B. J., Statistical Principles in Experimental Design McGraw-Hill, New York (1971).

TABLE 11
SUMMARY STATISTICS AND TEST OF SIGNIFICANCE

Analysis of Differences

	<u>Group I</u>	<u>Group II</u>
Mean	-1.26	1.66
Variance	0.513	1.133
Number	5	5
Pooled Variance		0.823
Est. Trt. Effect		-2.92
St. error		0.5737
t		-5.089
df		8
p		0.00094

Analysis of Sums

	<u>Group I</u>	<u>Group II</u>
Mean	4.46	5.50
Variance	1.933	2.485
Number	5	5
Pooled Variance		2.209
Est. trt. effect		-1.04
St. error		0.94
t		-1.1064
df		8
p		.30

TABLE 12
SUMMARY STATISTICS AND TESTS OF SIGNIFICANCE

Analysis of Period One Scores

	<u>Group I</u>	<u>Group II</u>
Mean	1.6	3.58
Var	.425	.992
n	5	5
Pooled Var	.7085	
Est. trt. effect	-1.98	
St. error	0.5323533	
t	-3.719	
df	8	
p	.006	

TABLE 13
PARTITION OF TREATMENT
SUM OF SQUARES

<u>Source of Variation</u>	<u>df</u>	<u>SS</u>	<u>MS</u>
Between Units	9	10.188	
Groups	1	1.352	
Units w. Groups	8	8.836	
Within Units	10		
P	1	0.20	
Treatments	1	10.658	10.658
Error	8	3.292	0.4115
Total	19		s_e^2

	<u>df</u>	<u>SS</u>	<u>R()-Notation</u>	
Direct Effects (eliminating residual)	1	9.801	$R(\phi \mu, \pi, \rho, \lambda)$	s_d^2
Residual Effects (ignoring direct effects)	1	2.209	$R(\lambda \mu, \pi, \rho)$	
Treatments (direct and residual)	2	12.01	$R(\phi, \lambda \mu, \pi, \rho)$	
OR				
Direct Effects (ignoring residuals)	1	10.658	$R(\phi \mu, \pi, \rho)$	$s_d'^2$
Residual Effects (eliminating direct)	1	1.352	$R(\lambda \mu, \pi, \rho, \phi)$	s_r^2
	2	12.01		

TABLE 14
A SUMMARY RELATING TESTS OF
SIGNIFICANCE FOR TWO APPROACHES TO
CROSSOVER ANALYSIS

1. Analysis of Differences

(Brown)

(Lucas)

$$t^2 = \left(\frac{-2.92}{.5737} \right)^2 = \underline{\underline{(-5.089)^2 = 25.9}} = \frac{10.658}{.4115} = \frac{s^2_d}{s^2_e}$$

2. Analysis of Sums

$$t^2 = \left(\frac{-1.04}{.94} \right)^2 = \underline{\underline{(-1.1064)^2 = 1.224}} = \frac{1.352}{(8.836/8)} = \frac{s^2_r}{MS \text{ units w. groups}}$$

NOTE: $3.286 = \frac{1.352}{.4115} = \frac{s^2_r}{s^2_e}$

**3. Analysis of First Period Scores
(residual effects present)**

$$t^2 = \left(\frac{-1.98}{[(.7085)(4/10)]^{1/2}} \right)^2 = \underline{\underline{(-3.719)^2 = 13.83}} = \frac{9.801}{.7085} = \frac{s^2_d}{s^2_e}$$

$$t^2 = \left(\frac{-1.98}{[(.4115)(4/10)]^{1/2}} \right)^2 = \underline{\underline{(-4.88)^2 = 23.82}} = \frac{9.801}{.4115} = \frac{s^2_d}{s^2_e}$$

A TIME SERIES ANALYSIS AND MODELING APPROACH OF
SENSE AND DESTROY ARMOR (SADARM) RADIOMETRIC
(ELECTROMAGNETIC RADIATION) NOISE DATA

Richard T. Maruyama
Probability and Statistics Branch
Ballistic Modeling Division
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

ABSTRACT. A series of tests were conducted in August 1978 to collect radiometric (electromagnetic radiation) data at the North East Test Site of the Rome Air Development Center in Rome, N.Y. This radiometric data was collected using the Sense and Destroy Armor (SADARM) target sensoring system. Both background (ground noise) and target (tanks) data were collected to investigate the signal characteristics of the SADARM weapon system. The data was recorded at equally spaced time intervals over five ranges. The objectives of the SADARM sensor are to detect and then aim the antiaarmor munition at the target.

This paper presents the Box and Jenkins time series modeling effort on the background radiometric data. This effort resulted in an Autoregressive-Moving Average (ARMA) model of order $p=1$ and $q=1$, where the autoregressive parameter ranged from 0.73 to 0.88 and the moving average parameter ranged from -0.59 to -0.64. This ARMA(1,1) model seems adequate for characterizing the background noise of the SADARM weapon system.

1. INTRODUCTION. The Sense and Destroy Armor (SADARM) is a "Fire and Forget" antiaarmor munition being developed and tested by the US Army Armament Research and Development Command (ARRADCOM). The lead laboratory for the program is ARRADCOM's Large Caliber Weapon System Laboratory (LCWSL). The systems analysis and sensor technology research is ongoing at the Ballistic Research Laboratory (BRL), Aberdeen Proving Ground, MD. As part of this BRL effort the SADARM sensor was tested in August 1978 at the North East Test Site of the Rome Air Development Center in Rome, NY. This electromagnetic radiation data collected in Rome, NY, is essentially the absolute temperature of the ground surface at the sensor's focus point, and is referred to as radiometric readings. The functional relationship between absolute temperature and the radiometric reading is given in equation (1):

$$\eta = 2kT/\lambda^2 \quad (1)$$

where

k is Boltzmann's constant

T is the absolute temperature

and λ is the wavelength.

The radiometric data was collected over field and tree terrains, at five different distances that ranged from 30 to 150 meters (see Figure 1). The SADARM sensor was mounted on a helicopter at a slant angle of 30° measured from perpendicular to the ground. The helicopter was then flown at a ground speed of 60 knots. The SADARM sensor recorded approximately two thousand equally spaced observations (data) per second and was instrumented to make four (4) revolutions per second (RPS).

The object of the sensor is to detect the target and then aim the weapon system under battlefield conditions. Hence, the response of the SADARM sensor to varying background conditions will effect the weapon system's ability to detect targets.

Presented in this report is the Box and Jenkins modeling approach for the radiometric non-target observations. The analysis demonstrates the ability of the ARIMA model to characterize the SADARM data.

2. THE BOX AND JENKINS TIME SERIES APPROACH. The Box and Jenkins¹ Autoregressive Integrated Moving Average (ARIMA (p,d,q)) model is used to characterize many types of business, economics and engineering observations. The need to develop a model of the SADARM sensors' responses to backgrounds (terrain) has been ongoing at BRL. In order to satisfy this requirement the Box and Jenkins approach was initiated.

A representative plot of the August 1978 SADARM data is presented in Figure 2.

The ARIMA (p,d,q) model is presented below:

$$\phi_p(B)(1-B)^d(Z_t - \mu) = \theta_q(B)a_t \quad (2)$$

where B is the backshift operator such that $BZ_t = Z_{t-1}$,

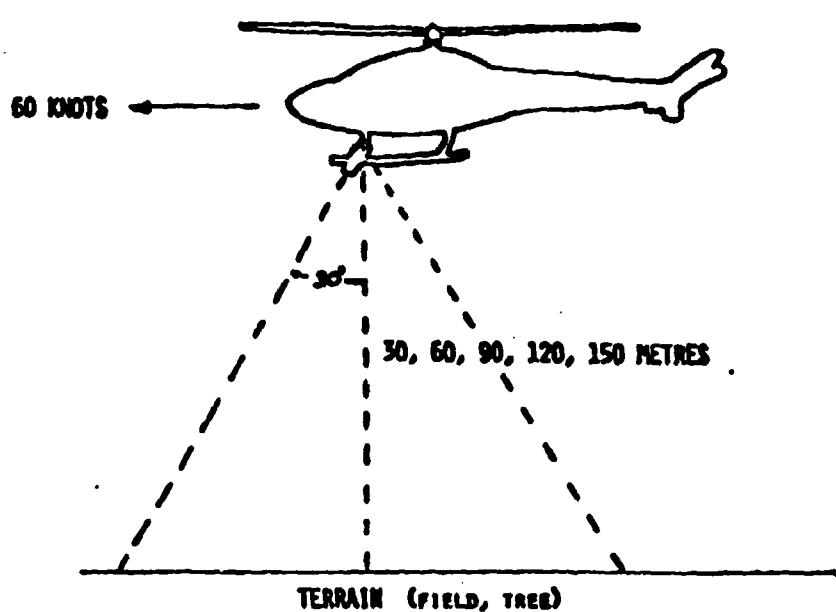
$\phi_p(B)$ is a polynomial in B of order p and ϕ_i are autoregressive parameters,

$\theta_q(B)$ is a polynomial in B of order q and θ_i are moving average parameters,

p,d,q are non-negative integers, and

a_t are random shocks (white noise) assumed to be independently distributed normal variates, $N(0, \sigma_a^2)$.

A series of five hundred observations were analyzed and used to estimate the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) out to 40 lags (see Figure 3 and Table 1).



NORTH EAST TEST SITE OF THE ROME AIR DEVELOPMENT CENTER IN ROME, N.Y.
AUGUST 1978

FIGURE 1

SADARM DATA (NO TARGET)

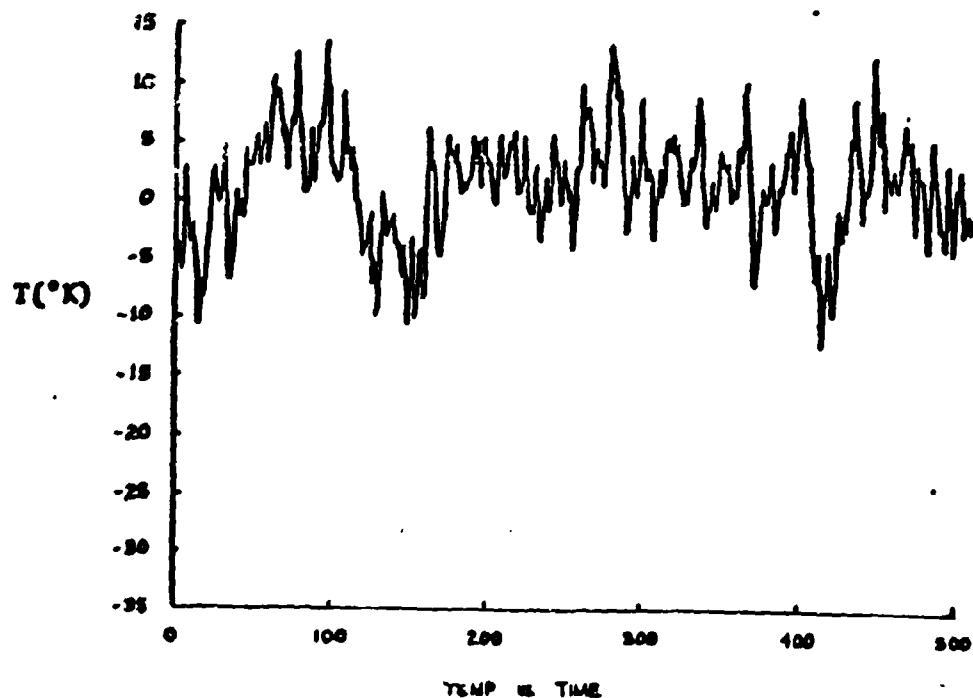


FIGURE 2

TABLE 1. ESTIMATED AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTION
RADAR RADIOMETRIC READING: 2000 OBSERVATIONS PER SECOND

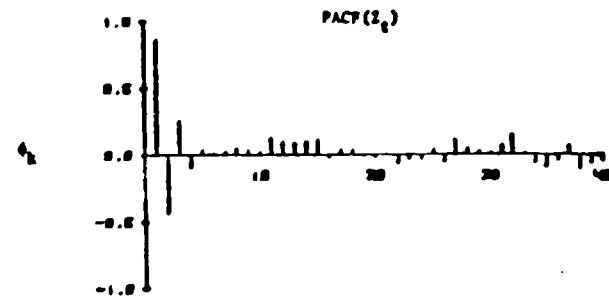
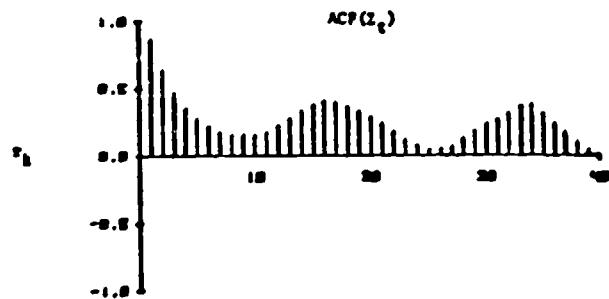
		Autocorrelations									
		500 Observations									
		1	2	3	4	5	6	7	8	9	10
I	Lags 1-10	.86	.63	.46	.38	.27	.21	.17	.15	.15	.15
	11-20	.17	.22	.27	.33	.37	.40	.39	.36	.32	.28
	21-30	.23	.17	.11	.07	.04	.04	.06	.12	.18	.23
	31-40	.24	.31	.36	.37	.31	.23	.16	.10	.04	-.03
V2	Lags 1-10	.34	-.21	-.22	-.11	-.08	-.07	-.08	-.06	-.02	-.06
	11-20	-.09	-.04	.02	.01	.00	.12	.06	.03	.04	.01
	21-30	.02	.01	-.03	-.10	-.06	-.11	-.11	-.02	.06	.04
	31-40	-.05	.01	.35	.22	.12	-.07	-.03	.04	-.00	-.12

		Partial Autocorrelations									
		500 Observations									
		1	2	3	4	5	6	7	8	9	10
I	Lags 1-10	.86	-.43	.26	-.09	.03	.01	.02	.04	.03	.02
	11-20	.12	.09	.08	.00	.11	-.02	.03	.03	-.01	-.02
	21-30	-.06	-.06	-.03	-.03	.03	-.05	.11	.04	.02	.01
	31-40	.06	.14	.01	-.06	-.09	-.04	.06	-.10	-.06	-.04
V2	Lags 1-10	.34	-.37	.01	-.12	-.06	-.09	-.11	-.08	-.07	-.15
	11-20	-.12	-.10	-.11	-.12	.02	-.05	-.04	-.00	.01	-.01
	21-30	.05	.01	.02	-.05	.04	-.14	-.05	-.04	-.01	-.00
	31-40	-.14	-.01	.05	.08	.04	-.08	.10	.04	.03	-.05

ESTIMATED AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS

$$\bar{v}_s = -1.603$$

$$\hat{\sigma}_s = 3.712$$



In building a dynamic time series model three steps are required; (1) tentatively identify the model, (2) fit the data to the model, and (3) perform a diagnostic check for lack of fit. The identification step requires an overall view of the data structure. In this case a damping sinusoidal structure for the ACF and two significant spikes at the first and second PACF are displayed (see Figure 3). The estimated mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) of this time series are -1.603 and 3.712, respectively.

This identification step implies that an autoregressive model of order $p=2$ be tentatively entertained. Hence, the ARIMA (2,0,0) model was fitted.

ARIMA (2,0,0) Tentatively Entertained

$$(1 - \phi_1 B - \phi_2 B^2)(Z_t - \hat{\mu}) = a_t \quad (3)$$

where the estimated parameters are $\hat{\mu} = -1.636$
 $\hat{\phi}_1 = 1.252$
 $\hat{\phi}_2 = -0.450$

The autocorrelation function of the residual, $a_t = Z_t - \hat{Z}_t$, of the ARIMA (2,0,0) model was then looked at for lack of fit. These residual ACF are listed in Table II, where the estimated residual mean and standard deviation are $\hat{\mu}_{a_t} = 0.00175$ and $\hat{\sigma}_{a_t} = 1.681$. The ACF of the residuals at lags $k=1$ and $k=2$ indicate some remaining residual structure. Also, the cutoff of the residual ACF demonstrates a possible need for a moving average term in the model.

Hence, the ARIMA (2,0,1) model was entertained to remove the spikes in the residual ACF. The ARIMA (2,0,1) model is as follows:

ARIMA (2,0,1)

$$(1 - \phi_1 B - \phi_2 B^2)(Z_t - \hat{\mu}) = (1 - \theta_1 B)a_t \quad (4)$$

where the estimated parameters are $\hat{\mu} = -1.622$
 $\hat{\phi}_1 = 0.886$
 $\hat{\phi}_2 = -0.138$
 $\hat{\theta}_1 = -0.506$.

TABLE II
ESTIMATED AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS OF RESIDUALS
(ARIMA (2,0,0))

Residual Mean = .001747

Residual Standard Deviation = 1.6809

ACF											
500 Observations		1	2	3	4	5	6	7	8	9	10
τ_1	Lags 1-10	.12	-.21	.61	.30	.07	.04	.01	.01	.00	.02
	11-20	-.02	.04	.09	.01	.11	.18	.06	.04	.00	.04
	21-30	.06	.05	.04	-.06	.06	-.01	-.06	.04	.10	.06

PACF											
500 Observations		1	2	3	4	5	6	7	8	9	10
θ_1	Lags 1-10	.12	-.22	.67	.06	.07	.06	.02	.02	.07	-.00
	11-20	-.00	.04	.06	-.00	.14	.10	.06	.06	.07	.01
	21-30	.06	.01	.03	-.08	.05	-.00	-.04	-.02	.06	.06

PLOT OF THE ESTIMATED ACF AND PACF OF RESIDUALS (ARIMA (2,0,0))

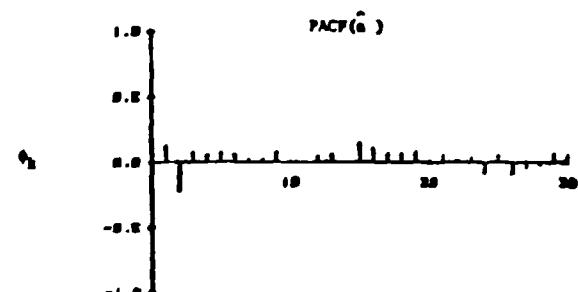
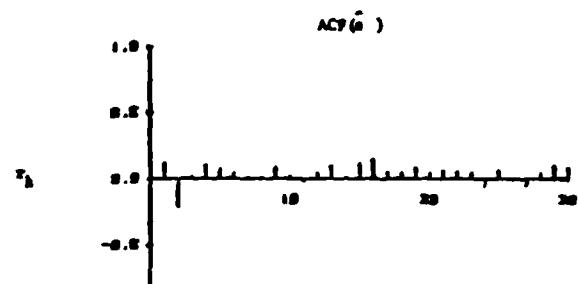


FIGURE 4

The addition of the moving average term, $\hat{\theta}_1$, did remove the autocorrelation spikes for lags $k = 1$ and $k = 2$ (see Table III). Inspection of the residuals with mean $\hat{\mu}_{a_t} = 0.0009$ and standard deviation $\hat{\sigma}_{a_t} = 1.616$ indicated a good fit (see Figure 5).

(see Figure 5). The 95% confidence interval for the second estimated autoregressive parameter, $\hat{\phi}_2 = -0.138$, overlapped the zero point (-0.285, 0.0086).

This suggested the possible removal of this term from the model. Based upon this information and the principle of parsimony in the use of parameters, the second autoregressive parameter was removed and the ARIMA (1,0,1) model was considered.

The ARIMA (1,0,1) model is as follows:

ARIMA (1,0,1)

$$(1 - \hat{\phi}_1 B)(z_t - \hat{\mu}) = (1 - \hat{\theta}_1 B)a_t \quad (5)$$

where the estimated parameters are $\hat{\mu} = -1.618$
 $\hat{\phi}_1 = 0.7553$
 $\hat{\theta}_1 = -0.5960$.

Both the ACF and PACF of the ARIMA (1,0,1) residuals indicated a lack of any remaining structure. Table IV and Figure 6 show that the residual $\{a_t = z_t - z_t\}$ no longer contains any structure. Further data analysis to better characterize this time series was unsuccessful.

A summary of the models and their estimated parameters are presented in Table V.

TABLE III
ESTIMATED AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS OF RESIDUALS
(ARIMA (2,0,1))

Residual Mean = 0.0009

Residual Standard Deviation = 1.6169

		ACF									
		500 Observations									
		1	2	3	4	5	6	7	8	9	10
$\hat{\epsilon}_k$	Lags 1-10	.00	-.02	-.03	.06	.01	.02	-.01	-.01	.05	-.00
	11-20	-.02	.01	.09	-.00	.11	.11	.06	.04	.09	.02
	21-30	.07	.01	.05	-.09	.05	-.05	-.04	.02	.06	.07

		PACF									
		500 Observations									
		1	2	3	4	5	6	7	8	9	10
$\hat{\epsilon}_k$	Lags 1-10	.00	-.02	-.03	.06	.01	.02	-.00	-.02	.05	-.01
	11-20	-.02	.02	.06	-.00	.12	.12	.06	.05	.09	.02
	21-30	.07	.01	.06	-.09	.04	-.06	-.06	-.01	.06	.06

PLOT OF THE ESTIMATED ACF AND PACF OF RESIDUALS (ARIMA (2,0,1))

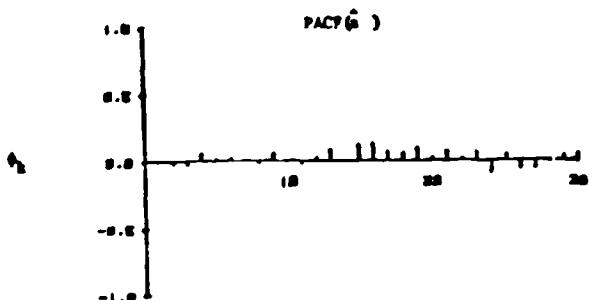
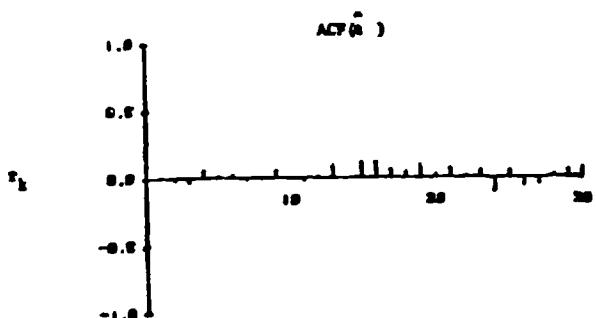


TABLE IV
ESTIMATED AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS OF RESIDUALS
(ARIMA (1,0,1))

Residual Mean = 0.00039

Residual Standard Deviation = 1.623

ACF

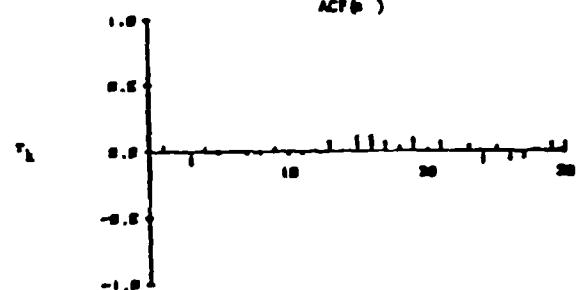
		300 Observations									
		1	2	3	4	5	6	7	8	9	10
τ_1	Lags 1-10	.04	.00	-.00	.03	-.02	.00	-.02	-.02	.03	-.01
	11-20	-.02	.00	.00	-.00	.11	.11	.07	.03	.00	.01
	21-30	.07	.00	.04	-.00	.03	-.07	-.05	.01	.04	.06

PACF

		300 Observations									
		1	2	3	4	5	6	7	8	9	10
θ_1	Lags 1-10	.04	.00	-.00	.04	-.02	-.00	-.02	-.02	.04	-.02
	11-20	-.03	.01	.07	-.01	.12	.11	.06	.05	.10	.02
	21-30	.00	.02	.06	-.06	.04	-.06	-.06	.00	.03	.03

ESTIMATED ACF AND PACF OF RESIDUAL (ARIMA (1,0,1))

ACF($\hat{\phi}_1$)



PACF($\hat{\theta}_1$)

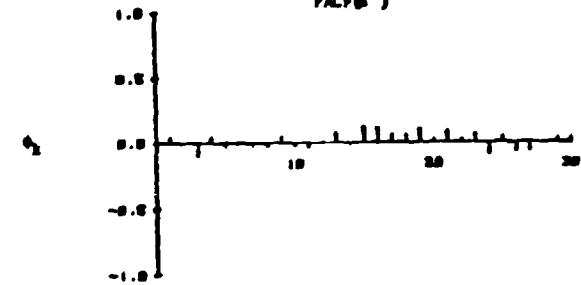


TABLE V
A SUMMARY OF ARIMA (•) MODELS ENTERTAINED

<u>MODEL</u>	<u>ESTIMATED PARAMETERS</u>	<u>RESIDUAL SUM SQ.</u>	<u>WHITE NOISE</u>
(1) $(1-\theta_1 B-\theta_2 B^2)(Z_t - u) = a_t$	$\hat{u} = -1.636$ $\hat{\theta}_1 = 1.281$ $\hat{\theta}_2 = -0.450$	1,612 497 df.	$\hat{\sigma}_t = 0.0017$ $\hat{\sigma}_a = 1.64$
(2) $(1-\theta_1 B-\theta_2 B^2)(Z_t - u) = (1-\theta_1 B)a_t$	$\hat{u} = -1.622$ $\hat{\theta}_1 = 0.886$ $\hat{\theta}_2 = -0.138$ $\hat{\theta}_3 = -0.506$	1,307 496 df.	$\hat{\sigma}_t = 0.0009$ $\hat{\sigma}_a = 1.616$
(3) $(1-\theta_1 B)(Z_t - u) = (1-\theta_1 B)a_t$	$\hat{u} = -1.618$ $\hat{\theta}_1 = 0.788$ $\hat{\theta}_2 = -0.294$	1,317 497 df.	$\hat{\sigma}_t = 0.0039$ $\hat{\sigma}_a = 1.62$

3. APPLICATION OF ARIMA '(1,0,1) MODEL TO THE REMAINING SADARM DATA.

In Section 2 the rationale for the selection of the ARIMA (1,0,1) model was presented. In this section the remaining data for the non-target cases are analyzed. The ARIMA (1,0,1) model was found to be adequate for modeling the non-target SADARM data collected in Rome, N.Y.

The first step is to investigate more of this data. In Table VI, both the ACF and PACF of additional samples of 500 observations at different distances (30, 60, 90, 120 and 150 meters) were estimated. Plots of both are presented in Figure 7. The similarity in the correlograms indicate the possibility of modeling all the SADARM non-target data with the same ARIMA model. Hence, Table VII was generated from other sets of SADARM data using the ARIMA (1,0,1) model.

This analysis suggested that the means (μ) are varying, but that the autoregressive parameter (ϕ_1) and the moving average parameter (θ_1) are not. The estimated parameters ($\hat{\mu}_a$, $\hat{\sigma}_a^2$) for white noise (random shocks) are consistent for those cases investigated. That is $\hat{\mu}_a \approx 0.0$ and $\hat{\sigma}_a^2 \approx (1.62 \text{ to } 1.79)$. A closer look at the residuals, a_t , indicates a lack of any consistent pattern after fitting the ARIMA (1,0,1) model. Figure 8 was constructed to demonstrate the lack of structure in the residuals, indicating the ability of the ARIMA (1,0,1) model to characterize all the SADARM non-target data. This ability to model the different cases by a white noise model is used to simulate the sensor's characteristics. The ARIMA (•) model used for this purpose is that of Equation (6).

$$z_t = (1 - \hat{\phi})\hat{\mu} + a_t - \hat{\theta}a_{t-1} + \hat{\phi}z_{t-1} \quad (6)$$

where $a_t = N(0, \hat{\sigma}_a^2)$ and
 $(\hat{\mu}, \hat{\phi}, \hat{\theta})$ are the estimated parameters.

A plot of one such simulated case is shown in Figure 9 as a comparison to the actual data plotted in Figure 2.

4. SUMMARY. The SADARM data collected in August 1978 was analyzed using the Box and Jenkins Time Series approach. This approach indicated an ARIMA (1,0,1) model. This particular ARIMA model characterizes the data remarkably well. What is more remarkable is the consistent behavior of both the estimated parameters ($\hat{\phi}_1$, $\hat{\theta}_1$) and the white noise parameters (\hat{a}_t , $\hat{\sigma}_a^2$). The indications are that the ARIMA (1,0,1) structure is adequate for describing this set of SADARM data.

REFERENCE

1. Box, G.E.P. and Jenkins, G.M., *Time Series Analysis: Forecasting and Control*, San Francisco, CA, Holden-Day, 1970.

TABLE VI
AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS OF Z_t

500 Observations		ACF									
		1	2	3	4	5	6	7	8	9	10
Z_t	Lags 1-10	.86	.63	.46	.35	.27	.21	.17	.15	.15	.15
	(150 m)	.11-20	.17	.22	.27	.33	.37	.40	.39	.36	.32
Z_t	Lags 1-10	.86	.64	.48	.40	.35	.30	.24	.19	.18	.21
	(120 m)	.11-20	.27	.32	.35	.37	.41	.47	.50	.47	.42
Z_t	Lags 1-10	.85	.60	.44	.35	.28	.20	.15	.13	.14	.18
	(90 m)	.11-20	.21	.25	.32	.41	.48	.53	.52	.46	.38
Z_t	Lags 1-10	.87	.66	.50	.39	.31	.24	.15	.09	.06	.07
	(60 m)	.11-20	.15	.22	.28	.32	.33	.33	.36	.36	.36
Z_t	Lags 1-10	.86	.64	.50	.42	.35	.27	.21	.17	.16	.17
	(30 m)	.11-20	.21	.26	.29	.33	.39	.43	.44	.41	.38

PACF											
	1	2	3	4	5	6	7	8	9	10	
Z_t	Lags 1-10	.86	-.43	.25	-.08	.03	.01	.02	.04	.05	.02
	(150 m)										
Z_t	Lags 1-10	.86	-.41	.26	-.01	.05	-.06	-.01	.00	.13	.09
	(120 m)										
Z_t	Lags 1-10	.85	-.43	.32	-.10	-.01	-.04	.07	.02	.09	.06
	(90 m)										
Z_t	Lags 1-10	.87	-.38	.20	-.06	.05	-.11	-.02	.00	.07	.06
	(60 m)										
Z_t	Lags 1-10	.86	-.38	.33	-.13	.02	-.04	.04	-.01	.09	.08
	(30 m)										

THE ACT AND PACF OF $\{z_t\}$

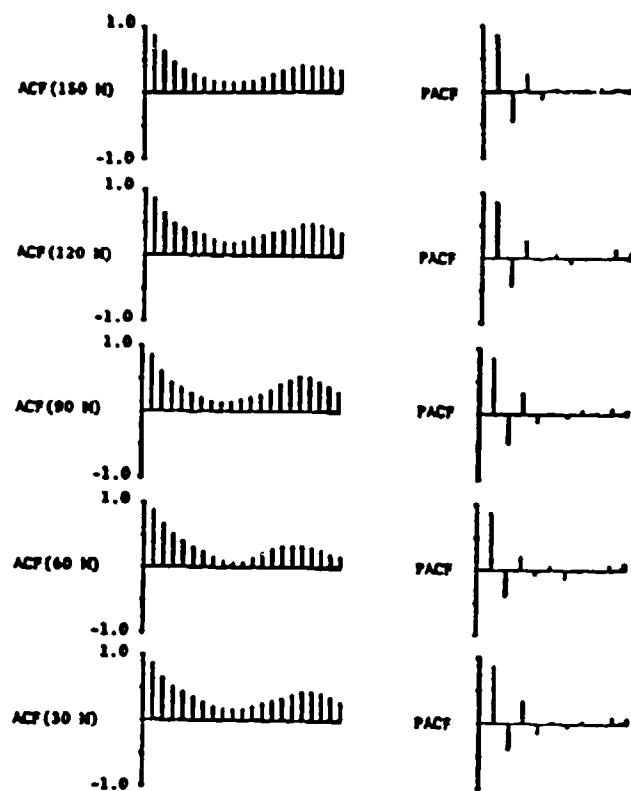


FIGURE 7

TABLE VII
SUMMARY OF ARIMA (1,0,1) MODEL FITTED TO SADARM FIELD AND TREE DATA

		Case A22 August 3, 1978 (Field)		ARIMA (1,0,1)	$(1-\alpha_1 B)(Z_t - u_t) + (1-\theta_1 B)u_t$		
		\hat{u}_t	$\hat{\alpha}_1$	$\hat{\theta}_1$	\hat{u}_t	$\hat{\alpha}_1$	$\chi^2_{\text{d.f.}}(27)$
30 Meters	(a)	-20.43	.7329	-.67395	-.00078	1.6542	49.21
	(b)	-24.12	.96376	-.59912	-.000895	1.6323	56.06
60 Meters	(a)	-17.185	.76154	-.54064	-.000674	1.5790	58.89
	(b)	-17.456	.79747	-.60273	-.000316	1.6816	47.83
90 Meters	(a)	-7.5867	.72707	-.62197	-.000283	1.6216	55.50
	(b)	-9.2541	.68369	-.67064	-.0003039	1.6276	56.78
120 Meters	(a)	-7.614	.75301	-.56238	-.0001220	1.5886	70.70
	(b)	-7.909	.7572	-.57466	-.0000809	1.5959	54.47
150 Meters	(a)	-1.618	.7552	-.5959	.000398	1.6231	45.00
	(b)	-1.7179	.68908	-.61738	.000671	1.6515	42.67
	(c)	-0.8284	.66975	-.67832	.0001374	1.5796	74.36
Mean		N/A	.732691	-.610632	-.0001333	1.618355	:
St. Dev.		N/A	.040613	.047867	.000507	0.037667	:
		{ .753697 }		{ .079614 }			

		Case A41 August 4, 1978 (Field)		ARIMA (1,0,1)			
		\hat{u}_t	$\hat{\alpha}_1$	$\hat{\theta}_1$			
30 Meters	(a)	-9.3985	.89142	-.57673	-.0002821	1.7345	42.50
	(b)	-7.8494	.79993	-.61042	-.0000485	1.7487	32.29
60 Meters	(a)	6.8121	.87228	-.56256	-.0005989	1.7910	45.25
	(b)	8.4699	.80589	-.66890	.0002418	1.8386	47.28
90 Meters	(a)	23.055	.88165	-.62222	-.0000369	1.8039	55.20
	(b)	23.037	.86254	-.56396	.00031859	1.7622	51.01
120 Meters	(a)	1.2813	.91947	-.53543	.0003456	1.7566	55.01
	(b)	1.4580	.92506	-.62257	.0000953	1.7982	38.34
150 Meters	(a)	-14.029	.83538	-.67237	-.000327	1.7767	39.40
	(b)	-14.726	.87666	-.62120	-.0004968	1.9116	55.95
Mean		N/A	.867027	-.605644	-.00007889	1.7922	:
St. Dev.		N/A	.042567	.048512	.00033697	0.051868	:

		Case A41 August 4, 1978 (Trees)		ARIMA (1,0,1)			
		\hat{u}_t	$\hat{\alpha}_1$	$\hat{\theta}_1$			
30 Meters	(a)	-1.2039	.73473	-.63486	-.0001325	1.6602	69.56
	(b)	-1.4959	.81564	-.59693	-.0000552	1.747	47.10
60 Meters	(a)	-2.6053	.82087	-.59368	-.0004588	1.7232	37.90
	(b)	-2.0881	.76270	-.63814	-.0001127	1.7423	58.36
90 Meters	(a)	16.156	.72748	-.68936	.0002787	1.7499	42.76
	(b)	15.704	.83907	-.63629	-.0004865	1.8846	74.76
120 Meters	(a)	-7.6494	.78241	-.62126	.0001219	1.8704	46.77
	(b)	-7.2249	.81786	-.62218	.000946	1.8263	70.69
150 Meters	(a)	8.5403	.85605	-.59373	-.0002017	1.8195	53.10
	(b)	9.0051	.79945	-.61361	-.001264	1.8421	77.85
Mean		N/A	.794626	-.624002	-.000136	1.78654	:
St. Dev.		N/A	.044026	.028726	.00057	0.07243	:

THE AUTOCORRELATION FUNCTION OF THE ARIMA (1,0,1) RESIDUALS

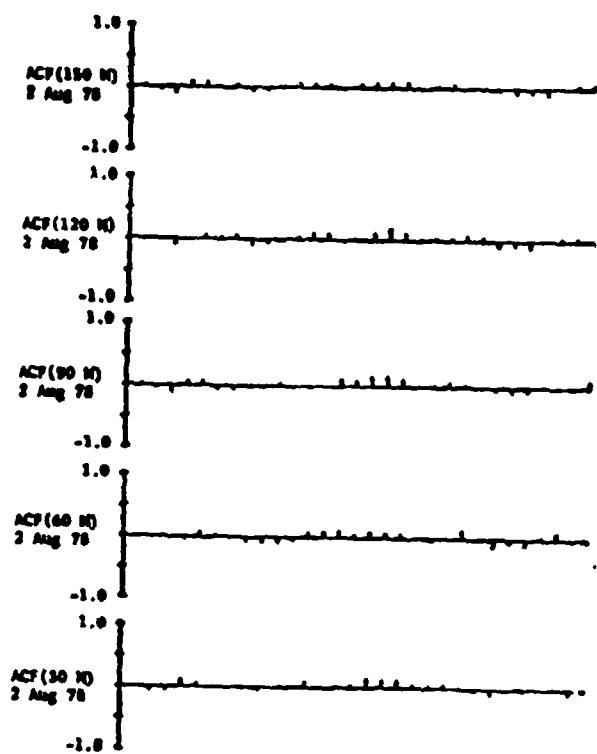
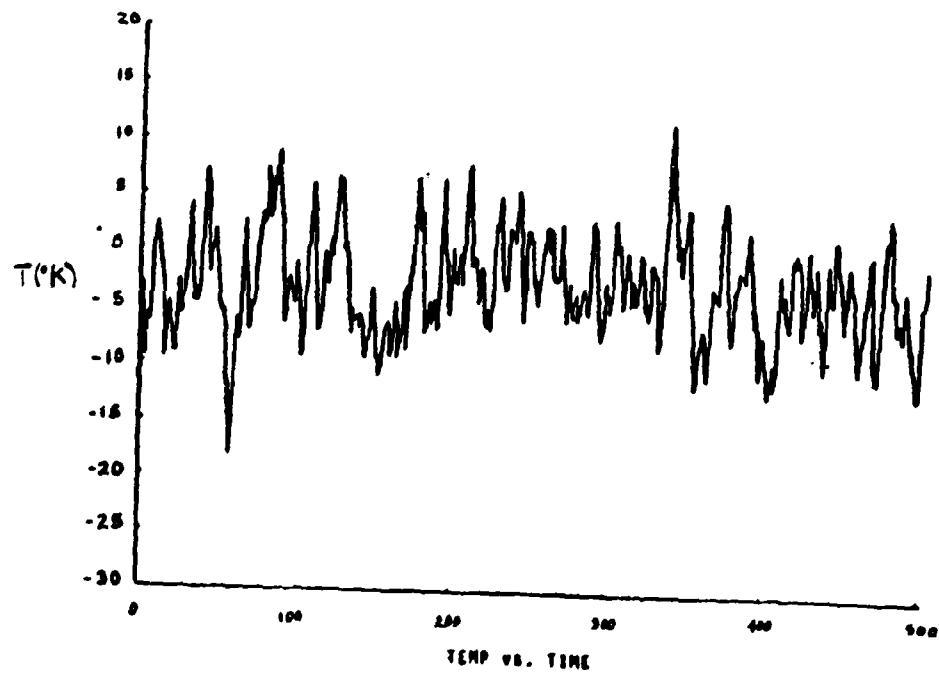


FIGURE 8

PREDICTED (MODELED) RESULTS



385

THE ROLE OF SPATIAL BANDWIDTH LIMITS IN THE MEASUREMENT AND INTERPRETATION
OF SECOND-ORDER STATISTICAL PROPERTIES

E. L. Church
Technology Branch
Armaments Division
Fire-Control and Small Caliber Weapon Systems Laboratory
US ARADCOM
Dover, New Jersey 07801

ABSTRACT. Many important physical processes depend on the second-order statistics of a random variable; that is, its autocovariance function or power spectral density. The measurement and specification of surface topographic finish is a case of particular interest to the Army. However, the spectra of the profiles of manufactured surfaces are frequently of the power-law form $W = p^{-s}$, where p is the spatial frequency and s is a number $\sim 1, 2$. Consequently, the classical spectral moments -- corresponding to the surface height, slope and curvature variances -- diverge when these spectral densities are extrapolated to very low and very high spatial frequencies, and one is forced to give up the idea of intrinsic finish parameters and to deal with bandwidth-limited values instead. There is therefore a critical need to understand the role of spatial bandwidth limits in the measurement and specification process. The present paper addresses this need: It discusses the role of second-order statistical functions and moments in the surface-characterization problem; the effects of bandwidth limits on the magnitudes of the spectral moments and the relationship between profile and area moments; and concludes with a discussion of the origin and magnitudes of the bandwidth limits in a number of generic measurement situations.

1.0 INTRODUCTION.

1.1 RELEVANCE AND PROBLEM STATEMENT. Surface-finish measurement, characterization and specification are an important problem for the military and industry. This new interest in an old area is due to a number of factors: The development of new manufacturing techniques for mechanical, electronic and optical surfaces; increasing emphasis on standardization, interchangeability and reliability; cost savings inherent in realistic as opposed to over-specification; and concerns for energy and material conservation.

The present standards for surface finish are stated in terms of the average deviation of the surface profile from its mean. This is inadequate because it is insensitive to the transverse character of the roughness and is only indirectly related to the surface area, which determines the functional properties of surfaces.

The statistical basis for a more comprehensive description of topography was developed many years ago by Longuet-Higgins in a famous series of papers concerning ocean waves,¹ and has been adapted more recently to mechanical surfaces.²

Unfortunately, however, this classical approach does not take account of two related real-world situations: First, that measurement techniques and functional properties of surfaces are sensitive to only limited ranges of surface spatial wavelengths; and second, that the spectral densities of real surfaces are such

that they do not "fit" into those wavelength windows. In fact, the spectra of interest are frequently of a form which diverges when extrapolated to very long or very short spatial wavelengths. The signature of the classical description is that it is expressed in terms of intrinsic surface properties, while in fact, one can only speak in terms of bandwidth-limited quantities -- the observed and the observation can no longer be separated.

We address three issues in this paper: 1) The classical generalization of the characterization problem to include second-order statistical properties. 2) The effects of bandwidth limits on the magnitudes of the profile- and surface-finish parameters and the relationship between them. And 3), the sources of bandwidth limits in various measurement situations. For simplicity we limit the discussion to surfaces whose roughness is purely random, isotropic, and weakly stationary and weakly ergodic.

2.0 CLASSICAL DESCRIPTION.

2.1 FIRST-ORDER STATISTICS. The most general probabilistic description of a random variable, Z , is given in terms of the N -th order joint probability distribution

$$P(z_1, \dots, z_n) \quad (1)$$

where $z_i = Z(x_i, y_i)$ is the surface height at the point i . The location and number of points, N , is chosen to provide adequate characterization for the purpose at hand.

The simplest case is $N = 1$; the first-order height distribution function

$$P(z_1) = \int dz_2 \dots \int dz_n P(z_1, \dots, z_n) \quad (2)$$

The simplest example of this is, of course, the Gaussian

$$P(z_1) = (1/\sqrt{2\pi}\sigma) \exp(-z_1^2/2\sigma^2) \quad (3)$$

where σ^2 is the height variance; that is, σ is the standard deviation or rms value of the surface height.

In practice the first-order height distribution function is characterized by a set of finish parameters -- the central moments:

$$M_n \triangleq \int_{-\infty}^{+\infty} dz P(z) |z|^n \quad (4)$$

where $n = 1, 2, 3, 4$ correspond to the average surface height, the height variance, its skew and kurtosis, respectively. The present US and international

standards for surface texture are stated only in terms of the average height:

$$R_a = M_1 = \langle |z| \rangle = \lim_{L \rightarrow \infty} (1/2L) \int_{-L}^{+L} dx |z(x)| \quad (5)$$

14

where $\langle \cdot \rangle$ denotes the ensemble average and $2L$ is the sample record length.

2.2 SECOND-ORDER STATISTICS. The description of Z in terms of first-order statistics alone is clearly limited: it conveys no information about the transverse character of the surface roughness. For example, the two surface profiles



(6)

have the same first-order statistics but very different functional properties. This forces one to consider the inclusion of higher-order distributions in the characterization process.

The next step in sophistication involves the second-order joint probability distribution function

$$P(z_1, z_2) = \int dz_3 \cdots \int dz_n P(z_1, \dots, z_n) . \quad (7)$$

The simplest example of this is again the well-known Gaussian result

$$P(z_1, z_2) = (1/2\pi\sqrt{\sigma^4 - C^2}) \exp[-(C^2 z_1^2 - 2Cz_1 z_2 + \sigma^2 z_2^2)/(2(\sigma^4 - C^2))] . \quad (8)$$

Here C is the height autocovariance function, which may be viewed as the simplest joint moment of the second-order probability distribution:

$$C(x_1, x_2) = M_{11} = \int dz_1 \int dz_2 P(z_1, z_2) z_1^1 z_2^1 . \quad (9)$$

In contrast with the first-order case such moments are functions rather than numbers; in this case a function of the two observation points, x_1 and x_2 . At the particular point $x_1 = x_2$, $C = \sigma^2$, the height variance. Further, if the surface height is a weakly stationary random variable, the autocovariance is only a function of the magnitude of the separation of the two observation points: the lag $\tau = |x_1 - x_2|$.

2.3 SECOND-ORDER FUNCTIONS. The height autocovariance function and its Fourier transforms form a family of functions that are used to describe the second-order statistical properties of a random variable. If the height Z is weakly stationary and weakly ergodic the covariance function can be written in the equivalent forms

$$C(\tau) = \langle z_1 z_2 \rangle = \lim_{L \rightarrow \infty} (1/2L) \int_0^{2L-\tau} dx z(x) z(x + \tau) . \quad (10)$$

The one-dimensional power spectral density -- which appears in the discussion of surface profile data -- is the cosine transform of the autocovariance function:

$$W_1(p) = (1/\pi) \int_0^\infty d\tau \cos(p\tau) C(\tau) = \quad (11a)$$

$$= \lim_{L \rightarrow \infty} (1/2\pi) \langle (1/2L) \left| \int_{-L}^{+L} dx \exp(ipx) Z(x) \right|^2 \rangle \quad (11b)$$

where the parameter p is the surface spatial wavenumber; that is, 2π times the spatial frequency. The two-dimensional power spectral density -- which appears in the discussion of isotropically rough surfaces -- is its zeroth-order Hankel transform:

$$W_2(p) = (1/2\pi) \int_0^\infty \tau d\tau J_0(p\tau) C(\tau) = \quad (12a)$$

$$= \lim_{R \rightarrow \infty} \langle (1/\pi R^2) \left[\int_0^R x dx J_0(px) Z(x) \right]^2 \rangle \quad (12b)$$

where J_0 is the ordinary Bessel function of zeroth order.

When Z is weakly stationary -- that is, the power spectra are derived from a common autocovariance function according to Eqs. (11a) and (12a) -- the one- and two-dimensional power spectra are related to each other through the Abel transforms

$$W_1(p) = 2 \int_p^\infty \frac{tdt}{\sqrt{t^2 - p^2}} W_2(t) \quad (13a)$$

and

$$W_2(p) = -(1/\pi) \int_p^\infty \frac{dt}{\sqrt{t^2 - p^2}} \frac{d}{dt} W_1(t) . \quad (13b)$$

These integral transforms are related to half-integral and half-derivative operations, respectively. Convenient tables of Cosine, Hankel and Abel transforms are given in the Bateman collection.³

Figure 1 is a sketch of the various interrelationships among the various quantities discussed above. A particular trio of functions which are relevant to the following discussions is given in the Appendix.

2.4 SECOND-ORDER FINISH PARAMETERS. The second-order statistical functions discussed above may themselves be characterized by a set of finish parameters corresponding to the central moments of the power spectral densities. Specifically, the even n -th order profile moments:

$$m_n \stackrel{\Delta}{=} 2 \int_0^\infty dp W_1(p) p^n = \langle \left| \frac{d^{n/2}}{dx^{n/2}} Z(x) \right|^2 \rangle . \quad (14)$$

These moments have direct physical meaning as indicated on the right: m_0 is the variance of the profile height, m_2 is the variance of the profile slope, and m_4 is the variance of the profile "curvature". A similar set of moments can be defined for the two-dimensional spectral density:

$$M_n \stackrel{\Delta}{=} 2\pi \int_0^{\infty} pdp W_2(p)p^n = \langle |\nabla_r^{(n/2)} z(r)|^2 \rangle . \quad (15)$$

As indicated on the right, M_0 is the variance of the surface height, M_2 is the variance of the surface gradient, and M_4 is the variance of the Gaussian curvature of the surface. The immediate value of these moments in determining the functional properties of surface roughness is obvious.

The relationship between the profile and area moments is of critical importance in the measurement and specification process since the traditional method of measuring surface finish is by means of a stylus gauge -- which measures the surface profile -- while the functional properties of surfaces clearly depend on their surface properties.

