ARO Report 78 - 2

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

Sponsored by

The Army Mathematics Steering Committee

on Behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT AND ACQUISITION

(9) Interim technical rept.

U. S. Army Research Office

(14) ARO-
Report No. 78-2

(11) July 1978

(12) 240 p.

(6)

PROCEEDINGS OF THE ~~TWENTY-THIRD~~ CONFERENCE ON THE DESIGN OF EXPERIMENTS (23rd)

Sponsored by the Army Mathematics Steering Committee

HOST

US Army Combat Developments Experimental Command

HELD AT

Naval Postgraduate School, Monterey, California

ON 19-21 October 1977

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

040 900

A

FOREWORD

Professor Karl Pearson wrote the following statement, "The field of science is unlimited; its material is endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science." If any one field of science exemplifies these remarks, it is the field of statistics. The papers in these Proceedings indicate a few areas where statistics and the design of experiments are helping the Army solve some of its many problems. Weapon system analysis is just one of those fields where statistics plays an important role. To bring this out, we quote a paragraph by Dr. Frank E. Grubbs which appears in the Engineering Design Handbook: DARCOM-P 706-101. "Chapter 21 brings us to the increasingly important topics of reliability, life testing, availability and maintainability of systems, and reliability growth. There is hardly any weapon system today which can or should escape analyses in terms of these fields of interest, and the analyst must be highly competent in evaluations associated with life-time or failure distributions such as the exponential, the Weibull, the lognormal, and the binominal probability distributions. Statistical testing for high reliability and safety of systems is introduced in Chapter 21, as well as a brief account of reliability growth. A major topic, and current effort, concerning systems today is that of being able to place confidence bounds on the true, unknown reliability of complex systems; accordingly, coverage of the more recent and accurate techniques is given for the practicing analyst. Finally, reliability now is often one of the major or sole, characteristics of some weapon systems, and hence may represent a prime activity for the systems analyst in many applications of his knowledge."

Except for the Nineteenth Conference on the Design of Experiments in Army Research, Development and Testing, which was conducted at Rock Island Arsenal, Rock Island, Illinois, the first twenty-two meetings of this series of symposia were held on the east coast. The concentration of Army installations in this area played a key role in selecting the hosts for these meetings. The Army Mathematics Steering Committee (AMSC) sponsors these conferences on behalf of the Office of the Chief of Research, Development and Acquisition. Members of the subcommittee on Probability and Statistics, whose responsibility it is to organize the Design Conferences for the AMSC, had some misgivings about holding the twenty-third meeting on the west coast. But these doubts were dispelled by the facts that the number of attendees as well as the number of contributed papers matched those of the east coast meetings. One anomaly did occur. Instead of having one fourth of the contributed papers classified as clinical, in the California meeting nearly one half were in this category.

The host for the twenty-third Design of Experiments Conference was the U. S. Army Combat Development Experimentation Command, Fort Ord, California. Excellent facilities for holding this meeting on 19-21 October 1977 were provided by the Naval Postgraduate School. Dr. Marion R. Bryson, acting for the host for the conference, served as Chairman on Local Arrangements. He was assisted in this task by Mr. John E. Banks and several other members of his staff. Those in attendance are grateful to them for so ably carrying out the many tasks that needed to be handled before and during the course of a meeting of this size.

The five nationally known invited speakers together with the titles of their addresses are listed below. These gentlemen gave those in attendance an opportunity to hear about recent developments in the field of statistics.

Speaker and Institution	Area of Talk
Prof. H. O. Hartley Texas A&M University	Analysis of Unbalanced Experiments
Prof. Norman Breslow University of Washington	Censored Data
Prof. Rupert Miller Stanford University	The Jackknife: Survey and Applications
Prof. Donald P. Gaver Naval Postgraduate School	Estimation of Complex System Availability
Prof. G. E. P. Box University of Wisconsin	Time Series Modelling

Dr. Churchill Eisenhart was recipient this year of the Samuel S. Wilks Memorial Medal. He richly deserves this honor for his scientific contributions. He has played many important roles in the conducting of these conferences. At this meeting there were forty-two contributed papers. Twenty-two of these were classified as technical and the rest were presented in clinical sessions. Ninety-six persons registered for the conference, but there were one hundred and eighteen individuals who attended the opening session.

The members of the AMSC are duly aware of all the effort that goes into making these conferences such memorable events. Their thanks go to all those in attendance. The speakers in particular need recognition for all the time they spent in preparing and delivering their interesting

papers. Dr. Frank E. Grubbs and Professor Herbert Solomon, who respectively served as Program Chairman and Chairman of the conference, are to be congratulated for guiding to conclusion another successful scientific meeting.