If the surface roughness is weakly stationary, the two power spectra are related -- by Eq. (13) -- and therefore, so are their moments. In particular,

$$\frac{m_n}{M_n} = \frac{\Gamma((1/2)[n+1])}{\Gamma(1/2)\Gamma((1/2)[n+2])} \quad (16)$$

where Γ is the gamma function. This result says, for example, that $m_0 = M_0$, which means that the profile roughness equals the area roughness; $m_2 = \frac{1}{2} M_2$, which means that the variance of the profile slope is half that of the surface gradient; and $m_4 = 3/8 M_4$, which means that the variance of the profile curvature is three-eighths of the Gaussian curvature of the surface. All this, of course, for an isotropically rough surface. Straightforward generalizations to anisotropic surfaces have also been given.¹

2.5 CLASSICAL SOLUTION. We call the above the classical solution to the characterization problem. It has a number of attractive features: 1) It represents the next step in the systematic generalization of the present finish standards which includes the transverse as well as the vertical character of the roughness. 2) It is stated in terms of a set of finish parameters which have direct physical meaning: the rms values of the surface height and its derivatives. 3) It provides a direct relationship between profile measurements -- which are easier to make -- and the properties of the surface area -- which determine its functional properties. And 4), the scheme on which it is based is readily generalizable to include higher-order statistical properties -- if and when required. In fact, if the roughness can be taken as a full Gaussian process, the probabilistic description to arbitrary order may be described solely in terms of its second-order properties.⁴ In that case, a wealth of results concerning the surface can be expressed in terms of the first three profile moments, m_0 , m_2 and m_4 discussed above. Examples are given by Longuet-Higgins and others.^{1,2}

3.0 REAL SURFACES.

3.1 MECHANICAL SURFACES. The traditional method of measuring surface roughness uses a stylus gauge to determine the surface profile $Z(x)$. This is usually used for "mechanical" surfaces; that is, surfaces generated by machining, grinding or lapping processes, which generally lead to surfaces with vertical roughness of > 1 microinch (25 nanometers).

Such surfaces very frequently exhibit power spectral densities of the form

$$W_1(p) \propto p^{-2} \quad (17)$$

where $p = 2\pi/d$ is the surface spatial wavenumber and d is the surface spatial wavelength.⁵ This result says that the profile power spectral density is proportional to the square of the spatial wavelength, or equivalently, that the roughness is scale invariant: it is statistically the same at all magnifications.⁶

The ubiquity of this "inverse-square" law is amazing, and a number of models have been invoked to explain it. One is that "machining" processes lead to surface finish involving numerous vertical edges.⁶ In other words, the surface profile can be viewed as a kind of telegrapher's signal. However, this same spectrum is also exhibited by Brownian, Markov and autoregressive processes, and its appearance in nature probably lies more in the multiplicity of the processes that exhibit this behavior than in a common physical origin.

3.2 OPTICAL SURFACES. Optical-quality surfaces generally have roughnesses of < 1 microinch and are difficult to measure by mechanical stylus techniques. An alternative method is to measure the angular distribution of the intensity of light scattered from the surface, which is a simple mapping of the two-dimensional power spectrum of the surface height when the vertical roughness is much less than the radiation wavelength.

Although this art is in its infancy, available data suggest that polished optical surfaces frequently exhibit spectra which, when translated into the one-dimensional form, correspond to

$$W_1(p) \propto p^{-1} . \quad (18)$$

This behavior has been interpreted in terms of a surface-tension model in which polishing is viewed not as a cutting process per se, but a smoothing operation which minimizes the excess surface area due to the residual surface roughness.^{6,7} Superficially, Eq. (18) resembles 1/f electrical noise; another ubiquitous form which also derives from a variety of physical processes in nature. Other inverse-power-law forms have been reported for polished surfaces as well.⁶

3.3 LIQUID SURFACES. Liquid surfaces exhibit a variety of height distributions depending on the nature of their excitation. Simple forms for the power spectral densities appear in two limiting cases: Capillary waves, which are governed by surface tension and lead to the hyperbolic form, Eq. (18); and the "fully aroused sea" which exhibits an inverse-cube profile spectrum.⁸

The fact that mechanical, polished and liquid surfaces all exhibit approximate inverse-power-law spectra with integral (or half-integral?) exponents suggests a deeper physical connection between surface roughness, the properties of the surface medium, and the roughening process. A first cut at understanding this interesting situation has been given in terms of a shot model of surface roughness coupled with the asymptotic properties of the Fourier integral.⁶

3.4 PRACTICAL IMPLICATIONS. The immediate significance of these results lies less in their specific algebraic forms or their physical origins, than in their ill-behavedness -- that is, they can cause the moment integrals, Eqs. (14) and (15) to blow up when the power spectra in the integrand are extrapolated to very small or very large spatial frequencies.

One can adopt two views of this: The classical view holds that this situation is an artifact of the measurement process; that if measurements were made over a sufficiently wide range of spatial frequencies the spectra would ultimately become well behaved and their moments finite. The radical view is that intrinsic surface parameters such as the classical spectral moments are operationally undefined -- if for no other reason than that the spatial wavelengths involved can't be smaller than atomic dimensions or larger than the size of the workpiece -- and the real world can only be discussed in terms of bandwidth-limited values of those parameters. Further, that even if "intrinsic" parameters could be defined and measured by some extrapolation procedure, the results would be of no practical value since the functional properties of surfaces -- as measurement processes themselves -- depend only on a limited range of spatial wavelengths. Other workers have described surfaces exhibiting such ill-behaved power spectra as non-stationary.⁵

Philosophy aside, the practical implications for surface-finish measurement and specification appear to be the following:⁹ 1) Surface-finish parameters are not intrinsic properties -- one cannot speak of a "1-micron surface" but only of a surface which exhibits a 1-micron roughness over a certain range of spatial wavelengths. 2) The effects of bandwidth limits upset the classical relationship between the profile and area moments. 3) One must understand the details of individual measurement procedures to be able to specify the range of wavelengths included in the measured values, both for specificity and to ensure valid comparison with other measurement processes. And 4), one must understand the functional properties of surfaces in enough detail to be able to estimate the range of spatial wavelengths to be included in any specifications for such surfaces.

Items 1, 2 and 3 are discussed at greater length in the following sections of this paper.

4.0 EFFECTS OF BANDWIDTH LIMITS.

4.1 BANDWIDTH-LIMITED MOMENTS. Bandwidth-limited values of the spectral moments are defined as

$$m_n(d_{\min}, d_{\max}) \stackrel{\Delta}{=} 2 \int_{p_{\min}}^{p_{\max}} dp W_1(p) p^n \quad (19a)$$

and

$$M_n(d_{\min}, d_{\max}) \triangleq 2\pi \int_{p_{\min}}^{p_{\max}} p dp W_2(p) p^n , \quad (19b)$$

where

$$p_{\min} = 2\pi/d_{\max} , \quad p_{\max} = 2\pi/d_{\min} \quad (20)$$

are the minimum and maximum spatial wavenumbers included in the definition or measurement. When these limits go to zero and infinity the bandwidth-limited moments become the classical results, Eqs. (14) and (15).

4.2 RING-SPECTRUM SURFACE. The properties of the bandwidth-limited moments can depend sensitively on the bandwidth limits and the shape of the power spectrum. To illustrate the dramatic effects that are possible we begin with an extreme example, that of a ring spectrum:

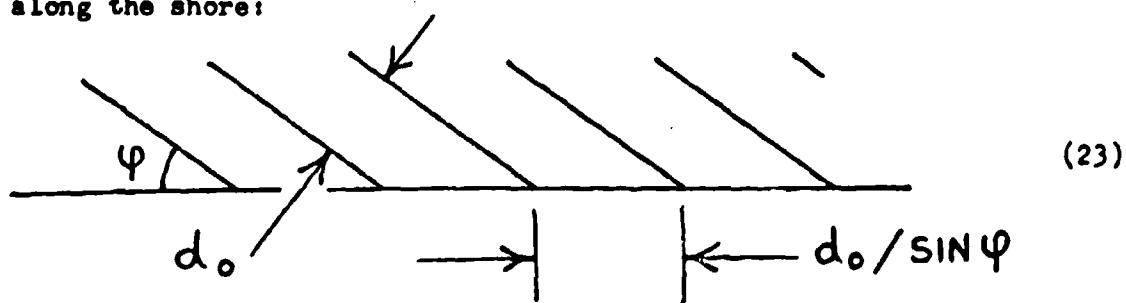
$$W_2(p) = \delta(p - p_0) \quad (21)$$

where δ is a delta function.* Such surfaces could be realized by superimposing a set of sinusoidal corrugations with a fixed spatial wavelength, $d_0 = 2\pi/p_0$, in random directions over the surface. A beam of light striking such a surface normally would then be diffracted into a "ring" of light at the polar angle given by the grating equation, $\sin \theta = \lambda/d_0$, where λ is the radiation wavelength; hence the name. The corresponding one-dimensional power spectrum is, from Eq. (13a):

$$W_1(p) = 2p_0/\sqrt{p_0^2 - p^2} \quad (22)$$

for $p < p_0$, but zero otherwise. These two spectra are sketched in Fig. 2.

The striking difference in their form is due to the wave-on-the-beach phenomenon: A surface wave of wavelength d_0 exhibits a longer wavelength, $d_0/\sin \varphi$, when viewed along the shore:



*For simplicity a normalizing factor with the dimensions of length³ has been suppressed on the right of Eq. (21).

The classical moments of the ring-spectrum surface are given by

$$M_n = 2\pi p_0^{n+1}, \quad (24)$$

plus the moment ratio, Eq. (16). However, the bandwidth-limited moments are very different.

Look at Fig. 2, and begin by considering the situation where the spectral window falls entirely above p_0 ; that is, $p_{\min} > p_0$. In that case both the profile and area moments vanish and the surface appears to be perfectly smooth. If the spectral window is now moved downward to include the point $p = p_0$, that is, $p_{\min} < p_0 < p_{\max}$, both the profile and the area moments will be nonvanishing, although they only satisfy the classical moment ratio when $p_{\min} = 0$. And finally, when the spectral window is moved entirely below p_0 , that is, $p_0 > p_{\max}$, the profile moments are still nonvanishing but the area moments are identically zero. In other words: the surface is rough to static measurements but smooth to scattering measurements even though both measurements are made over the same range of spatial wavelengths!

Although ring-spectra surfaces are not encountered in conventional manufacturing, the point is still well made that bandwidth-limited finish parameters can have very different properties from intrinsic parameters of the same surface: Both their magnitudes and the relationship between the profile and area properties are affected. Such effects are examined below for more realistic surface spectra.

4.3 POWER-LAW SURFACES. Suppose the profile power spectrum measured over a limited range of spatial frequencies has the form

$$W_1(p) = p^{-s} \quad (25)$$

where s is a number.* In order to obtain the corresponding form of W_2 we must know W_1 to infinitely high spatial frequencies. If we take the inverse power-law form shown to be valid into that unmeasured region, Eq. (13b) shows that the corresponding form of the area spectrum is

$$W_2(p) = F(s) \cdot p^{-s-1} \quad (26)$$

where

$$F(s) = \frac{\Gamma(\{1/2\}[s + 1])}{\Gamma(1/2)\Gamma(\{1/2\} s)} . \quad (27)$$

*For simplicity a normalizing factor with the dimensions of length^(3-s) has been suppressed on the right of Eq. (25).

Given these analytic forms for the power spectra we can easily calculate the magnitudes of the bandwidth-limited spectral moments, Eq. (19). Rather than giving the obvious algebraic forms, values are given in Table 1 for selected values of s and n .

In all cases the magnitudes of the moments depend explicitly on the minimum and maximum spatial wavelengths included, and diverge in the limit $p_{\min} \rightarrow 0$ and/or $p_{\max} \rightarrow \infty$. However, the precise dependences, and the sensitivity of the results to variations in the finite wavelength limits, depends on the case considered. For example, for machined surfaces, $s = 2$, the rms profile height, slope and curvature scale approximately as $d_{\max}^{-1/2}$, $d_{\min}^{-1/2}$ and $d_{\min}^{-3/2}$, respectively.¹⁰ While for polished surfaces, $s = 1$, they scale as $\log^2(d_{\max}/d_{\min})$, d_{\min}^{-1} and d_{\min}^{-2} . For the cases of interest the slope and curvature parameters are generally most sensitive to d_{\min} , as expected physically.

4.4 PROFILE-AREA RELATIONSHIP. The relationship between the profile and area parameters is determined by the moment ratio m_n/M_n . For power-law spectra this ratio is simply

$$\frac{m_n}{M_n} = \frac{\Gamma(s/2)}{\Gamma(1/2)\Gamma(\{1/2\}[s+1])} \quad (28)$$

which has the values of 1, $2/\pi$ and $1/2$ for $s = 1, 2$, and 3 . This result is to be compared with the corresponding classical expression, Eq. (16).

There are a number of interesting similarities and differences: Both results are independent of the bandwidth limits, although this is an "accident" in the case of the power-law spectra. The bandwidth-limited ratio depends only on the shape of the power spectrum, i.e. the parameter s , and is independent of the moment order -- further accidental properties of the power-law spectra -- while the classical result is independent of the spectral shape and depends only on the order, n .

The simplest way of displaying the magnitudes of these differences is to examine the ratio of the surface moments derived from a given set of profile moments using the classical and bandwidth-limited ratios, Eqs. (16) and (28):

$$\frac{M_n^{\text{classical}}}{M_n^{\text{bw-limited}}} = \frac{\Gamma(\{n+2\}/2)}{\Gamma(\{n+1\}/2)} \cdot \frac{\Gamma(s/2)}{\Gamma(\{s+1\}/2)} \quad . \quad (29)$$

Values of this ratio are given in Table 2 for particular cases of interest. They indicate that the numerical error introduced by using the classical recipe is moderate for the lower moments of machined surfaces, $s = 2$, but more significant in other cases. Once again, however, these results apply only to isotropically rough surfaces exhibiting power-law spectra to infinitely high frequency.

If the measured profile spectrum cannot be extrapolated to infinite frequency, the corresponding form of the surface spectrum cannot be determined, and the profile-area translation is impossible. This follows from the form of the

Abel transformations which relate W_1 and W_2 , Eq. (13): $W_2(p)$ depends on the form of $W_1(t)$ for all $t \geq p$, and conversely. This situation indicates the essential role that models play in the discussion of surface-finish measurements and specification, since it is only through faith in empirical or physical models that one can justify the extrapolation procedures required for translating profile data into area data, or vice versa. The ability to accomplish this is critical since present finish standards are given in terms of profile measurements.¹⁴

4.5 CLASSICAL SURFACES. The classical view is that all power spectra are naturally well behaved and only appear to be ill behaved over a limited range of spatial frequencies. Or to put it another way, the ill-behavedness lies in the use of improper extrapolation functions.

Roughly speaking, the high-frequency tail of the spectrum determines the properties of the higher moments while its low-frequency behavior determines the lower moments. In the case of power-law spectra discussed above a simple extrapolation to high frequencies was sufficient to establish the profile-area connection, but the classical moments still diverge either at their upper or lower limits. Such behavior is, of course, classically unacceptable.

A v -th order classical spectrum may be defined as one whose classical profile moments, m_n , are finite for all $n < 2v$. Thus, a first-order classical surface possesses a finite height variance, but no higher finite moments. A Cauchy spectrum, derived from a simple exponential autocovariance function, is a familiar example of this type. A Gaussian spectrum, on the other hand, is an infinite-order classical spectrum: all of its profile moments are finite.

To achieve finite classical moments the measured spectra must be extrapolated with rounding functions which kill the divergences, especially at low frequencies. However, those rounding functions will necessarily involve new length parameters -- such as correlation lengths -- in place of the window wavelengths in the bandwidth-limited case. The resulting classical moments, although finite, will then have different (larger) magnitudes, and will depend on different sets of physical parameters than the bandwidth-limited values.

The preceding discussions were principally concerned with moment ratios in which these additional parameters do not appear. A detailed discussion of a particular rounding of the low-frequency tail of the inverse-power-law spectra is given in the Appendix.

5.0 SOURCES OF BANDWIDTH LIMITS. The preceding Section illustrates the importance of bandwidth limits in the characterization of surfaces with ill-behaved power spectra. This Section discusses a number of sources of such limits in common measurement processes: to emphasize that all measurements are inherently bandwidth-limited, and to indicate how the magnitudes of those limits arise in practical situations.

5.1 ELECTRONIC FILTERING. Mechanical stylus measurements involve drawing a fine diamond-tipped stylus over the surface being measured, converting the tip motion into an electrical signal, and analyzing the resulting time series in terms of the surface profile, $Z(x)$. However, the output represents not the true profile

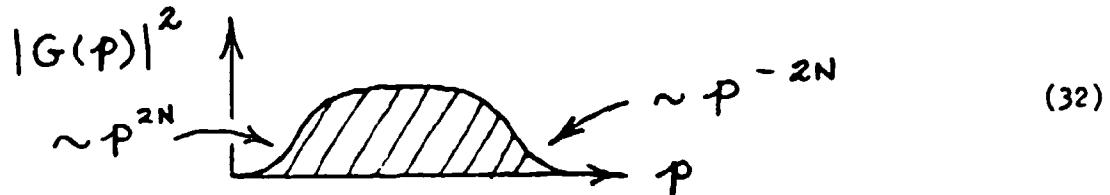
but the apparent profile $z'(x)$:

$$z'(x) = \mathcal{F}^{-1}\{G(p) \cdot \mathcal{F}\{z(x)\}\} \quad (30)$$

where \mathcal{F} denotes the Fourier transform and G is the transfer function of the measurement system, which is usually dominated by electronic filters or "cutoffs". The power spectral density of the apparent profile is then

$$W'(p) = |G(p)|^2 \cdot W_1(p) \quad (31)$$

where $W_1(p)$ is the true spectrum. The form of $G(p)$ depends on the nature of the filter. For example, for a simple N -stage RC bandpass filter the low- and high-frequency cutoffs behave asymptotically as $p^{\pm 2N}$, respectively:



5.2 STYLUS SMOOTHING. Mechanical styli act as a low-pass filter since they tend to ride over height fluctuations with spatial wavelengths smaller than the tip radius.

One measure of the high-frequency cutoff that appears frequently in the literature is obtained from the model of a circular tip riding over a set of sinusoidal corrugations. Simple geometry then shows that the requirement that the tip track into the valleys is equivalent to the statement that the maximum undistorted spatial frequency is

$$p_{\max} = (aR)^{-1/2} \quad (33)$$

where a is the amplitude of the corrugations (half the peak-to-valley distance), and R is the radius of the stylus tip. For example, $a = 1 \mu$ and $R = 1 \mu$ require that $d_{\min} \sim 6 \mu$.

This result can be generalized to a randomly rough surface by replacing the geometrical condition by the requirement:

$$m_4 < R^{-2}, \quad (34)$$

which states that the rms profile curvature must be less than that of the stylus tip. A sinusoidal profile has the spectrum $W = \frac{4\pi a^2}{\lambda} \delta(p - p_{\max})$, which leads precisely to the geometric result, Eq. (33). On the other hand, the spectrum

$W = p^{-2}$ gives

$$P_{\max} \approx \left[\frac{3}{2\pi} \cdot \frac{d_{\max}}{\sigma^2 R^2} \right]^{1/3} \quad (35)$$

where σ^2 is the height variance and d_{\max} is the maximum spatial wavelength included in its measurement. For example, $R = 1 \mu$, $\sigma = 1 \mu$ and $d_{\max} = 10^3 \mu$ give $d_{\min} \sim 8 \mu$. Similarly, for a spectrum of the form $W = p^{-1}$

$$P_{\max} \approx \left[\frac{4 \log}{\sigma^2 R^2} \right]^{1/4}, \quad (36)$$

where $\log = \ln(d_{\max}/d_{\min})$. For the same parameters as above, this gives $d_{\min} \sim 3 \mu$.

Interestingly, although each of these three examples gives values of the minimum undistorted spatial wavelengths which are of the same order of magnitude as the tip radius, none of the expressions derived predicts proportionality between d_{\min} and R . Also, the non-intuitive form of the factors in Eq. (36) is more apparent than real: the quantity \log/σ^2 is a simple constant, as seen from Table 1.

5.3 APERTURE SMOOTHING. Optical stylus measurements produce an apparent profile which is a smoothed version of the true profile. In one-dimensional terms

$$Z'(x) = \int dt \omega(t-x) Z(t) \quad (37)$$

where ω is the window function of the apparatus. The power spectral density of this smoothed profile is then

$$W'_1(p) = \Omega(p) \cdot W_1(p) \quad (38)$$

where W_1 is the true spectrum and

$$\Omega(p) = \left| \int dx e^{ipx} \omega(x) \right|^2 \quad (39)$$

is the window transfer function. Such smoothing, of course, acts as a low-pass filter with the nominal cutoff

$$P_{\max} \approx 2\pi/A \quad (40)$$

where A is the width of the smoothing window. The precise form of the transfer

function depends on the window shape: A Gaussian window gives a Gaussian Ω ; a rectangular window, a Sinc^2 function; a cylindrical window, an Airy function, and so on.

5.4 SURFBOARDING. Surfboarding arises from the fact that profile measurements are usually made with reference to a local rather than an absolute baseline. That is, one usually measures not the true profile, $Z(x)$, but

$$Z'(x) = Z(x) - (a + bx) \quad (41)$$

where the term in parentheses represents the least-squares-average line through the individual records. The apparent power spectrum is then a convolution of the true spectrum of the form^{11,12}

$$W'(p) = \frac{L}{\pi} \int_{-\infty}^{+\infty} dq [\xi - n - \zeta]^2 W_1(q) \quad (42)$$

where the three functions in the kernel are

$$\xi = j_0(pL - qL) \quad (43a)$$

$$n = j_0(pL)j_0(qL) \quad (43b)$$

$$\zeta = 3j_1(pL)j_1(qL) \quad (43c)$$

and the j 's are spherical Bessel functions. The ξ term corresponds to the finite record length, $2L$; the n term to the removal of the average from each record; and the ζ term to the removal of the least-squares slope from each individual record in the ensemble. In the limit of very large record length the kernel becomes a δ function and $W' \rightarrow W$ for $p > 0$, as expected. Numerical and analytic evaluation of Eq. (42) shows that surfboarding acts as a high-pass filter which cuts off spatial frequencies below

$$p_{\min} \approx \pi/L, \quad (44)$$

where the shape of the cutoff depends on the form of W .

The nature of this effect is sketched in Figure 3. It is called surfboarding because, in effect, the baseline for a finite record length "surfboards" over the surface profile and measurements relative to that baseline are insensitive to very long spatial wavelengths. In fact, it is readily seen from Eqs. (42) and (43) that the apparent power spectrum vanishes at zero spatial frequency.

5.5 SURFACE SCATTERING. Light and acoustic scattering offer a means for the direct measurement of the two-dimensional power spectra of the surface roughness, W_2 .⁶ In such measurements the scattering angle is related to the surface spatial wavelength, d , through the familiar grating equation:

$$|\sin\theta_s - \sin\theta_i| = \lambda/d \quad (45)$$

where θ_i and θ_s are the polar angles of incidence and scattering, respectively, and λ is the radiation wavelength. This equation follows from the spatial invariance of the electromagnetic or acoustic equations and is independent of the details of any specific scattering theory.

In practice, the scattering angle, θ_s , is limited to a maximum of $\pi/2$ — glancing scattering — and a minimum of $|\theta_s - \theta_i| \sim \lambda/L$, where $2L$ is the diameter of the illuminated surface area. Interestingly, there are at least three mechanisms that determine this minimum scattering angle: The diffraction limit corresponding to the finite illuminated surface aperture; the requirement that the maximum spatial wavelength be smaller than the record length to ensure statistical stability of the measurements; and finally, surfing, as described above. This last enters through the fact that the record mean and slope are automatically removed in aligning the scattering apparatus with reference to the centroid of the specular reflection for each sample area.

For the usual case of near-normal incidence the spatial bandwidth limits involved in scattering measurements are

$$\frac{2\pi}{\lambda} \leq p \leq \frac{2\pi}{L} \quad (46)$$

or equivalently, $d_{\min} \sim \lambda$ and $d_{\max} \sim L$. Practical limits generally fall well within these extremes. For example, for a beam of red HeNe laser light 1 mm in diameter, $d_{\min} \sim 1 \mu$ and $d_{\max} \sim 100 \mu$.

6.0 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS. Classical discussions of surface topography assume that the power spectral densities of the profile height $W_1(p)$, the profile slope $p^2 W_1(p)$, and the profile curvature $p^4 W_1(p)$, are integrable over spatial frequencies from zero to infinity. This leads to the satisfying picture that the rms surface height, slope and curvature are finite, intrinsic surface-finish parameters; and further, that there is a one-to-one connection between profile properties and those of the surface area.

However, there are two serpents in this Eden of simplicity: Measurements and functional properties of surfaces are sensitive to only limited ranges of surface spatial wavelengths, and many surfaces display spectra that lead to non-integrable power spectral densities when extrapolated to very low and/or very high spatial frequencies.

As a result:

- 1) The range of spatial wavelengths must be included in measured or specified finish parameters; for specificity, to ensure valid comparison with other measurements, and for functional applications.
- 2) Measurement techniques must be analyzed to determine the range of spatial wavelengths which they include; or better said, their frequency characteristics or transfer functions.
- 3) Functional properties of surfaces must be examined to determine the range of spatial wavelengths of importance. In the case of electromagnetic and acoustic scattering this function-finish relationship is well known, but in other cases, such as friction and wear, the connection is not as clear.
- 4) Physical models should be developed for the finish generated by various manufacturing processes. These models are necessary for the further understanding of the generation process, the simplification of the specification process, and the extrapolation of measured spectra into unknown regions. Such extrapolation is necessary for the broadening of the range of application of various measurement techniques and for relating profile and area specifications.
- 5) Bandwidth-limit effects discussed in this paper should be extended to other important issues involved in the measurement and specification processes which have been omitted in the present discussion: such as two-dimensional measurement techniques, anisotropic surfaces, sampled data, and statistical stability.¹³

7.0 APPENDIX. This Appendix describes a classical trio of second-order statistical functions which behave as the inverse power-law spectra discussed in the text at high spatial frequencies, $W_1 = p^{-s}$, but which possess finite spectral moments, m_n , for all $n < s - 1$. These functions are related by the Fourier, Hankel and Abel transformations illustrated in Fig. 1.

The autocovariance function is

$$C(\tau) = 2D_s \frac{\Gamma(1/2)}{\Gamma(s/2)} \left(\frac{\tau}{2a}\right)^{[s-1]/2} K_{[s-1]/2}(a\tau) \quad (47)$$

where Γ is the gamma function and K is the modified Bessel function. D_s is a constant having the dimensions of length to the $(3-s)$ power, and a is a constant with the dimensions of reciprocal length: $1/a$ is the "correlation length" for this class of functions.

The corresponding one-dimensional power spectral density -- the cosine transform of C -- is

$$W_1(p) = D_s (p^2 + a^2)^{-s/2} \quad (48)$$

The two-dimensional power spectral density -- the zeroth-order Hankel transform C -- is

$$W_2(p) = D_s F(s) \cdot (p^2 + a^2)^{-(s+1)/2} \quad (49)$$

where

$$F(s) = B^{-1}\left(\frac{1}{2}, \frac{s}{2}\right), \quad (50)$$

and B is the beta function

$$B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta). \quad (51)$$

In the limit $p \gg a$ -- that is, for spatial wavelengths much shorter than $2\pi/a$ -- these spectra become the simple power-law forms discussed in the text: Eqs. (25) - (27).

The classical moments of the spectra, Eqs. (48) and (49), are:

$$m_n = D_s \cdot B\left(\frac{n+1}{2}, \frac{s-n-1}{2}\right) \cdot a^{n+1-s} \quad (52)$$

and

$$M_n = \pi B^{-1}\left(\frac{1}{2}, \frac{n+1}{2}\right) \cdot m_n. \quad (53)$$

The ratio of the classical profile moments to the corresponding bandwidth-limited values is then

$$\frac{m_n^{\text{classical}}}{m_n^{\text{bw-limited}}} = B\left(\frac{n+1}{2}, \frac{s-n-1}{2}\right) \cdot \left(\frac{a_{\max}}{2\pi}\right)^{n+1-s}, \quad (54)$$

in the limit $a_{\max} \gg a_{\min}$. This ratio indicates the conjugate roles of the correlation length and the maximum spatial wavelength in determining the magnitudes of the classical and bandwidth-limited moments, and the fact that the former moments are much larger than the latter.

Two special cases of the above general results are of particular interest: $s = 1$ and $s = 2$.

For $s = 1$:

$$C(\tau) = 2D_1 K_0(a\tau) \quad (55a)$$

$$W_1(p) = D_1 [p^2 + a^2]^{-1/2} \quad (55b)$$

$$W_2(p) = (D_1/\pi) [p^2 + a^2]^{-1} \quad (55c)$$

and for $s = 2$:

$$C(\tau) = \pi(D_2/a)e^{-a\tau} \quad (56a)$$

$$W_1(p) = D_2 [p^2 + a^2]^{-1} \quad (56b)$$

$$W_2(p) = (D_2/2) [p^2 + a^2]^{-3/2} \quad (56c)$$

These two cases represent zeroth- and first-order classical surfaces, respectively, in the sense defined in Section 4.5.

8.0 REFERENCES.

1. M. S. Longuet-Higgins, "The Statistical Analysis of a Random Moving Surface", Phil. Trans. R. Soc. London, A249, 321-387 (1956-57); "Statistical Properties of an Isotropic Random Surface" (*ibid.*) A250, 157-174 (1957-58).
2. D. J. Whitehouse and J. F. Archard, "The Properties of Random Surfaces of Significance in Their Contact", Proc. R. Soc. London, A316, 97-121 (1970).
3. Tables of Integral Transforms (McGraw-Hill Book Company, New York, 1954), A. Erdelyi, Ed.
4. G. A. Korn and T. M. Korn, Mathematical Handbook for Scientists and Engineers (McGraw-Hill Book Company, New York, 1968), pages 654 et seq.
5. R. S. Sayles and T. R. Thomas, "Surface Topography as a Nonstationary Random Process", Nature 271, 431-434 (1978); 273, 573 (1978).
6. E. L. Church, H. A. Jenkinson and J. M. Zavada, "The Relationship between Surface Scattering and Microtopographic Features", Opt. Engineering 18, 125-136 (1979).
7. E. L. Church, "Sources of $1/\theta^2$ Scattering from Optical Surfaces", Jour. Opt. Soc. Amer. 68, 1426A (1978).

8. O. M. Phillips, The Dynamics of the Upper Ocean (Cambridge University Press, New York, 1966).
9. E. L. Church, "Surface Finish Specifications", Jour. Opt. Soc. Amer. 69, 11404A (1979).
10. T. R. Thomas and R. S. Sayles, "Some Problems in the Tribology of Rough Surfaces", Tribology Int. 11, 163-168 (1978).
11. E. L. Church, "Small-Angle Scattering from Smooth Surfaces", Jour. Opt. Soc. Amer. 70, 1592A (1980).
12. E. L. Church, "Interpretation of High-Resolution X-Ray Scattering Measurements", Proc. SPIE 257, 254-260 (1980).
13. E. L. Church, Unpublished results (1981).
14. American National Standard, Surface Textures (Surface Roughness, Waviness and Lay), ANSI B46.1, 1978.

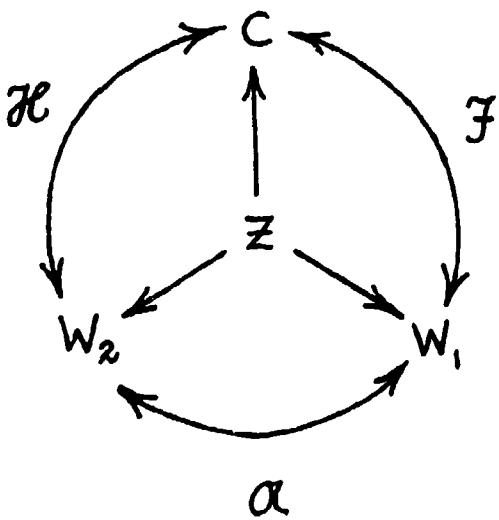


Figure 1. Relationships between the three second-order statistical functions C , W_1 and W_2 of the random variable Z . F , H and A stand for the Fourier, Hankel and Abel transforms discussed in the text.

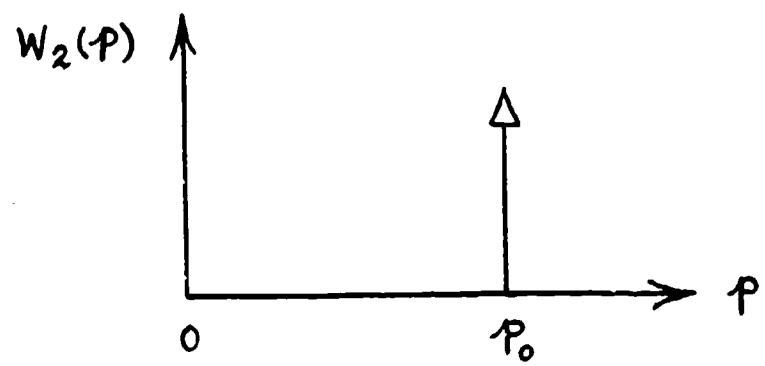
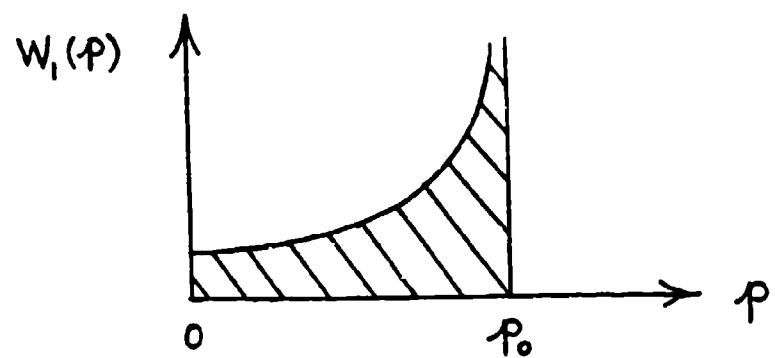


Figure 2. The one- and two-dimensional power spectral densities of a "ring-spectrum surface".

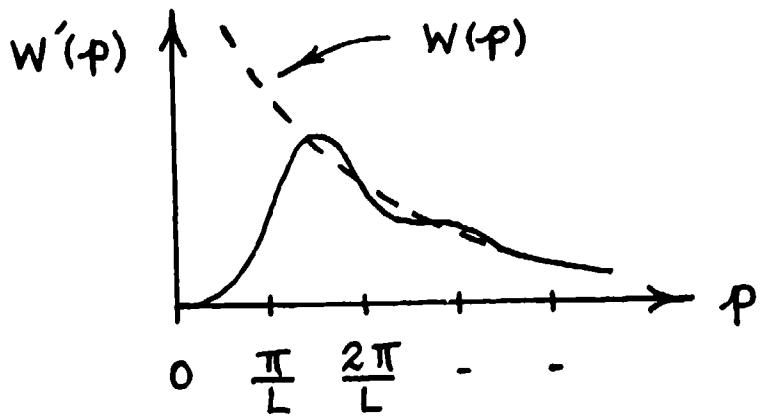
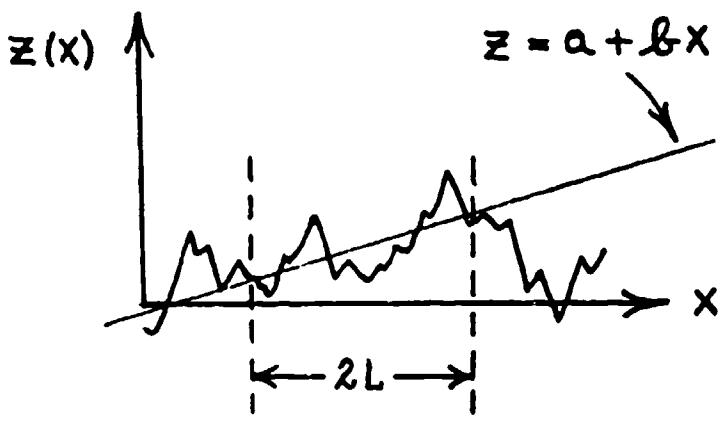


Figure 3. Surfboarding in configuration space (top), and in frequency space (bottom). The local reference "surfboards" over the surface profile and washes out spatial wavelengths longer than the record length, $2L$.

$W_1 =$	P^{-1}	P^{-2}	P^{-3}
m_0	$2 \log$	$2 \left(\frac{d_{\max}}{2\pi} \right)^1$	$\left(\frac{d_{\max}}{2\pi} \right)^2$
m_2	$\left(\frac{2\pi}{d_{\min}} \right)^2$	$2 \left(\frac{2\pi}{d_{\min}} \right)^1$	$2 \log$
m_4	$\frac{1}{2} \left(\frac{2\pi}{d_{\min}} \right)^4$	$\frac{2}{3} \left(\frac{2\pi}{d_{\min}} \right)^3$	$\left(\frac{2\pi}{d_{\min}} \right)^2$
M_0	$2 \log$	$\pi \left(\frac{d_{\max}}{2\pi} \right)^1$	$2 \left(\frac{d_{\max}}{2\pi} \right)^2$
M_2	$\left(\frac{2\pi}{d_{\min}} \right)^2$	$\pi \left(\frac{2\pi}{d_{\min}} \right)^1$	$4 \log$
M_4	$\frac{1}{2} \left(\frac{2\pi}{d_{\min}} \right)^4$	$\frac{\pi}{3} \left(\frac{2\pi}{d_{\min}} \right)^3$	$2 \left(\frac{2\pi}{d_{\min}} \right)^2$

Table 1. Bandwidth-limited values of the profile and area moments corresponding to the profile spectra shown. Results are given in the limit $d_{\max} \gg d_{\min}$. Here $\log = \ln(d_{\max}/d_{\min})$. Dimensional constants have been suppressed.

$s \backslash m$	0	2	4
1	$1 = 1.000$	$2 = 2.000$	$\frac{8}{3} = 2.667$
2	$\frac{2}{\pi} = 0.637$	$\frac{4}{\pi} = 1.273$	$\frac{16}{3\pi} = 1.698$
3	$\frac{1}{2} = 0.500$	$1 = 1.000$	$\frac{4}{3} = 1.333$

Table 2. Ratio of the area moments $M_n^{\text{classical}}/M_n^{\text{bw-limited}}$ calculated from Eq. (29) for different profile spectra of the form $W_1 = p^{-s}$.

Bounds for Optimal Confidence

Limits for Series Systems

Bernard Harris* and Andrew P. Soms**

Abstract

Lindstrom-Madden type approximations to the lower confidence limit on the reliability of a series system are theoretically justified by extending and simplifying the results of Sudakov (1973). Applications are made to Johns (1976) and Winterbottom (1974).

Numerical examples are presented.

Key words: Lindstrom-Madden approximation; Optimal confidence bounds; Reliability; Series system.

*University of Wisconsin-Madison.

**University of Wisconsin-Milwaukee. Research supported by the Office of Naval Research under Contract No. N00014-79-C-0321 and the United States Army under Contract No. DAAG29-75-C-0024.

1. Introduction and Summary

A problem of fundamental interest to practitioners in reliability is the statistical estimation of the reliability of a system using experimental data collected on subsystems. In this paper, the subsystem data available consists of a sequence of Bernoulli trials in which a "one" is recorded if the subsystem functions and a zero is recorded if the subsystem fails. Thus for each of the k subsystems composing the system, the data provided consists of the pair (n_i, Y_i) , $i=1,2,\dots,k$, where Y_i is binomially distributed (n_i, p_i) . We assume that Y_1, Y_2, \dots, Y_k are mutually independent random variables.

The magnitude of interest in this problem is easily evidenced by the extensive literature devoted to it. In this regard, see the survey paper by Harris (1977) and Section 10.4 of the book by Mann, Schafer, and Singpurwalla (1974). In addition, the Defense Advanced Research Projects Agency has recently issued a Handbook for the Calculation of Lower Statistical Confidence Bounds on System Reliability (1980).

Historically, the first significant work on this problem was produced by Buehler (1957). However, Buehler's method as described in that paper is difficult to implement computationally when $k > 2$.

We proceed by describing Buehler's method in Section 2. In Section 3 we specialize to series systems, that is, a system which fails whenever at least one subsystem fails. Sudakov's (1974) results are extended in Section 4 and employed to exhibit some optimality properties of the Lindstrom-Madden method (see Lloyd

and Lipow (1962)) for constructing lower confidence bounds for the reliability of series systems of stochastically independent subsystems. Some numerical examples are given in Section 5 and the results needed for this generalization of Sudakov's Theorem are provided in the Appendix to this paper.

2. Buehler's Method for Lower Confidence Bounds

A system composed of k independent subsystems is said to be a coherent system (with respect to the specified decomposition into subsystems), if the system fails when all subsystems fail and the system functions when all subsystems function; and replacing a defective subsystem by a functioning subsystem can not cause a functioning system to fail. Coherent systems are described in Birnbaum, Esary and Saunders (1961) and Barlow and Proschan (1975).

To any system one can associate a function, $h(\tilde{p}) = h(p_1, p_2, \dots, p_k)$, $0 \leq p_i \leq 1$, $i=1, 2, \dots, k$, where $h(\tilde{p})$ is the reliability of the system when p_i is the probability that the i^{th} subsystem functions. It is well-known that if the system is coherent,

$$0 \leq h(\tilde{p}) \leq 1 ,$$

$$h(0, \dots, 0) = 0, \quad h(1, \dots, 1) = 1 ,$$

and $h(p_1, \dots, p_k)$ is non-decreasing in each variable.

For coherent systems, Buehler's method may be described as follows: The observed outcome (y_1, \dots, y_k) can assume any of $N = \prod_{i=1}^k (n_i + 1)$ values, since $y_i = 0, 1, \dots, n_i$. For convenience, we denote $n_i - y_i$ by x_i , $i=1, 2, \dots, k$.

A partition (A_1, A_2, \dots, A_s) , $s \geq 1$, of the N possible outcomes is said to be a monotonic partition, that is, $A_1 < A_2 < \dots < A_s$ if $(0, 0, \dots, 0) \in A_1$, $(n_1, n_2, \dots, n_k) \in A_s$ and if $\tilde{x}_1 = (x_{11}, \dots, x_{1k})$, $\tilde{x}_2 = (x_{21}, \dots, x_{2k})$ with $x_{1i} \leq x_{2i}$, $i = 1, 2, \dots, k$, then $\tilde{x}_1 \in A_1$ implies $\tilde{x}_2 \in A_j$, $j \geq i$.

Let

$$f(\tilde{x}; \tilde{p}) = p_{\tilde{p}}(\tilde{x} = \tilde{x}) = \prod_{i=1}^k \binom{n_i}{x_i} p_i^{n_i - x_i} q_i^{x_i} = \prod_{i=1}^k \binom{n_i}{y_i} p_i^{y_i} q_i^{n_i - y_i} \quad (2.1)$$

and for $1 \leq n \leq s-1$, let

$$a_n = \inf \left\{ h(p) \mid \sum_{x_i \in A_i, i \leq n} f(\tilde{x}_i; \tilde{p}) = \alpha \right\} \quad (2.2)$$

and $a_s = 0$.

Each such partition may be identified with a function defined on the set of sample outcomes by defining the ordering function $g(\tilde{x})$, where

$$g(\tilde{x}) = n \text{ if } \tilde{x} \in A_n, \quad 1 \leq n \leq s; \quad (2.3)$$

obviously $g(\tilde{x})$ inherits the monotonicity properties of the partition.

Subsequently it will be convenient to use ordering functions $g(\tilde{x})$ such that the range of $g(\tilde{x})$ will be a finite set of real numbers, $r_1 < r_2 < \dots < r_s$. With no loss of generality, we can identify the sets A_i by defining $A_i = \{\tilde{x} \mid g(\tilde{x}) = r_i\}$, $i = 1, 2, \dots, s$. We can now establish the following theorems.

Theorem 2.1. Let \tilde{x} be distributed by (2.1). Then $a_{g(\tilde{x})}$ is a $(1-\alpha)$ lower confidence bound for $h(\tilde{p})$. If $b_{g(\tilde{x})}$ is also a $(1-\alpha)$ lower confidence bound for $h(\tilde{p})$, then $b_i \leq a_i$, $1 \leq i \leq s$.

Proof: Fix \tilde{p} and let $n(\tilde{p})$ be the smallest integer such that

$$P_{\tilde{p}} \left\{ \bar{x} \in \bigcup_{i=1}^{n(\tilde{p})} A_i \right\} \geq \alpha, \quad (2.4)$$

and

$$P_{\tilde{p}} \left\{ \bar{x} \in \bigcup_{i=n(\tilde{p})}^s A_i \right\} \geq 1-\alpha. \quad (2.5)$$

Let

$$D_n = \left\{ \tilde{p} \mid P_{\tilde{p}} \left\{ \bar{x} \in \bigcup_{i=1}^n A_i \right\} \geq \alpha \right\}. \quad (2.6)$$

Then $D_g(\bar{x})$ is a $1-\alpha$ confidence set for \tilde{p} , since

$$P_{\tilde{p}} \left\{ \tilde{p} \in D_g(\bar{x}) \right\} = P_{\tilde{p}} \left\{ g(\bar{x}) \geq n(\tilde{p}) \right\} \geq 1-\alpha. \quad (2.7)$$

This establishes the first part of the conclusion. Further, since $h(\tilde{p})$ is continuous and $0 \leq p_i \leq 1$, the infimum in (2.2) is attained.

Now assume that i_1 is the smallest index such that $b_{i_1} > a_{i_1}$, $1 \leq i_1 \leq s-1$. Then, for some \tilde{p}_o, \tilde{p}_1 ,

$$b_{i_1} > \inf \left\{ h(\tilde{p}) \mid \sum_{x_i \in A_i, i \leq i_1} f(\bar{x}; \tilde{p}) = \alpha \right\} = h(\tilde{p}_o),$$

and

$$\sum_{x_i \in A_i, i \leq i_1} f(\bar{x}; \tilde{p}_1) > \alpha, \quad h(\tilde{p}_1) < b_{i_1}.$$

Therefore

$$P_{\tilde{p}_1} \left\{ h(\tilde{p}_1) < b_{g(\bar{x})} \right\} \geq \sum_{x_i \in A_i, i \leq i_1} f(\bar{x}; \tilde{p}_1) > \alpha,$$

a contradiction.

Remark. Let $d_n = \sup \left\{ 1-h(\tilde{p}) \mid \sum_{x_i \in A_i, i \leq n} f(\bar{x}_i; \tilde{p}) = \alpha \right\}$. Then d_n is a $(1-\alpha)$ upper confidence bound for $1-h(\tilde{p})$, the unreliability.

Let $A = \left\{ \bar{x} \in E_k, 0 \leq x_i \leq a_i, i=1, 2, \dots, k \right\}$ and let $g(\bar{x})$ be continuous on \bar{A} (the closure of A) and strictly increasing in each

variable for \tilde{x} . $g(\tilde{x})$ is to be regarded as an ordering function as described immediately preceding Theorem 2.1. We require the following additional property of $g(\tilde{x})$.

Fix $\tilde{x}_0 \in A$. Let $g(\tilde{x}_0) < g(a_1, 0, \dots, 0) = g_1$. Then $g(y_1, 0, \dots, 0) = g(\tilde{x}_0)$ has a unique solution in y_1 . Proceeding recursively, let $i_1 \leq y_1$ and define $y_2 = y_2(i_1)$ as the solution of $g(\tilde{x}_0) = g(i_1, y_2, 0, \dots, 0)$. For each $1 \leq j \leq k$ and $i_{j-1} \leq y_{j-1}$, $i_{j-2} \leq y_{j-2}, \dots, i_1 \leq y_1$, let $y_j = y_j(i_1, i_2, \dots, i_{j-1})$ be the solution of

$$g(\tilde{x}_0) = g(i_1, i_2, \dots, i_{j-1}, y_j, 0, \dots, 0). \quad (2.8)$$

We require that the equations indicated in (2.8) have unique solutions for each y_j .

Then define

$$F(\tilde{x}_0; \tilde{p}) = \sum_{i_1=0}^{\{y_1\}} \sum_{i_2=0}^{\{y_2\}} \dots \sum_{i_k=0}^{\{y_k\}} f(\tilde{i}; \tilde{p}), \quad (2.9)$$

where, for $j > 1$, $y_j = y_j(i_1, i_2, \dots, i_{j-1})$. Let

$$f^*(\tilde{x}_0; a) = \sup_{h(\tilde{p})=a} F(\tilde{x}_0; \tilde{p}), \quad 0 < a < 1. \quad (2.10)$$

Then we have

Theorem 2.2. If \tilde{x}_0 satisfies $\inf_{0 < a < 1} f^*(\tilde{x}_0; a) = 0$, $\sup_{0 < a < 1} f^*(\tilde{x}_0; a) = 1$

and $f^*(x_0; a)$ is a strictly increasing function of a , and if

$\tilde{x}_0 \in A_n$ where $g(\tilde{x})$ determines (A_1, A_2, \dots, A_n) , and if

$$b = \inf \left\{ h(\tilde{p}) \mid \sum_{x \in A_1, 1 \leq n} f(\tilde{x}_1, \tilde{p}) = a \right\}, \quad (2.11)$$

then we have

$$f^*(\tilde{x}_0; b) = a.$$

Proof: Since the infimum in (2.11) is attained, there is a \tilde{p}_0 such that $b = h(\tilde{p}_0)$ and $F(\tilde{x}_0; \tilde{p}_0) = a$. Then $f^*(\tilde{x}_0, b) \geq a$. If $f^*(\tilde{x}_0, b) > a$, there exists \tilde{p}_a , with $a = h(\tilde{p}_a)$, $a < b$ and $f^*(\tilde{x}_0; a) = a$ contradicting (2.11).

Obviously, the above discussion can easily be modified to obtain upper confidence bounds on the unreliability $1-h(\tilde{p})$ by replacing inf by sup in (2.11) and requiring that $f^*(\tilde{x}_0; a)$ be a strictly decreasing function of a , $0 < a < 1$.

3. Applications to Series Systems

For a series system $h(\tilde{p}) = \prod_{i=1}^k p_i$. Further, throughout this section we assume that $g(\tilde{x})$ satisfies the conditions necessary to insure that the solutions for y_1, \dots, y_k indicated in (2.8) are unique. Then we have the following theorem.

Theorem 3.1. If $h(\tilde{p}) = \prod_{i=1}^k p_i$, then $\inf_{0 < a < 1} f^*(\tilde{x}_0; a) = 0$, $\sup_{0 < a < 1} f^*(\tilde{x}_0, a) = 1$ and $f^*(\tilde{x}_0; a)$ is strictly increasing in a , whenever $\tilde{x}_0 = (x_{01}, \dots, x_{0k})$ satisfies $x_{0j} < n_j$, $j = 1, 2, \dots, k$.

Proof. Since $h(\tilde{p}) = 1$ if and only if $p_i = 1$, $i = 1, 2, \dots, k$, it follows from (2.1) that

$$\lim_{a \rightarrow 1} \sup_{h(\tilde{p})=a} F(\tilde{x}_0; \tilde{p}) = 1.$$

Similarly, $h(\tilde{p}) = 0$ if and only if at least one $p_i = 0$, $i = 1, 2, \dots, k$. Since $F(\tilde{x}_0; \tilde{p}) \leq P_{\tilde{p}}\{X_i < n_i\} = 1 - P_{\tilde{p}}\{X_i = n_i\} = 1 - q_i^{n_i}$, we have

$$\lim_{a \rightarrow 0} \sup_{h(\tilde{p})=a} F(\tilde{x}_0; \tilde{p}) = 0.$$

To show that $f^*(\tilde{x}_0; a)$ is strictly increasing in a , consider

$0 < a < b < 1$ and let $\tilde{p}_a = (p_{a1}, \dots, p_{ak})$ satisfy $f^*(\tilde{x}_0; a) = F(\tilde{x}_0; \tilde{p}_a)$. Similarly, let \tilde{p}_b satisfy $f^*(\tilde{x}_0; b) = F(\tilde{x}_0; \tilde{p}_b)$. Let $I = \{i_1, i_2, \dots, i_r\}$ be any non-empty set of indices such that $p_{ai_j} (\frac{b}{a})^{1/r} < 1$ and let I^c be the remaining indices. Then

$$\left(\prod_{j \in I} p_{ai_j} \left(\frac{b}{a} \right)^{1/r} \right) \prod_{j \in I^c} p_{ai_j} = b . \quad (3.1)$$

From the monotone likelihood ratio property of the binomial distribution,

$$F(\tilde{x}_0; \tilde{p}_a) < F(\tilde{x}_0; \tilde{p}^*) ,$$

where the components of \tilde{p}^* are given by (3.1). Then

$$F(\tilde{x}_0; \tilde{p}^*) \leq \sup_{h(p)=b} F(\tilde{x}_0; \tilde{p}) = F(\tilde{x}_0; \tilde{p}_b) = f^*(x_0; b) .$$

4. Sudakov's Method

Let

$$I_p(r, s) = \frac{1}{B(r, s)} \int_0^p t^{r-1} (1-t)^{s-1} dt .$$

Then if y is an integer, $y < n$, we have

$$\sum_{i=0}^y \binom{n}{i} p^{n-i} q^i = I_p(n-y, y+1) .$$

For $0 \leq y < n$, real, define $u(n, y, a)$ by $a = I_{u(n, y, a)}(n-y, y+1)$.

Thus, for integer values of y , $u(n, y, a)$ is a 100(1-a) percent lower confidence limit for p . Sudakov (1973) showed that for

$$n_1 \leq n_2 \leq \dots \leq n_k \text{ and } g(\tilde{x}) = \prod_{i=1}^k (n_i - x_i) ,$$

$$u(n_1, y_1, a) \leq b \leq u(n_1, [y_1], a) ,$$

$$\text{where } y_1 = n_1 q_0, q_0 = 1 - \prod_{i=1}^k ((n_i - x_{0i})/n_i).$$

$u(n_1, y_1, \alpha)$ is called the Lindstrom-Madden method for determining lower confidence limits for the reliability of series systems (see Lloyd and Lipow (1962)).

Lipow and Riley (1959) used a different ordering function; nevertheless they noted that for "small" n_i , their tabulated values provided good agreement with the results using the Lindstrom-Madden method. For large values of n_i , the tabulated values that they provided are based on the Lindstrom-Madden method. Here we provide a further justification for the Lindstrom-Madden method by establishing that it provides conservative lower confidence limits (i.e. is a lower bound to b defined in (2.9)) using the ordering function $g(\bar{x})$ employed by Sudakov and we also obtain an upper bound for b , thus determining the possible error of the Lindstrom-Madden method.

Sudakov's proof is unnecessarily complicated and contains some incorrect assertions, which nevertheless do not affect the validity of the conclusion. In the Appendix we provide a simpler proof of some auxiliary results needed for the generalization of Sudakov's theorem given below.

Theorem 4.1. Let $g(\bar{x})$ satisfy the hypothesis of Theorem 3.1.

Then,

$$b \leq \min_{1 \leq i \leq k} u(n_i, [y_1^*], \alpha), \quad (4.1)$$

where b is given by (2.11) and $y_1^* = y_1(j_1, j_2, \dots, j_{i-1})$ is evaluated at $j_l = 0$, $l = 1, 2, \dots, i-1$. Note that $y_1 = y_1^*$. If we also have

$$\frac{y_{j+1} - i_j}{n_j - i_j} \geq \frac{y_{j+1}}{n_{j+1}}, \quad j=1, 2, \dots, k-1, \quad (4.2)$$

then

$$u(n_1, y_1, a) \leq b. \quad (4.3)$$

Proof: (4.1) is immediate from (2.11) upon setting $p_j = 1$, $j \neq i$ and solving $F(\tilde{x}_o; 1, \dots, 1, p_i, 1, \dots, 1) = a$. Recall that $n_1 \leq n_2 \leq \dots \leq n_k$ and

$$F(\tilde{x}_o; \tilde{p}) = \sum_{i_1=0}^{\lfloor y_1 \rfloor} b(n_1 - i_1; p_1, n_1) \dots \sum_{i_{k-1}=0}^{\lfloor y_{k-1} \rfloor} b(n_{k-1} - i_{k-1}; p_{k-1}, n_{k-1}) I_{p_k}(n_k - \lfloor y_k \rfloor, \lfloor y_k \rfloor + 1). \quad (4.4)$$

Now, apply Lemmas A1, A2, and A3 to the innermost sum in (4.4), to get

$$\begin{aligned} & \sum_{i_{k-1}=0}^{\lfloor y_{k-1} \rfloor} b(n_{k-1} - i_{k-1}; p_{k-1}, n_{k-1}) I_{p_k}(n_k - \lfloor y_k \rfloor, \lfloor y_k \rfloor + 1) \leq \\ & \sum_{i_{k-1}=0}^{\lfloor y_{k-1} \rfloor} b(n_{k-1} - i_{k-1}; p_{k-1}, n_{k-1}) I_{p_k}(n_k - y_k, y_k + 1) \leq \\ & \sum_{i_{k-1}=0}^{\lfloor y_{k-1} \rfloor} b(n_{k-1} - i_{k-1}; p_{k-1}, n_{k-1}) I_{p_k}(n_{k-1} - y_{k-1}, y_{k-1} - i_{k-1} + 1) \leq \\ & I_{p_{k-1} p_k}(n_{k-1} - y_{k-1}, y_{k-1} + 1). \end{aligned}$$

Repeated applications of the above establish that

$$F(\tilde{x}_o; \tilde{p}) \leq I_k(n_1 - y_1, y_1 + 1). \quad (4.5)$$

(4.3) follows immediately from (4.5), completing the proof.

Remarks. If (4.3) holds and y_1 is an integer, then $b = f(n_1, y_1, \alpha)$.

It has often been suggested (Lloyd and Lipow (1962), Winterbottom (1974), Bolshev and Loginov (1966), Mirny and Solov'yev (1964)) that the confidence level should depend only on n_1 , the smallest sample size. We now provide a numerical illustration to show that the bound in (4.1) may be improved by taking all the n_i 's into consideration.

Let $k=3$, $\alpha=.1$, $\bar{n} = (10, 12, 30)$, $\bar{x}_0 = (0, 3, 0)$. Then for
$$g(\bar{x}) = \prod_{i=1}^3 (n_i - x_i)$$
, $f(n_1, [y_1], \alpha) = .541$, $f(n_2, [y_2], \alpha) = .525$,
 $f(n_3, [y_3], \alpha) = .639$. The use of (4.3) establishes $.500 \leq b \leq .525$.

Note that if $x_{oi} = n_i$, for some i , $1 \leq i \leq k$, then $g(\bar{x}) = 0$ and $b=0$. It seems reasonable to use $b=0$ as the lower confidence limit whenever $x_{oi}=n_i$ for any monotone ordering function satisfying the conditions of Section 2.

We now show that if $g(\bar{x}) = \prod_{i=1}^k (n_i - x_i)$, then (4.2) is satisfied and Theorem 4.1 applies. This result will extend a result due to Winterbottom (1974), who established this fact for particular special cases. In addition, we will also show that (4.2) holds for a number of other ordering functions used in the literature.

Theorem 4.2. Let $g(\bar{x}) = \prod_{i=1}^k (n_i - x_i + \alpha_i)$, where $\alpha_i > 0$ and $n_{i+1}\alpha_i \geq \alpha_{i+1}n_i$, $i=1, 2, \dots, k-1$. Then (4.2) is satisfied.

Proof. If

$$(n_i - y_i + \alpha_i) \prod_{j=i+1}^k (n_j + \alpha_j) = c$$

and

$$(n_i - k_i + \alpha_i)(n_{i+1} - y_{i+1} + \alpha_{i+1}) \prod_{j=i+2}^k (n_j + \alpha_j) = c,$$

then we have

$$(n_i - y_i + \alpha_i)(n_{i+1} + \alpha_{i+1}) = (n_i - k_i + \alpha_i)(n_{i+1} - y_{i+1} + \alpha_{i+1}),$$

establishing

$$\frac{y_i - k_i}{n_i - k_i} = \frac{y_{i+1}}{n_{i+1}} \frac{n_{i+1}(n_i + \alpha_i - k_i)}{(n_{i+1} + \alpha_{i+1})(n_i - k_i)}.$$

Thus (4.2) holds if

$$\frac{n_{i+1}(n_i + \alpha_i - k_i)}{(n_{i+1} + \alpha_{i+1})(n_i - k_i)} \geq 1;$$

this last inequality will be true whenever $n_{i+1}\alpha_i \geq \alpha_{i+1}n_i$. In particular, this is valid when $\alpha_i = 0$, $i = 1, \dots, k$ which is Sudakov's ordering function.

Theorem 4.3. If $g(\bar{x}) = 1 - \sum_{i=1}^k x_i/n_i$, then (4.2) is satisfied.

Proof. If $1 - y_i/n_i = c = 1 - \frac{k_i}{n_i} - \frac{y_{i+1}}{n_{i+1}}$, then

$$\frac{y_i - k_i}{n_i} = \frac{y_{i+1}}{n_{i+1}}$$

or

$$\frac{y_i - k_i}{n_i - k_i} > \frac{y_{i+1}}{n_{i+1}}.$$

This type of ordering function has been employed by Pavlov (1973), for example.

Theorem 4.4. Let $g(\bar{x}) = \sum_{i=1}^k a_i x_i + z_\alpha (a_1^2 x_1)^{\frac{1}{2}}$, where z_α satisfies $1 - \phi(z_\alpha) = \alpha$ and $\phi(x)$ is the standard normal distribution function, $a_1 \geq a_2 \geq \dots \geq a_k$, and $a_i = (n_i \sum_{j=1}^k 1/n_j)^{-1}$. Then $g(\bar{x})$ satisfies (4.2) if and only if

$$(a_j - a_{j+1})y_j \geq (a_j - a_{j+1})z_\alpha^2 + a_j k_j - a_j k_j (z_\alpha^2 a_j + 2c - a_j(y_j + k_j)) . \quad (4.6)$$

Proof: If $g(\bar{x}_0) = c + \sum_{i=1}^{j-1} a_i k_i$, then defining $\sum_{i=1}^{j-1} a_i^2 k_i = c_1$,

$$a_j y_j + z_\alpha (c_1 + a_j^2 y_j)^{\frac{1}{2}} = c \quad (4.7)$$

and

$$a_j k_j + a_{j+1} y_{j+1} + z_\alpha (c_1 + a_j^2 k_j + a_{j+1}^2 y_{j+1})^{\frac{1}{2}} = c . \quad (4.8)$$

Equating the left hand sides of (4.7) and (4.8), we obtain (4.6).

If $k=2$, (4.6) holds for all cases of interest.

If (4.6) holds, then setting

$$1-\alpha = (\Gamma(x))^{-1} \int_0^{f(x, 1-\alpha)} t^{x-1} e^{-t} dt ,$$

a straightforward limiting argument shows that

$$\max_i a_i f([y_i] + 1, 1-\alpha) \leq b \leq a_1 f(y_1 + 1, 1-\alpha) . \quad (4.9)$$

This ordering function has been used by Johns (1976) and b in (4.7) is the value tabulated by Johns for $k=2$. The validity of the lower bound does not depend on (4.6). In Table 1 below, the lower and upper bounds given in (4.9) are tabulated along with the values given by Johns for $\alpha=.1$. These refer to upper confidence limits for the Poisson parameter combinations $a_1 \lambda_1 + a_2 \lambda_2$.

Note in particular that three of the values tabulated by Johns (indicated by asterisks) violate (4.9). Specifically consider 5.24, in which case $[y_1] = 5$, since $g^*(2, 5) = 4.78$, $g^*(5, 0) = 4.72$ and $g^*(6, 0) = 5.48$. Using the Poisson approximation we obtain the value 9.275 for the upper confidence limit to λ for $\alpha=.1$ and thus $a_1 \lambda_1 + a_2 \lambda_2 = 5.56$. Consequently the sup must exceed 5.56. An

alternative approach to the one suggested by Johns for $k \geq 3$ is to simply use $a_1 f(y_1 + 1, 1-a)$ for b.

Table 1

Comparison of Upper and Lower Bounds
With Values Tabulated by Johns for $\alpha=.1$

a_1	x_1	x_2	Lower Bound	Upper Bound	Johns' Tabled Value
.9	7	2	4.79	5.50	5.17
.9	3	0	2.07	2.27	2.16
.75	6	3	6.00	6.65	6.23
.75	12	3	7.90	8.29	7.91
.67	3	3	5.36	5.61	5.33*
.67	15	2	8.71	9.24	8.81
.60	5	2	5.56	5.62	5.24*
.60	7	6	9.24	9.53	9.18*

5. Numerical Examples and Concluding Remarks

Examples 1 and 2 illustrate the method we have described in this paper.

Example 1: Let $H(\tilde{x}) = \prod_{i=1}^k (n_i - x_i)$, $\alpha = .05$, $k = 5$, $\tilde{n} = (20, 30, 40, 25, 60)$, $\tilde{x} = (2, 6, 10, 8, 15)$. Then the 95% upper confidence limit for the failure probability is contained in (.86, .88).

Example 2: Let $H(\tilde{x}) = \prod (n_i - x_i)$, $\alpha = .05$, $k = 2$, $\tilde{n} = (10, 10)$, $\tilde{x} = (3, 2)$. Then the 95% upper confidence limit for the failure probability is contained in (.70, .73). The value given in Lipow and Riley (1959) is .70.

Remarks. In this paper we have showed that the Lindstrom-Madden technique is conservative for ordering functions satisfying (4.2).

Further, if y_1 is an integer, then the Lindstrom-Madden method is exact. We have also relaxed the conditions needed in Winterbottom (1974) and provided an alternative to the method of Johns (1976).

Appendix

The auxiliary results employed in the proof of Theorem 4.1 are provided here.

Lemma A1: $I_y(n-x, x+1)$, $0 \leq y \leq 1$, is a decreasing function of n and an increasing function of x . $I_y(np, nq+1)$, $p+q = 1$, $0 < p < 1$, is an increasing function of q .

Proof: The proof is immediate from the observation that the beta distribution with parameters α and β has monotone likelihood ratio in α and $-\beta$ and that if a probability distribution has monotone likelihood ratio in θ , $F_\theta(x)$ is a decreasing function of θ (Lehmann (1959), p. 68 and p. 74).

Lemma A2: If $\frac{y_i - k_i}{n_i - k_i} > \frac{y_{i+1}}{n_{i+1}}$ and $n_i \leq n_{i+1}$, then

$$I_y(n_i - y_i, y_i - k_i + 1) \geq I_y(n_{i+1} - y_{i+1}, y_{i+1} + 1). \quad (\text{A.1})$$

Proof: Rewriting the left and right hand sides of (A.1) as

$$\begin{aligned} I_y\left[(n_i - k_i)\left(1 - \frac{y_i - k_i}{n_i - k_i}\right), (n_i - k_i)\left(\frac{y_i - k_i}{n_i - k_i}\right) + 1\right] &\geq \\ I_y\left[n_{i+1}\left(1 - \frac{y_{i+1}}{n_{i+1}}\right), n_{i+1}\left(\frac{y_{i+1}}{n_{i+1}}\right) + 1\right], \end{aligned} \quad (\text{A.2})$$

Lemma A1 applies and the conclusion follows.

Lemma A3: Let $y_1 y_2 = y$, $0 \leq y_i \leq 1$, $i = 1, 2$. Then

$$I_{y_1 y_2}(n-x, x+1) \geq \sum_{k=0}^{\lfloor x \rfloor} b(n-k; y_1, n_1) I_{y_2}(n-x, x-k+1). \quad (\text{A.3})$$

Proof:

$$\begin{aligned}
 & \sum_{k=0}^{[x]} \binom{n}{k} y_1^{n-k} (1-y_1)^k \frac{\Gamma(n-k+1)}{\Gamma(n-x)\Gamma(x-k+1)} \int_0^{y_2} t^{n-x-1} (1-t)^{x-k} dt \\
 & = \frac{\Gamma(n+1)}{\Gamma(n-x)} \sum_{k=0}^{[x]} \frac{(1-y_1)^k y_1^{n-k}}{k! \Gamma(x-k+1)} \int_0^{y_1 y_2} \left(\frac{t}{y_1}\right)^{n-x-1} \left(\frac{y_1-t}{y_1}\right)^{x-k} \frac{dt}{y_1} \\
 & = \frac{\Gamma(n+1)}{\Gamma(n-x)} \int_0^{y_1 y_2} \sum_{k=0}^{[x]} \frac{(1-y_1)^k t^{n-x-1} (y_1-t)^{x-k}}{k! \Gamma(x-k+1)} dt .
 \end{aligned}$$

Thus (A.3) will hold whenever

$$\begin{aligned}
 & \frac{\Gamma(n+1)}{\Gamma(n-x)\Gamma(x+1)} \int_0^{y_1 y_2} t^{n-x-1} (1-t)^x dt \geq \\
 & \frac{\Gamma(n+1)}{\Gamma(n-x)} \int_0^{y_1 y_2} \sum_{k=0}^{[x]} \frac{(1-y_1)^k t^{n-x-1} (y_1-t)^{x-k}}{k! \Gamma(x-k+1)} dt
 \end{aligned}$$

or

$$\frac{\Gamma(n+1)}{\Gamma(n-x)\Gamma(x+1)} \int_0^{y_1 y_2} t^{n-x-1} (1-t)^x .$$

$$\left(1 - \sum_{k=0}^{[x]} \frac{\Gamma(x+1)}{k! \Gamma(x-k+1)} \left(\frac{1-y_1}{1-t} \right)^k \left(\frac{y_1-t}{1-t} \right)^{x-k} \right) dt \geq 0 . \quad (A.4)$$

Writing $\frac{y_1-t}{1-t} = \left(1 - \frac{1-y_1}{1-t} \right)$ and noting that $0 \leq t \leq y_1 y_2 \leq 1$, $t < y_1$ and $(1-t) > 1-y_1$, we observe that (A.4) holds and the lemma is proved.

References

- Barlow, Richard E. and Proschan, Frank (1975), Statistical Theory of Reliability and Life Testing, Probability Models, New York: Holt, Rinehart and Winston.
- Birnbaum, Z.W., Esary, James D., and Saunders, Sam C. (1961), "Multicomponent Systems and Their Reliability", Technometrics, 3, 55-77.
- Bol'shev, L.N. and Loginov, E.A. (1966), "Interval Estimates in the Presence of Noise", Theory of Probability and Its Applications, 11, 82-94.
- Buehler, Robert J. (1957), "Confidence Limits for the Product of Two Binomial Parameters", Journal of the American Statistical Association, 52, 482-93.
- Defense Advanced Research Projects Agency (1980), Handbook for the Calculation of Lower Statistical Confidence Bounds on System Reliability.
- Harris, Bernard (1977), "A Survey of Statistical Methods in Systems Reliability Using Bernoulli Sampling of Components", in Theory and Applications of Reliability: With Emphasis on Bayesian and Nonparametric Methods, eds. Chris P. Tsokos and I.N. Shimi, New York: Academic Press.
- Johns, M.V., Jr. (1976), "Confidence Bounds for Highly Reliable Systems", Unpublished technical report, Department of Statistics, Stanford University.
- Lehmann, E.L. (1959), Testing Statistical Hypotheses, New York: John Wiley and Sons.

- Lipow, M. and Riley, J. (1959), "Tables of Upper Confidence Bounds on Failure Probability of 1, 2, and 3 Component Serial Systems", Vols. I and II, Space Technology Laboratories.
- Lloyd, D.K. and Lipow, M. (1962), Reliability: Management, Methods, and Mathematics, Englewood Cliffs: Prentice Hall.
- Mann, Nancy R., Schafer, Roy E., and Singpurwalla, Nozer D. (1974), Methods for Statistical Analysis of Reliability and Life Data, New York: John Wiley and Sons.
- Mirniy, R.A. and Solov'yev, A.D. (1964), "Estimation of the Reliability of a System from the Results of Tests of its Components", Kibernetika Sluzhby Kommunizmy, 2, Energiya, Moscow.
- Pavlov, I.V. (1973), "A Confidence Estimate of System Reliability from Component Testing Results", Izvestiya Akad. Nauk. Tech. Kibernetiki, 3, 52-61.
- Sudakov, R.S. (1974), "On the Question of Interval Estimation of the Index of Reliability of a Sequential System", Engineering Cybernetics, 12, 55-63.
- Winterbottom, Alan (1974), "Lower Limits for Series System Reliability from Binomial Data", Journal of the American Statistical Association, 69, 782-8.

Optimal Upper Confidence Limits for Products of Poisson
Parameters with Applications to the Interval Estimation of the
Failure Probability of Parallel Systems

Bernard Harris* and Andrew P. Soms**

Abstract

The problem of obtaining optimal upper confidence limits for systems of independent parallel components is treated. Exact optimal upper confidence limits are obtained for an arbitrary number of components for specified failure combinations. For a small number of failures, bounds on the upper confidence limits are obtained. For an arbitrary number of failures an approximation is given which is justified numerically and asymptotically. The results of this paper are compared with the results given by Buehler (1957) and some numerical examples are presented.

Key words: Bounds; Optimal confidence limits; Parallel system;
Reliability.

*University of Wisconsin-Madison

**University of Wisconsin-Milwaukee

Research supported by the Office of Naval Research under Contract No. N00014-79-C-0321 and the United States Army under Contract No. DAAG29-80-C-0041.

1. Introduction and Summary

A problem of fundamental interest to practitioners in reliability is the statistical estimation of the reliability of a system using experimental data collected on subsystems. In this paper, the subsystem data available consists of a sequence of Bernoulli trials in which a "one" is recorded if the subsystem functions and a zero is recorded if the subsystem fails. Thus for each of the k subsystems composing the system, the data provided consists of the pair (n_i, Y_i) , $i=1,2,\dots,k$, where Y_i is binomially distributed (n_i, p_i) . We assume that Y_1, Y_2, \dots, Y_k are mutually independent random variables.

The magnitude of interest in this problem is easily evidenced by the extensive literature devoted to it. In this regard, see the survey paper by Harris (1977) and Section 10.4 of the book by Mann, Schafer, and Singpurwalla (1974). In addition, the Defense Advanced Research Projects Agency has recently issued a Handbook for the Calculation of Lower Statistical Confidence Bounds on System Reliability (1980).

Historically, the first significant work on this problem was produced by Buehler (1957). However, Buehler's method as described in that paper is difficult to implement computationally when $k > 2$.

In this paper, we examine the problem of obtaining upper confidence limits for products of Poisson parameters. This problem is studied by means of majorization methods and Schur-convexity, such as described in the book by Marshall and Olkin (1979). A significant application is the determination of confi-

dence limits for the reliability of systems of k parallel subsystems, a fundamental problem in the statistical analysis of reliability.

2. Exact Solutions for Products of Poisson Parameters for Small Failure Combinations

Let $\tilde{\mathbf{x}} = (x_1, x_2, \dots, x_k)$ be independent Poisson random variables with parameters $\lambda_1, \lambda_2, \dots, \lambda_k$, $k \geq 2$, and let $h(\tilde{\lambda}) = \prod_{i=1}^k \lambda_i$. Let

$$g(\tilde{\mathbf{x}}) = \prod_{i=1}^k (x_i + d), \quad 1 < d < 1.5, \quad x_i = 0, 1, \dots \quad (2.1)$$

and denote the ordered points in the range of $g(\tilde{\mathbf{x}})$ by $j_1 < j_2 < \dots < j_m < \dots$. Define

$$A_i = \left\{ \tilde{\mathbf{x}} \mid g(\tilde{\mathbf{x}}) = j_i \right\}. \quad (2.2)$$

Since x_i , $i=1, 2, \dots, k$, takes on non-negative integral values, we regard it as desirable to have d in (2.1) only assume non-integer values. This has the effect of making the partition defined in (2.2) finer than would be the case if d were an integer.

It is easily verified that

$$a_n = \sup \left\{ h(\tilde{\lambda}) \mid \tilde{\mathbf{x}}_i \in A_i, i \leq n \quad f(\tilde{\mathbf{x}}_i; \tilde{\lambda}) = \alpha \right\} \quad (2.3)$$

is a $(1-\alpha)$ upper confidence limit for $h(\tilde{\lambda})$, where

$$f(\tilde{\mathbf{x}}; \tilde{\lambda}) = e^{-\sum_{i=1}^k \lambda_i} \prod_{i=1}^k \frac{\lambda_i^{x_i}}{x_i!}, \quad \lambda_i > 0, \quad x_i = 0, 1, \dots. \quad (2.4)$$

The proof is identical with that given in Harris and Soms (1980).

Note that if $\tilde{\mathbf{x}}$ is fixed as $n_i \rightarrow \infty$, $i=1, 2, \dots, k$, then

$$a_n = \lim_{n \rightarrow \infty} \bar{q} \prod_{i=1}^k n_i \text{ where}$$

$$\bar{q} = \sup \left\{ \prod_{i=1}^k q_i \mid \tilde{x}_i \in A_i, i \leq n \sum_{j=1}^k \binom{n_j}{x_{ij}} p_j^{n_j - x_{ij}} q_i^{x_{ij}} = a \right\}.$$

Thus in practice $a_n / \prod_{i=1}^k n_i$ may be employed as an approximate $(1-\alpha)$ upper confidence limit for $\prod_{i=1}^k q_i$, $q_i = 1 - p_i$. In this sense the methods of this paper can be used as approximations for estimating the reliability of parallel systems when independent binomially distributed data is obtained for each component.

We proceed by showing that $g(\tilde{x})$ is a Schur-concave function and consequently

$$B_{\tilde{x}_o} = \left\{ \tilde{x} \mid g(\tilde{x}) \leq g(\tilde{x}_o) \right\}$$

is a Schur-convex set (see Marshall and Olkin (1974), pp. 1189-90 and Nevius, Proschan and Sethuraman (1977), p. 264). The Schur-concavity of $g(\tilde{x})$ follows immediately by noting that

$$(x_1 - x_2) \left(\frac{\partial g(\tilde{x})}{\partial x_1} - \frac{\partial g(\tilde{x})}{\partial x_2} \right) \leq 0.$$

Define $F(\tilde{x}_o; \tilde{\lambda})$ by

$$F(\tilde{x}_o; \tilde{\lambda}) = \sum_{\tilde{x}_i \in B_{\tilde{x}_o}} f(\tilde{x}_i; \tilde{\lambda}) = P_{\tilde{\lambda}}(B_{\tilde{x}_o}) \quad (2.5)$$

and let

$$u(\tilde{x}_o; a) = \sup_{h(\tilde{\lambda})=a} F(\tilde{x}_o; \tilde{\lambda}), \quad 0 < a < 1. \quad (2.6)$$

Since the Poisson distribution has a monotone likelihood ratio, $u(\tilde{x}_o; a)$ is a strictly decreasing function of a for fixed x_o . Hence for every c , $0 < c < 1$, there is a unique $a(c)$ such that

$$u(\tilde{x}_o; a(c)) = c. \quad (2.7)$$

Consequently, we also have that a_n (see (2.3)) is the solution in

a of

$$u(\tilde{x}_o; a) = a . \quad (2.8)$$

(2.8) is established exactly as in Harris and Soms (1980).

The methodology to be employed is as follows. If $F(\tilde{x}; \tilde{\lambda})$ is a Schur-concave function of $R_i = -\ln \lambda_i$, $i=1, 2, \dots, k$, then it follows that $u(\tilde{x}_o; a) = F(\tilde{x}_o; a^{1/k}\bar{1})$, where $\bar{1} = (1, 1, \dots, 1)$, and then the solution in a of $u(\tilde{x}_o; a) = a$ is an optimal upper confidence limit for $\prod_{i=1}^k \lambda_i$. This will entail verifying (for fixed \tilde{x}_o) that

$$(R_1 - R_2) \left(\frac{\partial F(\tilde{x}_o; \tilde{\lambda})}{\partial R_1} - \frac{\partial F(\tilde{x}_o; \tilde{\lambda})}{\partial R_2} \right) \leq 0 \quad (2.9)$$

(see Marshall and Olkin (1974), p. 1190). Accordingly we have the following theorem.

Theorem 2.1: Let $g(\tilde{x}) = \prod_{i=1}^k (x_i + d)$, $1 < d < 1.5$, $k \geq 3$. Define \tilde{o}_j as the j -vector all of whose components are zeros. Then let

$$x^{(1)} = \tilde{o}_k, \quad x^{(2)} = (1, \tilde{o}_{k-1}), \quad x^{(3)} = (2, \tilde{o}_{k-1}), \quad x^{(4)} = (1, 1, \tilde{o}_{k-2}),$$

$$x^{(5)} = (3, \tilde{o}_{k-1}), \quad x^{(6)} = (4, \tilde{o}_{k-1}), \quad x^{(7)} = (2, 1, \tilde{o}_{k-2}) \text{ and}$$

$$x^{(8)} = (5, \tilde{o}_{k-1}). \quad \text{The set } A_1 \text{ defined by (2.2) is the point } x^{(i)}$$

and the different permutations of its components, $i=1, 2, \dots, 8$.

Further, for $j=1, 2, \dots, 7$, $F(x^{(j)}; \tilde{\lambda})$ is Schur-concave in R_i , $i=1, 2, \dots, k$.

Proof: In the sense of the ordering given by (2.2), obviously $x^{(1)} < x^{(2)} < x^{(3)} < x^{(5)} < x^{(6)} < x^{(8)}$. Trivially, $d(2+d) < (1+d)^2$ and hence $(2+d)d^{k-1} = g(x^{(3)}) < (1+d)^2 d^{k-2} = g(x^{(4)})$. Similarly,

since $1 < d < 1.5$, $(1+d)^2 < d(3+d)$ and hence $g(\tilde{x}^{(4)}) < g(\tilde{x}^{(5)})$. In the same way $g(x^{(6)}) < g(x^{(7)})$, $g(x^{(7)}) < g(x^{(8)})$, $g(x^{(8)}) < g(1, 1, 1, \tilde{o}_{k-3})$ and $g(x^{(8)}) < g(2, 2, \tilde{o}_{k-2})$, establishing the first part of the conclusion.

In order to establish Schur-concavity, we must verify (2.9).

Thus consider

$$F(\tilde{x}^{(l)}; \tilde{\lambda}) = \sum_{\tilde{x}_j \in A_j, j \leq l} \prod_{i=1}^k e^{-\lambda_i} \lambda_i^{x_i^{(j)}} / x_i^{(j)}, \quad (2.10)$$

where $\lambda_i = e^{-R_i}$. Define

$$G(\tilde{x}^{(l)}; \tilde{R}) = \left(\frac{\partial F(\tilde{x}^{(l)}; \tilde{\lambda})}{\partial R_1} - \frac{\partial F(\tilde{x}^{(l)}; \tilde{\lambda})}{\partial R_2} \right) / e^{-\sum_{i=1}^k e^{-R_i}}. \quad (2.11)$$

Letting $\tilde{R} = (R_1, \dots, R_k)$, we obtain

$$G(\tilde{x}^{(1)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}),$$

$$G(\tilde{x}^{(2)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\sum_{i=1}^k e^{-R_i} \right),$$

$$G(\tilde{x}^{(3)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\sum_{i=3}^k e^{-R_i} + \sum_{i=1}^k \frac{e^{-2R_i}}{2} \right),$$

$$G(\tilde{x}^{(4)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\sum_{i=1}^k \frac{e^{-2R_i}}{2} + \sum_{i < j} e^{-R_i - R_j} \right),$$

$$G(\tilde{x}^{(5)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\sum_{i=3}^k \frac{e^{-2R_i}}{2} + \frac{e^{-R_1 - R_2}}{2} \right.$$

$$\left. + \sum_{i < j, (i,j) \neq (1,2)}^k e^{-R_i - R_j} + \sum_{i=1}^k \frac{e^{-3R_i}}{3!} \right),$$

$$G(\tilde{x}^{(6)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\sum_{i=3}^k \frac{e^{-2R_i}}{2} + \frac{e^{-R_1 - R_2}}{2} + \sum_{i < j, (i,j) \neq (1,2)}^k e^{-R_i - R_j} \right.$$

$$\left. + \sum_{i=3}^k \frac{e^{-3R_i}}{3!} - \frac{e^{-2R_1 - R_2} + e^{-R_1 - 2R_2}}{3!} + \sum_{i=1}^k \frac{e^{-4R_i}}{4!} \right),$$

and

$$G(\tilde{x}^{(7)}; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\frac{e^{-2R_1-R_2}}{3} + \frac{e^{-R_1-2R_2}}{3} + \sum_{3 \leq i \leq j} e^{-R_i-R_j} \right) \\ + \sum_{i=3}^k \frac{e^{-3R_i}}{3!} + \sum_{i=1}^k \frac{e^{-4R_i}}{4!} + \sum_{i \neq j, (i,j) \neq (1,2) \text{ or } (2,1)} \frac{e^{-2R_i-R_j}}{2!}.$$

Now $R_1 > R_2$ implies $e^{-R_1} < e^{-R_2}$ and thus $(e^{-R_1} - e^{-R_2})$ and $(e^{-R_1-2R_2})$ have opposite signs. Hence it follows that $F(\tilde{x}^{(i)}; \tilde{\lambda})$, $i=1,2,3, 4,5,7$ is Schur-concave in R_1 . The verification that $F(\tilde{x}^{(6)}; \tilde{\lambda})$ is Schur-concave may be accomplished by letting $k=2$, $e^{-R_2} = ce^{-R_1}$, $c>0$, and examining the discriminant.

To show that (2.9) need not be positive for all \tilde{x}_o , consider $k=2$ and $\tilde{x}_o = (7,0)$. Then

$$\tilde{G}(\tilde{x}_o; \tilde{R}) = (e^{-R_1} - e^{-R_2}) \left(\frac{e^{-6R_1} + e^{-6R_2}}{6!} \right. \\ \left. - \frac{e^{-4R_1-R_2} + e^{-3R_1-2R_2} + e^{-2R_1-3R_2} + e^{-R_1-4R_2}}{5!} \right. \\ \left. - \frac{e^{-3R_1-R_2} + e^{-2R_1-2R_2} + e^{-R_1-3R_2}}{4!} + \frac{e^{-2R_1-R_2} + e^{-R_1-2R_2}}{3!} \right),$$

and this is Schur-convex near $e^{-R_1} = e^{-R_2} = 4$.

Buehler provided an extensive discussion of this problem for the ordering function determined by the product of the upper confidence limits for the individual components. In particular, he provided some numerical tabulations for $k=2$. Asymptotically Buehler's ordering function is given by

$$g_B(\tilde{x}) = \prod_{i=1}^k (x_i + z_{\alpha}, x_i^{1/2}),$$

where $\alpha' = 1 - (1-\alpha)^{1/k}$, z_{α} satisfies $\Phi(z_{\alpha}) = 1-\alpha$ and $\Phi(x)$ is the standard normal cumulative distribution function. It is easy to see that $g_B(\tilde{x})$ is Schur-concave (see, e.g., Marshall and Olkin (1974), p. 1191).

3. Bounds on Confidence Limits

In this section we employ majorization techniques described in Proschan and Sethuraman (1977) and Nevius, Proschan and Sethuraman (1977) to obtain bounds for a_n . Throughout this section we assume only that the ordering function $g(\bar{x})$ is strictly increasing in each component and Schur-concave and thus the set $B_{\bar{x}_0}$ will be Schur-convex (see the discussion immediately preceding (2.5)).

In order to proceed, we need the preliminary results established below.

Theorem 3.1: Let c and a be given with $c > ka^{1/k}$ and consider the set $A(a,c)$ of vectors $\tilde{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_k)$, $\lambda_i \geq 0$, such that

$$\prod_{i=1}^k \lambda_i = a \quad \text{and} \quad \sum_{i=1}^k \lambda_i = c. \quad (3.1)$$

Let $s_j = \max_{\tilde{\lambda} \in A(a,c)} \sum_{i=1}^j \lambda_i$. Then there is a unique $\lambda^* \in A(a,c)$ of the form $\lambda_i = M_j$, $1 \leq i \leq j$, $\lambda_i = m_j$, $j+1 \leq i \leq k$, $M_j > m_j$, $s_j = jM_j$.

Proof: The condition $c > ka^{1/k}$ is a consequence of the arithmetic-geometric mean inequality and insures that $A(a,c)$ is non-trivial for $k \geq 3$. If $k=2$, there is only one solution of (3.1) with $\lambda_1 > \lambda_2$, and hence the Theorem is trivially true. Consequently, suppose $k \geq 3$. Then for fixed j , (3.1) requires that any solution of the required type satisfy

$$jM_j + (k-j)m_j = c, \quad M_j^j m_j^{k-j} = a$$

and hence setting $m_j = (c - jM_j)/(k-j)$, we consider

$$f_j(M) = M^j \left[(c-jM)/(k-j) \right]^{k-j}, \quad 1 \leq j \leq k-1, \quad 0 \leq M \leq c/j. \quad (3.2)$$

Note that $f_j(0) = f_j(c/j) = 0$, and

$$f'_j(M) = (c-Mk) \left(\frac{jM^{j-1}}{k-j} \right) \left(\frac{c-jM}{k-j} \right)^{k-j-1}. \quad (3.3)$$

Thus, $f_j(M)$ is increasing for $0 \leq M < c/k$ and decreasing otherwise, further $f_j(c/k) = (c/k)^k > a$. Hence there is exactly one solution M_j of $f_j(M)=a$ with $M_j > c/k$, and therefore $M_j > m_j$.

Now assume that for some j , $1 \leq j \leq k-1$, the vector

$\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_k^*)$ with $\sum_{i=1}^j \lambda_i^* = s_j$ is not of the form $(\underbrace{M_j, \dots, M_j}_j, \underbrace{m_j, \dots, m_j}_{k-j})$. Then let $\bar{\lambda}_{1j} = s_j/j$ and $\bar{\lambda}_{2j} = (c-s_j)/(k-j)$.

Define $\lambda'_j = (\lambda'_{j1}, \lambda'_{j2}, \dots, \lambda'_{jk})$ by $\lambda'_{ji} = \bar{\lambda}_{1j}$, $1 \leq i \leq j$, $\lambda'_{ji} = \bar{\lambda}_{2j}$, $j+1 \leq i \leq k$. Since the geometric mean of a set of positive numbers whose sum is fixed is a maximum when they are all equal, we have $\prod_{i=1}^k \lambda'_{ji} > a$. Now λ'_j is of the required form, however, from (3.2) and (3.3), $\prod_{i=1}^k \lambda'_{ji} > a$ implies that there is another solution of the required form with $\lambda_i > s_j/j$, $1 \leq i \leq j$, contradicting the maximality of s_j .

From (2.5) and (2.6), we can write

$$u(\tilde{x}_0; a) = \sup_{\substack{k \\ \prod_{i=1}^k \lambda_i = a}} P_{\lambda}(\tilde{B}_{\tilde{x}_0}) = \sup_c \sup_{\substack{k \\ \sum_{i=1}^k \lambda_i = c, \prod_{i=1}^k \lambda_i = a}} P_{\lambda}(\tilde{B}_{\tilde{x}_0}). \quad (3.4)$$

We state now the main result of this section, using Theorem 3.1.

Theorem 3.2: Let $v_1 = M_1$, $v_i = iM_i - (i-1)M_{i-1}$, $2 \leq i \leq k-1$, $v_k = c - \sum_{i=1}^{k-1} v_i$, where M_i is specified by Theorem 3.1. Then

$$u(\tilde{x}_o; a) \leq \sup_c P_{\tilde{v}}(B_{\tilde{x}_o}) . \quad (3.5)$$

Proof: Since $\sum_{i=1}^k v_i = s_j$, $1 \leq j \leq k-1$, $\sum_{i=1}^k v_i = c$, \tilde{v} majorizes every λ with $\sum_{i=1}^k \lambda_i = c$, $\prod_{i=1}^k \lambda_i = a$ (Theorem 3.1). Then (3.5) follows, since if $\tilde{\lambda}_1$ majorizes $\tilde{\lambda}_2$, then for any Schur-convex set A ,

$$P_{\tilde{\lambda}_1}(A) \geq P_{\tilde{\lambda}_2}(A) \quad (\text{Proschan and Sethuraman (1977) and Nevius, Proschan and Sethuraman (1971), p. 264 and pp. 267-9}).$$

The vector \tilde{v} may be interpreted as the best vector that majorizes all vectors $\tilde{\lambda}$ such that $\sum_{i=1}^k \lambda_i = c$ and $\prod_{i=1}^k \lambda_i = a$. More specifically, there is no vector $\tilde{w} \neq \tilde{v}$ such that \tilde{v} majorizes \tilde{w} and \tilde{w} majorizes all $\tilde{\lambda}$ satisfying the two conditions given above.

The following is a suggested method for employing Theorem 3.2. Find a_d such that

$$a = F(\tilde{x}_o; a_d^{1/k} \tilde{1}) .$$

Next calculate the smallest a , say a_m , such that $\sup_c P_{\tilde{v}}(B_{\tilde{x}_o}) \leq a$. If $a_m = a_d$, this is the exact solution. Otherwise $a_d < a_m$ and $\sup_c P_{\tilde{v}}(B_{\tilde{x}_o}) < a$ (here $a = a_m$) and the solution a_n satisfies $a_d \leq a_n \leq a_m$. The vector \tilde{v} may be calculated by any of a variety of numerical techniques. In the numerical examples presented here, interval bisection was employed.

Example 1: Let $k = 5$, $a = 25$, $c = 15$. Then the 4 vectors $\tilde{\lambda}_1$, $\tilde{\lambda}_2$, $\tilde{\lambda}_3$ and $\tilde{\lambda}_4$ of Theorem 3.1 are

$$\tilde{\lambda}_1 = (9.9660, 1.2585, 1.2585, 1.2585, 1.2585)$$

$$\tilde{\lambda}_2 = (6.2004, 6.2004, .8664, .8664, .8664)$$

$$\tilde{\lambda}_3 = (4.6696, 4.6696, 4.6696, .4955, .4955)$$

$$\text{and } \tilde{\lambda}_4 = (3.7172, 3.7172, 3.7172, 3.7172, .1309) ,$$

from which \tilde{v} is determined to be

$$\tilde{v} = (9.9660, 2.4349, 1.6079, .8601, .1309) .$$

Note that in the above example $v_1 \geq v_2 \geq \dots \geq v_k$. This in fact is always true, as the following theorem establishes.

Theorem 3.3: For \tilde{v} defined by Theorem 3.2, we have $v_1 \geq v_2 \geq \dots \geq v_k$.

Proof: It follows immediately that $v_1 \geq v_2$, since $M_1 \geq M_2$. Consider therefore v_j , $j \geq 2$. $v_j \geq v_{j+1}$, $j=2, 3, \dots, k-1$ holds if and only if

$$jM_j - (j-1)M_{j-1} \geq (j+1)M_{j+1} - jM_j$$

or

$$jM_j \geq ((j+1)M_{j+1} + (j-1)M_{j-1})/2 ,$$

where $M_k = c/k$ (satisfying the condition $S_k = c = kM_k$ of Theorem 3.1).

Let $\tilde{\lambda}_{A_j} = (1-\alpha_j)\tilde{\lambda}_{j-1} + \alpha_j\tilde{\lambda}_{j+1}$, $j=2, 3, \dots, k-1$, where $\alpha_j = (1/2) + (1/(2j))$ and

$$\tilde{\lambda}_j = (\lambda_{j1}, \lambda_{j2}, \dots, \lambda_{jk})$$

and

$$\lambda_{ji} = M_j, \quad 1 \leq i \leq j, \quad \lambda_{ji} = m_j, \quad j+1 \leq i \leq k .$$

It follows that

$$\prod_{i=1}^k \lambda_{A_j, i} \geq a ,$$

since $\sum_{i=1}^k \ln x_i$ is a concave function of x_1, \dots, x_k . Now let $\lambda_{B_j, i} = \left(\sum_{i=1}^j \lambda_{A_j, i} \right) / j$, $i=1, 2, \dots, j$, $\lambda_{B_j, i} = \left(\sum_{i=j+1}^k \lambda_{A_j, i} \right) / (k-j)$, $i=j+1, \dots, k$. Then

$$\prod_{i=1}^j \lambda_{B_j, i} \geq \prod_{i=1}^j \lambda_{A_j, i} , \quad j=1, 2, \dots, k-1 .$$

Thus, using the properties of M_i in Theorem 3.1,

$$jM_j \geq (j-1)[(1-\alpha_j)M_{j-1} + \alpha_j M_{j+1}] + (1-\alpha_j)m_{j-1} + \alpha_j m_{j+1},$$

yielding

$$jM_j \geq ((j+1)M_{j+1} + (j-1)M_{j-1})/2 + (1-\alpha_j)m_{j-1},$$

which establishes the theorem.

To illustrate the techniques of this paper, we compare numerical values obtained by the above method with those given in the examples from Mann, Schafer and Singpurwalla (1974, p. 505). From now on we assume $d = 1.1$.

Example 2: For $\tilde{x}_0 = (1, 2, 1)$ we obtain $a_d = a_n = 20.56$ for $\alpha = .10$. In Mann, Schafer and Singpurwalla, an AO non-randomized confidence bound of 20.7 is obtained.

Example 3: Let $\tilde{x}_0 = (2, 3, 5)$, $\alpha = .10$. Then we obtain $a_d = 135.46$. A summary of computer calculations which establishes $135.46 \leq a_n \leq 142.46$ is given below in Table 1. With the exception of the likelihood-ratio value of 133 and the AO non-randomized confidence bound of 129, all the other confidence bounds given in Mann, Schafer and Singpurwalla exceed the upper bound of 142.46. For $k=3$ it is possible to do a direct computer tabulation of $u(\tilde{x}_0; \alpha)$. This gives $a_n = 135.46$, the diagonal value. (See Table 1 on the following page.)

The two examples below are for four and five component systems for which there are no comparable numerical examples available.

Example 4: Let $\tilde{x}_0 = (2, 2, 2, 2)$ and $\alpha = .10$. Then $a_d = a_n = 150.63$.