TABLE OF CONTENTS

AGENDA

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

19-21 October 1977

Host: Combat Developments Experimentation Command

Held at: Naval Postgraduate School

***** Wednesday, 19 October *****

0800-0900 Registration - Lobby of Ingersoll Hall

0900-1015 GENERAL SESSION I -- Ingersoll Hall, Room 122

CALLING OF THE CONFERENCE TO ORDER

Dr. Marion R. Bryson, Chairman of Local Arrangements, U.S. Army Combat Developments Experimentation Command, Fort Ord, California

WELCOMING REMARKS

BG Donald F. Packard, Commander, U.S. Army Combat Developments Experimentation Command

RADM Isham W. Linder, Superintendent, U.S. Naval Postgraduate Academy

CHAIRMAN OF SESSION I

Dr. Frank E. Grubbs, Program Committee Chairman, Aberdeen Proving Ground, Maryland

ANALYSIS OF UNBALANCED EXPERIMENTS

Professor H.O. Hartley, Director, Institute of Statistics, Texas A&M University, College Station, Texas

1015-1045 BREAK

***** Wednesday *****

1045-1200 CLINICAL SESSION A -- Ingersoll Hall, Room 271*

CHAIRMAN

Robert L. Launer, U.S. Army Research Office, Research Triangle Park, North Carolina

PANELISTS

Gerald Andersen, U.S. Army Materiel Development and Readiness Command, Alexandria, Virginia

Frank E. Grubbs, Aberdeen Proving Ground, Maryland

H.O. Hartley, Institute of Statistics, Texas A&M University, College Station, Texas

MOEs FOR DIVISION LEVEL MODELS

John H. Shuford and CPT Fredrick H. Knack, White Sands Missile Range

ANALYSIS OF RATIO DATA FROM FIELD EXPERIMENTATION

CPT Brian Barr, Fort Ord

1045-1200 CLINICAL SESSION B -- Room 322

CHAIRMAN

Edward W. Ross, Jr., U.S. Army Natick Research and Development Command, Natick, Massachusetts

PANELISTS

Norman Breslow, Department of Biostatistics, University of Washington, Seattle, Washington

Walter D. Foster, Armed Forces Institute of Pathology, Washington, D.C.

Douglas B. Tang, Department of Biostatistics and Applied Mathematics Division, Walter Reed Army Institute of Research, Washington, D.C.

* All sessions will be held at Ingersoll Hall.

***** Wednesday *****

PHYSIOLOGICAL AND PERCEPTUAL ADAPTION TO SUSTAINED AND MAXIMAL WORK IN YOUNG WOMEN

D. Kowal, D. Horstman, and _. Vaughn, Army Research Institute of Environmental Medicine, Natick

ANALYSIS OF WELL BEING AND OPERATIONAL EFFICIENCY IN A LABORATORY SIMULATION OF A FIELD ARTILLERY FIRE DIRECTION CENTER

L.E. Banderet, LCOL J.W. Stokes, Army Research Institute of Environmental Medicine, Natick

1045-1200 TECHNICAL SESSION I -- Room 288 -- CURVE FITTING

CHAIRMAN

Norman L. Wykoff, U.S. Army Jefferson Proving Ground, Madison, Indiana

THEORY OF LEAST CHI-SQUARE FOR POLYNOMIALS: IMPLICATION FOR DESIGN OF EXPERIMENTS

Richard L. Moore, U.S. Army Armament Research and Development Command

SIMPLIFIED CONSTRUCTION OF BASIS FUNCTIONS FOR POLYNOMIAL SPLINES

J.J. Heimbold, Mark Resources Incorporated

VALT PARAMETER IDENTIFICATION FLIGHT TEST

Robert L. Tomaine, Wayne H. Bryant, Ward F. Hodge, Langley Directorate

1200-1300 LUNCH

1300-1500 CLINICAL SESSION C -- Room 271

CHAIRMAN

Harold Larson, U.S. Naval Postgraduate School, Monterey, California

PANELISTS

Donald P. Gaver, Operations Analysis Department, Naval Postgraduate School, Monterey, California

***** Wednesday *****

James R. Moore, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Malcolm Taylor, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

EXPERIMENTAL DESIGNS FOR SENSITIVITY EXPERIMENTS OF COMPUTER SIMULATION MODELS

Carl B. Bates, U.S. Army Concepts Analysis Agency

ON VALIDATING MISSILE SIMULATIONS: FIELD DATA ANALYSIS VIA TIME-SERIES TECHNIQUES

Donald W. Sutherlin, Redstone Arsenal, and Naim A. Kheir, University of Alabama

STATISTICAL VALIDATION OF PROJECTILE/MISSILE SIMULATION MODELS

Harold L. Pastrick, Redstone Arsenal

1300-1500 CLINICAL SESSION D -- Room 322

CHAIRMAN

Douglas B. Tang, Department of Biostatistics and Applied Mathematics Division, Walter Reed Army Institute of Research, Washington, D.C.