1. Summary of Calculations Used to
Obtain the Upper Bound for a_n in Example 3

<u>a</u>		\tilde{v}		$\sup_c P_{\tilde{v}}(B_{x_0})$
135.46	13.0680	4.7283	1.7108	.1101
136.46	13.0867	4.7409	1.7173	.1086
137.46	13.1053	4.7532	1.7240	.1071
138.46	13.1239	4.7656	1.7305	.1057
139.46	13.1423	4.7780	1.7370	.1042
140.46	13.1607	4.7902	1.7435	.1028
141.46	13.1789	4.8024	1.7500	.1014
142.46	13.3299	4.8057	1.7325	.9999

Example 5: Let $\bar{x}_o = (2, 2, 2, 2, 2)$ and $u = .10$. Then $a_d = 429.69$.

A summary of the computer calculations which establish $429.69 \leq a_n \leq 435.69$ is provided in Table 2. (See the following page for this table.)

As $a_d^{1/k}$ increases, the difference between a_d and a_n becomes wider. Thus the techniques of Section 3 are more useful for small x_o , or equivalently, small $a_d^{1/k}$. For example, for $\bar{x}_o = (5, 5, 5)$, $a_d = 387.18$, and it is not practical to compute a_m because it is much bigger than a_d . However, direct tabulation of $u(\bar{x}_o; a)$ reveals once more that $a_d = u(\bar{x}_o; a)$. A justification of why $a_d \approx u(\bar{x}_o; a)$ for large $a_d^{1/k}$ is given in the Appendix. This, together with the results of Section 2, suggests very strongly that for all practical purposes $a_d = a_n$.

Remarks: Note that Tables 1 and 2 are virtually linear in their behavior in the neighborhood of the solution. This suggests that solutions are obtainable by interpolation and then one should subject them to verification.

The calculations described above utilized two short FORTRAN programs for 2-10 components. Listings are obtainable from the authors.

4. Comparisons with Buehler's Tables

In order to provide an illustration of the performance of $g(\bar{x}_o) = \prod_{i=1}^k (x_i + d)$, $1 < d < 1.5$, when compared with the tables given by Buehler (1957), we chose $d=1.1$, $k=2$. For $k=2$, the values of a_n and a_d coincided for both the ordering based on $g(\bar{x})$ and Buehler's ordering and further were for all practical purposes

2. Summary of Calculations Used to
Obtain the Upper Bound for a_n in Example 5

<u>a</u>	$\sup_c P_v^-(B_{x_0}^-)$
429.69	.1016
430.69	.1013
431.69	.1010
432.69	.1007
433.69	.1004
434.69	.1001
435.69	.0998

equal for the two different orderings.

In Table 3 we give Buehler's upper confidence limit, Buehler's diagonal value and the exact upper confidence limit and diagonal value corresponding to g , denoting them by a_{nB} , a_{dB} , a_{ng} and a_{dg} , respectively. These values are provided for all failure combinations from $(0,0)$ to $(5,5)$ for $\alpha=.1$.

See next page for Table 3

An examination of Table 3 shows that differences between the four alternatives presented are small for the specific example ($k=2$, $\alpha=.1$).

5. Concluding Remarks

In this paper a procedure for obtaining bounds on an optimal upper confidence limit for the failure probability of a parallel system is given. The procedure employs the theory of majorization and is valid for an arbitrary number of components and gives the exact answer or narrow bounds when the observed number of failures is small for each component. In addition, numerical and asymptotic justification is given for using a_d as an approximation to a_n . Tables of a_d are in preparation for moderate numbers of failures for 3, 4 and 5 components and will be available in the near future.

3. Comparison of Exact and Diagonal Buchler's Values,
 a_{nB} and a_{dB} , Respectively, with the Exact and Diagonal Values
 a_{ng} and a_{dg} , Respectively, Corresponding to $g(\bar{x})$

x_1	x_2	a_{nB}	a_{dB}	a_{ng}	a_{dg}
5	5	60.7	60.70	60.70	60.70
5	4	51.8	51.89	51.89	51.89
5	3	41.2	41.21	41.22	41.21
5	2	31.9	31.91	31.91	31.90
5	1	23.3	23.34	23.34	23.34
5	0	12.3	12.32	12.32	12.32
4	4	44.3	44.40	44.40	44.40
4	3	35.7	35.73	35.74	35.73
4	2	27.2	27.23	27.23	27.23
4	1	18.8	18.77	18.77	18.76
4	0	9.05	9.05	9.05	9.05
3	3	28.9	28.89	28.89	28.89
3	2	22.0	22.04	22.04	22.03
3	1	15.1	15.08	15.08	15.08
3	0	8.24	8.24	8.24	8.24
2	2	16.8	16.80	16.80	16.79
2	1	11.8	11.85	11.85	11.85
2	0	5.59	5.59	5.59	5.59
1	1	7.09	7.08	7.08	7.08
1	0	3.86	3.78	3.78	3.78
0	0	1.33	1.33	1.33	1.33

Appendix

Theorem A1: Let x_{1i} , $1 \leq i \leq k$, be independent identically distributed normal random variables with means λ and variances λ . Let x_{2i} , $1 \leq i \leq k$, be independent normally distributed random variables with means τ_i and variances τ_i , where $\tau_i = \lambda + O(\lambda^c)$,

$c < 1$, as $\lambda \rightarrow \infty$, and $\prod_{i=1}^k \tau_i = \lambda^k$. Let β be given, $0 < \beta < 1$, let a be a specified positive real number, let $z_1 = \prod_{j=1}^k (x_{1j} + a)$,
 $z_2 = \prod_{j=1}^k (x_{2j} + a)$ and let $d(\lambda)$ satisfy

$$P[z_1 \leq d(\lambda)] = \beta. \quad (\text{A.1})$$

Then as $\lambda \rightarrow \infty$,

$$\beta - P[z_2 \leq d(\lambda)] = \begin{cases} O[(\ln \lambda)^{1.5} \lambda^{-1}] & , c \leq 0, \\ O[\lambda^{c-1}] & , 0 < c < 1. \end{cases} \quad (\text{A.2})$$

Proof: Throughout, let ϕ and Φ denote the density and distribution function of the standard normal. Clearly,

$$P[z_1 \leq d(\lambda)] - P[z_2 \leq d(\lambda)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (f_1(\tilde{x}) - f_2(\tilde{x})) d\tilde{x},$$

$$(\tilde{x}: \prod_{j=1}^k (x_j + a) \leq d(\lambda)) \quad (\text{A.3})$$

where $\tilde{x} = (x_1, x_2, \dots, x_k)$, f_1 is the probability density function of $x_{11}, x_{12}, \dots, x_{1k}$ and f_2 is that of $x_{21}, x_{22}, \dots, x_{2k}$. Now

$$P[X_{1j} \geq -a, j = 1, 2, \dots, k] \geq \left(1 - \frac{\lambda^{1/2}}{(\lambda+a)} \phi\left(\frac{\lambda+a}{\lambda^{1/2}}\right)\right)^k \quad (\text{A.4})$$

and

$$P[X_{2j} \geq -a, j = 1, 2, \dots, k] \geq \prod_{j=1}^k \left(1 - \frac{\tau_j^{1/2}}{\tau_j+a} \phi\left(\frac{\lambda_j+a}{\tau_j^{1/2}}\right)\right). \quad (\text{A.5})$$

Consequently, for λ sufficiently large, there exists a constant $m > 0$ such that

$$P[X_{ij} \geq -a, j = 1, 2, \dots, k] \geq 1 - e^{-m\lambda}, \quad i = 1, 2. \quad (\text{A.6})$$

Then, for $i = 1, 2$,

$$\begin{aligned} P[z_i \leq d(\lambda)] &= P[z_i \leq d(\lambda), X_{ij} \geq -a, j = 1, 2, \dots, k] \\ &\quad + P[z_i \leq d(\lambda), \bigcup_{j=1}^k (X_{ij} < -a)], \end{aligned}$$

and therefore

$$P[z_i \leq d(\lambda)] - P[z_i \leq d(\lambda), X_{ij} \geq -a, j = 1, 2, \dots, k] \leq e^{-m\lambda}. \quad (\text{A.7})$$

Next, we calculate

$$P[z_1 \leq d(\lambda), X_{1j} \geq -a, j = 1, 2, \dots, k] - P[z_2 \leq d(\lambda), X_{2j} \geq -a, j = 1, 2, \dots, k]$$

Now

$$\begin{aligned} P[z_1 \leq d(\lambda), X_{1j} \geq -a, j = 1, 2, \dots, k | X_{1j} = x_j, j = 2, 3, \dots, k] \\ &= P\left[X_{11} \leq \frac{d(\lambda)}{\sum_{j=2}^k (x_j+a)} - a\right] \\ &= \Phi\left(\frac{\frac{d(\lambda)}{\sum_{j=2}^k (x_j+a)} - a - \lambda}{\lambda^{1/2}}\right) = \frac{\Phi(b_-)}{\lambda}. \end{aligned} \quad (\text{A.8})$$

Therefore

$$P\{z_1 \leq d(\lambda), x_{1j} \geq -a, j = 1, 2, \dots, k\} \quad (\text{A.9})$$

$$= \int_{-a}^{\infty} \int_{-a}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(b_{-}) g_1(x_2, x_3, \dots, x_k) dx_2 dx_3 \cdots dx_k ,$$

where $g_1(x_2, x_3, \dots, x_k)$ is the probability density function of $x_{12}, x_{13}, \dots, x_{1k}$. From (A.6), we have that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(b_{-}) g_1(x_2, x_3, \dots, x_k) dx_2 dx_3 \cdots dx_k \quad (\text{A.10})$$

$$- \int_{-a}^{\infty} \int_{-a}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(b_{-}) g_1(x_2, x_3, \dots, x_k) dx_2 dx_3 \cdots dx_k \leq e^{-m\lambda}.$$

Hence we will estimate the first expression on the left hand side of (A.10). Similarly, for z_2 we will consider

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi \left[\left(\frac{d(\lambda)}{\prod_{j=2}^k (x_j + \tau_j)} - a - \tau_1 \right) / \tau_1^{1/2} \right] g_2(x_2, x_3, \dots, x_k) dx_2 dx_3 \cdots dx_k , \quad (\text{A.11})$$

where $g_2(x_2, x_3, \dots, x_k)$ is the probability density function of $x_{22}, x_{23}, \dots, x_{2k}$. In the first integral in (A.10), let

$$(y_i - \lambda) / \lambda^{1/2} = u_i \text{ and in (A.11) let } (y_i - \tau_i) / \tau_i^{1/2} = u_i ,$$

$i = 2, 3, \dots, k$, obtaining

$$\begin{aligned}
& \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Phi \left(\left| \frac{d(\lambda)}{\prod_{j=2}^k (x_j + a)} - a - \lambda \right| / \lambda^{1/2} \right) g_1(x_2, x_3, \dots, x_k) dx_2 dx_3 \cdots dx_k \\
& - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Phi \left(\left| \frac{d(\lambda)}{\prod_{j=2}^k (x_j + a)} - a - \tau_1 \right| / \tau_1^{1/2} \right) g_2(x_2, x_3, \dots, x_k) dx_2 dx_3 \cdots dx_k \\
& = \int_{-M}^M \int_{-M}^M \cdots \int_{-M}^M \left\{ \Phi \left(\left| \frac{d(\lambda)}{\prod_{j=2}^k (\lambda^{1/2} x_j + \lambda + a)} - a - \lambda \right| / \lambda^{1/2} \right) \right. \\
& \quad \left. - \Phi \left(\left| \frac{d(\lambda)}{\prod_{j=2}^k (\tau_1^{1/2} x_j + \tau_1 + a)} - a - \tau_1 \right| / \tau_1^{1/2} \right) \right\} \left(\prod_{j=2}^k \Phi(x_j) \right) dx_2 dx_3 \cdots dx_k \\
& + R_M,
\end{aligned} \tag{A.13}$$

where $M = (2 \ln \lambda)^{1/2}$ and $R_M \leq 4 \frac{(k-1)e^{-M^2/2}}{(2\pi)^{1/2} M} = O(\lambda^{-1})$.

Using $d(\lambda) = \lambda^k - k_d(\lambda)\lambda^{k-1/2}$, $k_d(\lambda) = O(1)$,

$$\left| \frac{d(\lambda)}{\prod_{j=2}^k (\lambda^{1/2} x_j + \lambda + a)} - a - \lambda \right| / \lambda^{1/2} = (\lambda^{1/2} - k_d(\lambda)) \prod_{j=2}^k (1 + x_j \lambda^{-1/2} + a \lambda^{-1})^{-1} - \lambda^{1/2} - a \lambda^{-1/2}.$$

Since $|x_1| \leq M$, we have

$$(1 + x_1 \lambda^{-1/2} + a \lambda^{-1})^{-1} = 1 - x_1 \lambda^{-1/2} + (-a + x_1^2) \lambda^{-1} + O((\ln \lambda)^{1.5} \lambda^{-1.5}).$$

Thus

$$\begin{aligned}
 & (\lambda^{1/2} - k_d(\lambda)) \prod_{j=2}^k (1 + x_j \lambda^{-1/2} + a \lambda^{-1})^{-1} - \lambda^{1/2} - a \lambda^{-1/2} \\
 & = - \sum_{i=2}^k x_i - k_d(\lambda) - k a \lambda^{-1/2} + k_d(\lambda) \left(\sum_{i=2}^k x_i \right) \lambda^{-1/2} + \left(\sum_{i=2}^k x_i^2 \right) \lambda^{-1/2} \\
 & \quad + \left(\sum_{2 \leq i < j} x_i x_j \right) \lambda^{-1/2} + O((\ln \lambda)^{1.5} \lambda^{-1}). \tag{A.14}
 \end{aligned}$$

Similarly, using $\tau_1 = \lambda^k / \prod_{j=2}^k \tau_j$, $\tau_j / \lambda = 1 + O(\lambda^{c-1})$,

$(\tau_j / \lambda)^{1/2} = 1 + O(\lambda^{c-1})$, $j = 1, 2, \dots, k$, $|x_i| \leq M$, we have

$$\begin{aligned}
 & \left(\frac{d(\lambda)}{\prod_{j=2}^k (\tau_j^{1/2} x_j + \tau_j + a)} - a - \tau_1 \right) / \tau_1^{1/2} \\
 & = (\tau_1 / \lambda_1)^{1/2} \left[(\lambda^{1/2} - k_d(\lambda)) \prod_{j=2}^k (1 + x_j \tau_j^{-1/2} + a \tau_j^{-1})^{-1} - \lambda^{1/2} \right] - a \tau_1^{-1/2} \\
 & = - \sum_{i=2}^k x_i - k_d(\lambda) - k a \lambda^{-1/2} + k_d(\lambda) \left(\sum_{i=2}^k x_i \right) \lambda^{-1/2} + \left(\sum_{i=2}^k x_i^2 \right) \lambda^{-1/2} \\
 & \quad + \left(\sum_{2 \leq i < j} x_i x_j \right) \lambda^{-1/2} + O(\lambda^{c-1}) + O((\ln \lambda)^{1.5} \lambda^{-1}). \tag{A.15}
 \end{aligned}$$

Combining (A.14) and (A.15) with (A.7), (A.9), (A.10) and (A.11) establishes the theorem.

For $c < \frac{1}{2}$ standard weak convergence arguments show that

$$\lim_{\lambda \rightarrow \infty} (\beta - P[Z_2 \leq d(\lambda)]) = 0.$$

In this case Theorem A1 provides additional information by specifying the rate of convergence.

By standardizing the first expression in (A.13) and applying the dominated convergence theorem the following result can be obtained.

Theorem A2: Let x_{1i} , $1 \leq i \leq k$, be independent identically distributed normal random variables with means λ and variances λ . Let x_{2i} , $1 \leq i \leq k$, be independent normally distributed random variables with means τ_i and variances τ_i , where $\tau_i = \lambda + O(\lambda^c)$, $c < 1$ and let β , z_1, z_2 and $d(\lambda)$ be specified as in Theorem A1.

Then

$$\lim_{\lambda \rightarrow \infty} (\beta - P[z_2 \leq d(\lambda)]) = 0.$$

References

- Buehler, Robert J. (1957), "Confidence Limits for the Product of Two Binomial Parameters," Journal of the American Statistical Association, 52, 482-93.
- Harris, Bernard (1977), "A Survey of Statistical Methods in Systems Reliability Using Bernoulli Sampling of Components," in Theory and Applications of Reliability: With Emphasis on Bayesian and Nonparametric Methods, eds. Chris P. Tsokos and I.N. Shimi, New York: Academic Press.
- Harris, Bernard and Soms, Andrew P. (1980), "Bounds for Optimal Confidence Limits for Series Systems," University of Wisconsin-Madison Department of Statistics technical report.
- Mann, Nancy R., Schafer, Ray E., and Singpurwalla, Nozer D. (1974), Methods for Statistical Analysis of Reliability and Life Data, New York: John Wiley and Sons.
- Marshall, Albert W. and Olkin, Ingram (1974), "Majorization in Multivariate Distributions," The Annals of Statistics, 2, 1189-1200.
- _____, and Olkin, Ingram (1979), Inequalities: Theory of Majorization and Its Applications, New York: Academic Press.
- Proschan, Frank and Sethuraman, Jayaram (1977), "Schur Functions in Statistics, I. The Preservation Theorem," The Annals of Statistics, 5, 256-262.
- Nevius, Edward S., Proschan, Frank, and Sethuraman, Jayaram (1977), "Schur Functions in Statistics, II. Stochastic Majorization," The Annals of Statistics, 5, 263-273.

RELIABILITY BASED SAFETY FACTORS
FOR CONCRETE STRUCTURES SLIDING STABILITY

Paul F. Mlakar
Structures Laboratory
U. S. Army Engineer Waterways Experiment Station
Vicksburg, MS

SYNOPSIS. A safety factor criteria for sliding stability, which explicitly quantifies load and resistance uncertainties, is developed through the probabilistic analysis of a simplified illustrative problem. Certain points of this analytical development are noted to warrant further study.

I. INTRODUCTION. A concrete gravity structure, such as a dam, will fail by sliding along a critical foundation surface when the resultant effect of externally applied forces exceeds the total resistance developed along this surface. (US Army Corps of Engineers, 1958). (US Bureau of Reclamation, 1976). A practical means of assuring the safety of such structures is to require that the ratio of available resistance to the effect of the applied load, termed the factor of safety, exceeds a minimum value greater than unity. Heretofore, this value has been established somewhat subjectively on the basis of engineering judgement. The implicit purpose of this factor is to account for uncertain definition of the loading and resistance. Accordingly, probability theory could be used to supplement engineering judgement in objectively establishing a minimum safety factor. Such a procedure is detailed herein for a simplified example problem to facilitate a discussion at this clinical session. This development will indicate items requiring further research before a reliability based sliding safety factor can be practically implemented.

II. MECHANISM OF SLIDING FAILURE. Figure 1 is a free body diagram of a concrete gravity dam cross section under its normal operating condition.

In this diagram, H represents the resultant of the hydrostatic pressure exerted on the upstream face of the cross section by the reservoir when filled to the spillway elevation and W is the total weight of the dam. For illustrative purposes only, the critical foundation surface is shown to be oriented horizontally and the total shearing and normal forces acting on this plane are designated by Q and N respectively. If this structure is stable against a sliding failure, the conditions of translational equilibrium require that

$$Q = H \quad (1)$$

and

$$N = W \quad (2)$$

Now, the maximum total shearing resistance R which can be developed along the foundation surface is approximated by (Lambe and Whitman, 1969)

$$R = C + N \cdot \tan \phi \quad (3)$$

In this expression, C represents the total cohesive resistance which can be developed in the absence of any normal force and ϕ is termed the friction angle of the foundation material. Figure 2 indicates that the shearing resistance available from equation (3) to counteract externally applied loads is a function of the foundation material properties, C and ϕ , and the normal force N induced by these loads through equation (2).

The safety factor is usually taken to be ratio of available resistance to load effect or in this illustration

$$F = \frac{R}{Q} \quad (4)$$

Accordingly, the structure is stable against sliding for values of F not less than unity. As neither Q nor R can be known with certainty, prudent practice has traditionally required that the calculated F exceed unity by a comfortable margin. The magnitude of this margin has traditionally been

determined through a professional consideration of the uncertainty in load and resistance definition as well as the consequences of structural instability in specific situations.

III. PROBABILISTIC ANALYSIS. Probability theory provides a means of rationally quantifying the various uncertainties associated with sliding stability as follows. First, for computational convenience in this illustrative problem, assume that the load is described by the value of a lognormal random variable having a median m_Q and a coefficient of variation V_Q . Similarly, model the resistance by the value of a lognormally distributed random variable with median m_R and coefficient of variation V_R . If load and resistance are further presumed to be independent of one another, it then follows that the safety factor is lognormally distributed with median

$$m_F = \frac{m_R}{m_Q} \quad (5)$$

and coefficient of variation V_F satisfying

$$\ln(V_F^2 + 1) = \ln(V_R^2 + 1) + \ln(V_Q^2 + 1) \quad (6)$$

A further consequence of these conditions is that the reliability against sliding failure is

$$r = \phi \left(\frac{m_F}{\sqrt{\ln(V_F^2 + 1)}} \right) \quad (7)$$

where $\phi(\cdot)$ is the cumulative distribution function of the standard unit normal variate. Theoretically, one could use this distribution to compute the probability of sliding instability (Prendergast, 1979). However, practical difficulties exist with this concept because there is insufficient justification for the assumed lognormal distribution of load and resistance, because an explicit criteria for an acceptable failure probability is difficult to

establish and because many practicing civil engineers are unschooled in probabilistic calculations.

IV. RELIABILITY BASED SAFETY FACTOR. Similar difficulties in the structural design of buildings have recently been circumvented by establishing safety factors based on probabilistic computations which can then be used in the traditional deterministic manner (Ellingwood *et al*, 1980). A corresponding approach in the case of sliding stability would proceed by estimating the level of reliability implied by current design practice. A recent examination (Baecher *et al*, 1980) of modern United States dams of all types disclosed a failure rate from all failure mechanisms of 2×10^{-4} per dam-year of which approximately 10% were attributed to slides. A preliminary estimate for the sliding reliability implied by current design practice for the normal operating condition might then be $r = 1 - 0.10 \times 2 \times 10^{-4} = 0.99998$. Equations (5) and (7) and the tabled values of the normal distribution then imply that an appropriate median safety factor is given by

$$\left(\frac{m_F}{4.11} \right)^2 = \ln(V_R^2 + 1) + \ln(V_Q^2 + 1) \quad (8)$$

In Figure 3, it is seen that this criteria desirably requires a higher safety factor as the uncertainty about load or resistance increases. Surely, the appropriateness of the lognormal model assumed in the foregoing warrants further investigation. Data in addition to (Baecher *et al*, 1980) should also be examined to refine the reliability level implicit in equation (8) before implementing this procedure. The result of these studies would be a relation among m_F , V_R and V_Q that more realistically described the sliding stability of a concrete structure.

V. APPLICATION. The above criteria would be applied in the following step-by-step procedure:

1. Estimate the load effect's coefficient of variation V_Q . This estimate should encompass not only the uncertainty of the variables from which the load is computed but also the accuracy of the analytical model used in the computation.
2. Similarly estimate V_R considering not only the sample variability of test data but also any differences between measured and in situ values as well as the accuracy of the resistance model adopted through equation (3).
3. Enter Figure 3 to obtain the median safety factor required consistent with the load and resistance uncertainties determined in steps 1 and 2.
4. If a structure's safety factor, computed from median values, equals or exceeds the required value, the design is acceptable as in traditional deterministic practice.
5. If a structure's computed median safety factor is less than the required value, the design has not been shown to be acceptable. It can now be modified to an acceptable level as in traditional practice. Alternately, further design studies can be conducted in hope of reducing V_Q or V_R and thereby requiring a lower m_F which the original design may satisfy.

Notice that this procedure allows the engineer to quantify his uncertainty about load and resistance rather than relying on only nominal deterministic values for these variables. However, some guidance on this quantification must be developed for practical use.

For example, suppose that the normal operating load on a dam is characterized by $m_Q = 300$ kip and a relatively small $V_Q = 0.05$ since this load

effect results from a well understood hydrostatic pressure. From the dam's weight and preliminary estimates of foundation material properties the resistance is thought to be described by $m_R = 550$ kip and $V_R = 0.50$. The corresponding $m_F = \frac{550}{300} = 1.8$ which is less than the value of 2.0 required in Figure 3 to be consistent with the $V_R = 0.50$ and $V_Q = 0.05$.

Now suppose further foundation investigations are performed to qualify this design which revise the estimated resistance parameters to be $m_R = 500$ kip and $V_R = 0.40$. The updated median safety factor becomes $m_F = \frac{500}{300} = 1.7$ which exceeds the value of 1.6 now required in Figure 3 to be consistent with the improved estimate of V_R .

VI. CONCLUSION. A reliability based safety factor for a simplified example of sliding stability has been developed which quantifies the engineer's uncertain knowledge of load and resistance. Simplifying assumptions in both the deterministic and probabilistic analyses leading to this criteria should be critically examined. Statistical data are also needed to refine the calibration of this procedure with current deterministic practice.

VII. ACKNOWLEDGEMENT. This development was supported by the Office of the Chief of Engineers.

REFERENCES

1. Baecher, G., Pate, M. and de Neufville, R., 1980. "Dam Failure in Benefit/Cost Analysis," Journal of the Geotechnical Engineering Division, ASCE, Vol 106, No GT1, pp 101-5.
2. Ellingwood, B., Galambos, R. V., MacGregor, J. G. and Cornell, C.A., 1980. Development of a Probability Based Load Criterion for American National Standard A58. National Bureau of Standards Special Publication 577, Washington, DC.
3. Lambe, T. W. and Whitman, R. V.; 1969. Soil Mechanics. John Wiley & Sons Inc., New York.
4. Prendergast, J. D., 1979. "Probabilistic Concept for Gravity Dam Analysis," Special Report M-265, US Army Construction Engineering Research Laboratory, Champaign, IL.
5. US Army Corps of Engineers, 1958. "Gravity Dam Design," EM 1110-2-2200, Washington, DC.
6. US Bureau of Reclamation, 1976. Design of Gravity Dams. US Government Printing Office, Denver, CO.

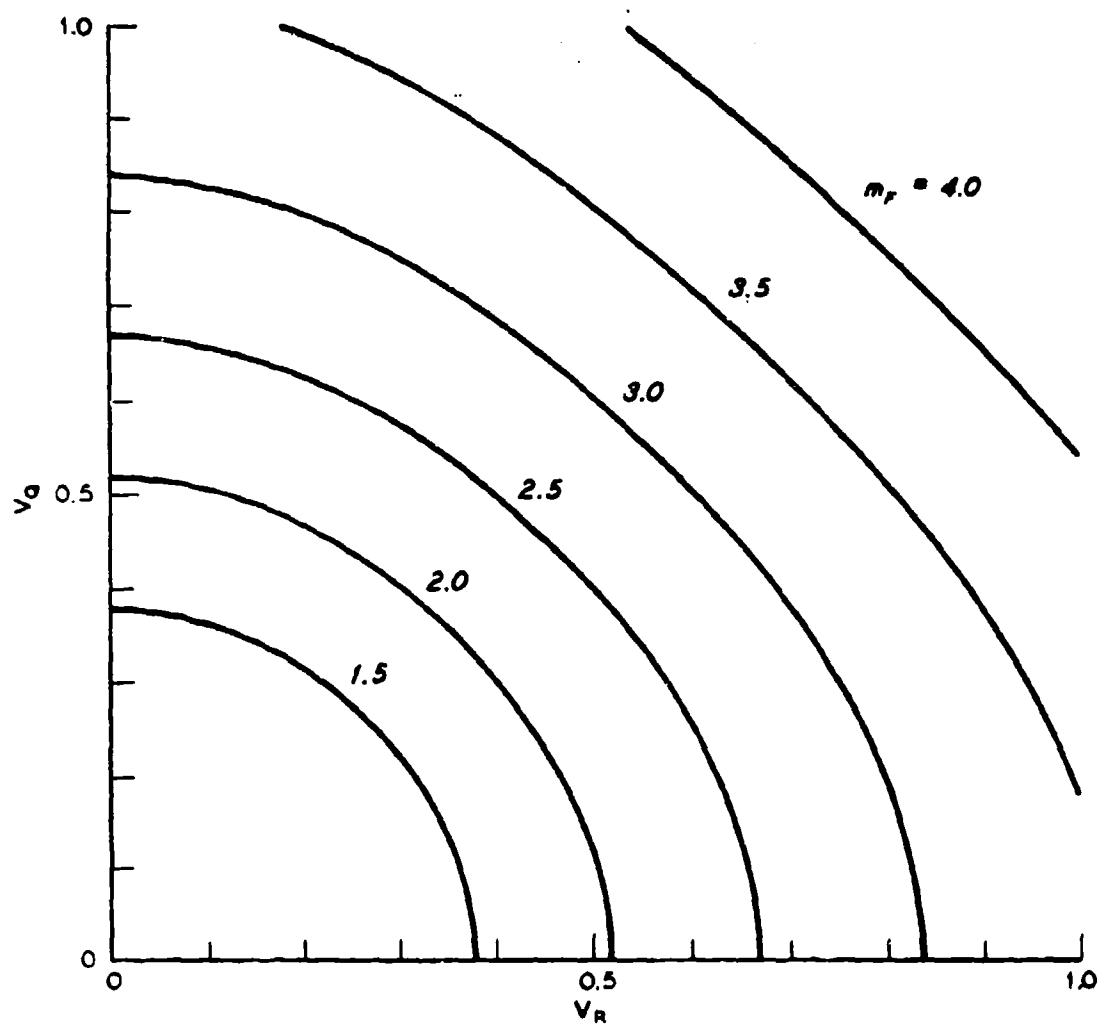


Figure 3. Reliability Based Sliding Safety Factor

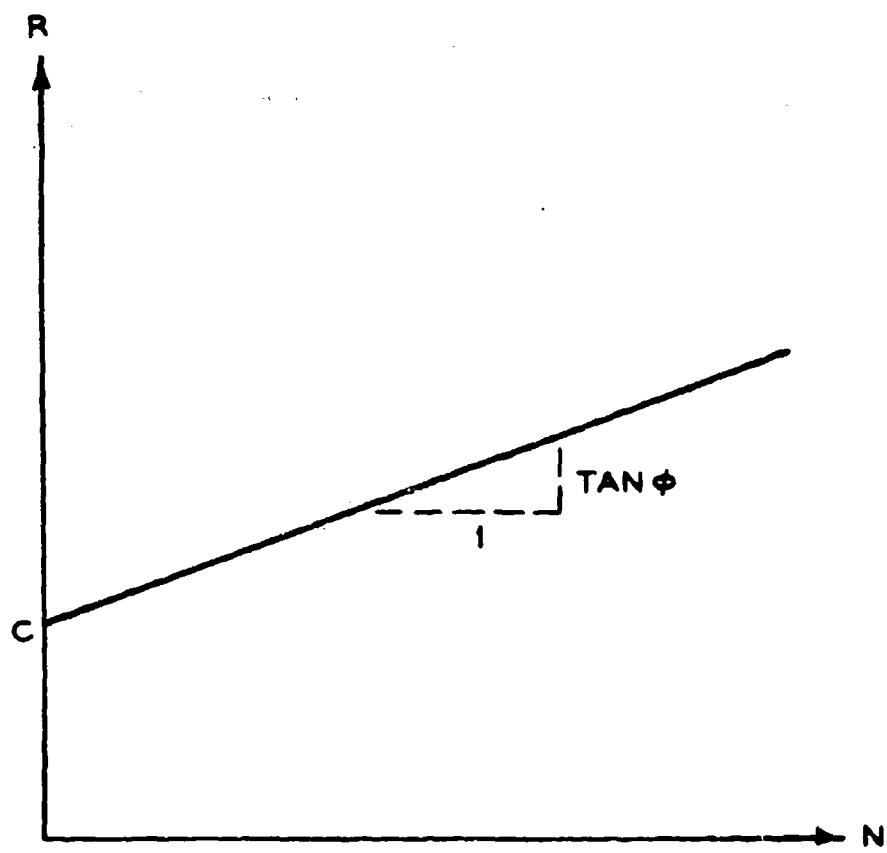


Figure 2. Sliding Resistance of Foundation

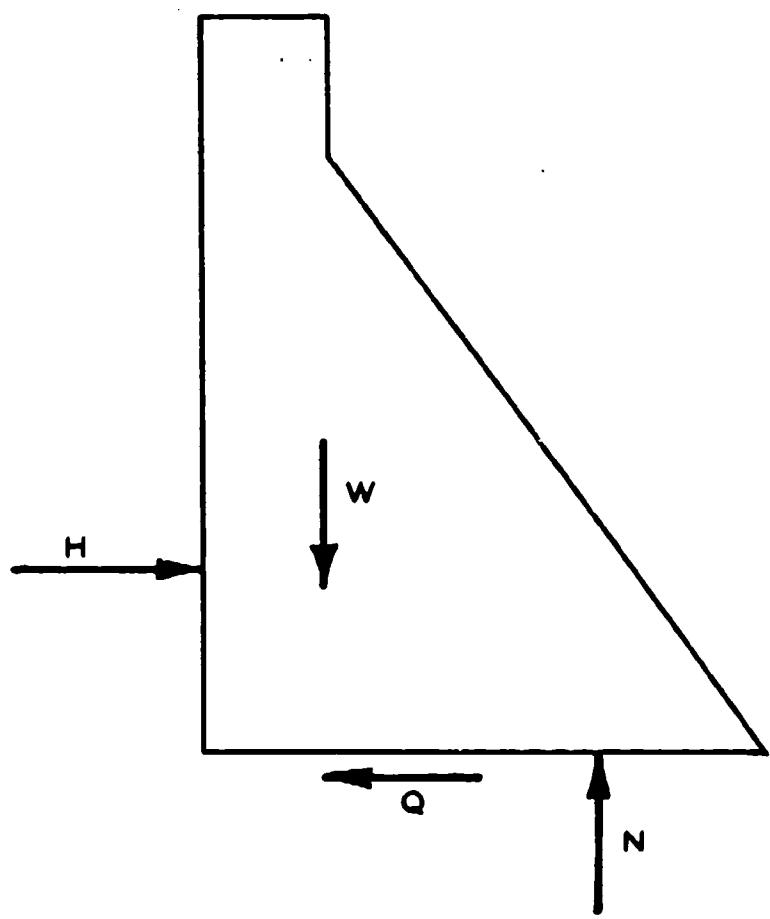


Figure 1. Concrete Gravity Dam Under Normal Operating Conditions

A UNIFIED AND UNBIASED ANALYSIS
FOR
DECISION MAKING IN PATERNITY DISPUTES

Paul H. Thrasher
10400 Omicron Place
El Paso, Texas 79924

ACKNOWLEDGEMENT. The author wishes to thank the Conference on the Design of Experiments in Army Research, Development, and Testing for hearing and publishing this paper on a non-Army topic. Recognition is also due to the University of Texas at El Paso for paid but appreciated computer time.

ABSTRACT. The analysis of genetic testing in paternity disputes uses several probabilities. The commonly used probabilities, first developed in Europe, are discussed. Two additional probabilities, based on (1) the possible fathers of the child when maternity is assumed and (2) the possible children of the mother and putative father, are introduced to measure the putative father's and mother's risks. A sample calculation is made to clarify the different probabilities. The results are organized in an analysis table to present an unbiased and comprehensive summary. A hypothesis test is proposed for decision making. This hypothesis test is designed for the United States legal system as opposed to the European system. A brief discussion of the attitudes and interactions of analysts and courts is included.

I. INTRODUCTION. Statisticians and people in other professions who deal with probabilities are often asked to interpret data from experiments whose design cannot be controlled. When this data can be interpreted with several techniques, great care must be taken to present an unbiased and complete analysis. When the analysis is to be used by people not trained in statistics, a logical and uncomplicated presentation is extremely important.

The topic of paternity testing is a challenge to statisticians, judges, and juries. It has great social and economic impact for the mother, child, putative father, and all tax payers who contribute to state funded child support. The intense personal involvement creates biased viewpoints. The antagonistic legal system, as it is practiced in the United States, makes the presentation of an unbiased analysis difficult. However, the people who must decide paternity disputes need all the information they can obtain in an unbiased and unconfusing manner.

Courts in the United States have been hesitant to assign much weight to statistical analyses which have been developed and used in Europe. In the United States, there may well be two primary reasons that the use of probabilities in paternity disputes has been unpopular and infrequent. First, the United States legal system starts with the assumption that the putative father is not the true father; but the European system begins with the opposite hypothesis.¹ Second, most proposed analyses have been so incomplete and fragmented that opposing lawyers could easily stress biased viewpoints and create confusion.²

II. MEASUREMENTS. There are several genetic polymorphisms which yield information about paternity. The first measurements were made on blood groups such as the ABO system. Now there are 20 to 25 genetic marker systems¹ including blood groups, protein groups, enzyme groups, and HLA types.¹

All available systems³ fit into the model of the hypothetical systems of Appendices A, B, and C. Any tested person is classified according to phenotype. This phenotype includes one or more genotypes. Each genotype contains two genes which may be identical. The laboratory testing on any person cannot measure the genotype unless the measured phenotype has only one genotype. The phenotype, gene, and genotype frequencies are the fractions of the relevant population that have those phenotypes, genes, and genotypes. These three types of frequencies each naturally sum to unity. The numerical values are independent of sex but depend on the racial subgroup and geographical area of the population.⁴ Since there are three adults that must be considered in a paternity dispute, three populations must be characterized by gene frequencies. Although these three may be identical, this paper will consider the general case; for the first hypothetical genetic system, Appendix A describes the random man and Appendix B describes the populations from which the mother and putative father are taken. Appendix C describes all three populations for the second hypothetical genetic system. Although both hypothetical systems have obviously been given exact frequencies for all three populations, actual measurement of gene frequencies uses sampling and the desired gene frequencies do not come from an exact algebraic solution. One approach is to use the least squares technique as suggested in Appendix A.

In addition to gene frequencies describing the populations of the random man, mother, and putative father, the exact phenotypes of the mother, child, and putative father are measured. It is assumed that no errors are made by the medical laboratory which types the mother, child, and putative father.

The child's and mother's genetic systems are further measured by their relationship. This restriction of the laboratory measured types occurs because the child gets one gene of each genotype from the mother and one from the true father. For example, if the mother is phenotype R and the child is phenotype Q, the mother must be genotype ab because she had to give gene a to the child. Also, if the mother is phenotype Q and the child is phenotype R, the child must be genotype ab because the mother could not have given gene b to the child. Occasionally both the mother's and child's genotypes are restricted; for example, an R mother and S child must be ab and ac, respectively. This assumption of true maternity between the mother and child is normally valid when the mother is attempting to establish paternity of the putative father; and it will be made for all calculations in this paper. However, it must be considered very suspect in cases such as a man and wife attempting to establish paternity of children for purposes of immigration.

III. EXCLUSIONS. One possible result of genetic testing is to exclude the putative father from any possibility of paternity.⁵ An example of an unquestionable exclusion in the hypothetical genetic system of Appendix C occurs when the child is phenotype JK but the mother and putative father are both

phenotype JJ; the assumption of maternity requires that the child's J gene was transmitted by the mother but the child also has a K gene which the putative father does not possess and could not transmit. Medical authorities are justly concerned with the possibility of false exclusions.⁵ A few examples are given here just to underscore the fact that any statistical analysis relies on a medical foundation. An apparent exclusion is obtained if the mother and child are JJ while the putative father is KK in the hypothetical genetic system of Appendix C. This exclusion could be false, however, if (1) a rare L gene existed in the system, (2) the antiserum necessary to detect the L gene is unavailable, and (3) the child is really JL and the putative father is really KL. A second type of false exclusion occurs if the child is not old enough for the genetic system to be fully developed. An example using the hypothetical system of Appendix A would occur if both the mother and putative father were U = bc and the child tested as V = bd but would become U = bc when his or her genetic system fully developed. For this paper, the genetic systems will be assumed to be completely developed, known, and tested.

If no exclusion is obtained, two probabilities may be calculated as a rough indication of paternity.⁶ These are:

$$Z' \equiv P(\text{a random man could be the father}) \text{ and}$$

$$Z \equiv P(\text{a random man tests as if he could be the father}).$$

An alternate way of viewing the same information is to consider:

$$P'_{\text{ex}} = P(\text{a random man could not be the father}) \text{ and}$$

$$P_{\text{ex}} = P(\text{a random man would be excluded by testing}).$$

Appendix D shows an example of these calculations based on gene frequencies of Appendices A and C. Z' will always be less than or equal to Z . Neither Z' or Z depend on the result of putative father's genetic test although the calculation loses its significance if the putative father is excluded. For both Z' and Z , the only gene frequencies considered are those of a random man; the information about the mother's or putative father's racial and/or geographic backgrounds are not used. The combination of Z' or Z values from separate systems assumes that these systems are independent.

The discrimination of genetic systems may be measured by average values of the probabilities of excluding falsely accused putative fathers. This calculation, which is illustrated in Appendix E, utilizes the probabilities of occurrence of random mother-child combinations, the gene frequencies of the mother's population, and the gene frequencies of a random man.

IV. CONDITIONAL PROBABILITIES. The existence of the putative father's phenotype may be used in the calculation of conditional probabilities which give a crude likelihood of paternity.⁷ These are defined as:

$Y' = P(\text{the putative father's phenotype occurs in random men})$ and

$X' = P(\text{the putative father's phenotype occurs in possible fathers having the putative father's gene frequencies}).$

Appendix F shows this calculation for the case started in Appendix D using the genetic systems of Appendices A, B, and C.

Conditional probabilities containing more information may be calculated on the basis of gene transmission.⁸ These probabilities, which also use the putative father's phenotype and the gene frequencies of his population, are defined as:

$Y = P(\text{the necessary genes to produce the child were transmitted})$ and

$X = P(\text{the necessary genes to produce the child were transmitted to the mother if the putative father was the father}).$

Appendix G shows this calculation for the example considered in Appendices D and F. The calculation of gene transmissions is illustrated in the demonstration that the probability of a random man transmitting a gene is just the gene frequency of a random man. The consideration of the genetically possible ways that the mother and a random man or putative father can transmit genes makes Y and X more informative than Y' and X' .

The paternity index is sometimes defined⁹ as X'/Y' and sometimes¹ as X/Y . As the example of Appendices F and G show, these are not the same. This illustrates that users of probabilities in paternity disputes must be very careful with their definitions.

The combination of X_i 's and Y_i 's from different genetic systems into a composite X and Y requires independence of the systems in order for the multiplication procedure,

$$X = X_1 X_2 \dots X_{\text{final system}} \text{ and}$$

$$Y = Y_1 Y_2 \dots Y_{\text{final system}},$$

to be valid. When systems are not independent, they must be combined into a single system as indicated in Appendix H.

V. BAYES FORMULA. The likelihood of paternity for the putative father is given by

$W = P(\text{the putative father is the father given that the necessary set of genes have been transmitted to the mother to produce the child}).$

Since the child exists and accurate genetic testing is assumed, W reduces the probability that the putative father is the father. Examination of X and Y shows that W , X , and Y are related by Bayes Formula.¹ By using P and $1-P$ to denote a priori probabilities of paternity for the putative father and random man, W may be written as

$$W = \frac{X P}{X P + (1-P) Y}$$

The a priori probability P is an integral part of the calculation of W . As seen in Appendix I, which continues the example started in Appendix D, the selection of a number for P influences the numerical value of W . In some paternity calculations, P is set at one-half²; the argument for this substitution is that nothing is really known other than the genetic test results. In other analyses, some values such as seven-tenths is used on the assumption that this is the proportion of valid paternity claims which women bring to court.³ Obviously, the judge and/or jury should consider the non-genetic evidence in each case and subjectively establish the a priori value P .

VI. STATISTICAL RISKS. The putative father may be a random man whose genetic characteristics just happen to yield a high W value.¹⁰ The likelihood of this happening to a man who is not the father is

$$\alpha_{PUF,min} = P(\text{finding a possible father whose likelihood of paternity is as great or greater than } W_{PUF} \text{ when the putative father and possible fathers are random men}).$$

A calculation of this minimum, putative father's risk is illustrated in Appendix J for the example started in Appendix D.

The putative father may be the father but the genetic characteristics of the mother, child, and father may just happen to yield a low W value.¹⁰ The likelihood of this happening is

$$\beta_{M,min} = P(\text{the mother and putative father having a child which results in the putative father's likelihood of paternity being as low or lower than } W_{PUF} \text{ when the putative father is the father}).$$

A calculation of this minimum mother's risk is illustrated in Appendix K for the example started in Appendix D. This risk becomes the state's risk if the state is to pay child support when neither the mother or the putative father provide it.

VII. HYPOTHESIS TESTING. A comprehensive and combined use of the probabilities in a paternity dispute may be made with a hypothesis test.¹⁰ The null and alternate hypothesis with the associated risks are:

H_0 : The putative father is not the father

α

H_A : The putative father is the father

β

where

$\alpha = P(\text{rejecting the null hypothesis, that the putative father is not the father, when he really is not the father})$ and

$\beta = P(\text{accepting the null hypothesis, that the putative father is not the father, when he really is the father}).$

This hypothesis test is structured according to the traditional assumption of the United States legal system. Innocence (i.e., non-paternity) is assumed as long as there is reasonable doubt of guilt (i.e., paternity). This hypothesis test may be performed by using the analysis table of Appendix L; this completes the numerical example of Appendices D, F, G, I, J, K, and L which uses the hypothetical genetic systems of Appendices A, B, and C whose discrimination in an average paternity dispute is calculated in Appendix E.

A court's use of the analysis table certainly does not relieve the court of its judgment. In fact, the court must make knowledgeable and judicious determinations to use the procedure. This seven-step procedure is described in Appendix M. The result of the procedure may very well be that a decision cannot be made without either obtaining more genetic information or lowering the risks that the court is willing to take. This need for more definitive tests may occur for the reasons listed in Appendix N. The hypothetical example calculated in this paper, which culminated in Appendix L, is not very definitive. The results of a more definitive example is shown in Appendix O. Appendix P shows possible uses of this analysis table.

The number of actual cases that have been analyzed by the technique proposed in this paper is quite small. For these cases, Appendix Q presents (1) the line of the analysis table which has the putative father's posterior probabilities of paternity and (2) the entry in the α column which is either on or just above the putative father's line. The values were calculated from six red cell antigen systems recommended by the Texas Society of Pathologists; these are ABO, MNSs, Rh, Kell, Duffy, and Kidd systems. These analyses were done before the availability of HLA testing so they are not as definitive as currently possible. Although a lack of small α and β risks may disturb the analyst, the results in Appendix Q are useful when the non-genetic evidence indicates a high value of P ; for each case, a hypothesis test may be used in ruling against the putative father. The fact that a test has a high β risk has little meaning when the ruling favors the mother anyway; the important risk is that of the party who is not favored by the ruling.

VIII. DISCUSSION. The basic philosophy behind the use of probabilities in paternity disputes is that all available information should be used in paternity judgments. Unless probabilities are used, genetic tests add nothing to the decision making process when the putative father is not excluded.

Although probabilities can never result in certainty of the necessary decision, they can certainly add information when no exclusion is obtained. Even if an apparent exclusion is obtained in one genetic system, probabilities based on other systems can be used to get an indication of the possibility of a false exclusion.¹

Since genetic tests are common and even required in many states before a paternity dispute can be brought to court, all possible information from the tests should logically be available for the court. The difficulty of presenting probabilistic information is that many different probabilities can be discussed and confusion easily results. This is the reason for organizing the results in an analysis table as shown in Appendices L and O. The court should have an unbiased presentation of probabilities; the analysis table provides such a presentation.

Judges and/or juries should be hesitant to use partial probabilistic arguments which can easily be given biased interpretations by lawyers in an antagonistic legal system. Since no information is normally provided about the putative father's and mother's risks denoted in this paper by α and β , the courts are justifiably hesitant to apply much weight to the probabilities that are presented. This hesitancy is especially justified when the probabilistic argument does not let the court choose the a priori probability denoted by P in this paper. A useful analysis must include all genetic and non-genetic information. It also must describe all risks involved without unduly emphasizing any one of them.

There is occasionally a hesitancy to use a decision making process that admits any uncertainty in the final decision. This attitude can never be completely eliminated; but the person performing and/or presenting a probability analysis must stress that certainties are very rare and analytical consideration of uncertainties is essential to intelligent decision making.

There is occasionally a reluctance to perform a long tedious calculation. This attitude can influence the preparation of the procedure proposed in this paper because the number of calculations expands rapidly when the lists of possible fathers and possible children lengthens as more genetic systems are added to the calculation. There is an obvious and successful method of avoiding the tedium and possibility of numerical mistakes; the whole procedure may be computerized. To write the necessary program, it has naturally been necessary to use both logical and arithmetic programming techniques.

There is a hesitancy of some analysts to use any hypothesis test which does not have very low values for both the α and β risks. The analyst's job in paternity disputes is not to set risks. The analyst must explain the meaning of the risks to the judge and/or jury; the court must then assign the numerical values to all parameters that affect the use of the analysis table. The court may very well wish to assign different limits on the risks of the putative father and the mother or state. This decision involves the relative cost of making a mistake; and only the holders of ultimate responsibility should set risks of making mistakes.

There is a reasonable hesitancy of analysts to use imprecise data. This leads to two concerns in the calculation of paternity probabilities. First, the combination of the results from different genetic systems requires the different systems to be independent. If this is not true, a combined system such as illustrated in Appendix H must be used. If the gene frequencies for this single system are not available, the only prudent courses of action are (1) ignore both systems or (2) present two analyses with each considering one of the two non-independent systems. It is extremely dangerous to use the system which the analyst considers to yield the most informative results; this is done in some European analyses¹ but it amounts to selecting data which will yield a result of predetermined bias. Second, the gene frequencies of the genetic systems must be known for all involved populations. Although the medical profession must collect the data for these populations, there is no real medical application for gene frequencies. This is especially true for a combined system made from two or more non-independent systems. When the analyst is not given accurate population gene frequencies, he must repeat the calculation for several sets of gene frequencies which surround the true population frequencies. Selecting the frequencies to use, interpreting the results for each system, and combining the results of the different systems requires more analytical judgment than merely inputting data to a computer and reading the final result. This is the real job of the analyst.

Perhaps the most important prerequisite for widespread use of probabilities in paternity cases is a standardized format. There are many probabilities involved and many people need to use them. Confusion will be both possible and likely until some standardized presentation is accepted. This needed standardization is one important reason for the use of the analysis table shown in Appendices L and O. One method of obtaining standardization would be for state legislatures to prescribe a format for information to be presented in court. If the state legislature did this, they might very well want to provide guidance in the numerical values to be used for the limiting risks α_L and β_L and the decision level W_C that are required in the procedure of Appendix M. This guidance would, of course, have to provide a list of feasible genetic systems and a procedure to be followed if this list were exhausted without yielding sufficiently low values of test risks α_T and β_T to result in a decision.

Finally, state legislatures might very well want to direct state organizations such as medical schools to compile gene frequencies for genetic systems and combinations of systems that are not well known to be independent.

IX. PROSPECTS. The use of probabilities in paternity disputes is very fragmentary in the United States. It is much more prevalent in European courts. Differences in the European and United States legal systems have prevented the spread of the technique to the United States.

The use of probabilities will undoubtedly grow in the United States. One reason is that scientific methods and statistical analyses are continuously becoming more understood. Another is that future availability of the putative father's and mother's risks, which are not used in European courts, should

satisfy objections that an incomplete and even erroneous description can be obtained by focusing all of the court's attention on a single and non-comprehensive probability.

X. REFERENCES.

- (1) Salmon, D. and Salmon C.; Blood Groups and Genetic Markers Polymorphism and Probability of Paternity, Transfusion, 20: 684-694, 1980.
- (2) Ellman, I. M. and Kaye, D; Probability and Proof: Can HLA and Blood Group Testing Prove Paternity?, New York University Law Review, 54: 1131-1162, 1979.
- (3) American Medical Association Committee on Transfusion and Transplantation, and American Bar Association Committee on Standards for the Judicial Use of Scientific Evidence in the Ascertainment of Paternity; Joint AMA - ABA Guidelines: Present Status of Serologic Testing in Problems of Disputed Parentage, Family Law Quarterly, VOL X, No. 3: 247-285, Fall 1976.
- (4) Mourant, A. E., Kopeć, Ada C., and Domaniewska-Sobczak, Kazimiera; The Distribution of the Human Blood Groups and other Polymorphisms, Oxford University Press, London, 1976.
- (5) Issitt, P. and Issitt, C.; Applied Blood Group Serology, Spectra Biologicals of Becton, Dickinson and Company, 1975.
- (6) Chakraborty, R., Shaw, M., and Schull, W.; Exclusion of Paternity: The Current State of the Art, Am J Hum Genet, 26: 477-488, 1974.
- (7) Solomon, H.; Jurimetrics, Research Papers in Statistics, John Wiley and Sons, New York, 1966.
- (8) Beautyman, M.; Paternity Actions - A Matter of Opinion or a Trial of the Blood?, J of Legal Medicine, 4: 17-25, 1976.
- (9) Lee, C. and Henry, J.; Laboratory Evaluation of Disputed Paternity, Immunology and Immunopathology, 44: 1507-1548.
- (10) Thrasher, P.; Risks Inherent in Paternity Testing, Trial Lawyers Form of the Texas Trial Lawyers Association, 12:4: 41-43, 1978.

APPENDIX A
FIRST HYPOTHETICAL GENETIC SYSTEM FOR RANDOM MAN

<u>PHENOTYPES [I]</u>	<u>POSSIBLE GENOTYPES [ij]</u>	<u>PHENOTYPE FREQUENCIES [P(I)]</u>
Q	aa	$P(Q) = .16$
R	bb or ab	$P(R) = .33$
S	cc, cd, or ac	$P(S) = .24$
T	dd or ad	$P(T) = .09$
U	bc	$P(U) = .12$
V	bd	$P(V) = .06$

RELATIONS USED TO FIND GENE FREQUENCIES [P(i)]:

$$P(Q) = [P(a)]^2$$

$$P(R) = [P(b)]^2 + 2P(a)P(b)$$

$$P(S) = [P(c)]^2 + 2P(c)P(d) + 2P(a)P(c)$$

$$P(T) = [P(d)]^2 + 2P(a)P(d)$$

$$P(U) = 2P(b)P(c)$$

$$P(V) = 2P(b)P(d)$$

LEAST SQUARES SOLUTION FOR P(i)'s MINIMIZES δ^2 :

$$\begin{aligned} \delta^2 &= \{P(Q) - [P(a)]^2\}^2 + \{P(R) - [P(b)]^2 - 2P(a)P(b)\}^2 \\ &\quad + \{P(S) - [P(c)]^2 - 2P(c)P(d) - 2P(a)P(c)\}^2 \\ &\quad + \{P(T) - [P(d)]^2 - 2P(a)P(d)\}^2 + \{P(U) - 2P(b)P(c)\}^2 \\ &\quad + \{P(V) - 2P(b)P(d)\}^2 \end{aligned}$$

GENE FREQUENCIES:

i	a	b	c	d
P(i)	.4	.3	.2	.1

GENOTYPE FREQUENCIES [P(ij)]:

	a	b	c	d
a	.16			
b		.09		
c	.16	.12	.04	
d	.08	.06	.04	.01

APPENDIX B
FIRST HYPOTHETICAL GENETIC SYSTEM FOR ADULTS IN THE CASE

MOTHER

GENE FREQUENCIES:

i	a	b	c	d
P(i)	.42	.31	.19	.08

GENOTYPE FREQUENCIES:

	a	b	c	d
a	.1764			
b	.2604	.0961		
c	.1596	.1178	.0361	
d	.0672	.0496	.0304	.0064

PHENOTYPE FREQUENCIES:

I	Q	R	S	T	U	V
P(I)	.1764	.3565	.2261	.0736	.1178	.0496

PUTATIVE FATHER

GENE FREQUENCIES:

i	a	b	c	d
P(i)	.38	.29	.21	.12

GENOTYPE FREQUENCIES:

	a	b	c	d
a	.1444			
b	.2204	.0841		
c	.1596	.1218	.0441	
d	.0912	.0696	.0504	.0144

PHENOTYPE FREQUENCIES:

I	Q	R	S	T	U	V
P(I)	.1444	.3045	.2541	.1056	.1218	.0696

APPENDIX C
SECOND HYPOTHETICAL GENETIC SYSTEM

PHENOTYPES: JJ JK KK

POSSIBLE GENOTYPES: JJ JK KK

GENE FREQUENCIES:

	i	P(i)
RM = RANDOM MAN	J	.3
	K	.7
M = MOTHER	J	.4
	K	.6
PUF = PUTATIVE FATHER	J	.2
	K	.8

GENOTYPE FREQUENCIES:

RM	J	K	M	J	K	PUF	J	K
J	.09		J	.16		J	.04	
K	.42	.49	K	.48	.36	K	.32	.64

PHENOTYPE FREQUENCIES:

	I	P(I)
RM	JJ	.09
	JK	.42
	KK	.49
M	JJ	.16
	JK	.48
	KK	.36
PUF	JJ	.04
	JK	.32
	KK	.64

APPENDIX D
PROBABILITIES OF EXCLUSION IS A SPECIFIC CASE

DEFINITIONS

M	Mother
C	Child
POF	Possible Father
+	Transmits
RM	Random Man
G	Gene POF Must +
Z'	$P(RM \text{ could } + \text{ G})$
Z	$P(RM \text{ tests as if he could } + \text{ G})$
P'_{ex}	$P(RM \text{ could not } + \text{ G})$
P_{ex}	$P(RM \text{ is Excluded by Tests})$

SYSTEM 1

$M = S = cc, dc, \text{ or } ac$
 $C = T = dd \text{ or } ad$
 $M = cd \text{ or } ac$
 $POF + a \text{ or } d$
 $POF = aa, ab, ac, ad, bd, cd, \text{ or } dd$
 $POF = Q, R, S, T, \text{ or } V$
 $Z' = P(RM = aa, ab, ac, ad, bd, cd, \text{ or } dd)$
 $= .16 + .24 + .16 + .08 + .06 + .04 + .01$
 $= .75$
 $Z = P(RM = Q, R, S, T, \text{ or } V)$
 $= .16 + .33 + .24 + .09 + .06$
 $= .88$

SYSTEM 2

$M = JJ$
 $C = JK$
 $POF + K$
 $POF = JK \text{ or } KK$
 $Z' = P(RM = JK \text{ or } KK)$
 $= .42 + .49$
 $= .91$
 $Z = Z'$
 $= .91$

COMBINATION

$$\begin{aligned}
Z' &= Z'_1 Z'_2 = (.75)(.91) = .6825 \\
Z &= Z_1 Z_2 = (.88)(.91) = .8008 \\
P'_{ex} &= 1 - Z' = 1 - .6825 = .3175 \\
P_{ex} &= 1 - Z = 1 - .8008 = .1992
\end{aligned}$$

APPENDIX E
AVERAGE PROBABILITY OF EXCLUSIONS

S Y S T E M 2

M	P(M)	C	P(C M)	POF	Z
JJ	.16	JJ	(1)(.3)	JJ, JK	.09 + .42
JJ	.16	JK	(1)(.7)	JK, KK	.42 + .49
JK	.48	JJ	(1/2)(.3)	JJ, JK	.09 + .42
JK	.48	JK	(1/2)(.3) + (1/2)(.7)	JJ, JK, KK	.09 + .42 + .49
JK	.48	KK	(1/2)(.7)	JK, KK	.42 + .49
KK	.36	JK	(1)(.3)	JJ, JK	.09 + .42
KK	.36	KK	(1)(.7)	JK, KK	.42 + .49

$$z_{2,\text{ave}} = \sum_i^{\text{POC IN SYSTEM 2}} P_i(M) P_i(C|M) z_i = .8404$$

S Y S T E M 1

M	P(M)	C	M	C	P(C M)
Q	.1764	Q	aa	aa	(1)(.4)
Q	.1764	R	aa	ab	(1)(.3)
Q	.1764	S	aa	ac	(1)(.2)
Q	.1764	T	aa	ad	(1)(.1)
R	.3565	Q	ab	aa	(.2604/.3565)(1/2)(.4)
R	.3565	R	bb, ab	bb, ab	(.0961/.3565)(1)(.3+.4) + (.2604/.3565)[(1/2)(.3) +(1/2)(.3+.4)]
R	.3565	S	ab	ac	(.2604/.3565)(1/2)(.2)
R	.3565	T	ab	ad	(.2604/.3565)(1/2)(.1)
R	.3565	U	bb, ab	bc	(.0961/.3565)(1)(.2) + (.2604/.3565)(1/2)(.2)
R	.3565	V	bb, ab	bd	(.0961/.3565)(1)(.1) + (.2604/.3565)(1/2)(.1)

M	P(m)	C	M	C	P(C M)
S	.2261	Q	ac	aa	$(.1595/.2261)(1/2)(.4)$
S	.2261	R	ac	ab	$(.1596/.2261)(1/2)(.3)$
S	.2261	S	cc,cd, ac	cc,dc ac	$(.0361/.2261)(1)(.2+.1+.4)$ + $(.0304/.2261)[(1/2)(.2+.1+.4)$ + $(1/2)(.2)]$ + $(.1596/.2261)[(1/2)(.2)$ + $(1/2)(.2+.1+.4)]$
S	.2261	T	cd,ac	dd,ad	$(.0304/.2261)(1/2)(.1+.4)$ + $(.1596/.2261)(1/2)(.1)$
S	.2261	U	cc,cd, ac	bc	$(.0361/.2261)(1)(.3)$ + $(.0304/.2261)(1/2)(.3)$ + $(.1596/.2261)(1/2)(.3)$
S	.2261	V	cd	bd	$(.0304/.2261)(1/2)(.3)$
T	.0736	Q	ad	aa	$(.0672/.0736)(1/2)(.4)$
T	.0736	R	ad	ab	$(.0672/.0736)(1/2)(.3)$
T	.0736	S	dd,ad	cd,ac	$(.0064/.0736)(1)(.2)$ + $(.0672/.0736)[(1/2)(.2)$ + $(1/2)(.2)]$
T	.0736	T	dd,ad	dd,ad	$(.0064/.0736)(1)(.1+.4)$ + $(.0672/.0736)[(1/2)(.1)$ + $(1/2)(.1+.4)]$
T	.0736	V	dd,ad	bd	$(.0064/.0736)(1)(.3)$ + $(.0672/.0736)(1/2)(.3)$
U	.1178	R	bc	bb,ab	$(1/2)(.3+.4)$
U	.1178	S	bc	cc,cd,ac	$(1/2)(.2+.1+.4)$
U	.1178	U	bc	bc	$(1/2)(.2)+(1/2)(.3)$
U	.1178	V	bc	bd	$(1/2)(.1)$
V	.0496	R	bd	bb,ab	$(1/2)(.3+.4)$
V	.0496	S	bd	cd	$(1/2)(.2)$
V	.0496	T	bd	dd,ad	$(1/2)(.1+.4)$
V	.0496	U	bd	bc	$(1/2)(.2)$
V	.0496	V	bd	bd	$(1/2)(.1)+(1/2)(.3)$

GENES THE FATHER MUST PROVIDE
AS INDICATED BY .4, .3, .2,
AND .1 FOR a, b, c, AND d,
RESPECTIVELY

	Z'	Z
.4	.16+.24+.16+.08=.64	.16+.33+.24+.09=.82
.3	.24+.09+.12+.06=.51	.33+.12+.06=.51
.2	.16+.12+.04+.04=.36	.24+.12=.36
.1	.08+.06+.04+.01=.19	.24+.09+.06=.39
.4 and .3	.16+.24+.16+.08 +.09+.12+.06=.91	.16+.33+.24+.09 +.12+.06=1.0
.4 and .1	.16+.24+.16+.08 +.06+.04+.01=.75	.16+.33+.24+.09 +.06=.88
.3 and .2	.24+.09+.12+.06 +.16+.04+.04=.75	.33+.12+.06 +.24=.75
.3 and .1	.24+.09+.12+.06 +.08+.04+.01=.64	.33+.12+.06 +.24+.09=.84
.4, .2, and .1	.16+.24+.16+.08 +.12+.04+.04 +.06+.01=.91	.16+.33+.24+.09 +.12 +.06=1.0

$$z'_{1,\text{ave}} = \sum_i^{\text{POC IN SYSTEM 1}} P_i(M) P_i(C|M) z'_i = .6771$$

$$z_{1,\text{ave}} = \sum_i^{\text{POC IN SYSTEM 1}} P_i(M) P_i(C|M) z_i = .7646$$

COMBINATION

$$Z' = z'_1 z'_2 = (.8404)(.6771) = .569$$

$$Z = z_1 z_2 = (.8404)(.7646) = .643$$

$$P'_{\text{ex}} = 1 - Z' = 1 - .569 = .431$$

$$P_{\text{ex}} = 1 - Z = 1 - .643 = .357$$

APPENDIX F

CONDITIONAL PROBABILITIES BASED ON EXISTENCE OF PHENOTYPES

ADDITIONAL DEFINITIONS

F	Father
PUF	Putative Father
Y'	$P(\text{PUF's Phenotype} \text{M's and C's Phenotypes and PUF} \neq \text{F})$
X'	$P(\text{PUF's Phenotype} \text{M's and C's Phenotypes and PUF} = \text{F})$

SYSTEM 1		SYSTEM 2
M	$S = cc, cd, \text{ or } ac$	$M = JJ$
C	$T = dd \text{ or } ad$	$C = JK$
M	$cd \text{ or } ac$	$POF \rightarrow K$
P	$\rightarrow a \text{ or } d$	$POF = JK \text{ or } KK$
POF	$= Q, R, S, T, \text{ or } V$	$PUF = JK$
PUF	$= S$	$Y' = P(\text{PUF} = JK \text{PUF} = RM)$ $= .42$
Y'	$= P(\text{PUF} = S \text{PUF} = RM)$ $= .24$	$X' = P(\text{PUF} = JK \text{PUF} = JK \text{ or } KK)$ $= .32 / (.32 + .64)$ $= .3333$
X'	$= P(\text{PUF} = S \text{PUF} = Q, R, S, T, \text{ or } V)$ $= .2541 / (.1444 + .3045 + .2541 + .1056 + .0696)$ $= .2893$	$Y'/X' = .4200 / .3333$ $= 1.260$
Y'/X'	$= .2400 / .2893$ $= .8296$	

COMBINATION

$$Y' = Y'_1 Y'_2 = (.2400)(.4200) = .1008$$

$$X' = X'_1 X'_2 = (.2893)(.3333) = .09642$$

$$Y'/X' = .1008 / .0964 = 1.05$$

APPENDIX G
CONDITIONAL PROBABILITIES BASED ON GENE TRANSMISSION

ADDITIONAL DEFINITIONS

G_{any}	Any Genes That Could Be →
Y	$P(G \text{ was } \rightarrow RM \rightarrow G_{\text{any}})$
X	$P(G \text{ was } \rightarrow PUF \rightarrow G_{\text{any}})$
<hr/>	
SYSTEM 1	SYSTEM 2
$M = S = cc, cd, ac$	$M = JJ$
$C = T = dd \text{ or } ad$	$C = JK$
$M = cd \text{ or } ac$	$POF \rightarrow K$
$POF \rightarrow a \text{ or } d$	$PUF = JK$
$PUF = S = cc, cd, \text{ or } ac$	$P(RM \rightarrow J) = P(RM=JJ)P(JJ \rightarrow J) + P(RM=JK)P(JK \rightarrow J)$ $= (.09)(1) + (.42)(1/2)$ $= .3$ $= P(J) \text{ as expected}$
$P(RM \rightarrow a) = P(RM=aa)P(aa \rightarrow a) + P(RM=ab)P(ab \rightarrow a) + P(RM=ac)P(ac \rightarrow a) + P(RM=ad)P(ad \rightarrow a)$ $= (.16)(1) + (.24)(1/2)$ $+ (.16)(1/2) + (.08)(1/2)$ $= .40$ $= P(a) \text{ as it should}$	$P(M=cd) = .0304 / (.0304 + .1596) = .1600$ $P(M=ac) = .1596 / (.0304 + .1596) = .8400$ $Y = P(M=cd)P(cd \rightarrow d)[P(RM \rightarrow d) + P(RM \rightarrow a)] + P(M=ac)P(ac \rightarrow a)P(RM \rightarrow d)$ $= (.1600)(1/2)(.1 + .4) + (.8400)(1/2)(.1)$ $= .0820$
$P(POF \rightarrow d) = P(POF=cd)P(cd \rightarrow d)$ $= (.0504/.2541)(1/2)$ $= .0992$	$P(M=JJ) = 1$ $Y = P(JJ \rightarrow 1)P(RM \rightarrow K)$ $= (1)(.7)$ $= .7$
$P(POF \rightarrow a) = P(POF=ac)P(ac \rightarrow a)$ $= (.1596/.2541)(1/2)$ $= .3140$	$P(PUF \rightarrow K) = 1/2$

SYSTEM 1 (Continued)

$$\begin{aligned} X &= P(M=cd)P(cd+d)[P(PUF \rightarrow d) + (PUF \rightarrow a)] \\ &\quad + P(M=ac)P(ac \rightarrow a) + (PUF \rightarrow d) \\ &= (.1600)(1/2)(.0992 + .3140) \\ &\quad + (.8400)(1/2)(.0992) \\ &= .0747 \end{aligned}$$

$$Y/X = .0820/.0747 = 1.10$$

SYSTEM 2 (Continued)

$$\begin{aligned} X &= P(JJ+J)P(PUF \rightarrow K) \\ &= 1(1/2) \\ &= .500 \end{aligned}$$

$$Y/X = .7/.5 = 1.40$$

COMBINATION

$$Y = Y_1 Y_2 = (.0820)(.7) = .0574$$

$$X = X_1 X_2 = (.0747)(.5) = .0374$$

$$Y/X = .0574/.0374 = 1.53$$

APPENDIX H

TABLE OF COMBINED SYSTEM TO REPLACE
THE HYPOTHETICAL SYSTEMS 1 AND 2 WHEN THEY ARE NOT INDEPENDENT

	Q	R	S	T	U	V
JJ	PH ₁₁	PH ₁₂	PH ₁₃	PH ₁₄	PH ₁₅	PH ₁₆
JK	PH ₂₁	PH ₂₂	PH ₂₃	PH ₂₄	PH ₂₅	PH ₂₆
KK	PH ₃₁	PH ₃₂	PH ₃₃	PH ₃₄	PH ₃₅	PH ₃₆

The number of phenotypes in a combined system is the product of the number of phenotypes in the individual systems.

APPENDIX I
APPLICATION OF BAYES THEOREM

ADDITIONAL DEFINITIONS

W	$P(PUF \rightarrow G_{any} \rightarrow G \text{ was } \rightarrow)$
---	---

P	$P(PUF = F \text{ on the basis of all non-genetic information})$
---	--

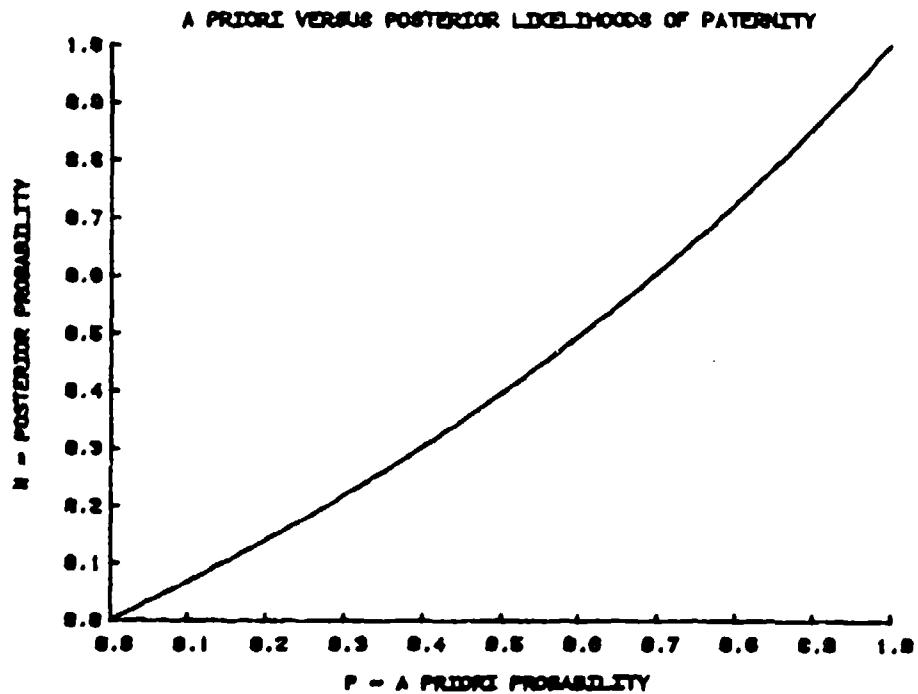
$$P(PUF \rightarrow G_{any} | G \text{ was } \rightarrow) = \frac{P(G \text{ was } \rightarrow | PUF \rightarrow G_{any}) P}{P(G \text{ was } \rightarrow | PUF \rightarrow G_{any}) P + P(G \text{ was } \rightarrow | RM + G_{any}) (1-P)}$$

$$W' = \frac{X P}{X P + Y(1-P)} = \frac{1}{1 + \frac{Y}{X} \left(\frac{1}{P} - 1 \right)}$$

For $Y/X^t = 1.53$ (from Appendix G),

P	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
W	0	.068	.140	.219	.303	.395	.495	.604	.723	.855	1.00

An alternate presentation of the P - W Table is a graph of P versus W:



APPENDIX J
MINIMUM PUTATIVE FATHER'S RISK

ADDITIONAL DEFINITIONS

x_{PUF}	X of the PUF
$x^{POF=I}$	X of a POF of I'th Phenotype from RM Population
w_{PUF}	W of the PUF
w_{POF}	W of a POF
$\alpha_{PUF, \min}$	$P(w_{POF} \geq w_{PUF} PUF = RM)$

SYSTEM 1	SYSTEM 2
$M = S = cc, cd, \text{ or } ac$	$M = JJ$
$C = T = dd \text{ or } ad$	$C = JK$
$M = cd \text{ or } ac$	
$P(M=cd) = .16; P(M=ac) = .84$	
$Y = .0820$	$Y = P(M+J) P(RM+K) = (1)(.7)$
$Y/X_{PUF} = 1.10$	$Y/X_{PUF} = 1.4 \text{ (from Appendix G)}$
$POF \rightarrow a \text{ or } d$	$POF \rightarrow K$
$POF = Q, R, S, T, \text{ or } V$	$POF = JK \text{ or } KK$
$x^{POF=I} = P(M=cd)P(cd \rightarrow d)[P(RM_I \rightarrow d) + P(RM_I \rightarrow a)] + P(M=ac)P(ac \rightarrow a)P(RM_I \rightarrow d)$	$x^{POF=I} = P(M+J)P(RM_I \rightarrow K)$
$x^{POF=Q} = (.16)(1/2)(0+1) + (.84)(1/2)(0) = .0800$	$x^{POF=JK} = (1)(1/2) = .500$
$x^{POF=R} = (.16)(1/2)[0 + (.24/.33)(1/2)] + (.84)(1/2)(0) = .0291$	$x^{POF=KK} = (1)(1) = 1.00$
$x^{POF=S} = (.16)(1/2)[(.04/.24)(1/2) + (.16/.24)(1/2)] + (.84)(1/2)(.04/.24)(1/2) = .0683$	
$x^{POF=T} = (.16)(1/2)[(.01/.09)(1) + (.08/.09)(1/2) + (.08/.09)(1/2)] + (.84)(1/2)[(.01/.09)(1) + (.08/.09)(1/2)] = .313$	

SYSTEM 1 (Continued)

$$X^{POF=V} = (.16)(1/2)(.5+0) + (.84)(1/2)(.5) \\ = .250$$

$$Y/X^{POF=Q} = .082/.0800 = 1.02$$

$$Y/X^{POF=R} = .082/.0291 = 2.82$$

$$Y/X^{POF=S} = .082/.0683 = 1.20$$

$$Y/X^{POF=T} = .082/.313 = .262$$

$$Y/X^{POF=V} = .082/.250 = .328$$

SYSTEM 2 (Continued)

$$Y/X^{POF=JK} = .7/.500 = 1.40$$

$$Y/X^{POF=KK} = .7/1.00 = .700$$

COMBINATION

$$Y/X_{PUF} = (Y_1/X_{1,PUF}) (Y_2/X_{2,PUF}) = 1.53$$

POF	$P(POF) = P(I_1)P(I_2)$	$(Y/X) = (Y/X)_1(Y/X)_2$	<u>IS $(Y/X) \leq (Y/X_{PUF})$</u>
Q and JK	$(.16)(.42) = .0672$	$(1.02)(1.40) = 1.43$	YES
Q and KK	$(.16)(.49) = .0784$	$(1.02)(.700) = .714$	YES
R and JK	$(.33)(.42) = .1386$	$(2.82)(1.40) = 3.95$	NO
R and KK	$(.33)(.49) = .1617$	$(2.82)(.700) = 1.97$	NO
S and JK	$(.24)(.42) = .1008$	$(1.20)(1.40) = 1.68$	NO
S and KK	$(.24)(.49) = .1176$	$(1.20)(.700) = .840$	YES
T and JK	$(.09)(.42) = .0378$	$(.262)(1.40) = .367$	YES
T and KK	$(.09)(.49) = .0441$	$(.262)(.700) = .183$	YES
V and JK	$(.06)(.42) = .0252$	$(.328)(1.40) = .459$	YES
V and KK	$(.06)(.49) = .0294$	$(.328)(.700) = .230$	YES
	$Z = Z_1Z_2 = .8008$		

$$\alpha_{PUF,min} = \sum_i^{\text{all } i \text{ for which } (Y/X)_i \leq (Y/X_{PUF})} P_i(POF) = .3997$$

APPENDIX K
MINIMUM MOTHER'S RISK

ADDITIONAL DEFINITIONS

POC $Y_{C=I}$ $X_{C=I}^{PUF=J}$ $P(POC=I)$ W_{POC} $\beta_{M,\min}$	Possible Child of M and PUF Y for a Child of Phenotype I X for a Child of Phenotype I when the PUF has Phenotype J Probability that M and PUF will have Child of Phenotype I W for PUF if PUF and M had POC $P(W_{POC} \leq W_{PUF} PUF = F)$
--	--

SYSTEM 1	SYSTEM 2
$P = S = cc, cd, \text{ or } ac$	$M = JJ$
$C = T = dd \text{ or } ad$	$C = JK$
$M = cd \text{ or } ac$	
Unless M has C ≠ T, $P(M=ac) = .84$ from $P(M=cd) = .16$ Appendix G	
$PUF = S = cc, cd, \text{ or } ac$	$PUF = JK$
$POC = aa, ac, ad, cc, cd, \text{ or } dd$	$POC = JJ \text{ or } JK$
$POC = Q, S, \text{ or } T$	
If $POC = Q, M = ac$ $P(M=ac) = 1$	
If $POC = S, M = cd \text{ or } ac$ $P(M=cd) = .16; P(M=ac) = .84$	
If $POC = T, M = cd \text{ or } ac$ $P(M=cd) = .16; P(M=ac) = .84$	
$P(PUF \rightarrow a) = (.1596/.2541)(1/2) = .3140$	$P(PUF \rightarrow J) = 1/2$
$P(PUF \rightarrow c) = (.1596/.2541 + .0504/.2541)(1/2) + (.0441/.2541)(1) = .5868$	$P(PUF \rightarrow K) = 1/2$
$P(PUF \rightarrow d) = (.0504/.2541)(1/2) = .0992$	

SYSTEM 1 (Continued)

$$Y_{C=Q} = P(M=ac)P(ac \rightarrow a)P(RM \rightarrow a) \\ = (1)(1/2)(.4) = .200$$

$$X_{C=Q}^{PUF=S} = P(M=ac)P(ac \rightarrow a)P(PUF \rightarrow a) \\ = (1)(1/2)(.3140) = .157$$

$$Y_{C=Q}/X_{C=Q}^{PUF=S} = .200/.157 = 1.27$$

$$P(POC=Q) = P(M=ac)P(ac \rightarrow a)P(PUF \rightarrow a) \\ = (.84)(1/2)(.314) \\ = .132$$

$$Y_{C=S} = P(M=ac)[P(ac \rightarrow a)P(RM \rightarrow c) \\ + P(ac \rightarrow c)P(RM \rightarrow a, c, \text{ or } d)] \\ + P(M=cd)[P(cd \rightarrow d)P(RM \rightarrow c) \\ + P(cd \rightarrow c)P(RM \rightarrow a, c, \text{ or } d)] \\ = (.84)[(1/2)(.2) + (1/2)(.4+.2+.1)] \\ + (.16)[(1/2)(.2) + (1/2)(.4+.2+.1)] \\ = .450$$

$$X_{C=S}^{PUF=S} = P(M=ac)[P(ac \rightarrow a)P(PUF \rightarrow c) \\ + P(ac \rightarrow c)P(PUF \rightarrow a, c, \text{ or } d)] \\ + P(M=cd)[P(cd \rightarrow d)P(PUF \rightarrow c) \\ + P(cd \rightarrow c)P(PUF \rightarrow a, c, \text{ or } d)] \\ = (.84)[(1/2)(.5868) + (1/2)(1)] \\ + (.16)[(1/2)(.5868) + (1/2)(1)] \\ = .793$$

$$Y_{C=S}/X_{C=S}^{PUF=S} = .450/.793 = .567$$

$$P(POC=S) = X_{C=S}^{PUF=S} = .793$$

$$Y_{C=T} = P(M=ac)P(ac \rightarrow a)P(RM \rightarrow d) \\ + P(M=cd)P(cd \rightarrow d)[P(RM \rightarrow d) + P(RM \rightarrow a)] \\ = (.84)(1/2)(.1) + (.16)(1/2)(.1+.4) \\ = .0820$$

$$X_{C=T}^{PUF=S} = P(M=ac)P(ac \rightarrow a)P(PUF \rightarrow d) \\ + P(M=cd)P(cd \rightarrow d)[P(PUF \rightarrow d) + P(PUF \rightarrow a)] \\ = (.84)(1/2)(.0992) \\ + (.16)(1/2)(.0992 + .3140) \\ = .0747$$

$$Y_{C=T}/X_{C=T}^{PUF=S} = .0820/.0747 = 1.10$$

$$P(POC=T) = X_{C=T}^{PUF=S} = .075$$

SYSTEM 2 (Continued)

$$Y_{C=JJ} = P(JJ \rightarrow J)P(RM \rightarrow J) \\ = (1)(.3) = .300$$

$$X_{C=JJ}^{PUF=JK} = P(JJ \rightarrow J)P(PUF \rightarrow J) \\ = (1)(1/2) = .500$$

$$Y_{C=JJ}/X_{C=JJ}^{PUF=JK} = .3/.5 = .600$$

$$P(POC=JJ) = X_{C=JJ}^{PUF=JK} \\ = .500$$

$$Y_{C=JK} = P(JJ \rightarrow J)P(RM \rightarrow K) \\ = (1)(.7) = .700$$

$$X_{C=JK}^{PUF=JK} = P(JJ \rightarrow J)P(PUF \rightarrow K) \\ = (1)(1/2) = .500$$

$$Y_{C=JK}/X_{C=JK}^{PUF=JK} = .7/.5 = 1.40$$

$$P(POC=JK) = X_{C=JK}^{PUF=JK} = .500$$

COMBINATION

$Y/X_{PUF} = 1.53$ from Appendix G or J or K

POC	$P(POC) = P(POC_1)P(POC_2)$	$(Y/X) = (Y_1/X_1)(Y_2/X_2)$	IS $(Y/X) \geq (Y/X_{PU})$
Q and JJ	$(.132)(.500) = .066$	$(1.27)(.600) = .762$	NO
Q and JK	$(.132)(.500) = .066$	$(1.27)(1.40) = 1.78$	YES
S and JJ	$(.793)(.500) = .397$	$(.567)(.600) = .340$	NO
S and JK	$(.793)(.500) = .397$	$(.567)(1.40) = .794$	NO
T and JJ	$(.075)(.500) = .037$	$(1.10)(.600) = .660$	NO
T and JK	$(.075)(.500) = .037$	$(1.10)(1.40) = 1.53$	YES
	$P(\text{all POC}) = 1.000$		

$$B_{M/\min} = \sum_i^{\text{all } i \text{ for which } (Y/X)_i \geq (Y/X_{PU})} P_i(\text{POC}) = .103$$

APPENDIX L
ANALYSIS TABLE

ADDITIONAL DEFINITIONS

α	Putative Father's Risk
β	Mother's Risk
W_P	Likelihood of Paternity for a priori value P

INPUT INFORMATION		ANALYSIS TABLE					
ORDERED Y/X VALUES FROM APPENDICES J AND K	α INPUT FROM APPENDIX J	β INPUT FROM APPENDIX K	α	β	$W_{1/10}$	$W_{1/2}$	$W_{9/10}$
			0.0000	1.000			
.183	.0441		0.0441		.378	.845	.980
.230	.0294		0.0735		.326	.813	.975
.340		.397		1.000	.246	.746	.964
.367	.0378		0.1113		.232	.732	.961
.459	.0252		0.1365		.195	.685	.951
.660		.037		0.603	.144	.602	.932
.714	.0784		0.2149		.135	.583	.926
.762		.066		0.566	.127	.568	.922
.794		.397		0.500	.123	.557	.919
.840	.1176		0.3325		.117	.543	.915
1.43	.0672		0.3997		.072	.412	.863
1.53		.037		0.103	.068	.395	.855
1.68	.1008		0.5005		.062	.373	.843
1.78		.066		0.066	.059	.360	.835
1.97	.1617		0.6622		.053	.337	.820
3.95	<u>.1386</u>	<u></u>	0.8008		.027	.203	.696
	<u>Z = .8008</u>	<u>1.000</u>	1.0000	0.000			

The putative father's line is the one containing $W_{1/2} = .395$

APPENDIX M
HYPOTHESIS TEST PROCEDURE

1. Set a numerical value of α_L = the largest allowable risk of erroneously deciding that the putative father is the father.
2. Set a numerical value of β_L = the largest allowable risk of erroneously deciding that the putative father is not the father.
3. Set a numerical value of W_C = the lowest probability of paternity which implies that the putative father is the father.
4. Set a numerical value of P = the "a priori" probability of paternity for the putative father.
5. From the Analysis Table, obtain α_T = the risk in this particular test of erroneously deciding that the putative father is the father. This is the number at the intersection of the alpha column and the row of the smallest number in the appropriate W_p column which is not smaller than W_C ; if this location is vacant, use the number immediately above the intersection.
6. From the Analysis Table, obtain β_T = the risk in this particular test of erroneously deciding that the putative father is not the father. This is the number at the intersection of the beta column and the row of the largest number of the appropriate W_p column which is smaller than W_C ; if this location is vacant, use the number immediately below the intersection.
- 7a. If (1) α_T is not larger than α_L , (2) β_T is not larger than β_L , and (3) W_p of the putative father is as large or larger than W_C ; conclude that the putative father is the father.
- 7b. If (1) α_T is not larger than α_L , (2) β_T is not larger than β_L , and (3) W_p of the putative father is smaller than W_C ; conclude that the putative father is not the father.
- 7c. If (1) α_T is larger than α_L and/or (2) β_T is larger than β_L , recognize that the test results are inconclusive and more genetic tests are necessary to reduce α_T and/or β_T .

APPENDIX N

REASONS MORE GENETIC TESTS MAY BE REQUIRED TO MAKE THE ANALYSIS TABLE MORE DEFINITIVE

1. The court may be overly stringent and set α_L and/or β_L too low.
2. The court may set w_C high and thus cause β_T to be high.
3. The court may set w_C low and thus cause α_T to be high.
4. The court may set P high and thus cause α_T to be high. In the limit as P approaches 1.0, α_T approaches 1.0; this statistically states that a putative father who really is not the father has no chance if the court is certain that he is the father.
5. The court may set P low and thus cause β_T to be high. In the limit as P approaches 0.0, β_T approaches 1.0; this statistically states that a putative father who is the true father will not be judged to be the father if the court is certain that he is not the father.
6. Finally, the genetic tests may not be sufficiently definitive. Performing genetic tests using more marker systems will, on the average, tend to either (a) lower the alpha and beta risks of the tests and (b) drop the probability of paternity to exactly 0.0 if the putative father is not the father or raise the probability of paternity toward 1.0 if he is the father.

APPENDIX O
EXAMPLE OF STANDARD THREE PARTY PATERNITY DISPUTE ANALYSIS

GENETIC SYSTEMS ANALYZED: ABO and Kell

PHENOTYPES: Mother = O and kk, Child = A₂ and kK, Putative Father = A₂B and kK

RACIAL BACKGROUNDS: Mother = Mexican-American; Putative Father = Anglo

GEOGRAPHIC AREA: El Paso, Texas

<u>ALPHA</u>	<u>BETA</u>	<u>W_{1/10}</u>	<u>W_{1/3}</u>	<u>W_{1/2}</u>	<u>W_{2/3}</u>	<u>W_{9/10}</u>
0.00000	1.00					
	1.00	0.948	0.988	0.994	0.997	0.999
	0.88	0.948	0.988	0.994	0.997	0.999
0.00013		0.851	0.963	0.981	0.990	0.998
0.00068		0.795	0.946	0.972	0.986	0.997
0.00072		0.786	0.943	0.971	0.985	0.997
0.00848		0.741	0.928	0.963	0.981	0.996
0.04049		0.661	0.898	0.946	0.972	0.994
0.04057		0.653	0.894	0.944	0.971	0.993
0.04281		0.648	0.892	0.943	0.971	0.993
0.04312		0.621	0.881	0.937	0.967	0.993
	0.74	0.532	0.836	0.911	0.953	0.989
0.04758		0.485	0.809	0.894	0.944	0.987
	0.63	0.482	0.807	0.893	0.944	0.987
0.06552		0.451	0.787	0.881	0.937	0.985
	0.50	0.387	0.740	0.850	0.919	0.981
	0.38	0.386	0.739	0.850	0.919	0.981
	0.24	0.038	0.150	0.260	0.413	0.760
	0.13	0.031	0.126	0.224	0.366	0.722
1.00000	0.00	0.000	0.000	0.000	0.000	0.000

The putative Father's W_p scores are in the row with $W_{1/2} = 0.893$

APPENDIX P

FOUR POSSIBLE USES OF ANALYSIS TABLE OF APPENDIX N

INPUT DATA				INFORMATION FROM ANALYSIS TABLE			RESULT
α_L	β_L	w_C	P	α_T	β_T	w_{PUF}	
.10	.25	.900	2/3	.06552	.24	.944	$\alpha_T \neq \alpha_L$ $\beta_T \neq \beta_L$ $w_{PUF} \geq w_C$ Therefore $PUF = F$
—	—	—	—	—	—	—	—
.05	.25	.900	2/3	.06552	.24	.944	$\alpha_T > \alpha_L$ Therefore no decision without more information or altered risk
—	—	—	—	—	—	—	—
.01	.25	.900	1/10	.00000	.74	.482	$\beta_T > \beta_L$ Therefore no decision without more information or altered risk
—	—	—	—	—	—	—	—
.05	.75	.850	1/3	.04312	.74	.807	$\alpha_T \neq \alpha_L$ $\beta_T \neq \beta_L$ $w_{PUF} < w_C$ Therefore $PUF \neq F$

APPENDIX Q
SUMMARIES OF RESULTS IN FOUR ACTUAL PATERNITY DISPUTES

CASE	$\alpha_{PUF,min}$	$\beta_{M,min}$	$W_{1/10}$	$W_{1/3}$	$W_{1/2}$	$W_{2/3}$	$W_{9/10}$
1	.077	Not Available	.30	.66	.79	.88	.97
2	.018	.49	.41	.76	.86	.93	.98
3	.022	.85	.75	.93	.96	.98	.99
4	.084	.99	.31	.67	.80	.89	.97

For a hypothesis test to indicate a ruling against the putative father:

(1) P , the a priori or non-genetic probability of paternity of the putative father, and W_C , the critical or decision level of the posterior likelihood of paternity, must be set such that $W_P \geq W_C$; and

(2) α_T , the risk of falsely rejecting the assumption that the putative father is not the father, must be set such that $\alpha_T > \alpha_{PUF,min}$.

For a hypothesis test to indicate a ruling against the mother:

(1) P and W_C must be set such that $W_P < W_C$; and

(2) β_T , the risk of falsely accepting the assumption that the putative father is not the father must be set such that $\beta_T > \beta_{M,min}$.

AN ALGORITHM FOR TRILATERATION

James T. Hall
Atmospheric Sciences Laboratory
US Army Electronics Research and Development Command
White Sands Missile Range, New Mexico 88002

ABSTRACT

A vector algorithm is presented for determining spatial position of one or more objects with range-only information from three noncollinear stations. An error analysis shows the dependence of spatial position uncertainty on the geometry of the measurement array.

INTRODUCTION

The operations associated with the objectives of a National Range produce technically complex problems. A prime example is the requirement for time-space-position-information (TSPI) which is acquired by instrumentation radars. This information is required if weapon systems effectiveness is to be evaluated. The complexity of an evaluation is exemplified by the fact that these weapons are referred to as "smart" or "dumb" depending largely on their inherent ability to maneuver themselves to a predetermined target which may also be moving.

Radar likewise have moved from the category of dumb to smart, having learned the basic laws of physics and acquired the ability to selectively filter incoming information. This filtering is based on prior knowledge of the physical constraints of the target involved.

The computer also points and drives the antenna based on the physical laws of motion and uses the normal radar tracking signal to periodically verify that its past prediction of the current target position is accurate. This is a giant step in the basic philosophy of radar operation; it does nevertheless, require that the antenna be pointed. This single property, i.e., pointing the antenna, produces virtually all the requirements and problems with precision pedestals, calibration, and some of the propagation errors including refractive bending.

This report is concerned with an analytical scheme which uses three noncollinear stations for range-only measurements on any number of targets to provide TSPI. There is, of course, no pointing antenna required in such a system. The type of station to perform this task has come to be known as distance measurement equipment (DME) stations. This is not a new idea, but one that has become workable with the advent of the high-speed computer. The principle and existence of pointing and ranging radars may never be replaced in most applications; there are, however, many applications which require great accuracy in spatial determination of one or more objects. For these

applications, pointing radars become very expensive, and calibration procedures and checks consume increasingly more time.

The early DME analytical methodology and analysis was done in the 1960's [1, 2, 3]. During the early 1970's IBM, Cubic and General Dynamics developed systems which used these analytic procedures to perform a variety of tracking and position location problems. Current information on these systems can be acquired from those companies.

Most analytical developments have followed the form given here with the exception of the angle dependence given in equation (21) and its resulting error analysis. This vector solution is believed to offer computational advantages and insight into the geometric needs of the triad.

METHOD I (INTERSECTING SPHERES)

The first and most obvious analytic method of determining the spatial position, given three ranges from three known points, is the simultaneous solution of three distance equations. Each distance equation can be written to define a sphere of radius R_i , $i = 1, 2, 3$, equal to the measured ranges from three known points to the unknown point (x, y, z) . Three intersecting spheres define two points separated in this case by the plane of observation; therefore, if the known points which define this plane are contained on the earth's surface, only one point will be of real concern, i.e., that above the plane of observation.

The location of the plane of observation is arbitrary and alternations of the above conclusion are apparent. The only analytic requirement is that the three known points not be collinear.

Let the observation points have locations $(0, 0, 0)_1$, $(e, 0, 0)_2$, and $(g, h, 0)_3$, where the subscripts denote station number. The three equations are then:

$$R_1^2 = x^2 + y^2 + z^2 \quad (1)$$

$$R_2^2 = (x - e)^2 + y^2 + z^2 \quad (2)$$

$$R_3^2 = (x - g)^2 + (y - h)^2 + z^2 \quad (3)$$

with resultant solutions for (x, y, z) as:

$$x = \frac{1}{2e} (R_1^2 - R_2^2 + e^2) \quad (4)$$

$$y = \frac{1}{2h} (R_2^2 - R_3^2 + 2x(e - g) + h^2 + g^2 - e^2) \quad (5)$$

and

$$z = (R_1^2 - x^2 - y^2)^{1/2} \quad (6)$$

Observe that in choosing the locations of the stations, (1) was made the origin, and the line connecting station (1) and (2) was made the x-axis of an orthogonal coordinate system. The three stations also define the plane of $z = 0$. This, or a similar selection greatly simplifies the solution while remaining semigeneral, i.e., this choice can always be made when accompanied by a proper coordinate transform.

The requirement for nonlinearity of the three stations is apparent in the solution since neither e nor h can be zero. There are two conditions of station geometry which permit equation (5) to have the same number of terms as well as the form of equation (4); this will be seen to be important in the section on Error Analysis.

The first condition is: if $g = e$, then equation (5) may be written as

$$y = \frac{1}{2h} (R_2^2 - R_3^2 + h^2) \quad (5a)$$

and the second condition is: if $g = o$, equation (5) will be

$$y = \frac{1}{2h} (R_1^2 - R_3^2 + h^2) \quad (5b)$$

These conditions require only that the choice of station positions constitute an orthogonal array.

METHOD II (VECTOR)

This method is somewhat more complex in appearance, but for a general application of DME data to acquire a solution for spatial position it is remarkably simple when compared to that required using Method I. It also gives a better insight into the geometry requirements of trilateration, thus providing knowledge necessary to achieve the measurement configuration for best accuracy with a given number of stations. It provides directly a vector solution which lends itself to "tracking" problems very well. For these and other reasons which become apparent, this method constitutes the major effort of this report.

Given three noncollinear stations F, M, and C with three correspondingly measured ranges R_1 , R_2 , and R_3 to any unknown single point, one can derive the position vector \vec{P} in terms of two base line vectors \vec{M} and \vec{C} and their cross-product as follows (Figure 1).

The vector \vec{P} can always be written as a linear combination of the vectors \vec{M} , \vec{C} , and $(\vec{M} \times \vec{C})$ as:

$$\vec{P} = a\vec{M} + b\vec{C} + d(\vec{M} \times \vec{C}). \quad (7)$$

where $\vec{M} \times \vec{C}$ denotes the cross-product of \vec{M} and \vec{C} .

The vectors \vec{M} and \vec{C} are defined by the geometry of the station array; therefore, the problem is to find the set of scalars (a, b, d).