PANELISTS

Robert E. Bechhofer, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York

Badrig Kurkjian, Mathematical Sciences, University of Alabama

William S. Mallios, BDM Services Company

ANALYSIS OF VARIANCE OF MULTIVARIABLE FLIGHT TEST DATA

James S. Hayden, U.S. Army Aviation Engineering Flight Activity

***** Wednesday *****

TOPICS IN THE ANALYSIS OF VARIANCE: SELECTION OF A MODEL AND APPROPRIATE SUMMARY STATISTICS

Frederick H. Steinheiser, Jr., and Kenneth I. Epstein, Army Research Institute for the Behavioral and Social Sciences

STATISTICAL DESIGN AND ANALYSIS OF UNDERGROUND STRUCTURES TESTS

Jon D. Collins and Eugene Sevin, Defense Nuclear Agency

1300-1500 TECHNICAL SESSION 2 -- Room 288 -- EXPERIMENTAL DESIGN

CHAIRMAN

Beatrice Orleans, Naval Sea Systems Command, Washington, D.C.

EXPERIMENTAL DESIGN FOR TESTING EFFECT OF INGESTING CRUDE-FIBER ON PLASMA ZINC LEVELS IN HUMAN VOLUNTEERS

Walter D. Foster, Armed Forces Institute of Pathology, and Barbara F. Harland, Food and Drug Administration

THE ANALYSIS OF PARTIALLY FACTORIAL EXPERIMENTS

John R. Burge, Walter Reed Army Institute of Research

CONSIDERATIONS IN DESIGNING MANPOWER EXPERIMENTS

Gus W. Haggstrom, Rand Corporation

1500-1530 BREAK

1530-1700 CLINICAL SESSION E -- Room 271

CHAIRMAN

James R. Moore, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

PANELISTS

Robert E. Bechhofer, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York

Norman Breslow, Department of Biostatistics, University of Washington, Seattle, Washington

***** Wednesday *****

H.O. Hartley, Institute of Statistics, Texas A&M University, College Station, Texas

FIELD EXPERIMENT DESIGN RISK UNDER PRACTICAL CONSTRAINTS

MAJ Lawrence T. Sughrue, Fort Ord

FIELD VERIFICATION OF RADIATION CHARACTERISTICS OF RADARS

J.L. Harris, U.S. Army Missile R&D Command

CONSTRUCTION OF CONFIDENCE LIMITS IN NONLINEAR REGRESSION

Charles Maxson Greenland, Systems Assessment Office Edgewood Arsenal, Maryland

1530-1700 CLINICAL SESSION F -- Room 322

CHAIRMAN

James Banks, Army Research Institute, Arlington, Virginia

PANELISTS

George E.P. Box, R.A. Fisher Professor of Statistics, University of Wisconsin, Madison, Wisconsin

Bernard Harris, Mathematics Research Center, Madison, Wisconsin

Malcolm Taylor, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

COMPUTING THE DEFINITE INTEGRAL ON A PROGRAMMABLE CALCULATOR

D.W. Rankin, White Sands Missile Range

A FRESHMAN ERROR CAN BE FATAL OR I'M NOT SURE ABOUT BEING 95 PERCENT SURE

Norman Wykoff, U.S. Army Jefferson Proving Ground

LASER VELOCIMETER DATA INTERPRETATION BY HISTOGRAM AND SPECTRAL ANALYSIS

Warren H. Young, Jr., James F. Meyers, and Danny R. Hoad, Langley Directorate, Hampton, Virginia

***** Wednesday *****

1530-1700 TECHNICAL SESSION 3 -- Room 288 -- STATISTICAL MODELLING I

CHAIRMAN

Carl B. Bates, U.S. Army Concepts Analysis Agency, Bethesda, Maryland

RANK ANALYSIS OF A CONSTRAINED GROUND-TO-AIR DETECTABILITY EXPERIMENT

Carl T. Russell, U.S. Army Operational Test and Evaluation Agency

METHODS OF RESOLVING UNDER-IDENTIFICATION IN STRUCTURAL DESIGN

William S. Mallios, BDM Services Company

1830- SOCIAL HOUR AND BANQUET -- Herrmann Hall, El Prado Room

***** Thursday, 20 October *****

0830-1000 CLINICAL SESSION G -- Room 271

CHAIRMAN

Walter D. Foster, Armed Forces Institute of Pathology, Washington, D.C.