Since

$$R_1 = |\vec{P}|,$$

and $R_2 = |\vec{P} - \vec{M}|$

$$R_3 = |\vec{P} - \vec{C}|$$

squaring these gives:

$$R_1^2 = \vec{P} \cdot \vec{P} \quad (8)$$

$$R_2^2 = \vec{P} \cdot \vec{P} - 2\vec{P} \cdot \vec{M} + \vec{M} \cdot \vec{M} \quad (9)$$

$$R_3^2 = \vec{P} \cdot \vec{P} - 2\vec{P} \cdot \vec{C} + \vec{C} \cdot \vec{C} \quad (10)$$

If, in equation (9) we let $P \cdot M = \alpha$ and equation (10) $P \cdot C = \beta$, they can be rewritten using equation (8) as:

$$R_2^2 = R_1^2 - 2\alpha + |\vec{M}|^2 \quad (9a)$$

and

$$R_3^2 = R_1^2 - 2\beta + |\vec{C}|^2 \quad (9b)$$

Solving for α and β gives:

$$\alpha = \frac{R_1^2 - R_2^2 + |\vec{M}|^2}{2} \quad (11)$$

and

$$\beta = \frac{R_1^2 - R_3^2 + |\vec{C}|^2}{2} \quad (12)$$

These scalars, α and β , are quantities which will be used extensively in this method. They are geometrically the projection of the position vector \vec{P} onto the base line vectors \vec{M} and \vec{C} respectively. Taking the dot or scalar product of equation (7) with \vec{M} gives:

$$\vec{P} \cdot \vec{M} = a\vec{M} \cdot \vec{M} + b\vec{C} \cdot \vec{M} + d(\vec{M} \times \vec{C}) \cdot \vec{M}$$

or

$$\alpha = a|\vec{M}|^2 + b\vec{C} \cdot \vec{M}. \quad (13)$$

Dotting equation (7) again with \vec{C} gives:

$$\vec{P} \cdot \vec{C} = a\vec{M} \cdot \vec{C} + b\vec{C} \cdot \vec{C} + d(\vec{M} \times \vec{C}) \cdot \vec{C}$$

or

$$\beta = a\vec{C} \cdot \vec{M} + b|\vec{C}|^2 \quad (14)$$

Solving equations (13) and (14) for a and b results in:

$$a = \frac{\begin{vmatrix} \alpha & \vec{C} \cdot \vec{M} \\ \beta & \vec{C} \cdot \vec{C} \end{vmatrix}}{\begin{vmatrix} \vec{M} \cdot \vec{M} & \vec{C} \cdot \vec{M} \\ \vec{M} \cdot \vec{C} & \vec{C} \cdot \vec{C} \end{vmatrix}} = \frac{\begin{vmatrix} \alpha & \vec{C} \cdot \vec{M} \\ \beta & \vec{C} \cdot \vec{C} \end{vmatrix}}{|\vec{M} \times \vec{C}|^2} = \frac{\alpha|\vec{C}|^2 - \beta\vec{C} \cdot \vec{M}}{|\vec{M} \times \vec{C}|^2}$$

and

$$b = \frac{\begin{vmatrix} \vec{M} \cdot \vec{M} & \alpha \\ \vec{C} \cdot \vec{M} & \beta \end{vmatrix}}{|\vec{M} \times \vec{C}|^2} = \frac{\beta|\vec{M}|^2 - \alpha\vec{C} \cdot \vec{M}}{|\vec{M} \times \vec{C}|^2}$$

The final scalar is found from

$$\vec{P} \cdot (\vec{M} \times \vec{C}) = d|\vec{M} \times \vec{C}|^2 \quad (15)$$

and

$$\vec{p} \cdot \vec{p} = a\vec{M} \cdot \vec{P} + b\vec{C} \cdot \vec{P} + d(\vec{M} \times \vec{C}) \cdot \vec{P} \quad (16)$$

combining (15) and (16) as:

$$R_1^2 = aa + bb + d^2 |\vec{M} \times \vec{C}|^2$$

or

$$d^2 = \frac{R_1^2 - aa - bb}{|\vec{M} \times \vec{C}|^2}.$$

Finally

$$d = \frac{(R_1^2 - aa - bb)^{\frac{1}{2}}}{|\vec{M} \times \vec{C}|}.$$

Therefore, the expressions for the scalars are:

$$a = \frac{\alpha |\vec{C}|^2 - \beta (\vec{C} \cdot \vec{M})}{\psi^2}, \text{ where } \psi = |\vec{M} \times \vec{C}|, \quad (17)$$

$$b = \frac{\beta |\vec{M}|^2 - \alpha (\vec{C} \cdot \vec{M})}{\psi^2} \quad (18)$$

and

$$d = \frac{(R_1^2 - aa - bb)^{\frac{1}{2}}}{\psi}. \quad (19)$$

The position vector may now be written as

$$\begin{aligned} \vec{p} = & \frac{\alpha |\vec{C}|^2 - \beta (\vec{C} \cdot \vec{M})}{\psi^2} \vec{M} + \frac{\beta |\vec{M}|^2 - \alpha (\vec{C} \cdot \vec{M})}{\psi^2} \vec{C} \\ & + \frac{(R_1^2 - aa - bb)^{\frac{1}{2}} (\vec{M} \times \vec{C})}{\psi}. \end{aligned} \quad (20)$$

By noting that:

$$\frac{\vec{M}}{|\vec{M}|} = \hat{i}, \frac{\vec{C}}{|\vec{C}|} = \hat{j} \text{ and } \frac{(\vec{M} \times \vec{C})}{\psi} = \hat{k}$$

Where \hat{i} , \hat{j} , \hat{k} are unit vectors in the \vec{M} , \vec{C} , $(\vec{M} \times \vec{C})$ directions respectively and using the expressions for the dot and cross-products, \vec{P} can be rewritten as:

$$\begin{aligned} \vec{P} &= \left(\frac{\alpha}{|\vec{M}|} - \frac{\beta}{|\vec{C}|} \cos\theta \right) \csc^2\theta \hat{i} + \left(\frac{\beta}{|\vec{C}|} - \frac{\alpha}{|\vec{M}|} \cos\theta \right) \csc^2\theta \hat{j} \\ &+ \left[R_1^2 - \left(\frac{\alpha}{|\vec{M}|} \right)^2 + \left(\frac{\beta}{|\vec{C}|} \right)^2 \right] \csc^2\theta + \frac{2\alpha\beta\cot\theta}{\psi} \right]^{\frac{1}{2}} \hat{k}, \end{aligned} \quad (21)$$

where θ is the angle between \vec{M} and \vec{C} .

Since the magnitudes of \vec{M} and \vec{C} , that is $|\vec{M}|$, $|\vec{C}|$, and θ are fixed for a given station geometry, the only quantities to be recurrently calculated are α and β as given by equations (11) and (12). For the general case the simplicity of this procedures as compared to the intersecting spheres method is apparent.

For the special but desirable case that i normal to j , equation (21), reduces to:

$$\vec{P} = \frac{\alpha}{|\vec{M}|} \hat{i} + \frac{\beta}{|\vec{C}|} \hat{j} + \left(R_1^2 - \frac{\alpha^2}{|\vec{M}|} - \frac{\beta^2}{|\vec{C}|} \right)^{\frac{1}{2}} \hat{k} \quad (21a)$$

which gives directly, if desired, the Cartesian coordinates in the i , j , k system as:

$$\vec{P} \cdot \hat{i} = x = \frac{\alpha}{|\vec{M}|}, \quad \vec{P} \cdot \hat{j} = y = \frac{\beta}{|\vec{C}|}$$

and

$$\vec{P} \cdot \hat{k} = z = \left(R_1^2 - \frac{\alpha^2}{|\vec{M}|} - \frac{\beta^2}{|\vec{C}|} \right)^{\frac{1}{2}}.$$

The correlation between these values and those of equations (4), (5b), and (6) is obvious.

In the general case, the transform to a chosen set of orthogonal coordinates, \hat{i}^* , \hat{j}^* , \hat{k}^* , is given by:

$$\begin{aligned}\hat{i}^*\cdot\hat{i} &= a_{11}, \hat{i}^*\cdot\hat{j} = a_{12}, \hat{i}^*\cdot\hat{k} = a_{13} \\ \hat{j}^*\cdot\hat{i} &= a_{21}, \hat{j}^*\cdot\hat{j} = a_{22}, \hat{j}^*\cdot\hat{k} = a_{23} \\ \hat{k}^*\cdot\hat{i} &= a_{31}, \hat{k}^*\cdot\hat{j} = a_{32}, \hat{k}^*\cdot\hat{k} = a_{33}\end{aligned}$$

i.e.,

$$v_n^* = \sum_{l=1}^3 A_{nl} v_l, \quad n = 1, 2, 3 \quad (22)$$

where v_l are the vector components of \vec{P} and v^* are Cartesian coordinates $x = v_1^*$, $y = v_2^*$, and $z = v_3^*$. The A_{nl} are the direction cosines of the transform.

ERROR ANALYSIS

The error analysis of Method I and Vector Method will be accomplished using the propagation of error principle which states for a quantity(s) which is a function of two or more measured quantities, i.e., $s = s(p, q, u)$, the uncertainty in s , δs , is:

$$\delta s = \left[\left(\frac{\partial s}{\partial p} \right)^2 (\delta p)^2 + \left(\frac{\partial s}{\partial q} \right)^2 (\delta q)^2 + \left(\frac{\partial s}{\partial u} \right)^2 (\delta u)^2 \right]^{\frac{1}{2}}$$

where δp , δq , and δu are the uncertainties in the measured quantities. This expression assumes a symmetric distribution of the measurement uncertainties, i.e., positive and negative errors are equally probable in the measured ranges.

Method I

Using equation (4), where $x = x(R_1, R_2)$, gives for the partial derivatives:

$$\frac{\partial x}{\partial R_1} = \frac{R_1}{e}, \quad \frac{\partial x}{\partial R_2} = -\frac{R_2}{e}, \quad \text{and the resulting uncertainty in } x, \delta x, \text{ is:}$$

$$\delta x = \frac{1}{e} [R_1^2 (\delta R_1)^2 + R_2^2 (\delta R_2)^2]^{\frac{1}{2}}.$$

Letting $t_{12} = R_1^2 (\delta R_1)^2 + R_2^2 (\delta R_2)^2$ this expression will be rewritten as:

$$\delta x = \frac{(t_{12})^{\frac{1}{2}}}{e}.$$

From equation (5), where $y = y(R_2, R_3, x)$ and the resulting partial derivatives are:

$$\frac{\partial y}{\partial R_2} = \frac{R_2}{h}, \quad \frac{\partial y}{\partial R_3} = -\frac{R_3}{h} \quad \text{and} \quad \frac{\partial y}{\partial x} = \frac{(e - g)}{h}$$

gives for the uncertainty in y , $\delta y = \left[\left(\frac{R_2}{h} \right)^2 (\delta R_2)^2 + \left(\frac{R_3}{h} \right)^2 (\delta R_3)^2 + \frac{(e - g)^2}{h^2} (\delta x)^2 \right]^{\frac{1}{2}}$.
where again $\delta x = \frac{(t_{12})^{\frac{1}{2}}}{e}$ giving:

$$\delta y = \frac{1}{h} [t_{23} + t_{12} (1 - 2(g/e) + (g/e)^2)]^{\frac{1}{2}}.$$

This appears to be a curious result since if $g = 0$, $\delta y = \frac{1}{h} [t_{23} + t_{12}]^{\frac{1}{2}}$,
but if $g = e$, $\delta y = \frac{(t_{23})^{\frac{1}{2}}}{h}$.

This result, however, indicates that for minimizing the uncertainty in any one component, x , y , or z , that two of the stations should be aligned in that direction. This is also apparent in the subscript change on the value of R , in equation (5b) from that in equation (5a). Pursuing this philosophy would result in the addition of a fourth station to optimize the uncertainty of a spatial position. The location of this fourth station would necessarily be directly above one of the other ground-based stations so that the line formed by these two would be normal to the plane $z = 0$, i.e., that formed by the three ground-based stations. That is, of course, not an absolute requirement but one to optimize, i.e., minimize the spatial uncertainty.

This argument also implies that you should expect the uncertainty in the z component to be the largest of the three components when using only ground-based stations. This will be shown, in general, to be true since:
 $z = z(R_1, x, y)$ from equation (6)

$$\frac{\partial z}{\partial R_1} = \frac{R_1}{z}, \quad \frac{\partial z}{\partial x} = \frac{-x}{z}, \quad \frac{\partial z}{\partial y} = -y/z \quad \text{and}$$

$$\delta z = \frac{1}{z} [R_1^2 (\delta R_1)^2 + x^2 (\delta x)^2 + y^2 (\delta y)^2]^{\frac{1}{2}}.$$

A significant difference in this expression for the uncertainty in z is the fact that it is inversely dependent on the value of z . Producing the condition that as z approaches zero ($z \rightarrow 0$) the uncertainty in z approaches infinity ($\delta z \rightarrow \infty$).

Minimizing the uncertainty in z would require small values in both x and y reaching a limit when $x = 0$, $y = 0$ and resulting in

$$\delta z = \frac{R_1}{z}(\delta R_1)$$

but for this condition $R_1 = z$; therefore, $\delta z = \delta R_1$. Since this is the smallest uncertainty that can be obtained in any one component, δz can obviously range from this value to the largest and will in general, be greater than δx or δy .

Vector Method

Using the uncertainties in α and β the uncertainty in the scalars a , b , and d can be evaluated. These then give the uncertainty in the position vector \vec{P} , i.e., $\delta\vec{P}$.

From equations (11) and (12), the $\delta a = t_{12}^{\frac{1}{2}}$ and $\delta b = t_{13}^{\frac{1}{2}}$. The first scalar $A = a(\alpha, \beta)$ and its partial derivatives are:

$$\frac{\partial a}{\partial \alpha} = \frac{\csc^2 \theta}{|\vec{M}|^2} \text{ and } \frac{\partial a}{\partial \beta} = \frac{-\cos \theta \csc^2 \theta}{|\vec{M}| |\vec{C}|} .$$

Since

$$(\vec{P} + \delta\vec{P}) \cdot \hat{i} = (a + \delta a) |\vec{M}| \hat{i} \cdot \hat{i}$$

or

$$\delta\vec{P} \cdot \hat{i} = \delta a |\vec{M}|$$

i.e., the uncertainty in the i th component is:

$$\delta\vec{P}_i = \left[\frac{\csc^4 \theta}{|\vec{M}|^4} t_{12} + \frac{\cos^2 \theta \csc^4 \theta}{|\vec{M}|^2 |\vec{C}|^2} t_{13} \right]^{\frac{1}{2}} |\vec{M}|$$

and finally rewritten as:

$$\delta\vec{P}_i = \csc^2 \theta \left(\frac{t_{12}}{|\vec{M}|^2} + \frac{\cos^2 \theta t_{13}}{|\vec{C}|^2} \right)^{\frac{1}{2}} .$$

There is of course, one very obvious difference in this uncertainty, i.e., it is a function of the angles θ . This expression and the ones to follow show the strong dependence of the triad geometry on the resultant accuracy of spatial position determination or position uncertainty.

Again the need for noncollinear stations, since if $\theta = 0^\circ$, $\delta \vec{P}_i = \infty$. It is also apparent that to minimize $\delta \vec{P}_1$, θ must be equal to $\pi/2$. Using the same procedure gives for the uncertainty in b ,

$$\delta b = \frac{\csc^2 \theta}{|\vec{C}|} \left(\frac{t_{13}}{|\vec{C}|^2} + \frac{\cos^2 \theta t_{12}}{|\vec{M}|^2} \right)^{\frac{1}{2}}$$

and since

$$\delta \vec{P} \cdot \hat{j} = \delta b |\vec{C}| \hat{j} \cdot \hat{j},$$

the uncertainty in the j th component of \vec{P} is

$$\delta \vec{P}_j = \csc^2 \theta \left(\frac{t_{13}}{|\vec{C}|^2} + \frac{\cos^2 \theta t_{12}}{|\vec{M}|^2} \right)^{\frac{1}{2}}.$$

The final scalar d is a little more complicated since

$$d = d(R_1, \alpha, \beta)$$

The partial derivatives are:

$$\frac{\partial d}{\partial R_1} = \frac{R_1}{\psi(R_1^2 - a\alpha - b\beta)^{\frac{1}{2}}}$$

$$\frac{\partial d}{\partial \alpha} = \frac{-a}{2\psi(R_1^2 - a\alpha - b\beta)^{\frac{1}{2}}}$$

and

$$\frac{\partial d}{\partial \beta} = \frac{-b}{2\psi(R_1^2 - a\alpha - b\beta)^{\frac{1}{2}}}.$$

Letting $\tau = R_1^2 - a\alpha - b\beta$, the uncertainty in d may be written as

$$\delta d = \frac{1}{2\psi\tau^{\frac{1}{2}}} [4R_1^2(\delta R_1)^2 + a^2 t_{12} + b^2 t_{13}]^{\frac{1}{2}}.$$

Rewriting this equation in terms of α , β , and θ with consideration given to the fact that

$$\delta \vec{P}_k = (\delta d)\psi$$

results in the kth component of \vec{P} to be

$$\begin{aligned}\delta\vec{P}_k &= \frac{1}{2\tau} \left[4R_1^2 (\delta R_1)^2 + \frac{\csc^4 \theta}{|\vec{M}|^2} \left(\frac{\alpha^2}{|\vec{M}|^2} - \frac{2\alpha\beta\cos\theta}{|\vec{M}||\vec{C}|} + \frac{\beta^2\cos^2\theta}{|\vec{C}|^2} \right) t_{12} \right. \\ &\quad \left. + \frac{\csc^4 \theta}{|\vec{C}|^2} \left(\frac{\beta^2}{|\vec{C}|^2} - \frac{2\beta\alpha\cos\theta}{|\vec{M}||\vec{C}|} + \frac{\alpha^2\cos^2\theta}{|\vec{M}|^2} \right) t_{13} \right]^{\frac{1}{2}}.\end{aligned}$$

This expression is greatly simplified if $\theta = \pi/2$. For this desirable geometry:

$$\delta\vec{P}_k = \frac{1}{2\tau} \left[4R_1^2 (\delta R_1)^2 + \frac{\alpha^2 t_{12}}{|\vec{M}|^4} + \frac{\beta^2 t_{13}}{|\vec{C}|^4} \right]^{\frac{1}{2}}.$$

DISCUSSION

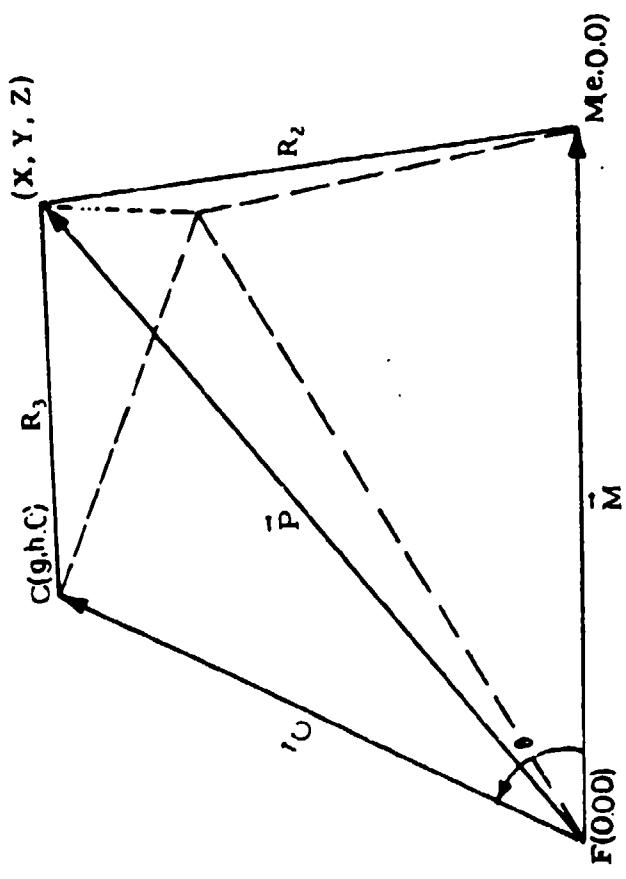
If the methodology of tracking systems move closer to computer controlled pointing antennas or nonpointing systems implicit in this report, software methods will become increasingly critical with regard to quality of solution. The ultimate accuracy of this solution is a function of target position with respect to triad location and geometry [4]. Although only three stations are required for a position solution, in practice many stations would be deployed. The analytical method presented here permits the selection of those triads which optimize the accuracy. To illustrate this method, baseline magnitudes were selected, i.e., $|\vec{C}|$ and $|\vec{M}|$, uncertainties in the measured ranges δR_1 were set, and a spatial point (x, y, z) determined.

The angle θ between \vec{M} and \vec{C} was varied to determine the effect of triad geometry on accuracy. Figures 2 through 4 are plots of position uncertainty versus θ , where the units of uncertainty are the same as those of the base lines. The uncertainties in the x and y components, δ_i and δ_j respectively, have the same values in the vector method and may be indicated as either one on the plots.

Observations and conclusions have been made in the body of this report where they seemed the most appropriate and will not be repeated.

REFERENCES

1. Armijo, L., "Determination of Trajectories Using Range Data from Three Noncollinear Radar Stations," Technical Memo 766, US Army Signal Missile Support Agency, White Sands Missile Range, NM, September 1960.
2. Alvarez, E. S., "An Analysis of Trilateration Systems for Near Launch Tracking," RISO-2-65, US Army Test and Evaluation Command, Range Instrumentation Systems Office, White Sands Missile Range, March 1965.
3. Green, R. E., "Effects of Geodetic Measurement Errors on Trajectory Data," RE-S-66-1, Systems Development Directorate, Deputy for National Range Engineering, White Sands Missile Range, NM, June 1966.
4. Sivazlian, B. D. and R. E. Green, "Optimal Instrument Siting for Stationary Target Tracking," Proceedings of the IEEE National Aerospace and Electronics Conference, 1975.



Triad geometry

FIG. 1

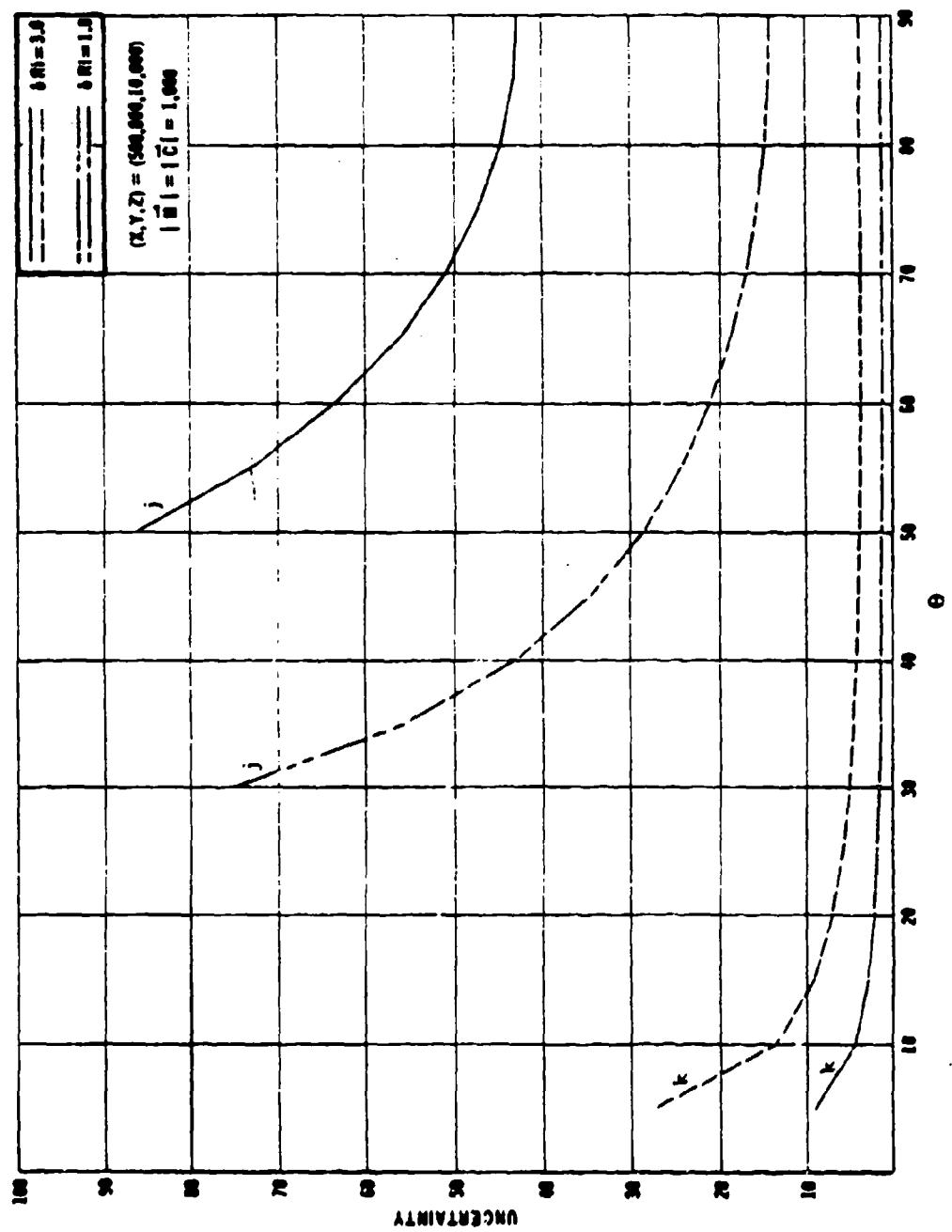
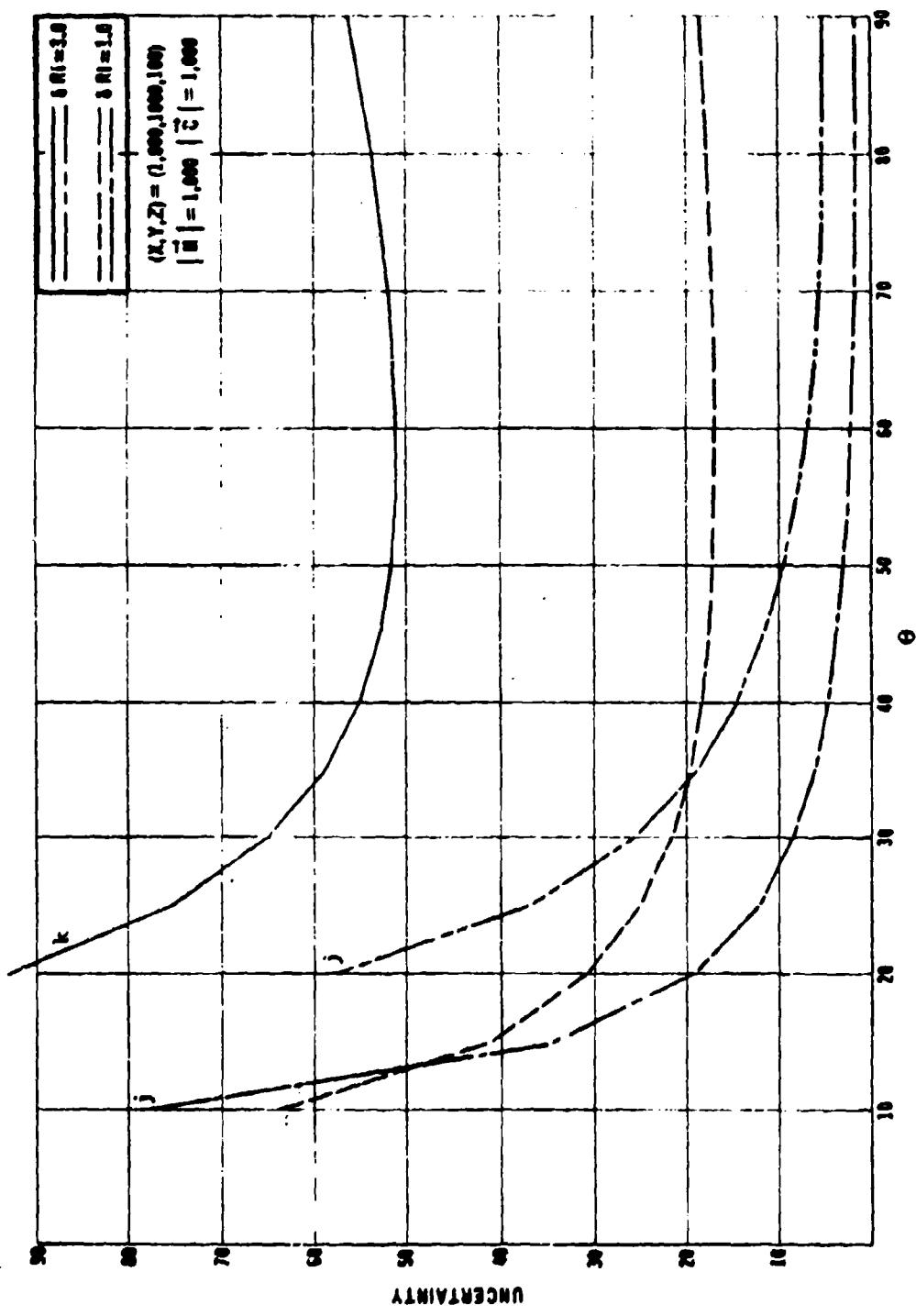


FIG. 2

FIG. 3



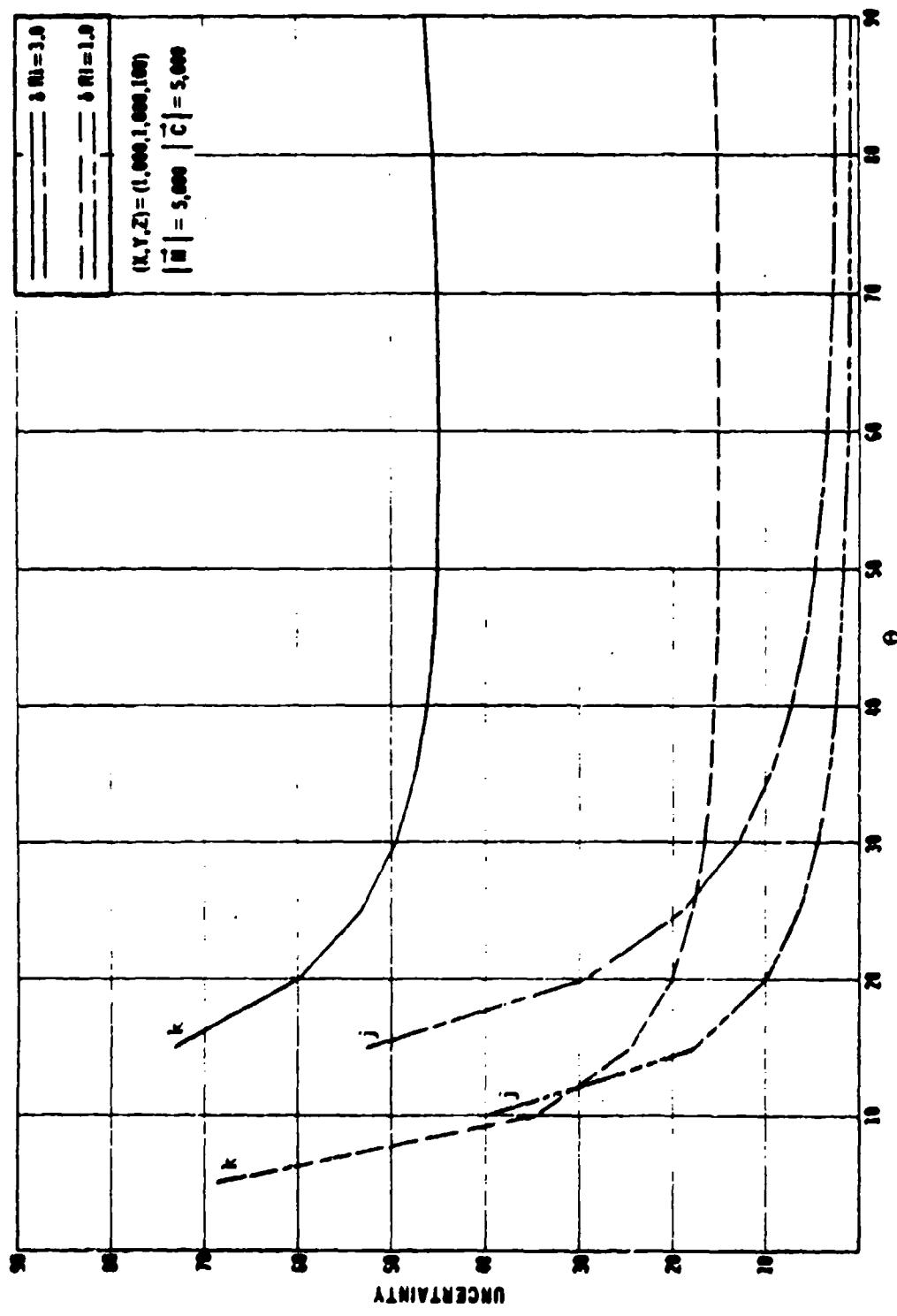


FIG. 4

SOCIAL SCIENTIST TECHNIQUE
THE CATALYST FOR OBTAINING
OBJECTIVITY FROM SUBJECTIVITY

RONALD L. JOHNSON
US Army Mobility Equipment Research and
Development Command, Ft. Belvoir, Virginia 22060

ABSTRACT

This study involved statistical efforts by a social scientist in the field of camouflage development. Statistical techniques were used to transform subjective data into objective results as a basis of statistical inferences. The specific task was to interpret and coordinate subjective, verbal data generated by 94 pairs of image interpreters viewing aerial film strips of tactically emplaced military equipment. Target detections were accomplished on film scaled 1:9,400. Target identification, if a correct detection was made, was accomplished on film scaled 1:5,000. Visual cues leading to target detection and identification were tabulated. Using the statistical technique of minimum contrasts, visual cues for detection and identification were objectively differentiated at the significance level of $\alpha = 0.025$.

1.0 INTRODUCTION

The social science techniques, when applied in a research setting are becoming more valuable as requirements for quantitative data increase. This is particularly true when the raw data is subjective in nature. This report investigated visual cues for detecting and identifying a ground-to-air weapon system. The visual cues were obtained by the social science technique of the open-ended interview of the test subjects. This paper describes the method of obtaining this subjective data, and the processing of it into objective definitive results.

2.0 TEST SITE AND EQUIPMENT

2.1 TEST SITE

The test site was at Fort Lewis, Washington. The exact area was referred to as the Merrill Drop Zone. The drop zone was approximately 1.75 x 0.65 km. in size and was located in a temperate climate zone. The existing vegetation was comprised of grasses, shrubs, and pine trees.

2.2 TEST EQUIPMENT

The test equipment consisted of a single tracked vehicle. It was camouflage pattern painted in a woodland (US/Europe) fall/winter color scheme consisting of approximately 45% each Forest Green and Field Drab, and 5% each Sand and Black colors.

3.0 TEST IMAGERY

The test item was tactically sited and photographed, using 9 inch strip color, aerial film, at scales of 1:5,000 and 1:9,400 each with 60% forward overlap. The 1:9,400 test strip contained 15 frames of imagery; the strip scaled 1:5,000 contained 5 frames. Each frame scaled 1:9,400 covered a ground area of approximately one square km. The camera used was a ZEISS RMK-15-23 mounted in an Aero Commander aircraft under contract to MERADCOM.

4.0 TEST PROCEDURES

The cut, strip imagery were given to ninety-four (94) pairs of operational image interpreting (II's). The term "operational" is used to indicate the subjects carry an II military occupational specialty code and are assigned in II positions. All had received service training in interpretation methods and procedures. The II's were read a briefing in which they were told to look for possible military equipment on the strip of film scaled 1:9,400. Detailed item analysis would be accomplished on the film scaled 1:5,000. The social scientist conducted an in-depth open-ended interview. From each II team, statements were independently extracted as to the physical features about the surrounding and the target that enabled detection and identification. Their responses, known as visual cues, were then tabulated to form a frequency distribution. These visual cues, subjective in nature, were then statistically analyzed using the method of minimum contrast.^{2,3} By employing this method, the visual cues were objectively ranked ($\alpha=0.025$) as to first order, second order, etc. This data is presented in the next section.

5.0 RESULTS

The refined visual cues for target detection and identification are found in tables one and two respectively.

TABLE I

Significant Differences Between Visual Cues For Target Detection.

	A	B	C	D	E	F	G	H	I	J	K	Frequency
A												83
B	XX											71
C	XX	XX										38
D	XX	XX										37
E	XX	XX	XX	XX	XX							21
F	XX	XX	XX	XX	XX							12
G	XX	XX	XX	XX	XX	XX						7
H	XX					2						
I	XX					1						
J	XX					1						
K	XX					1						

Cell Size 92 (Two II teams did not detect the target) xx-Significant $\alpha = 0.025$.

Key

- A-Target appears geometric (rectangular)
- B-Target appears lighter in color
- C-Track activity seen
- D-Target is set in the open
- E-Shadow seen
- F-Reflection seen
- G-Turret seen
- H-Target has height
- I-Gun barrel seen
- J-Target placed next to road
- K-Target appears smooth in texture

Conclude that the visual cues are ordered as follows:

- First Order - A
- Second Order - B
- Third Order - C and D
- Fourth Order - E and F
- Fifth Order - G, H, I, J, and K

TABLE 2
Significant Difference Between Visual Cues for Target Identification.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Frequenc.
A																61
B	XX															46
C	XX															42
D	XX															33
E	XX	XX														32
F	XX	XX	XX													25
G	XX	XX	XX													24
H	XX	XX	XX	XX	XX											16
I	XX									12						
J	XX								8							
K	XX								8							
L	XX						5									
M	XX					2										
N	XX				2											
O	XX			1												

Cell Size 77 (Fifteen II teams that made a correct detection did not identify the target) xx-Significant $\alpha=0.025$.

Key

- A-Primary lower unit and Environment Control unit (with space between) seen
- B-Headlight covers seen
- C-Sureveillance radar seen
- D-Front Part of target boat shaped (semi-pointed)
- E-Missile launchers seen
- F-Target has a stepped front
- G-Target has a length-to-width ratio
- H-Tracking radar seen
- I-Target has a cluttered turret
- J-Target has a height-to-width ratio
- K-Turret position is toward the back of the target
- L-Sloped front end
- M-Target has flat top
- N-Driver hatch seen
- O-Side edge seen

Conclude that the visual cues are ordered as follows:

- First Order - A
- Second Order - B and C
- Third Order - D, E, F, and G
- Fourth Order - H and I
- Fifth Order - J, K, L, M, N, and O

6.0 DISCUSSION

Tables one and two of the result section have shown how social science techniques have transferred a hodgepodge of verbal responses into a quantitatively ranked ($\alpha=0.025$) order of visual cues for target detection and identification. The camoufleur now has an statistical base from which he can make scientific decisions as to where to address his efforts. The addition of disrupters and or nets to the test equipment would effect the first order visual cue for both detection and identification. The processing of the subjective data into objective decision making data will save much time and money! This will be accomplished by identifying type and position placement of the prototype camouflage.

7.0 SUMMARY

Color aerial imagery containing a ground-to-air military weapon system were given to 94 pairs of II to determine visual cues for detection and identification. These subjective, verbal, responses were obtained by a social scientist. Through the application of social science techniques, the subjective data was transformed into objective decision making data for camouflage requirements. The application of social science techniques has therefore served as the catalyst for obtaining objectivity from subjectivity.

REFERENCES

1. Naval Reconnaissance and Technical Support Center, Image Interpretation Handbook vol 1, TM 30-245, December 1967.
2. Dixon and Massey, Introduction to Statistical Analysis, 3rd Edition, Mc Graw Hill, 1969.
3. Natrella, Mary G., Experimental Statistics, National Bureau of Standards, Handbook 91, US Department of Commerce, Washington, D.C. 1966.

SOME ASPECTS OF ENGINEERING TIME SERIES ANALYSIS

Victor Solo
Department of Statistics, Harvard University
Cambridge, Massachusetts 02138

0. Introduction

There are a number of characteristics of the type of time series problems met by civil, mechanical, electrical engineers that distinguish them from say those of econometrics, business. Firstly, there is usually much more data available - the engineer has great control over the choice of sampling interval. Secondly, there is often an interest in calculations performed in real time (rather than at leisure). This is especially so in adaptive control and forecasting (viz. of electric power demand). Thirdly, since engineers deal with physical processes many engineering time series are distributed (spatially) viz. the determination of thermal conductivity by measurement of temperature in a conducting solid. There has, however, been recent interest in spatial time series in geography. A fourth area concerns the engineering interest in transfer function relations between "input" and "output" series as opposed to analyzing the structure of "noise" processes. Of course, econometrics has a similar interest. Finally, since classical control theory makes a great use of spectral methods (gain and phase plots) these considerations are always in evidence even in time domain studies. In hydrology, economics, business time domain methods have recently been more popular. The aim of the present discussion is to consider a fundamental problem in time series analysis, namely the forecasting problem. In particular, it is pointed out how a very efficient algorithm (known in the control engineering literature) for computing exact finite data linear forecasts for an ARMA time series model has been passed by in the statistical times series literature. This algorithm also gives an efficient method for computing the exact likelihood.

1. The Linear Least Squares Filter

Consider the estimation of one process $\underline{x}(t)$ from measurements on another related process $\underline{y}(t)$ by measurements at times t_1, \dots, t_N giving $\underline{y}(t_1), \dots, \underline{y}(t_N)$ which we collect into a vector \underline{y} or \underline{Y} . Also write $\underline{y}(t_i) = \underline{y}_i = \underline{y}$. $\underline{x}_t = \underline{x}(t)$. We look at linear estimates of the form

$$\hat{\underline{x}}_t = \sum_1^N H_{t_i} \underline{y}_i .$$

The linear least squares filter (l.l.s.f.) chooses H_{t_i} to ensure $E(\underline{x}_t - \hat{\underline{x}}_t)^T (\underline{x}_t - \hat{\underline{x}}_t)$ is minimized. This best estimate is defined by the orthogonality condition

$$E(\underline{x}_t - \hat{\underline{x}}_t) \underline{y}'_s = 0, \quad s = 1, \dots, N .$$

Proof. Let \underline{x}_t^* be any other linear estimate. Then

$$\begin{aligned} E(\underline{x}_t - \underline{x}_t^*)^T (\underline{x}_t - \underline{x}_t^*) &= E\|\underline{x}_t - \hat{\underline{x}}_t + \hat{\underline{x}}_t - \underline{x}_t^*\|^2 \\ &= E\|\underline{x}_t - \hat{\underline{x}}_t\|^2 + E\|\hat{\underline{x}}_t - \underline{x}_t^*\|^2 + 2E(\underline{x}_t - \hat{\underline{x}}_t)^T (\hat{\underline{x}}_t - \underline{x}_t^*) . \end{aligned}$$

The third term vanishes by orthogonality. Thus, the mean square error of \underline{x}_t^* is larger than that of $\hat{\underline{x}}_t$.

Remark. If we allow nonlinear functions of the past then $\hat{\underline{x}}_t = E(\underline{x}_t | \underline{y})$ so we often write the linear estimate as

$$\hat{\underline{x}}_t = \tilde{E}(\underline{x}_t | \underline{y})$$

a wide sense conditional expectation. (Note $E(\underline{x}_t | \underline{y})$ is defined too by an orthogonality.) In fact, $\underline{x}_t - E(\underline{x}_t | \underline{y})$ is orthogonal to any linear or non-linear combination of the past.

We can apply the orthogonality condition to solve our problem thus,
substituting for \hat{x}_t implies

$$E(\hat{x}_t - \sum_1^N H_{ti} Y_i) Y_s' = 0, \quad s = 1, \dots, N$$

or

$$E(\hat{x}_t y') - H_t E(y y') = 0$$

or

$$H_t = E(\hat{x}_t y') E(y y')^{-1}.$$

Thus

$$\hat{x}_t = \tilde{E}(\hat{x}_t | y) = E(\hat{x}_t y') E(y y')^{-1} y.$$

The problem with this form of the solution is that it involves a matrix inversion so we are lead to a second approach. First process the data y to whiten it i.e. uncorrelate it then the inversion is easy, since the matrix is diagonal.

We perform a Cholesky factorization of $E(y y') = \tilde{U} \tilde{D} \tilde{U}^T$ where \tilde{U} is upper diagonal with I's on the leading diagonal. \tilde{D} is block diagonal. Consider

$$\tilde{v} = \tilde{U}^{-1} y = (\tilde{v}_1', \dots, \tilde{v}_N') .$$

Then

$$E(\tilde{v} \tilde{v}^T) = \tilde{U}^{-1} \tilde{U} \tilde{D} \tilde{U}^T \tilde{U}^{-T} = \tilde{D}$$

i.e. $E(v_i v_j^T) = 0, i \neq j, i, j = 1, \dots, N$ i.e. the v_i sequence is a white sequence. Since \tilde{U}^{-1} is lower triangular the v_i are linearly, causally, invertibly related to the y_i . The v_i is called the linear innovations sequence.

Because of the equivalence between \underline{v} , \underline{y} the orthogonality condition can be rewritten

$$E(\underline{x}_t - \hat{\underline{x}}_t) \underline{v}_s' = 0, \quad s = 1, \dots, N.$$

Then as before

$$\hat{\underline{x}}_t = \tilde{E}(\underline{x}_t | \underline{v}) = E(\underline{x}_t \underline{v}'') E(\underline{v} \underline{v}^T)^{-1} \underline{v}$$

or

$$\hat{\underline{x}}_t = \sum_1^N E(\underline{x}_t \underline{v}_i') R_i^{-1} \underline{v}_i$$

also

$$\hat{\underline{x}}_{t|N} = \hat{\underline{x}}_{t|N-1} + E(\underline{x}_t \underline{v}_N') R_N^{-1} \underline{v}_N.$$

Taking variances gives

$$E\|\underline{x}_t - \hat{\underline{x}}_{t|N}\|^2 = p_{t|N} = p_{t|N-1} - E(\underline{x}_t \underline{v}_N') R_N^{-1} E(\underline{v}_N \underline{x}_t').$$

Remark.

$$\underline{v}_N = \underline{y}_N - \hat{\underline{y}}_{N|N-1}$$

$$\therefore E(\underline{y}_N - (\underline{y}_N - \underline{v}_N)) \underline{v}_s' = E(\underline{v}_N \underline{v}_s') = 0, \quad s = 1, \dots, N-1.$$

It follows then by the uniqueness of \underline{v}_s that $\underline{y}_N - \underline{v}_N = \hat{\underline{y}}_{N|N-1}$.

To complete the algorithm we need recursive formulae for the Cholesky factoring. Now

$$\underline{v}_t = \underline{y}_t - \hat{\underline{y}}_{t|t-1} = \underline{y}_t - \sum_1^{t-1} E(\underline{y}_t \underline{v}_s') R_s^{-1} \underline{v}_s$$

so

$$E(\underline{y}_{m+1} \underline{v}_t') = E(\underline{y}_{m+1} \underline{y}_t') - \sum_1^{t-1} E(\underline{y}_{m+1} \underline{v}_s') R_s^{-1} E(\underline{v}_s \underline{y}_t'), \quad t = 2, \dots, m+1,$$

$$E(\underline{y}_{m+1} \underline{v}_1') = E(\underline{y}_{m+1} \underline{y}_1').$$

Also,

$$R_m = E(v_m v_m') = E(Y_m Y_m') .$$

We will return to this algorithm below.

2. Relation to Wiener Filtering

Take (i) x_t, y_t jointly stationary (so t_i are equispaced).

(ii) Suppose an "infinite past" of y data is available. We want $\hat{x}_k|_{-\infty} = \tilde{E}(x_k|y_k y_{k-1} \dots)$. Again we first whiten the data y_k . Suppose y_k has spectrum

$$\begin{aligned}\phi_{YY}(Z) &= \sum_{-\infty}^{\infty} E(Y_k Y_0') Z^k \\ &= Z\{E(Y_k Y_0')\} .\end{aligned}$$

Here $Z\{\cdot\}$ denotes the Z transform. If $\phi_{YY}(Z)$ is nonsingular on $|Z| = 1$ it has a factoring $\phi_{YY}(Z) = W(Z)W(Z^{-1})$ where $W(Z), W^{-1}(Z)$ are analytic on $|Z| \geq 1$ and $\lim_{Z \rightarrow \infty} W(Z) < \infty$, $W(Z)$ has no Z^{-1} powers. (This is an "infinite" analogue of the finite data Cholesky factor.)

Consider now $v_k = W^{-1}(Z)Y_k$. Then

$$\begin{aligned}Z\{E(v_k v_0')\} &= \phi_{VV}(Z) \\ &= \sum_{-\infty}^{\infty} W^{-1}(Z)E(Y_k Y_0')W^{-1}(Z^{-1})Z^k \\ &= W^{-1}(Z)\phi_{YY}(Z)W^{-1}(Z^{-1}) = I .\end{aligned}$$

So v_k is a white noise sequence of variance I and v_k are causally linearly equivalent to y_k . So we must be able to calculate $\hat{x}_k|_{k-1}$ as

$$\begin{aligned}
\hat{x}_{k|k-1} &= \tilde{E}(x_k | v_k v_{k-1} \dots) \\
&= E(x_k v_k) v_k + E(x_k v_{k-1}) v_{k-1} + \dots \\
&= E(x_0 v_0) v_k + E(x_0 v_{-1}) v_{k-1} + \dots \\
&= \sum_0^{\infty} D_j z^{-j} v_k = D_+(z^{-1}) v_k
\end{aligned}$$

where $D_j = E(x_0 v_{-j})$

$$\begin{aligned}
D(z^{-1}) &= \sum_{-\infty}^{\infty} D_j z^j = \sum_0^{\infty} D_j z^j + \sum_{-\infty}^{-1} D_j z^j \\
&= D_+(z) + D_-(z)
\end{aligned}$$

Thus the transform from v_k to $\hat{x}_{k|k-1}$ is $D_+(z^{-1})$. So the transform linking y_k to $\hat{x}_{k|k-1}$ is

$$\hat{x}_{k|k-1} = D_+(z^{-1}) v_k = D_+(z^{-1}) W^{-1}(z) y_k .$$

Finally observe

$$\begin{aligned}
D(z) &= \sum_{-\infty}^{\infty} E(x_0 v_{-j}) z^{-j} \\
&= \sum_{-\infty}^{\infty} E(x_0 W^{-1}(z) y_{-j}) z^{-j} \\
&= W^{-1}(z) \phi_{xy}(z) .
\end{aligned}$$

Thus

$$\hat{x}_{k|k-1} = \left\{ \frac{\phi_{xy}(z)}{W(z)} \right\}_+ \frac{1}{W(z)} y_k ,$$

which is the well known formula for the Wiener filter.

3. Exact Likelihoods Via the Linear Least Squares Filter

Under a Gaussian assumption on the data \underline{y} the likelihood or joint density of the \underline{y} data is

$$\begin{aligned}\ln L &= \text{constant} - \frac{1}{2} \ln |\underline{R}| - \frac{1}{2} \underline{y}^T \underline{R}^{-1} \underline{y} \\ &\quad \text{where } \underline{R} = E(\underline{y}\underline{y}^T) \\ &= \text{constant} - \frac{1}{2} \sum_1^N \ln \sigma_i^2 - \frac{1}{2} \sum_1^N v_i^2 / \sigma_i^2.\end{aligned}$$

This follows since $\underline{R} = \underline{U}\underline{D}\underline{U}^T$ implies

$$\underline{y}^T \underline{R}^{-1} \underline{y} = \underline{y}^T \underline{U}^{-1} \underline{D}^{-1} \underline{U}^{-1} \underline{y} = \underline{v}^T \underline{D}^{-1} \underline{v}$$

and

$$|\underline{R}| = |\underline{U}| |\underline{D}| |\underline{U}^T| = |\underline{D}| = \prod_1^N \sigma_i^2.$$

Alternatively, the second expression follows by writing L as an iterated conditional density.

So to get the exact likelihood we need only generate the v_i . Consider this process for the ARMA model

$$Y_n + a_1 Y_{n-1} + \dots + a_{n_a} Y_{n-n_a} = \epsilon_n + c_1 \epsilon_{n-1} + \dots + c_{n_a} \epsilon_{n-n_a}$$

with initial conditions chosen to ensure \underline{y} is stationary. Observe

$$E(Y_t Y_n) = - \sum_1^{r-1} a_s E(Y_{t-s} Y_n), \quad t \geq n+r, \quad r = n_a + 1. \quad (1)$$

Next

$$\hat{y}_{t|t-1} = \sum_{\tau=1}^{t-1} E(y_t v_\tau) R_\tau^{-1} v_\tau.$$

However,

$$v_\tau = \sum_1^T a_{\tau n} Y_n \text{ for some } a_{\tau n} \quad (2)$$

so if $t \geq \tau + r$ so that $t \geq n + r$ then using (2) in (1) will give

$$E(y_0 v_\tau) = - \sum_1^{r-1} a_s E(y_{t-s} v_\tau), \quad t \geq \tau + r \quad (3)$$

So write

$$\begin{aligned} \hat{y}_{t|t-1} &= \sum_1^{t-r-1} E(y_t v_\tau) R_\tau^{-1} v_\tau + \sum_{t-r}^{t-1} E(y_t v_\tau) R_\tau^{-1} v_\tau \\ &= \sum_1^{t-r-1} \left(- \sum_1^{r-1} a_s E(y_{t-s} v_\tau) \right) R_\tau^{-1} v_\tau + \sum_{t-r}^{t-1} E(y_t v_\tau) R_\tau^{-1} v_\tau \\ &= - \sum_1^{r-1} a_s \left[\sum_1^{t-r-1} E(y_{t-s} v_\tau) R_\tau^{-1} v_\tau \right] + \sum_{t-r}^{t-1} E(y_t v_\tau) R_\tau^{-1} v_\tau \\ &= - \sum_1^{r-1} a_s \left[\sum_1^{t-1} E(y_{t-s} v_\tau) R_\tau^{-1} v_\tau \right] + \sum_{t-r}^{t-1} [E(y_t v_\tau) + \sum_1^{r-1} a_s E(y_{t-s} v_\tau)] R_\tau^{-1} v_\tau \\ \hat{y}_{t|t-1} &= - \sum_1^{r-1} a_s y_{t-s} + \sum_{t-r}^{t-1} E(w_t v_\tau) R_\tau^{-1} v_\tau \end{aligned}$$

where $w_t = y_t + \sum_1^{r-1} a_s y_{t-s}$. Also, $R_\tau = E(v_\tau v_\tau) = E(y_\tau v_\tau)$. We can generate too, a finite recursive algorithm for $E(w_\tau v_\tau)$. For details (and an alternate derivation of the above filter) see Kaillath and Aasnes (1974). This reference also discusses the p step ahead forecasts, $p > 1$.

4. Exact Likelihood for Continuous Discrete Models

Suppose

$$\ddot{s} + a_1 \dot{s} + a_2 s = w + aw + a_2 \ddot{w}$$

where $\dot{s} = ds/dt$ etc., $s = s(t)$ etc., $y = s + v$ and w, v are "white" noises and y is sampled at times t_1, \dots, t_N . The log likelihood is as before

$$\text{log likelihood} = \text{const.} - \frac{1}{2} \sum_i^N \ln \sigma_i^2 - \frac{1}{2} \sum_i^N v_i^2 / \sigma_i^2$$

To generate the v_i observe

$$1. \hat{s}(t|N) = \hat{s}(t|N-1) + E(s(t)v_N') R_N^{-1} v_N. \text{ So}$$

$$\hat{s}_{N|N} = \hat{s}_{N|N-1} + E(s(t_N)v_N') R_N^{-1} v_N.$$

$$2. \frac{d^2}{dt^2} \hat{s}(t|N-1) + a_1 \frac{d\hat{s}}{dt}(t|N-1) + a_2 \hat{s}(t|N-1) \\ = \sum_{k=1}^{N-1} E[(\ddot{s} + a_1 \dot{s} + a_2 s)v_k] R_k^{-1} v_k \\ = 0, \quad t_{N-1} < t < t_N$$

with initial condition $\hat{s}_{N-1|N-1}$

$$3. E(s(t)v_N') = Es(t)[(y(t_N) - \hat{s}(t_N|N-1)] \\ = E[s(t)s(t_N)] - \hat{P}(t, t_N|N-1)$$

$$\hat{P}(t, t_N|N-1) = E[\hat{s}(t|N-1)s(t_N|N-1)]$$

$$4. P(N|N) = P(t_N, t_N|N)$$

$$= P(N|N-1) - E(s(N)v_N') R_N^{-1} E(v_N s_N')$$

5. Derive a differential equation for $P(t, t|N-1)$ for $t_{N-1} < t < t_N$ by differentiating

$$\hat{P}(t, t|N-1) = E \|\hat{s}(t|N-1)\|^2 = \sum_0^{N-1} E(s(t)v_N) R_N^{-1} E(v_N s(t))$$

to yield

$$\frac{d^2\hat{P}}{dt^2} + a_1 \frac{d\hat{P}}{dt} + a_2 \hat{P} = 0, \quad t_{N-1} < t < t_N.$$

Further details of this procedure will be discussed by the author elsewhere; however, compare this with the usual approach where

- (i) data must be equispaced;
- (ii) an equivalent discrete model is formed and a discrete likelihood computed so that the original parameters occur very nonlinearly.

5. Conclusion

The use of a recursive form of the linear least squares filter for estimating one time series from another has been illustrated in two cases. This idea has promising use in the statistical solution of inverse problems where in the signal plus noise model the signal is the solution to an integral equation.

References:

1. H. Aasnes and T. Kailath (1974), "Initial Condition Robustness of Linear Least Squares Filtering Algorithms, IEEE Trans. Auto. Control.

STRESS-STRENGTH MODELS FOR RELIABILITY:
OVERVIEW AND RECENT ADVANCES

G. K. Bhattacharyya

and

Richard A. Johnson

University of Wisconsin, Madison

ABSTRACT

A stress-strength model of reliability is relevant for the situation where random environmental stresses tend to interfere with the functioning of a device. Interest in these models abound in numerous disciplines of engineering and life sciences. In this article, we present an overview of the principal aspects of these models and our contributions to the recent advances in statistical analyses of the stress and strength data.

Research supported by Office of Naval Research
Grant N00014-78-C-0722.

1. INTRODUCTION

A stress-strength model of reliability is relevant for the situation where random environmental stresses tend to interfere with the functioning of a device. Interest in these models abound in numerous disciplines of engineering and life sciences. In this article, we present an overview of the principal aspects of these models and our contributions to the recent advances in statistical analyses of the stress and strength data.

Most engineered products must operate in environments which are not controlled and the possibility of random shocks usually prohibits any strictly deterministic formulation of the stresses. Stress-strength models use random variables to represent both

- (i) the variation in the ability (strength) to perform and
- (ii) the variation in the stress imposed by the environment.

Let

X = maximum stress

Y = strength of unit.

In this context, we define Reliability (R) = the probability that the unit performs its task satisfactorily, that is, the unit is strong enough to overcome the stress.

$$R = P[\text{Strength} > \text{stress}] = P[Y > X]$$

When X and Y are independent,

$$R = P[Y > X] = \int_{-\infty}^{\infty} F(y)dG(y) = 1 - \int_{-\infty}^{\infty} G(x)dF(x) \quad (1.1)$$

where F and G are the continuous cumulative distribution functions (cdf's) of X and Y , respectively.

The following examples help to delineate the diversity of applications.

Example 1. [Rocket engines] Let X represent the maximum chamber pressure generated by ignition of a solid propellant, and Y be the strength of the rocket chamber. Then R is the probability of a successful firing of the engine.

Example 2. [Comparing two treatments] A standard design for the comparison of two drugs is to assign Drug A to one group of subjects and Drug B to another group. Denote by X and Y the remission times with Drug A and B, respectively. Inferences about $R = P[Y > X]$ based on the remission times data X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m , are of primary interest to the experimenter. Although the name 'stress-strength' is not appropriate in the present context, our target of inference is the parameter R which has the same structure as in Example 1.

Example 3. [Threshold response model] A unit, say a receptor in the human eye, operates only if it is stimulated by a source whose random magnitude, Y , is greater than a (random) lower threshold for the unit. Here

$$P[Y > X] = P[\text{unit operates}]$$

is again of the form described above in stress-strength context.

2. ESTIMATION OF RELIABILITY $P[Y > X]$

We briefly review estimation of R when samples are available from both the strength and stress distributions. Specifically, let X_1, X_2, \dots, X_n be a random sample from F independent of Y_1, Y_2, \dots, Y_m , a random sample from G .

2.1 Nonparametric approach

A number of authors Birnbaum (1956) Birnbaum and McCarty (1958) Owen, Craswell and Hansen (1964) have proposed nonparametric estimators of $R = P[Y > X]$. An estimator, based on the count is

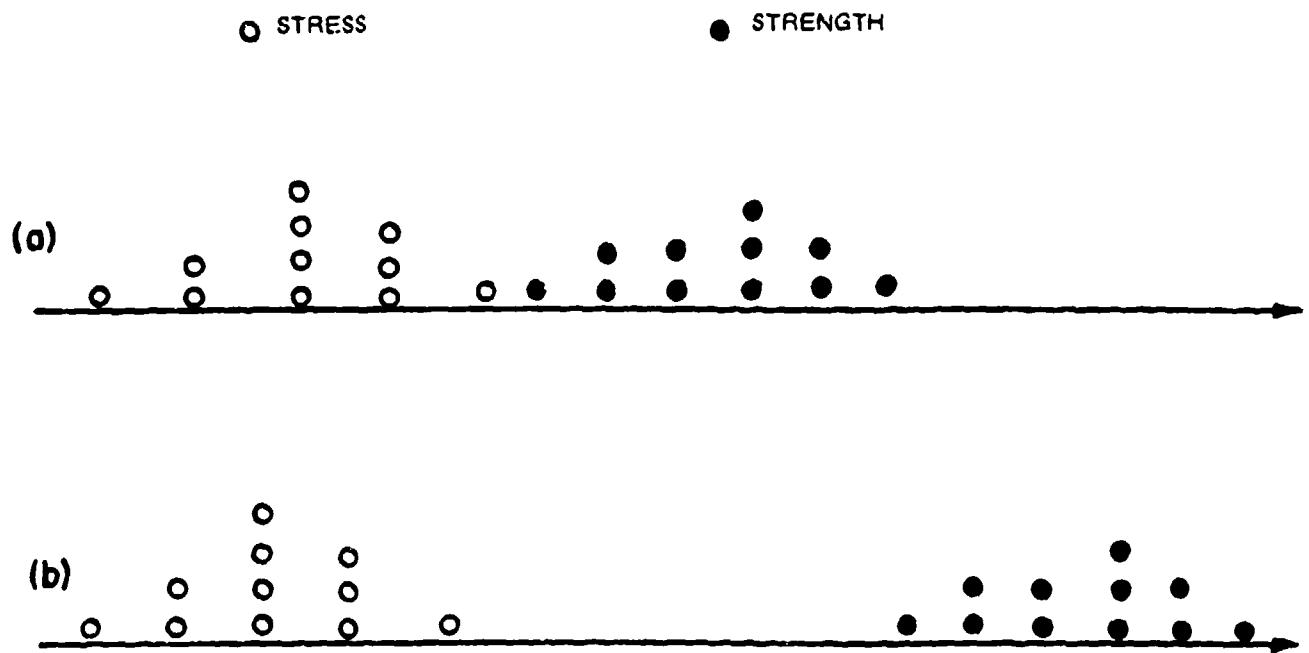
$$U = \# \text{ pairs } (X_i, Y_j) \text{ with } Y_j > X_i,$$

$$\hat{R} = \frac{U}{mn} = \int_{-\infty}^{\infty} F_n(y) dG_m(y) \quad (2.1)$$

where F_n, G_m are the empirical cdf's from the X_i 's and Y_j 's, respectively. Approximate confidence bounds can be obtained from the large sample distribution of \hat{R} (see also Govindarajulu (1968), Sen (1967)).

A difficulty with the nonparametric approach

With small or even moderate sample sizes, high reliability cannot be verified. For instances, with $m = n = 11$, if the strength observations were all larger than the stresses, we obtain $.77 < R$ with confidence .95. The same confidence bound is obtained when the strength measurements are only moderately larger than the stresses (Figure 1(a)) as when they are considerably larger (Figure 1(b)). The nonparametric method fails to discriminate between the two situations.



2.2 Parametric approaches

When X is $N(\mu_1, \sigma_1^2)$ and Y is $N(\mu_2, \sigma_2^2)$

$$R = P[Y > X] = \Phi\left(\frac{\mu_2 - \mu_1}{(\sigma_1^2 + \sigma_2^2)^{1/2}}\right) \quad (2.2)$$

where Φ denotes the cdf of $N(0,1)$. Church and Harris (1970) obtain large sample confidence bounds when the parameters of the stress distribution are known. (see also Mazumdar (1970)).

When $\sigma_1^2 = \sigma_2^2$, confidence bounds can be obtained from the non-central t-distribution of

$$\frac{\bar{Y} - \bar{X}}{\left\{ \left(\frac{1}{m} + \frac{1}{n} \right) \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \right\}^{1/2}}$$

as in Owen, Craswell, and Hansen (1964).

Difficulty with the parametric model

If a small fraction of the population of the units contain major defects of material or workmanship, a small or moderate sample of strengths will not show these 'rare' sources of failure. This is illustrated in Figure 2.

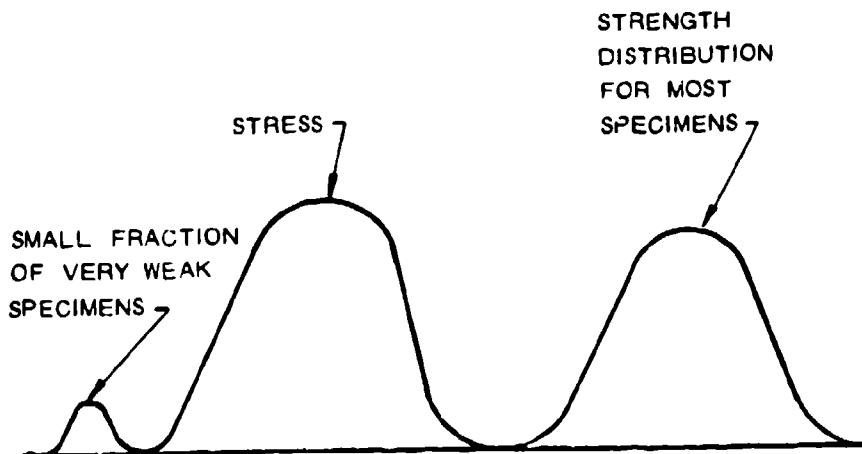


Figure 2. A strength distribution that is a mixture of two failure sources.

In this situation, use of an assumed parametric form for the stress distribution will, typically, lead to estimates of $P[Y>X]$ which are incorrectly very high.

Even without such extreme departures from the postulated models, tail areas remain very difficult to estimate. The choice between normal, Weibull or lognormal tails can change the estimated reliability by several orders of magnitude (when R is very high).

2.3 Bayesian approach

Enis and Geisser (1971) consider estimation of $P[Y>X]$ when X , Y are negative exponential, and also in several normal theory situations.

3. STRESS-STRENGTH MODELS WITH COVARIATES

Recently, we have encountered stress-strength analysis where covariates play an important role.

Example 4. A 2×4 used in the frame of a house has strength Y which can be obtained only by destructive testing. Yet, the stiffness z is easily measured.

The data of Figure 3 suggest that the strength

$$Y = \alpha_2 + \beta_2 z + e_2$$

where e_2 is $N(0, \sigma_2^2)$. For a specimen of stiffness z .

$$P[Y>X|z] = \Phi\left(\frac{\alpha_2 + \beta_2 z - \mu_1}{(\sigma_1^2 + \sigma_2^2)^{1/2}}\right)$$

Example 5. Let Y be the strength (in p.s.i.) of a glass fiber reinforced rocket motor case, and let X denote the operating pressure. The data of Figure 4 suggest that the stress depends on ambient temperature z according to a linear model

$$X = \alpha_1 + \beta_1 z + e_1$$

where e_1 is $N(0, \sigma_1^2)$.

For a given temperature z ,

$$P[Y > X | z] = \Phi\left(\frac{\mu_2 - \alpha_1 - \beta_1 z}{(\sigma_1^2 + \sigma_2^2)^{1/2}}\right) \quad (3.2)$$

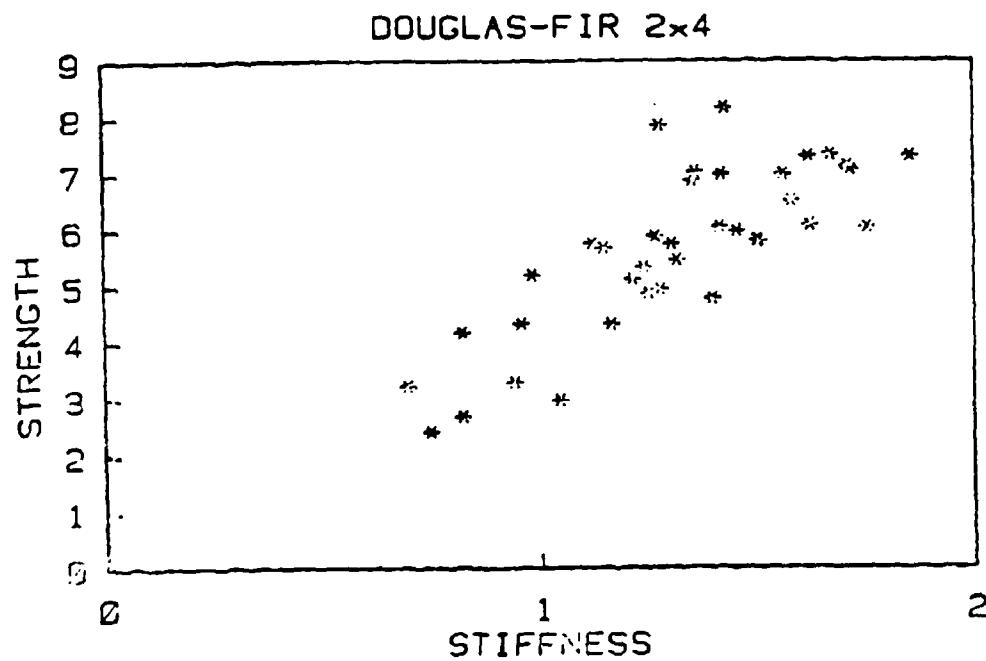


Figure 3. Stiffness of wood as a covariate.

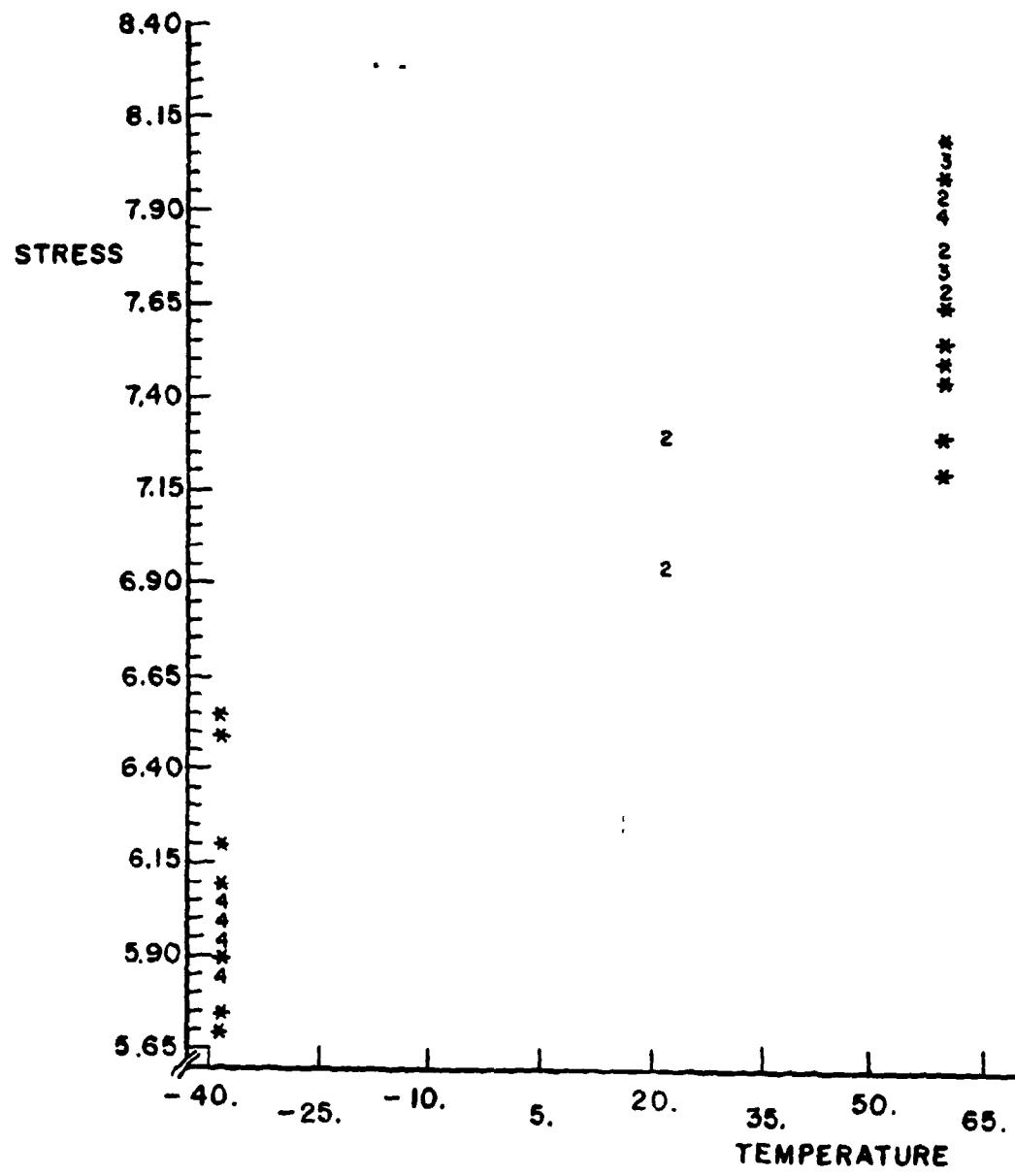


Figure 4. Ambient temperate as a covariate

Example 6. Comparison of Drug A and Drug B when age is a covariate:

X = remission time with Drug A

Y = remission time with Drug B

Z = age

Data structure:

$$\begin{bmatrix} X_1 \\ Z_{11} \end{bmatrix}, \begin{bmatrix} X_2 \\ Z_{21} \end{bmatrix}, \dots, \begin{bmatrix} X_m \\ Z_{m1} \end{bmatrix} : \text{Drug A}$$

$$\begin{bmatrix} Y_1 \\ Z_{12} \end{bmatrix}, \begin{bmatrix} Y_2 \\ Z_{22} \end{bmatrix}, \dots, \begin{bmatrix} Y_n \\ Z_{n2} \end{bmatrix} : \text{Drug B}$$

can be used to estimate the linear regressions

$$E(X|z) = \alpha_1 + \beta_1 z \quad \text{and} \quad E(Y|z) = \alpha_2 + \beta_2 z .$$

For a new subject of age z , we may wish to provide information about

$$P[Y > X|z].$$

In Example 6, if we assume that X and Y are independent normal random variables having the same variance σ^2 , we obtain the expression

$$\begin{aligned} P[Y > X | z] &= P\left[\frac{Y-X-(\alpha_2-\alpha_1)-(\beta_2-\beta_1)z}{\sqrt{2}\sigma} > \frac{-(\alpha_2-\alpha_1)-(\beta_2-\beta_1)z}{\sqrt{2}\sigma}\right] \\ &= \Phi\left(\frac{(\alpha_2-\alpha_1)+(\beta_2-\beta_1)z}{\sqrt{2}\sigma}\right) \end{aligned} \quad (3.3)$$

The resulting reliability is similar in form to the previous cases.

Allowing both X and Y to depend on the covariate does not lead to further complexities.

Under the assumptions

$$X|z_1 \sim N(\alpha_1 + \beta_1 z_1, \sigma^2)$$

$$Y|z_2 \sim N(\alpha_2 + \beta_2 z_2, \sigma^2).$$

We are interested in making inferences about the reliability

$$P[Y > X | z_1, z_2] = P\left[\frac{Y-\alpha_2-\beta_2 z_2}{\sqrt{2}\sigma} > \frac{\alpha_1-\alpha_2+\beta_1 z_1-\beta_2 z_2}{\sqrt{2}\sigma}\right]$$

at z_1, z_2 .

Classical Approach

Confidence bounds for $R(z_0)$. In order to determine a lower confidence bound for $R(z_0)$, we note that $\bar{Y} - \bar{X} - \hat{\beta}(z_0 - \bar{z})$ is normally distributed with mean $= \mu_2 - \alpha_1 - \beta_1(z_0 - \bar{z})$ and standard deviation σc_0 , where

$$c_0^2 = \frac{1}{m} + \frac{1}{n} + \frac{(z_0 - \bar{z})^2}{\sum_j (z_j - \bar{z})^2}.$$

Also, $(m+n-3)s^2$ is a $\sigma^2 \chi_{m+n-3}^2$ and is independent of $\bar{Y} - \bar{X} - \hat{\beta}(z_0 - \bar{z})$.

Consequently,

$$t = \frac{\bar{Y} - \bar{X} - \hat{\beta}(z_0 - \bar{z})}{s c_0}$$

has a non-central t-distribution with $m+n-3$ d.f. and the noncentrality parameter

$$\eta = \frac{\mu_2 - \alpha - \beta(z_0 - \bar{z})}{\sigma c_0}$$

A lower 95% confidence bound, $\underline{\eta}$, is obtained by solving $F_{\eta}(t_{\text{obs}}) = .95$.
Consequently, a

95% lower bound for $R(z_0) = P[Y > X | z_0]$ is

$$R(z_0) = \phi(\underline{\eta}/2^{1/2} c_0). \quad (3.4)$$

Refer to the data of Example 5. For temperature $z_0 = 30^\circ$, $t_{\text{obs}} = 12.487$, and a computer calculation gives $\underline{\eta} = 10.02$, and

$$R(30) = \phi(2.02) = .978.$$

Inverse problem

We want a confidence bound on z for which $R(z) \geq .99$. This time we consider

$$T(\rho) = \frac{\bar{Y} - \bar{X} - \hat{\beta}(\rho - \bar{z})}{s \left[\frac{1}{m} + \frac{1}{n} + \frac{(\rho - \bar{z})^2}{\sum_j (t_j - \bar{t})^2} \right]^{1/2}}$$

which is distributed as a non-central t with $m+n-3$ d.f. and the non-centrality parameter

$$\eta = \frac{(\mu_2 - \alpha - \beta(\bar{\rho} - \bar{z}))/\sigma}{[\frac{1}{m} + \frac{1}{n} + \frac{(\rho - \bar{z})^2}{\sum(z_j - \bar{z})^2}]^{1/2}}$$

The numerator of η is constrained by the relation

$$\Phi\left(\frac{\mu_2 - \alpha - \beta(\bar{\rho} - \bar{z})}{\sqrt{2}\sigma}\right) = .99.$$

The confidence region for ρ consists of all values of ρ_0 for which $H_0: \rho \leq \rho_0$ would not be rejected in favor of $H_1: \rho > \rho_0$.

The 95% upper bound for ρ is the largest $\bar{\rho}$ satisfying

$$T(\bar{\rho}) \geq t_{.05}(\bar{\rho})$$

or

$$F_{\bar{\rho}}(T(\bar{\rho}_{\text{obs}})) \geq .05.$$

For the data of Example 5, a computer calculation gives

$$\bar{\rho} = 46.8^\circ.$$

4. STRESS-STRENGTH MODELS FOR SYSTEM RELIABILITY

The formulation of a stress-strength reliability model has been extended to multicomponent systems and several problems of statistical inferences on the system reliability are addressed in Bhattacharyya and Johnson (1974, 1975, 1977). Here we only include a brief description of the models and the approach to inferences. Details of the technical results are available

in the cited references.

Model 1

The primary extension of a single component stress-strength model was focused on an s out of k system of identical components. This is a system of k components whose strengths Y_1, \dots, Y_k are independent and identically distributed (iid) random variables with a continuous cdf $G(x)$. The system successfully performs its mission if at least s ($1 \leq s \leq k$) components are operative.

First we consider the situation where all k components of the system encounter a common random stress X whose cdf is denoted by $F(x)$. The reliability $R_{s,k}$ of this system is then given by

$$\begin{aligned} R_{s,k} &= P[\text{at least } s \text{ of } Y_1, \dots, Y_k > X] \\ &= \sum_{l=s}^k \binom{k}{l} \int_{-\infty}^{\infty} [1-G(x)]^l G(x)^{k-l} dF(x) \\ &= 1 - \int_{-\infty}^{\infty} B[G(x)] dF(x), \end{aligned} \tag{4.1}$$

where B is the beta distribution with pdf $\propto u^{k-s}(1-u)^{s-1}$.

A parametric approach. Assuming the exponential distributions

$$\begin{aligned} F(x) &= 1 - \exp(-\theta_1 x) \\ G(x) &= 1 - \exp(-\theta_2 x), \quad \lambda = \theta_1/\theta_2, \end{aligned}$$

the system reliability is given by

$$1 - R_{s,k} = \frac{B(s+\lambda, k-s+1)}{B(s, k-s+1)} = \frac{k!}{s!} \frac{1}{\prod_{j=s}^k (\lambda+j)} \tag{4.2}$$

The uniformly minimum variance unbiased (UMVU) estimator of $R_{s,k}$ and exact confidence bounds are derived in Bhattacharyya and Johnson (1974). This work also includes a study of the bias and mean squared error of the maximum likelihood estimator.

Nonparametric estimation. Let X_1, \dots, X_{n_1} and Y_1, \dots, Y_{n_2} denote independent random samples from F and G , respectively. The corresponding empirical cdf's are denoted by

$$F_{n_1}(x) = \frac{\#X_i \leq x}{n_1}, \quad G_{n_2}(a) = \frac{\#Y_j \leq a}{n_2}.$$

A plausible estimator \tilde{R}^* of $R_{s,k}$ can then be obtained by replacing F and G in (4.1) by F_{n_1} and G_{n_2} , respectively. After some simplifications, this leads to

$$\tilde{R}^* = \frac{1}{n_1} \sum_{\ell=1}^{n_2} [B(\frac{\ell}{n_2}) - B(\frac{\ell-1}{n_2})] (s_{(\ell)} - \ell) \quad (4.3)$$

where $s_{(1)} < \dots < s_{(n_2)}$ are the ordered ranks of the Y -values in the combined sample. Employing the idea of a generalized U-statistic, the UMVU estimator \tilde{R} of $R_{s,k}$ is derived, and it is given by

$$\tilde{R} = \frac{1}{n_1 \binom{n_2}{k}} \sum_{\ell=k-s+1}^{n_2-s+1} \binom{\ell-1}{k-s} \binom{n_2-\ell}{s-1} (s_{(\ell)} - \ell) . \quad (4.4)$$

The statistical properties of these estimators, and large sample confidence intervals for the system reliability, $R_{s,k}$, are explored in Bhattacharyya and Johnson (1974, 1975, 1977).

A few other models, considered in Bhattacharyya and Johnson (1974) bare mention.

Model 2. s out of k system with standbys.

Operating.

k_1 independent components with strength distribution G_1

Standbys.

$k_2 = k - k_1$ independent components with strength distribution G_2

A single stress X is applied to all components.

Model 3. Subsystems with independent stresses.

Subsystem q.

k_q independent components with strength distribution G_q .

Stress $X^{(q)}$ with cdf F_q is applied to each component.

Subsystem operates if s_q out of k_q components operate.

Model 4. [Binomial Counts] Here we consider the structure of Model 1 with the variation that the test conditions do not permit independent sets of strength and stress measurements. Rather, the prototype components are tested under random stress conditions that prevail, and all that is recorded

are z_i , $i = 1, \dots, n$ where

$$z_i = \begin{cases} 1, & \text{if component } i \text{ functions} \\ 0, & \text{if component } i \text{ fails.} \end{cases}$$

In the case of a system with c subsystems where each subsystem conforms to the structure of a single-component stress-strength model, the problem reduces to one of estimating the system reliability from the binomial counts. See Myhre and Saunders (1968), Madansky (1965) and Easterling (1972) for discussions about inference methods. However, certain complications arise when groups of components are simultaneously tested, each under a random stress, and the group size is different from the system (or subsystem) size. In the context of Model 1, Bhattacharyya (1977) discusses nonparametric estimation of $R_{s,k}$ when groups of m components are tested under independent stresses and only the failure count is recorded for each group.

REFERENCES

- [1] G. K. Bhattacharyya and R. A. Johnson (1974). Estimation of reliability in a multicomponent stress-strength model. J. Amer. Statist. Assoc., 69, 966-70.
- [2] G. K. Bhattacharyya and R. A. Johnson (1975). Stress-strength models for system reliability. Proc. Symp. on Reliability and Fault-tree Analysis. SIAM, 509-32.
- [3] G. K. Bhattacharyya and R. A. Johnson (1977). Estimation of system reliability by nonparametric techniques. Bulletin of the Mathematical Society of Greece (Memorial Volume), 94-105.
- [4] G. K. Bhattacharyya (1977). Reliability estimation from survivor count data in a stress-strength setting. IAPQR Transactions—Journal of the Indian Association for Productivity, Quality and Reliability, 2, 1-15.
- [5] Z. W. Birnbaum (1956). On a use of the Mann-Whitney statistic. Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 13-17.
- [6] Z. W. Birnbaum and R. C. McCarty (1958). A distribution free upper confidence bound for $P(Y < X)$ based on independent samples of X and Y . Ann. Math. Statist., 29, 558-62.
- [7] J. D. Church and B. Harris (1970). The estimation of reliability from stress-strength relationship. Technometrics, 12, 49-54.
- [8] R. Easterling (1972). Approximate confidence limits for system reliability. J. Amer. Statist. Assoc., 67, 220-2.
- [9] P. Enis and S. Geisser (1971). Estimation of the probability that $Y < X$. J. Amer. Statist. Assoc., 66, 162-8.
- [10] Z. Govindarajulu (1968). Distribution-free confidence bounds for $P(X < Y)$. Ann. Inst. Statist. Math., 20, 229-38.
- [11] A. Madansky (1965). Approximate confidence limits for the reliability of series and parallel systems. Technometrics, 7, 495-503.
- [12] M. Mazumdar (1970). Some estimates of reliability using interference theory. Naval Res. Log. Quart., 17, 159-65.
- [13] J. M. Myhre and S. C. Saunders (1968). On confidence limits for the reliability of systems. Ann. Math. Statist., 39, 1463-72.
- [14] D. B. Owen, K. J. Craswell, and D. L. Hanson (1964). Nonparametric upper confidence bounds for $P(Y < X)$ and confidence limits for $P(Y < X)$ when X and Y are normal. J. Amer. Statist. Assoc., 59, 906-24.
- [15] P. K. Sen (1967). A note on asymptotically distribution-free confidence bounds for $P(X < Y)$ based on two independent samples. Sankhyā, A, 29, 95-102.

ON THE INTERPOLATION OF GRAVITY ANOMALIES AND
DEFLECTIONS OF THE VERTICAL IN MOUNTAINOUS TERRAIN

H. Baussus von Luetzow
U.S. Army Topographic Laboratories
Fort Belvoir, Virginia 22060

ABSTRACT: The paper first addresses the interpolation of gravity anomalies in mountainous terrain, to be represented as the sum of a "signal" variable with a quasi-stationary estimation structure and a computable "noise" variable without a stationary character. It then develops the particular solution of the boundary value problem of physical geodesy which permits a similar representation of deflections of the vertical and draws some conclusions concerning the inapplicability of Molodensky's series approach and of the collocation method for the accurate determination of vertical deflections from unmodified gravity anomalies in mountainous terrain. Thereafter, it discusses the estimation of signal-type deflections of the vertical by means of spatial covariance functions, i.e., by a linear regression technique called statistical collocation in physical geodesy, and provides first order expansions of planar covariance functions.

1. INTRODUCTION. Deflection of the vertical components ξ and η play a role in the adjustment of geodetic networks, in the computation of height anomaly differences, and in the transformation of local coordinates into terrestrial coordinates. Short of a three-dimensional solution of the geodetic boundary value problem under consideration of mountainous terrain, deflection components and gravity anomalies Δg are also desirable for the numerical upward continuation of the first order derivatives of the anomalous gravity potential. The interpolation or estimation of gravity vector components in flat terrain is not inherently difficult. In mountainous terrain, gravity anomalies Δg are profitably modified to Faye anomalies Δg_F by means of terrain corrections C , to be followed by a transformation

to Bouguer anomalies Δg_B which permit an approximate two-dimensional interpolation. Isostatic gravity anomalies Δg_I , to be corrected by the indirect effect, would require a three-dimensional interpolation technique in the case of high accuracy. The problem of deflection estimation has been discussed by Heiskanen and Moritz [1967] and others, including the method by Molodensky et al. [1962] for the calculation of deflection differences in flat terrain, and the difficulty to interpolate ξ, η in rough mountainous terrain. Baussus von Luetzow [1980] addressed the optimal densification of deflections of the vertical in flat terrain with and without consideration of gravity anomalies and extended Molodensky's approach. Badekas and Mueller [1968] utilized Eötvös torsion balance measurement together with appropriate terrain corrections for the interpolation of vertical deflections, a time-consuming procedure and soon to be replaced by the employment of moving base gravity gradiometers. Regardless of these efforts, an effective ξ, η -estimation method applicable in mountainous terrain will still be valuable and may also aid deflection estimation under consideration of a series of discrete inertial measurements. Section 2 of this study addresses the interpolation of gravity anomalies in mountainous terrain. In section 3, the appropriate solution of the boundary value problem for vertical deflections is presented and reformulated for optimal deflection estimation of "signal" components of ξ and η and computation of topographic "noise" terms. The estimation of signal-type components by means of spatial collocation and the development of first order approximations of spatial covariance functions is the subject of section 4.

2. INTERPOLATION OF GRAVITY ANOMALIES. It is well known that an accurate analytical representation of free-air anomalies in pronounced mountainous terrain can only be achieved by a polynomial of high degree by means of Δg -data available in a network of high resolution. As a consequence, satisfactory linear interpolation requires small mesh sizes $\Delta x, \Delta y$. The following modified anomalies have been

useful for geodetic applications and the purpose of interpolation:

$$\Delta g_F = \Delta g + C \quad (1)$$

where C is the terrain correction, is called Faye anomaly.

$$\Delta g_B = g_F - bh + \frac{3\delta T}{2R} \quad (2)$$

is the modified Bouguer or complete topographic anomaly where $b = 0.112 \text{ mgal m}^{-1}$ is the Bouguer gradient, h is the elevation of terrain, δT is the potential of topographic masses, and $R = 6371 \text{ Km}$ is the earth's mean radius.

$$\Delta g_I = \Delta g_B + C_I + a\delta\zeta = \Delta g + C - bh + C_I + a\delta\zeta + r \quad (3)$$

is the isostatic anomaly valid for the compensated geoid with $a = 0.3086 \text{ mgal m}^{-1}$, $\delta\zeta$ as the vertical separation between geoid and cogeoid, and r as a random error.

Equation (3) may be further written as

$$\Delta g = \Delta g_I + C_t + r \quad (4)$$

where C_t represents the aggregate of terms computable from the known topography.

In a more general form, also applicable to the optimal estimation of vertical deflections, equation (4) is reformulated as

$$m = s + n + r \quad (5)$$

In this equation, m is a "message" variable, s is a "signal" variable, n is deterministic or computable "noise," and r is random-type noise.

Under consideration of a linear signal estimation structure, a signal can then be optimally estimated as

$$\hat{s}_e = L(m_I - n_I - r_I) \quad (6)$$

where L denotes a linear operator and the subscripts e and I refer to the

estimation point P_e and measurement points P_f , respectively. The optimal measurement at P_e results as

$$\hat{m}_e = \hat{s}_e + n_e + r_e = L(m_i - n_i) + r_e - L(r_i) + n_e \quad (7)$$

The estimation error is

$$e(\hat{m}_e) = e(\hat{s}_e) + e[r_e - L(r_i)] \quad (8)$$

The corresponding estimation error resulting from the utilization of topographically unmodified measurements m_i is

$$\begin{aligned} e(m_e) &= \hat{s}_e - L(s_i) + r_e - L(r_i) + n_e - L(n_i) \\ &= e(\hat{s}_e) + e[r_e - L(r_i)] + n_e - L(n_i) \end{aligned} \quad (9)$$

Comparison of equation (9) with equation (8) shows that the non-optimal interpolation process is associated with a "topographic" estimation error $n_e - L(n_i)$ which becomes in general intolerable in moderate to rough mountainous terrain and thus induces the requirement of a fine mesh data grid.

The interpolation of isostatic anomalies by means of spatial collocation will be treated in conjunction with the interpolation of isostatic deflections of the vertical in section 4.

3. FORMULATION OF A VERTICAL DEFLECTION SOLUTION SUITABLE FOR OPTIMAL INTERPOLATION. Gravimetric-topographic solutions for the anomalous gravity potential and deflections of the vertical which inherently permit a "signal-noise" separation according to equation (5) have been established by Pellinen [1969], Moritz [1969], and Baussus von Luetzow [1971]. The latter emphasized that these essentially identical solutions are almost equivalent to those of Molodensky et al. [1962] and Br [1964], but are less data dependent, more direct from the computational view, and more advantageous for the utilization of artificial satellite data. The notations to

be used are the following:

$\zeta = T \gamma^{-1}$	height anomaly
T	anomalous gravity potential
γ	normal gravity
$\xi = -\left(\frac{\partial \zeta}{\partial x}\right)_{h=\text{const.}}$	prime vertical deflection
$\eta = -\left(\frac{\partial \zeta}{\partial y}\right)_{h=\text{const.}}$	meridian vertical deflection
$h = h_p$	elevation of terrain referring to moving point P
h_A	elevation of terrain referring to fixed computation point A
$B_1 = \text{arc } \frac{\partial h}{\partial x}$	northern terrain inclination
$B_2 = \text{arc } \frac{\partial h}{\partial y}$	eastern terrain inclination
$\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$	derivatives taken along the local horizon in a northern and eastern direction
G	global mean gravity
α	azimuth angle counted clockwise from north
ψ	angle between the radius vectors \vec{r}_A and \vec{r}_P originating at the earth's spherical center
$S(\psi)$	Stokes' function
$k = 6.67 \cdot 10^{-8} \text{ cm}^3 \text{ g}^{-1} \text{ sec}^{-2}$	gravitational constant
$\rho \approx 2.67 \text{ g cm}^{-3}$	standard density
R = 6371 Km	earth's mean radius
$l_0 = 2R \sin \frac{\psi}{2}$	see Figure 1
$l = (r_A^2 + r_P^2 - 2r_A r_P \cos \psi)^{\frac{1}{2}}$	see Figure 1
σ	unit sphere (full solid angle)
g	measured gravity
$\Delta g = g - \gamma$	gravity anomaly
C	terrain correction
$\Delta g_F = \Delta g + C$	Faye anomaly

$$b = 0.112 \text{ mgal m}^{-1}$$

Bouguer gradient

$$\delta T$$

potential of topographic masses

$$\Delta g_B = \Delta g_F - bh + \frac{3\delta T}{2R}$$

modified Bouguer or complete topographic anomaly

The geometry involving h_A , $h=h_p$, ψ , R , l_0 , and l is evident from Figure 1 below.

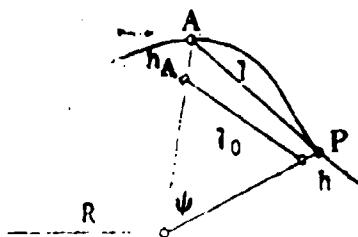


Figure 1

The established first order solution for the deflection components is

$$\left\{ \begin{array}{l} \varepsilon \\ \eta \end{array} \right\} = \frac{R}{4\pi G} \iint_{\sigma} (\Delta g_F + \delta g_1 + G_1) \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \frac{ds(\psi)}{d\psi} d\sigma + \left\{ \begin{array}{l} \delta \varepsilon \\ \delta \eta \end{array} \right\} - \frac{\Delta g + G_1}{G} \left\{ \begin{array}{l} \tan \beta_1 \\ \tan \beta_2 \end{array} \right\} \quad (10)$$

where

$$\delta g_1 = \frac{3}{2} k_p R \iint_{\sigma} \left(\ln \frac{l_1 + h - h_A}{l_0} - \frac{h - h_A}{l_0} \right) d\sigma, \quad l_1 = \sqrt{l_0^2 + (h - h_A)^2} \quad (11)$$

$$G_1 = \frac{R^2}{2\pi} \iint_{\sigma} \frac{\Delta g_B (h - h_A)}{l_0^3} d\sigma \quad (12)$$

$$\left\{ \begin{array}{l} \delta \varepsilon \\ \delta \eta \end{array} \right\} = \frac{k_p R^3}{G} \iint_{\sigma} \frac{h - h_A}{l_0^2} \left(\frac{1}{l_0} - \frac{1}{l_1} \right) \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \sin \psi d\sigma \quad (13)$$

It should be noted that δg is in general very small and that the computation of δg_1 , $\delta \epsilon$, and δn requires only integrations over $\sigma_1 \ll 4\pi$. It has further to be emphasized that, according to Baarda [1979], the inclination angles β_1 and β_2 should not exceed 70° .

Equation (10) is now reformulated under consideration of

$$\Delta g_F = \Delta \hat{g} + (\Delta g_F - \Delta \hat{g}) = \Delta \hat{g} + \delta g_2 \quad (14)$$

$$\Delta g_B = \Delta \hat{g} + (\Delta g_B - \Delta \hat{g}) = \Delta \hat{g} + \delta \hat{g}_3 \quad (15)$$

In these equations, $\Delta \hat{g}$ is a signal variable, profitably the isostatic anomaly defined in equation (3). In comparison with δg_2 , $\delta \hat{g}_3$ is a relatively smooth topographic quantity.

The substitutions (14) and (15) transform equation (10) into

$$\begin{aligned} \{\xi\} &= \frac{R}{4\pi G} \iint_{\sigma} [\Delta \hat{g} + G_1(\Delta \hat{g})] \left\{ \begin{array}{l} \cos \sigma \\ \sin \sigma \end{array} \right\} \frac{dS(\psi)}{d\psi} d\sigma - \frac{\Delta \hat{g} + G_1(\Delta \hat{g})}{G} \left\{ \begin{array}{l} \operatorname{tn} \beta_1 \\ \operatorname{tn} \beta_2 \end{array} \right\} \\ &+ \frac{R}{4\pi G} \iint_{\sigma} [\delta g_1 + \delta g_2 + G_1(\delta \hat{g}_3)] \left\{ \begin{array}{l} \cos \sigma \\ \sin \sigma \end{array} \right\} \frac{dS(\psi)}{d\psi} d\sigma + \left\{ \begin{array}{l} \delta \epsilon \\ \delta n \end{array} \right\} + \frac{C - \delta g_2 - G_1(\delta \hat{g}_3)}{G} \left\{ \begin{array}{l} \operatorname{tn} \beta_1 \\ \operatorname{tn} \beta_2 \end{array} \right\} \quad (16) \end{aligned}$$

The first two terms of equation (16), involving the anomaly $\Delta \hat{g}$, represent the "signal" components of ξ and n . The following three terms constitute computable topographic "noise." Permitting for random-type errors r_ξ and r_n , equation (16) can be written in analogy with equations (4) and (5) as

$$\{\xi\} = \{\hat{\xi}\} + \left\{ \begin{array}{l} \delta \epsilon \\ \delta n \end{array} \right\} + \{r_\xi\} \quad (17)$$

The numerical determination of the three topographic terms of equation (16) is a complex task, which can, however, be accomplished without inherent difficulties by means of high-speed computers. In this respect, the integration

area relating to the first topographic term can be considerably restricted. It appears that the last two topographic terms are particularly subject to rapid changes in mountainous terrain. Accurate interpolation is further favored if given and estimated deflections refer to points associated with small terrain inclinations.

In accordance with Moritz [1969], the second order correction for the height anomaly is

$$\delta \zeta^{(2)} = \frac{R}{4\pi} \iint_{\sigma} G_2(\Delta \hat{g} + \delta \hat{g}_3) S(\psi) d\sigma - \frac{R^2}{4\pi} \iint_{\sigma} (\Delta \hat{g} + \delta \hat{g}_3) \frac{(h-h_A)^2}{T_0^3} d\sigma \quad (18)$$

where

$$G_2 = \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_A}{T_0^3} G_1(\Delta \hat{g} + \delta \hat{g}_3) d\sigma + (\Delta \hat{g} + \delta \hat{g}_3) \tan^2 \beta_m \quad (19)$$

Here, β_m represents the maximal terrain inclination.

The second order deflection corrections are then

$$\begin{aligned} \left\{ \begin{array}{l} \delta \xi^{(2)} \\ \delta n_t^{(2)} \end{array} \right\} &= \left\{ \begin{array}{l} \delta \xi^{(2)}(\Delta \hat{g}) \\ \delta n_t^{(2)}(\Delta \hat{g}) \end{array} \right\} + \frac{R}{4\pi} \iint_{\sigma} G_2(\delta \hat{g}_3) \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \frac{dS(\psi)}{d\psi} d\sigma \\ &\quad + \frac{3R^3}{4\pi} \iint_{\sigma} \delta \hat{g}_3 \frac{(h-h_A)^2}{T_0^4} \left\{ \begin{array}{l} \cos \alpha \\ \sin \alpha \end{array} \right\} \frac{dI_0}{d\psi} d\sigma \end{aligned} \quad (20)$$

Designating the integral terms of equation (20) as second order topographic corrections $\delta \xi_t^{(2)}$ and $\delta n_t^{(2)}$, equation (17) assumes the modified form

$$\left\{ \begin{array}{l} \xi_n^{(2)} \\ \hat{\eta}^{(2)} \end{array} \right\} = \left\{ \begin{array}{l} \hat{\xi}^{(2)} \\ \hat{\eta}^{(2)} \end{array} \right\} + \left\{ \begin{array}{l} \delta \xi_t^{(2)} \\ \delta n_t^{(2)} \end{array} \right\} + \left\{ \begin{array}{l} \delta \xi_t^{(2)} \\ \delta n_t^{(2)} \end{array} \right\} + \left\{ \begin{array}{l} r_\xi^{(2)} \\ r_\eta^{(2)} \end{array} \right\} \quad (21)$$

Higher order topographic correction terms are not warranted because of a decreasing convergence radius in connection with higher derivatives, the assumption of a standard density or density uncertainties, respectively, and imperfect isostatic equilibrium. The structure of equation (21) clearly exhibits the fact that a highly accurate computation of $\xi^{(2)}$ and $n^{(2)}$ cannot be achieved by the exclusive utilization of free-air anomalies Δg . For the same reason, iterative solutions of the integral equations for generalized surface densities by Molodensky et al. [1962] and Brovar [1964] and the series solution by Molodensky et al. [1962] in general do not converge in mountainous terrain. The latter permits for auxiliary boundary surfaces under utilization of a shrinking parameter $k \leq 0$ and thus implies the possibility of analytical continuation with $\rho=0$. For the same reason, collocation solutions would only satisfactorily apply with respect to signal variables ξ and n . The analytical upward continuation of the first derivatives of the anomalous gravity potential in mountainous terrain would require a supplemental approach.

4. SIGNAL ESTIMATION BY STATISTICAL COLLOCATION AND FIRST ORDER EXPANSIONS OF PLANAR COVARIANCE FUNCTIONS. As indicated by Baussus von Luetzow [1980], deflection differences in flat terrain may be advantageously determined by a combination of statistical collocation and Vening Meinesz formulae provided gravity anomalies are also available in sufficient density within a limited region. Four point deflection estimation errors with mesh sizes $\Delta x = 5$ km, 8 km, and 24 km were found to be, respectively, of the order 0.1 arcsec, 0.2 arcsec, and 1.0 arcsec in the case of estimators free of errors. Astrogeodetically determined deflections are, however, presently associated with errors of the order of 0.25 arcsec. In accordance herewith, it is advantageous to employ a relatively great number of estimators if this is feasible.

The signal variable to be estimated and representing either ξ or n may be

\hat{x}_e , and the estimators may be written $x_1 + \delta_1$ with δ_1 as a correlated measurement error independent of x_1 . Under the assumption of an existing signal and noise covariance structure the following linear regression equation can be formulated:

$$\hat{x}_e = \sum a_i (\hat{x}_1 + \delta_1) = A_1 (\hat{x}_1 + \Delta_1) \quad (22)$$

It is then in matrix form, with bars indicating covariances,

$$\overline{\hat{x}_e \hat{x}_k} = A_1 (\overline{\hat{x}_1 \hat{x}_k} + \overline{\Delta_1 \Delta_k}) = A_1 N_{ik}, \quad \left\{ \begin{array}{l} i \\ k \end{array} \right\} = 1, 2, \dots, n \quad (23)$$

The solution for the regression coefficient matrix follows as

$$A_1 = \overline{\hat{x}_e \hat{x}_k} N_{ik}^{-1} \quad (24)$$

In the case of given astrogeodetic vertical deflections, δ_1 may be composed of astrogeodetic errors with a variance $(0.25 \text{ arcsec})^2$ and a correlated error partially caused by imperfect isostatic equilibrium.

With respect to the basis for the statistical collocation approach in physical geodesy, reference is made to Bjerhammar [1973], Grafarend [1973], Krarup [1969], Lauritzen [1971], Moritz [1970], and Tscherning [1973]. Of significance is that the spatial covariance function for the disturbing gravity potential has to satisfy Laplace's equation. Baussus von Luetzow [1973] emphasized the necessity to treat $\zeta - \bar{\zeta}$ as a correlated random variable where ζ is a deterministic development of τ in spherical harmonics of at least degree and order 15. In accordance herewith, the requirement of homogeneity prescribes and at least permits in practice a restriction to the planar approach in physical geodesy. Accordingly, $\frac{\partial T}{\partial z} = -g$, $\frac{\partial T}{\partial x}$, and $\frac{\partial T}{\partial y}$ are supposed to satisfy Laplace's equation. It is realized that the convenient requirements of homogeneity and quasi-flat terrain are only approximately satisfied.

Moritz [1976] and Nash and Jordan [1978] established specific T-covariance functions which can be expanded into space in a closed form. As has been shown by the latter authors, the spatial covariance function is

$$\phi_{TT}(r, z_1, z_2) = \int_0^\infty \rho F(\rho) e^{-\rho(z_1+z_2)} J_0(r\rho) d\rho \quad (25)$$

$$F(\rho) = \int_0^\infty r \phi_{TT}(r) J_0(r\rho) dr \quad (26)$$

In equations (25) and (26), J_0 is the zero-order Bessel function, r is the variable planar distance, z_1 and z_2 are the elevations of two points, and $\phi_{TT}(r)$ is the planar T - covariance function.

The spatial vertical deflection covariances may be derived from equation (25) in the form

$$\begin{Bmatrix} \phi_{EE}(r, z_1, z_2) \\ \phi_{nn}(r, z_1, z_2) \end{Bmatrix} = -(\gamma_1 \gamma_2)^{-1} \begin{Bmatrix} \frac{\partial^2}{\partial x^2} \\ \frac{\partial^2}{\partial y^2} \end{Bmatrix} \phi_{TT}(r, z_1, z_2) \quad (27)$$

where $\gamma_1 = \gamma(z_1), \gamma_2 = \gamma(z_2)$.

For ϕ_{TT} -functions which permit the derivation of realistic vertical deflection covariance functions, the Hankel transforms (26) and (25) cannot be evaluated in closed form. As an example, Jordan's [1972] third-order Markov model

$$\phi_{TT}(r) = \text{var } T \left(1 + \frac{r}{D} + \frac{r^2}{3D^2}\right) e^{-\frac{r}{D}} \quad (28)$$

leads to the hypergeometric function when introduced in equation (26). Thereafter, $\phi_{TT}(r, z_1, z_2)$ only can be obtained by an extremely lengthy numerical integration. For this reason, it appears to be advantageous to develop first order approximations for spatial vertical deflection covariance functions under consideration of Jordan's [1972] planar results. In this respect it has to be emphasized that Jordan

interchanged the conventional partial differentiations $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$.

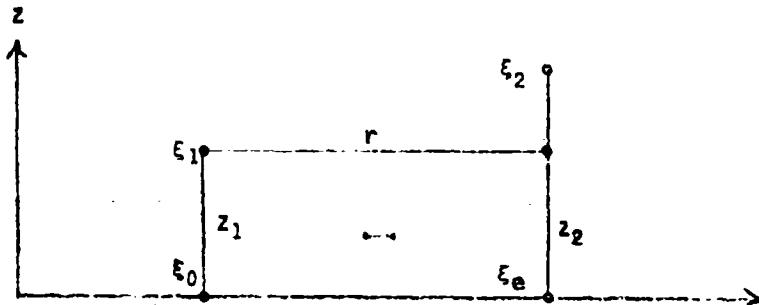


Figure 2

Under consideration of Figure 2, it is to the first order

$$\xi_1 = \xi_0 + \left(\frac{\partial \xi}{\partial z}\right)_0 z_1 \quad ; \quad \xi_2 = \xi_e + \left(\frac{\partial \xi}{\partial z}\right)_e z_2 \quad (29)$$

It is then, to the first order,

$$\overline{\xi_1 \xi_2} = \overline{\xi_0 \xi_e} + \overline{\xi_0 \frac{\partial \xi}{\partial z} z_2} + \overline{\xi_e \frac{\partial \xi}{\partial z} z_1} = \overline{\xi_0 \xi_e} + \overline{\xi_0 \frac{\partial \xi}{\partial z}} (\overline{z_1} + \overline{z_2}) \quad (30)$$

It is further at level $z = 0$

$$\frac{\partial \xi}{\partial z} = - \frac{\partial}{\partial y} \frac{\partial \xi}{\partial z} = - \frac{\partial}{\partial y} \left(- \frac{\Delta g}{G} \right) = \frac{1}{G} \frac{\partial \Delta g}{\partial y} \quad (31)$$

so that

$$\overline{\xi_0 \frac{\partial \xi}{\partial z}} = \overline{\xi_0} \cdot \frac{1}{G} \overline{\frac{\partial \Delta g}{\partial y}} = \frac{1}{G} \frac{\partial}{\partial y} \overline{\xi_0 \Delta g_e} \quad (32)$$

Under consideration of

$$\overline{\xi_0 \Delta g_e} = - \overline{\xi_e \Delta g_0} = - \frac{3\alpha \xi^{\sigma} g}{\sqrt{2}} h(r) \frac{y}{r} \quad (33)$$

it is

$$\frac{\partial}{\partial y} \left[h(r) \frac{y}{r} \right] = \frac{\partial h(r)}{\partial y} \frac{y}{r} + h(r) \left(\frac{1}{r} - \frac{y^2}{r^3} \right) = \frac{\partial h}{\partial r} \cos^2 \alpha + \frac{h}{r} \sin^2 \alpha \quad (34)$$

The final results are hereafter

$$\overline{\varepsilon_1 \varepsilon_2}(r, \alpha, z_1, z_2) = \phi_{\varepsilon\varepsilon} - \frac{3\sigma_{\varepsilon}\sigma_g}{\sqrt{2}G} \left(\frac{h}{r} \sin^2\alpha + \frac{\partial h}{\partial r} \cos^2\alpha \right) (z_1 + z_2) \quad (35)$$

$$\overline{n_1 n_2}(r, \alpha, z_1, z_2) = \phi_{nn} - \frac{3\sigma_n\sigma_g}{\sqrt{2}G} \left(\frac{h}{r} \cos^2\alpha + \frac{\partial h}{\partial r} \sin^2\alpha \right) (z_1 + z_2) \quad (36)$$

where $\phi_{\varepsilon\varepsilon}$ and ϕ_{nn} represent the planar covariance functions $\overline{\varepsilon_0 \varepsilon_e}$ and $\overline{n_0 n_e}$, respectively, and where $\sigma_g = \sigma_{\Delta g} = (\text{Var} \Delta g)^{1/2}$.

In analogy with equation (30), it is

$$\overline{\Delta g_1 \Delta g_2} = \overline{\Delta g_0 \Delta g_e} + \overline{\Delta g_0 \frac{\partial \Delta g_e}{\partial z}} (z_1 + z_2) \quad (37)$$

Under utilization of the planar approximation

$$\frac{\partial \Delta g}{\partial z} = -G \left(\frac{\partial \varepsilon}{\partial y} + \frac{\partial n}{\partial x} \right) \quad (38)$$

it is

$$-\overline{G \Delta g_0 \left(\frac{\partial \varepsilon_e}{\partial y} + \frac{\partial n_e}{\partial x} \right)} = -G \left(\frac{\partial}{\partial y} \overline{\Delta g_0 \varepsilon_e} + \frac{\partial}{\partial x} \overline{\Delta g_0 n_e} \right) \quad (39)$$

With the aid of equation (33), equation (39) can be formulated as

$$-G \overline{\Delta g_0 \left(\frac{\partial \varepsilon_e}{\partial y} + \frac{\partial n_e}{\partial x} \right)} = -G \frac{3\sigma_{\varepsilon}\sigma_g}{\sqrt{2}} \left(\frac{h}{r} + \frac{\partial h}{\partial r} \right) \quad (40)$$

Accordingly,

$$\overline{\Delta g_1 \Delta g_2}(r, z_1, z_2) = \phi_{gg} - \frac{3G\sigma_{\varepsilon}\sigma_g}{\sqrt{2}} \left(\frac{h}{r} + \frac{\partial h}{\partial r} \right) (z_1 + z_2) \quad (41)$$

where $\phi_{gg} = \overline{\Delta g_0 \Delta g_e}$.

It is evident from equations (35), (36), and (41) that these represent convenient closed approximations of the three spatial covariance functions of particular interest. In general, the planar covariance functions should apply to the lowest z-level in a particular area of application.

5. CONCLUSION. The immediate accurate interpolation of gravity anomalies and deflections of the vertical in mountainous terrain is only possible from data provided in a grid of high resolution. Optimal interpolation from data given at points separated by distances of the order 5-10 km or from multiple data incorporating measurement noise with shorter spacing can be accomplished by an appropriate representation of gravity anomalies and deflections as a signal-noise process with nonstationary noise computable from the earth's topography. In the case of deflections, a special solution of the geodetic boundary value problem is required. As a first approximation, Faye anomalies may be used as signal variables. Isostatic anomalies modified by the indirect effect provide a greater degree of homogeneity and isotropy. Implementation of the theory requires the utilization of existing "isostatic" computer programs and the establishment of a supplemental program under consideration of furnished analytical solutions. Signal estimation has to be facilitated by the use of spatial covariance functions first order approximations of which may be computed relatively easily from planar covariance expressions. The optimal interpolation method in conjunction with the special solution for deflections indicates that iterative or series solutions of the boundary value problem of physical geodesy cannot be expected to converge in mountainous terrain. The method developed is of practical significance for the densification of gravity anomaly and astrogeodetic deflection networks in mountainous terrain and is also valuable or indispensable, respectively, for the optimal estimation of gravity anomalies and deflections from astrogeodetic and inertial data in mountainous areas.

REFERENCES

- Baarda, W. 1979. A Connection Between Geometric and Gravimetric Geodesy, A First Sketch. Netherlands Geodetic Commission Public. on Geodesy, New Series, Vol. 6, Nr. 4.
- Badekas, J., and I. Mueller. 1968. Interpolation of the Vertical Deflection from Horizontal Gravity Gradients. J. Geoph. Research, Vol. 73, No. 22, pp. 6869-6878.
- Baussus von Luetzow, H. 1971. New Solution for the Anomalous Gravity Potential. J. Geoph. Research, Vol. 76, No. 20, pp. 4884-4891.
- Baussus von Luetzow, H. 1973. A Review of Collocation Methods in Physical Geodesy and Meteorology. Paper, presented at the National Fall Meeting of the Am. Geoph. Union, San Francisco, CA, Dec 1973.
- Baussus von Luetzow, H. 1980. Optimal Densification of Deflections of the Vertical by Means of Astrogeodetic and Gravity Anomaly Data. Paper, presented at the Spring Meeting of the Am. Geophysical Union, Toronto, Canada, May 1980.
- Bjerhammar, A. 1973. A General Model for Optimal Prediction and Filtering in the Linear Space. Methoden und Verfahren der mathematischen Physik, Band 14, Bibliogr. Inst., Mannheim, W. Germany.
- Brovar, V. V. 1964. On the Solutions of Molodensky's Boundary Value Problem, Bull. Geod., No. 72, Paris.
- Grafarend, E. 1973. Geodetic Stochastic Processes. Methoden und Verfahren der mathematischen Physik, Band 14. Bibliogr. Inst., Mannheim, W. Germany.
- Heiskanen, W., and H. Moritz. 1967. Physical Geodesy. W. H. Freeman and Co., San Francisco and London.
- Jordan, S. K. 1972. Self-Consistent Statistical Models for the Gravity Anomaly, Vertical Deflections, and Undulations of the Geoid. J. Geoph. Research, Vol. 77, No. 20, pp. 3660-3670.
- Krarup, T. 1969. A Contribution to the Mathematical Foundation of Geodesy. Public. No. 44. Geodetic Institute, Copenhagen, Denmark.
- Lauritzen, S. 1971. The Probabilistic Background of Some Statistical Methods in Physical Geodesy. Public. No. 48. Geodetic Institute, Copenhagen, Denmark.
- Molodensky, M. S., V. F. Eremeev, and M. I. Yurkina. 1962. Methods for study of the external gravitational field and figure of the earth. Transl. from Russian (1960). Israel Program for Scientific Translations, Jerusalem.
- Moritz, H. 1969. Nonlinear Solutions of the Geodetic Boundary Value Problem. Rep. 126, Dept. of Geodetic Science, Ohio State University,

Columbus, Ohio 43212.

Moritz, H. 1970. Least-Squares Estimation in Physical Geodesy. Rep. 130, Dept. of Geodetic Science, Ohio State University, Columbus, Ohio, 43212.

Moritz, H. 1976. Covariance Functions in Least-Squares Collocation. Rep. 240, Dept. of Geodetic Science, Ohio State University, Columbus, Ohio 43212.

Nash, R. A., and S. K. Jordan. 1978. Statistical Geodesy-An Engineering Perspective. Proc. IEEE, Vol. 66, No. 5, pp. 532-550.

Pellinen, L. P. 1969. On the Computations of Plumbline Deflections and Quasigeoidal Heights in Highlands (in Russian with English abstract), Tr. ZNIGAiK, 176, Nedra Press, Moscow.

Tscherning, C. C. 1973. Application of Collocation Determination of a Local Approximation to the Anomalous Potential of the Earth Using "Exact" Astro-Gravimetric Collocation. Methoden und Verfahren der mathematischen Physik, Band 14, Bibliogr. Inst., Mannheim, W. Germany.

A Sequential k-Group Random Allocation Method
with Applications to Simulation

Andrew P. Soms*

Abstract

A sequential method of random allocation is given and it is shown how it can be used to estimate the observed significance levels of k-sample nonparametric tests. The sequential technique is compared to the standard random allocation technique and shown to be more efficient. An application is made to the Dunn-Bonferroni method of multiple comparisons.

*University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the Office of Naval Research under Contract No. N00014-79-C-0321.

The author of this paper presented it at the 25th Conference on the Design of Experiments.

1. The Sequential Allocation Method

Bebbington (1975) showed that if there were N objects (such as file cards) from which it was desired to select (without replacement here and throughout) a random sample of size k without numbering the N objects, then one could proceed sequentially by selecting the first object with probability k/N and if at the T^{th} stage s have been selected, then the $T+1^{\text{st}}$ object is selected with probability $(k-s)/(N-T)$, $T = 1, 2, \dots, N-1$.

We now state and prove the extension to an arbitrary number of groups. Suppose there are N objects and it is desired to sequentially divide them randomly into r groups of size k_1, k_2, \dots, k_r , $\sum_{i=1}^r k_i = N$, i.e., each allocation has probability $1/\binom{N}{k_1, \dots, k_r}$. Let s_{iT}, \dots, s_{rT} be the number of objects selected for groups $1, 2, \dots, r$ at the T^{th} stage and let $p_{i,T+1}$ denote the selection probability for group i at the $T+1^{\text{st}}$ stage. Then if

$$p_{i,T+1} = (k_i - s_{iT})/(N-T), \quad T = 0, 1, \dots, N-1, \quad (1.1)$$

the selection is random. Note that $p_{i,1} = k_i/N$ and $\sum_{i=1}^r p_{i,T+1} = 1$. The randomness follows immediately by noting that the probability of a particular assignment is

$$\left(\prod_{i=1}^r k_i! \right) / N! = 1/\binom{N}{k_1, \dots, k_r}.$$

Bebbington's (1975) result is a special case of the above when $r = 2$.

As an example, suppose $r = 3$, $k_1 = 2$, $k_2 = 2$, $k_3 = 3$ and $N = 7$. In order to make the sequential allocation given by (1.1) we take

7 independent random numbers U_i , $i = 1, 2, \dots, 7$. Let

$$Q_{0,T} = 0 \text{ and } Q_{i,T} = \sum_{j=1}^i P_{j,T}, \quad i = 1, 2, \dots, r, \quad T = 1, 2, \dots, N.$$

Then the m^{th} object, $m = 1, 2, \dots, N$, is assigned to group n , where n is the unique integer such that

$$Q_{n-1,m} < U_m \leq Q_{n,m}.$$

Suppose the 7 random numbers are .79039, .01850, .99744, .81812, .93169, .22705, and .97709. The selection process is summarized in Table 1.

1. Selection Process

<u>Stage</u>	<u>Random Digit</u>	<u>P_{1T}</u>	<u>P_{2T}</u>	<u>P_{3T}</u>	<u>Group Selected</u>
1	.79039	2/7	2/7	3/7	3
2	.01850	2/6	2/6	2/6	1
3	.99744	1/5	2/5	2/5	3
4	.81812	1/4	2/4	1/4	3
5	.93169	1/3	2/3	0	2
6	.22705	1/2	1/2	0	1
7	.97709	0	1	0	2

Note that if all the k_i 's are one, a random permutation is produced if we think of the group as denoting position.

7 independent random numbers U_i , $i = 1, 2, \dots, 7$. Let

$$Q_{0,T} = 0 \text{ and } Q_{i,T} = \sum_{j=1}^i P_{j,T}, \quad i = 1, 2, \dots, r, \quad T = 1, 2, \dots, N.$$

Then the m^{th} object, $m = 1, 2, \dots, N$, is assigned to group n , where n is the unique integer such that

$$Q_{n-1,m} < U_m \leq Q_{n,m}.$$

Suppose the 7 random numbers are .79039, .01850, .99744, .81812, .93169, .22705, and .97709. The selection process is summarized in Table 1.

1. Selection Process

<u>Stage</u>	<u>Random Digit</u>	<u>P_{1T}</u>	<u>P_{2T}</u>	<u>P_{3T}</u>	<u>Group Selected</u>
1	.79039	2/7	2/7	3/7	3
2	.01850	2/6	2/6	2/6	1
3	.99744	1/5	2/5	2/5	3
4	.81812	1/4	2/4	1/4	3
5	.93169	1/3	2/3	0	2
6	.22705	1/2	1/2	0	1
7	.97709	0	1	0	2

Note that if all the k_i 's are one, a random permutation is produced if we think of the group as denoting position.

2. Applications to Simulation

In k -sample nonparametric tests the observed significance level of the test is obtained by considering all possible partitions M of the (possibly tied) observed values or (possibly average) ranks into r groups, computing the value of the test statistic, and counting the number of times m it is equal to or greater than the observed value. The observed significance level $\hat{\alpha}$ is then m/M . When the number of partitions is large this is prohibitive and $\hat{\alpha}$ is estimated either by simulation (taking a large random sample of the allocations) or by asymptotics. The advantage of simulation is that one can control the accuracy of the estimate (by taking a large or small random sample) depending on the importance of the situation, unlike asymptotics which each time it is used forces one into the straight-jacket of committing a usually unknown error. Since it is (perhaps regrettably) a well known fact that different actions will be taken for close values of $\hat{\alpha}$, one above and the other below some fixed level (e.g., .01, .05, or .1) of the decision-maker, the use of simulation at least prevents approximating error in $\hat{\alpha}$ to be the determining factor.

If it is decided to use simulation, then a possible procedure is to make the random assignment as described in Section 1 many times by using a computer. The commonly used method is to produce a random permutation by ordering a random sample of uniform numbers and choosing the first k_1 indexes for group 1, the next k_2 for group 2, and so on. If all the k_i 's are one, then this is more efficient than Section 1. However, as soon as the k_i 's depart even moderately from 1, the method of Section 1 becomes much more efficient. As an example, if $k_1 = k_2 = k_3 = k_4 = 10$ and it

is desired to make 2000 random assignments using a UNIVAC 1110 computer, a FORTRAN program using the methods of Section 1 uses 4.71 seconds of CPU time while a FORTRAN program using the random permutation method takes 9.17 seconds.

The Appendix contains a listing of the FORTRAN subroutine RANDOM that uses the theory of Section 1 to make random assignments. This may be tied in with any specific simulation problem, e.g., the case treated in Section 3...

3. Applications to the Dunn-Bonferroni Method of Multiple Comparisons

The D-B (Dunn-Bonferroni) method is described in Dunn (1964). Briefly, let Y_{ij} , $i = 1, 2, \dots, r$, $j = 1, 2, \dots, n_i$, be continuous (this assumption is not important and is removed later) random variables with distribution function F_i , $H_0: F_1 = F_2 = \dots = F_r$, $H_a:$ for at least one pair (i, j) , $F_i \neq F_j$ (in the sense of producing larger or smaller values), and the test must identify which, if any, pairs are different. Denote by z_α the upper α^{th} point of the standard normal. The D-B test declares all those pairs (i, j) , $i < j$, different for which

$$z_{ij} = |\bar{R}_i - \bar{R}_j| / \left[\frac{(N)(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right]^{1/2} \geq z_\alpha / (k(k-1)) , \quad (3.1)$$

where \bar{R}_i denotes the average of the ranks of the i^{th} group in the joint ranking. The nominal significance level of this procedure is α . The actual significance level α_A is

$$\alpha_A = P_0 \left(\max_{i < j} z_{ij} \geq z_\alpha / (k(k-1)) \right) , \quad (3.2)$$

and may be obtained by simulation based on Section 1. Table 2 gives some comparisons of nominal with actual, using Section 1 and 10,000 simulations.

2. Comparison of Actual to Nominal α

<u>r</u>	<u>Common Group Size</u>	<u>Nominal α</u>	<u>Actual α</u>
3	5	.05	.037
3	10	.05	.040
3	15	.05	.043
3	30	.05	.045
3	5	.01	.0030
3	10	.01	.0077
3	15	.01	.0077
3	30	.01	.010
5	5	.05	.026
5	10	.05	.036
5	5	.01	.0030
5	10	.01	.0067

The Appendix contains a listing of the program used for Table 2. It thus appears that D-B is conservative and we can remove the conservatism by substituting for $z_\alpha/(k(k-1))$ $d_{(i)}$, where $d_{(i)}$, $i = 1, 2, \dots, r(r-1)/2$, is the i^{th} largest observed values of z_{ij} , $i < j$, to obtain by simulation the $r(r-1)/2$ possible observed significance levels.

The K-S (Kruskal-Scheffé) method is also sometimes used in this situation (see, e.g., Miller, 1966, p. 166) and consists of replacing $z_\alpha/(r(r-1))$ in (3.1) with $h_\alpha^{1/2} = (x_{\alpha;r-1}^2)^{1/2}$, where $x_{\alpha;r-1}^2$ is the upper α^{th} point of χ^2 with $r-1$ degrees of freedom. The comparison of the critical constants in Table 3 shows that this is even more conservative than D-B.

3. Comparison of D-B and K-S Critical Constants

r	$z_{.05}/(r(r-1))$	$(\chi^2_{.05;r-1})^{1/2}$	$z_{.01}/(r(r-1))$	$(\chi^2_{.01;r-1})^{1/2}$
3	2.39	2.79	2.94	3.36
4	2.50	3.08	3.02	3.65
5	2.58	3.33	3.09	3.89
6	2.64	3.55	3.15	4.10
7	2.69	3.75	3.19	4.30
8	2.74	3.94	3.23	4.48
9	2.77	4.11	3.26	4.66
10	2.81	4.28	3.29	4.82

If the data is discrete, the D-B method can be modified as in Dunn (1964) and the random assignment done on average ranks. Thus ties present no problems in this approach.

The third method discussed in Miller (1964), the Steel many-one rank statistics, is too time-consuming for the simulation approach. For all practical purposes the exact D-B (use of the $d_{(i)}$ and simulation) seems the best method to use.

References

- Bebbington, A. C. (1975), "A Simple Method of Drawing a Sample Without Replacement", Applied Statistics, 24, 136.
- Dunn, Olive J. (1964), "Multiple Comparisons Using Rank Sums", Technometrics, 6, 241-52.
- Miller, Rupert J. (1966), Simultaneous Statistical Inference, New York: McGraw-Hill Book Company.

Appendix

```

C EXAMPLE OF MATV PROGRAM FOR SEQUENTIAL RANDOM ALLOCATIONS
C NR IS THE NUMBER OF GROUPS, NSIM THE NUMBER OF SIMULATIONS, K'S THE GROUP
C SIZES, -X'S- THE NUMBERS TO BE ALLOCATED
      DIMENSION X(1000),Z(20,100),K(20)
100 FORMAT ( )
99 READ 100,NR,NSIM
      IF (NR.EQ. 0) GO TO 101
      READ 100,(K(J),J=1, "R")
      KK=0
      DO 1 J=1,NR
1      KK=KK+K(J)
      AA 203 I=1,KK
203 X(I)=I
      DO 2 I=1,NSIM
      CALL RAND((NR,K,KK,X,7))
2      CONTINUE
      GO TO 99
101 STOP
      END
*FOR,IS .SUB1
      SHROUTINE RAND((NR,K,KK,X,Z)
C NR NUMBER OF GROUPS, K ARRAY OF GROUP SIZES, KK NUMBER OF ELEMENTS, X ARE
C OF KK ELEMENTS, Z RANDOM ALLOCATION OF X
      DIMENSION X(1:20),Z(20,100),SC(20),E(20),AC(70),U(1700),K(20)
      DO 3 I2=1,NR
3      NSC(I2)=0
      AA 333 I2=1,KK
333 U(I2)=RAND(X)
      DO 4 I3=1,NR
4      AC(I3)=0
      4 0(I3)=K(I3)
      MAXI=NR-1
      DO 5 I1=1,MAXI
      DO 5 III=1,II
5      AC(II)=AC(II)+0(III)
      AA 6 I4=1,KK
6      U(I4)=(KK-[4+1].1*U(I4))
      IF (U(I4).LT.0(1)) GO TO 61
      AA 7 I5=2,MAXI
7      IF (U(I4).LE.AC(I5)) GO TO 62
      IF (U(I4).GT.AC(NR-1)) INDEX=NR
      AA TR 64
61 INDEX=1
      GO TO 64
62 INDEX=I5
64 NSC(INDEX)=NSC(INDEX)+1
      NN=NSC(INDEX)
      Z(INDEX,NN)=X(I4)
C UPDATE
      DO 8 I6=1,NR
8      AC(I6)=0
      8 0(I6)=K(I6)-NSC(I6)
      DO 9 II=1,MAXI
9      AC(II)=AC(II)+0(II)
      9 0(II)=0
      6 CONTINUE
      AA TR 64
      END

```


Twenty-Sixth Conference on the Design of Experiments
in Army Research, Development, and Testing (22-24 Oct 80)

PARTICIPANTS

Mr. William Agee
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mr. Thomas W. Alexander
US Army White Sands Missile Range
ATTN: STEWS-QA-E
White Sands Missile Range, NM 88002

Mr. William Anderson
US Army White Sands Missile Range
ATTN: STEWS-NR-AD-A
White Sands Missile Range, NM 88002

Mr. Francis J. Anscombe
Box 2179, Yale Station
New Haven, CT 06520

Mr. Angelo Arcaro
US Army White Sands Missile Range
ATTN: STEWS-ID-ER
White Sands Missile Range, NM 88002

Mr. Elton P. Avara
US Army Electronics Research and
Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-BE
White Sands Missile Range, NM 88002

Mr. William E. Baker
US Army Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005

Mr. Carl B. Bates
US Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20014

LTC James S. Blesse
US Army Operational Test and
Evaluation Agency
5600 Columbia Pike
Falls Church, VA 22041

Mr. Douglas Bulgiano
Department of Experimental Statistics
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. Thomas C. Bumgarner
US Army White Sands Missile Range
ATTN: STEWS-QA-E
White Sands Missile Range, NM 88002

Mr. Robert J. Burge
Department of Biostatistics
Walter Reed Army Institute of Research
Washington, D.C. 20012

Mr. Donald Buttz
US Army White Sands Missile Range
ATTN: STEWS-TE-PC
White Sands Missile Range, NM 88002

Ms. Angelina Buttz
7736 Iroquois
El Paso, TX 79912

Mr. Patrick D. Cassidy
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TB
White Sands Missile Range, NM 88002

Mr. Cesar Castillo
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mr. Oscar J. Castro
US Army White Sands Missile Range
ATTN: STEWS-TE-LD
White Sands Missile Range, NM 88002

Mr. Danny C. Champion
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TFC
White Sands Missile Range, NM 88002

Mr. Jagdish Chandra
US Army Research Office
Mathematics Division
PO Box 12211
Research Triangle Park, NC 27709

Mr. German Chavez
US Army White Sands Missile Range
ATTN: STEWS-MR-PR
White Sands Missile Range, NM 88002

Mr. E. L. Church
US Army Armament Research & Development Command
Dover, NJ 07801

Mr. W. J. Conover
Texas Tech University
Lubbock, TX 79409

Mr. Davis Cope
Center for Naval Analyses
2000 N. Beauregard
Alexandria, VA 22311

Mr. Paul C. Cox
2930 Huntington Drive
Las Cruces, NM 88001

Mr. Larry H. Crow
US Army Materiel Systems Analysis Activity
ATTN: AMSAA-DRXSY-RM
Aberdeen Proving Ground, MD 21005

Mr. G. A. Culpepper
Mesilla Park, NM 88047

Mr. Richard Dale
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mr. Oren Dalton
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mr. Carl Diegert
Sandia National Laboratory
Albuquerque, NM 87115

Mr. Leonard D. Donahe
US Army White Sands Missile Range
ATTN: STEWS-NR-CF
White Sands Missile Range, NM 88002

Mr. Frank T. Dylla
US Army White Sands Missile Range
ATTN: STEWS-TE
White Sands Missile Range, NM 88002

Mr. Robert Easterling
Sandia National Laboratory
Albuquerque, NM 87115

Mr. Vince Ercolano
US Army White Sands Missile Range
ATTN: STEWS-PA
White Sands Missile Range, NM 88002

Mr. Oskar Essenwanger
US Army Missile Command
ATTN: DRSMI-RRA
Redstone Arsenal, AL 35898

Mr. Roberto Fierro
US Army White Sands Missile Range
ATTN: STEWS-NR-AS
White Sands Missile Range, NM 88002

Mr. Morris Finkner
Department of Experimental Statistics
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. Andrus Garay
US Army White Sands Missile Range
ATTN: STEWS-ID-DH
White Sands Missile Range, NM 88002

Mr. Charles R. Garcia
US Army White Sands Missile Range
ATTN: STEWS-NR-CF
White Sands Missile Range, NM 88002

Mr. Gerald Garfinkel
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mrs. Virginia Gildengorin
Letterman Army Institute of Research
ATTN: SGRD-ULZ-I
Presidio San Francisco, CA 94129

Mr. Robert Gilmore
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, VA 22202

Mr. Richard Glaze
Department of Experimental Statistics
New Mexico State University
Las Cruces, NM 88003

Mr. Jose Gomez
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mr. James Graves
US Army White Sands Missile Range
ATTN: STEWS-NR-C
White Sands Missile Range, NM 88002

Mr. Robert E. Green
US Army White Sands Missile Range
ATTN: STEWS-ID
White Sands Missile Range, NM 88002

Mr. Frank Grubbs
US Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, MD 21005

Mr. Douglas Gladden
US Army White Sands Missile Range
ATTN: STEWS-NR-AD-A
White Sands Missile Range, NM 88002

Mr. Gavin Gregory
University of Texas at El Paso
El Paso, TX 79958

Mr. Carroll Hall
Department of Experimental Statistics
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. James T. Hall
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-EO
White Sands Missile Range, NM 88002

Mr. Bernard Harris
University of Wisconsin
Department of Statistics
1210 W. Dayton Street
Madison, Wisconsin 53706

Mr. Larry Hatch
National Security Agency
Fort Meade, MD 20755

Mr. Glenn Herman
US Army White Sands Missile Range
ATTN: STEWS-NR-CF
White Sands Missile Range, NM 88002

Mr. George Hoffman
US Army White Sands Missile Range
ATTN: STEWS-TE-LG
White Sands Missile Range, NM 88002

Ms. Peggy Hoffer
US Army White Sands Missile Range
ATTN: STEWS-PL
White Sands Missile Range, NM 88002

Mr. Glenn Hoidal
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-AS
White Sands Missile Range, NM 88002

Mr. Herbert Holman
National Security Agency
Fort Meade, MD 20755

Mr. Donald Hoock
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-EO
White Sands Missile Range, NM 88002

Mr. Mohammed Hussain
Department of Experimental Statistics
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. Herbert Hamilton
US Army White Sands Missile Range
ATTN: STEWS-NR-AC
White Sands Missile Range, NM 88002

Mr. Mark Houldsworth
New Mexico State University
Physical Science Laboratory
Las Cruces, NM 88003

Mr. Woody Jenkins
US Army White Sands Missile Range
ATTN: STEWS-QA-E
White Sands Missile Range, NM 88002

Mr. James L. Jernigan
US Army White Sands Missile Range
ATTN: STEWS-ID-T
White Sands Missile Range, NM 88002

Mr. Richard A. Johnson
University of Wisconsin
Madison, Wisconsin 53706

Mr. Ronald L. Johnson
US Army MERADCOM
Countersurveillance Laboratory
Fort Belvoir, VA 22060

Ms. Karen A. Joyce
US Army White Sands Missile Range
ATTN: STEWS-TE-PD
White Sands Missile Range, NM 88002

Ms. Elizabeth Junkins
Directorate of Mathematics and Statistics
Department of National Defense
Ottawa, Canada K1AOK2

Mr. D. G. Kabe
New Mexico State University
Las Cruces, NM 88003

Mr. W. D. Kaigh
Department of Mathematical Sciences
University of Texas at El Paso
El Paso, TX 79968

Mr. Roger A. King
US Army White Sands Missile Range
ATTN: STEWS-TE-MH
White Sands Missile Range, NM 88002

Mr. James R. Knaub
US Army White Sands Missile Range
ATTN: STEWS-TE-PD
White Sands Missile Range, NM 88002

Mr. Kenneth E. Kunkel
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-AS-P
White Sands Missile Range, NM 88002

Mr. P. S. Ladouceur
Director of Mathematics and Statistics
Operational Research and Analysis Establishment
Department of Defense
Ottawa, Canada K1AOK2

Mr. Robert L. Launer
US Army Research Office
PO Box 12211
Research Triangle Park, NC 27709

Mr. Eldin Leighton
New Mexico State University
Box 3-I
Las Cruces, NM 88003

Mr. Peter D. J. Leong
US Army White Sands Missile Range
ATTN: STEWS-MS
White Sands Missile Range, NM 88002

Mr. Larry L. Little
US Army White Sands Missile Range
ATTN: STEWS-TE-RE
White Sands Missile Range, NM 88002

Mr. Lonnie C. Ludeman
Electrical and Computer Engineering Department
New Mexico State University
Las Cruces, NM 88003

Mr. William McCanna
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Mr. Dale R. McLaughlin
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

Georges J. McLaughlin
Defense Research Establishment Valcartier
P.O. Box 830 Courcellette
Quebec G0A 1R0, Canada

Ms. Gloria Maese
Department of Experimental Statistics
New Mexico State University
Box 3130
Las Cruces, NM 88003

Ms. Mary Anne Maher
Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003

Mr. Kenneth R. Manion
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TG-1
White Sands Missile Range, NM 88002

LTC Jack McGrath
US Army Operational Test and Evaluation Agency
5600 Columbia Pike
Falls Church, VA 22041

Mr. John T. Marrs
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-BE-C
White Sands Missile Range, NM 88002

Mr. Richard T. Maruyama
US Army Ballistics Research Laboratory
ATTN: DRDAR-BLB
Aberdeen Proving Ground, MD 21005

Ms. Barbara M. Matteson
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TG-P
White Sands Missile Range, NM 88002

Mr. Toby J. Mitchell
Union Carbide Nuclear Division
Oak Ridge, TN 37830

Mr. Paul Mlaker
US Army Waterways Experimental Station
Vicksburg, MS 39180

Mr. J. Richard Moore
US Army Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005

Mr. Larry Moore
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-AS
White Sands Missile Range, NM 88002

Mr. Donald Neal
US Army Materials and Mechanics Research Center
Arsenal Street
Watertown, MA 02172

Mr. Roger D. Odom
US Army White Sands Missile Range
ATTN: STEWS-TE
White Sands Missile Range, NM 88002

Mr. Clem Ota
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. Ricardo Pena
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-BE-A
White Sands Missile Range, NM 88002

Ms. Karen D. Pettigrew
Theoretical Statistics and Mathematics Branch
NIMH, NIH, Building 36, Room ID-19
Bethesda, MD 20205

Mr. William S. Phoebus
Aberdeen Proving Ground
ATTN: STEAP-MT-G
Aberdeen Proving Ground, MD 21005

Mr. Charles A. Pollard
US Army White Sands Missile Range
ATTN: STEWS-TE-RE
White Sands Missile Range, NM 88002

LTC Richard Porter
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TCB
White Sands Missile Range, NM 88002

Ms. Constance Kelly Preston
University of Texas at El Paso
El Paso, TX 79968

Mr. Donald W. Rankin
US Army White Sands Missile Range
ATTN: STEWS-TE-RE
White Sands Missile Range, NM 88002

Mr. Charles D. Revie
US Army White Sands Missile Range
ATTN: STEWS-TE-PD
White Sands Missile Range, NM 88002

Mr. Winston Richards
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. Ruben D. Rodriguez
US Army Electronics Research and Development Command
Office of Missile Electronic Warfare
ATTN: DELEW-M-ADL
White Sands Missile Range, NM 88002

Mr. Gerald Rogers
Department of Mathematical Science
New Mexico State University
Las Cruces, NM 88003

Mr. Nathaniel Roman
US Army White Sands Missile Range
ATTN: STEWS-TE-MC
White Sands Missile Range, NM 88002

Mr. Carl T. Russell
Cold Regions Test Center
Seattle, WA 98111

Mr. Ruben Rede
US Army White Sands Missile Range
ATTN: STEWS-NR-AC
White Sands Missile Range, NM 88002

Mr. Ernesto Sanchez
US Army White Sands Missile Range
ATTN: STEWS-NR-A
White Sands Missile Range, NM 88002

1LT Juan Santiago-Marini
US Army Air Defense Center-Ft Bliss
ATTN: DCD, USAADS
Fort Bliss, TX 79916

Mr. Roger Schulz
US Army White Sands Missile Range
ATTN: STEWS-TE-MF
White Sands Missile Range, NM 88002

Mr. Eugene F. Schuster
University of Texas at El Paso
El Paso, TX 79968

Mrs. Mary Ann Seagraves
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-EO-MQ
White Sands Missile Range, NM 88002

Mr. William L. Shepherd
2104 Atlanta
El Paso, TX 79930

Mr. Dan Singleton
US Army White Sands Missile Range
ATTN: STEWS-TE-LG
White Sands Missile Range, NM 88002

Ms. Jill H. Smith
US Army Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005

Ms. Lounell Snodgrass
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TH
White Sands Missile Range, NM 88002

Mr. Victor Solo
Department of Statistics
Harvard University
Cambridge, MA 02143

Mr. Morris Southward
Department of Experimental Statistics
New Mexico State University
Las Cruces, NM 88003

Mr. Floyd Spencer
Sandia National Laboratory
Albuquerque, NM 87115

Mr. Donald M. Swingle
1765 Pomona Drive
Las Cruces, NM 88001

Mr. Douglas B. Tang
Department of Biostatistics
Walter Reed Army Institute of Research
Washington, DC 20012

Mr. Malcolm S. Taylor
US Army Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005

Mr. Jerry Thomas
US Army Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005

Mr. James R. Thompson
Rice University
Houston, TX 77001

Mr. Paul Thrasher
US Army White Sands Missile Range
ATTN: STEWS-QA-E
White Sands Missile Range, NM 88002

Mr. Robert Turner
US Army White Sands Missile Range
ATTN: STEWS-NR-AM
White Sands Missile Range, NM 88002

N. Scott Urquhart
Department of Experimental Statistics
New Mexico State University
Box 3130
Las Cruces, NM 88003

Mr. Robert Valencia
US Army White Sands Missile Range
ATTN: STEWS-NR-CF
White Sands Missile Range, NM 88002

Mr. Ernest Vigil
US Army White Sands Missile Range
ATTN: STEWS-NR-AS
White Sands Missile Range, NM 88002

Mr. Donald L. Walters
US Army Electronics Research and Development Command
Atmospheric Sciences Laboratory
ATTN: DELAS-AS
White Sands Missile Range, NM 88002

Mr. M. A. Weinberger
Directorate of Mathematics and Statistics
Operational Research and Analysis Establishment
Department of National Defense
Ottawa, Canada K1AOK2

Mr. Roger F. Willis
US Army TRADOC Systems Analysis Activity
ATTN: ATAA-TG-P
White Sands Missile Range, NM 88002

Mr. Tasmen A. Yauney
US Army White Sands Missile Range
ATTN: STEWS-TE-LG
White Sands Missile Range, NM 88002

Mr. H. Allen Wallis
Chancellor
University of Rochester
942 Wilson Blvd
Rochester, NY 14627

UNCLASSIFIED

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report 81-2	2. GOVT ACCESSION NO. AD-101442	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and subtitle) PROCEEDINGS OF THE TWENTY-SIXTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report
7. AUTHOR(s)	6. PERFORMING ORG. REPORT NUMBER	
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on Behalf of the Chief of Research, Development and Acquisition		12. REPORT DATE June 1981
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) U. S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709		13. NUMBER OF PAGES 587
		15. SECURITY CLASS. (of this report) Unclassified
		16. DECLASSIFICATION/DOWNGRADING SCHEDULE
17. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. The findings in this report are not to be construed as official Department of the Army position, unless so designated by other authorized documents.		
18. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
19. SUPPLEMENTARY NOTES This is a technical report from the Twenty-Sixth Conference on the Design of Experiments in Army Research, Development and Testing. It contains most of the papers presented at that meeting. These treat various Army statistical and design problems. Subject matter include:		
19. KEY WORDS (Continuation reverse side if necessary and identify by block number) bridge regression, response surface fitting, order statistics, radar signal improvement, digital filters, life time estimates, risks, forecasting, linear regression models, adaptive median smoothing, fitting an ellipse, approximating functions, selection criteria, armor combat models, experimental designs, probability densities, testing hypotheses, quantile estimation, rank transformation, design of field tests, crossover experiments, time series analysis, spectral limits, confidence limits, reliability, camouflage development, gravity anomalies, and allocation methods.		