PANELISTS

George E.P. Box, Mathematics Research Center, Madison, Wisconsin

Churchill Eisenhart, Senior Research Fellow, National Bureau of Standards, Washington, D.C.

Bernard Harris, Mathematics Research Center, Madison, Wisconsin

THREE DIMENSIONAL CURVE FITTING TECHNIQUES TO EXPRESS SUPPRESSION AS A FUNCTION OF RANGE AND ASPECT ANGLE

MAJ Chaunchy F. McKearn and SP5 David Brown, Fort Ord

PREDICTION BOUNDS IN LINEAR CALIBRATION: HETEROSCEDASTIC CASE

C.C. Peck and L.A. Hopkins, Letterman Army Institute of Research

0830-1000 CLINICAL SESSION H -- Room 322

CHAIRMAN

Langhorne P. Withers, U.S. Army Operational Test and Evaluation Agency, Falls Church, Virginia

PANELISTS

Gerald Andersen, U.S. Army Materiel Development and Readiness Command, Alexandria, Virginia

Donald P. Gaver, Operations Analysis Department, Naval Postgraduate School, Monterey, California

William S. Mallios, BDM Services Company

ESTIMATING PRODUCT RELIABILITY IN A DYNAMIC MARKET SITUATION WHEN ONLY FAILURES ARE REPORTED

Leonard R. Lamberson, Wayne State University

CRITERION-REFERENCED JOB PROFICIENCY TESTING: A LARGE SCALE APPLICATION

Milton H. Maier and Stephen F. Hirshfeld, U.S. Army Research Institute for the Behavioral and Social Sciences

0830-1000 TECHNICAL SESSION 4 -- Room 288 -- MAN-MACHINE INTERFACE

CHAIRMAN

J. Bart Wilburn, Jr., I&M Branch, U.S. Army Electronics Proving Ground, Ft. Huachuca, Arizona

ANALYSIS OF MAN-MACHINE INTERFACE INFORMATION IN CURRENT COMMUNICATIONS SYSTEMS

R.J. D'Accardi and H.S. Bennett, U.S. Army Electronics Command, C.P. Tsokos, University of South Florida

***** Thursday *****

HIGHLIGHTS OF CASE STUDIES IN MILITARY OPERATIONS RESEARCH

William S. Mallios, R.D. Batesole, D.R. Leal, and T.Q. Tran, BDM Services Company

1000-1030 BREAK

1030-1200 TECHNICAL SESSION 5 -- Room 288 -- STATISTICAL MODELLING II

CHAIRMAN

Diane Brown, Combat Developments Experimentation Command, Fort Ord, California

ERRORS IN LINEAR FITS DUE TO FUNCTION MISMATCH AND NOISE WITH SPLINE APPLICATION

G.W. Lank, W.B. Kendall, and P.A. Gartenberg, Mark Resources Incorporated

THE FACT'S OF LIFE

S. Goodman, A. McGoldrick, K. Heulitt, U.S. Army Armament Research and Development Command

1030-1200 TECHNICAL SESSION 6 -- Room 277 -- RELIABILITY I

CHAIRMAN

John Robert Burge, Walter Reed Army Institute of Research, Washington, D.C.

CONFIDENCE INTERVALS FOR RELIABILITY GROWTH ANALYSIS

Larry H. Crow, U.S. Army Materiel Systems Analysis Activity

A SENSITIVITY EVALUATION OF LARGE SCALE TACTICAL SYSTEM AVAILABILITY UNDER VARYING SUPPORT RESOURCE LEVELS

R.A. Hall and H.M. Bratt, Ames Research Center

1030-1200 TECHNICAL SESSION 7 -- Room 322 -- RELIALBILITY II

CHAIRMAN

Donald Leal, BDM Services Company

***** Thursday *****

USE OF LOGNORMAL CONFIDENCE OF THE RELIABILITY FUNCTION IN THE TRUE LIFE DISTRIBUTION IS NOT LOGNORMAL

Eugene E. Coppola, Watervliet Arsenal

COMPARISON OF ESTIMATORS OF THE RELIABILITY FUNCTION IN THE EXPONENTIAL DISTRIBUTION

Jerome P. Keating, Bell Helicopter Textron

DOUBLE TESTING IN BINOMIAL DATA

G.R. Andersen, U.S. Army Materiel Development and Readiness Command

1200-1300 LUNCH

1300-1515 GENERAL SESSION II -- Ingersoll Hall, Room 122

CHAIRMAN OF SESSION II

Dr. Marion R. Bryson, Combat Developments Experimentation Command, Fort Ord, California

CENSORED DATA

Dr. Norman Breslow, Department of Biostatistics, University of Washington, Seattle, Washington

THE JACKKNIFE: SURVEY AND APPLICATIONS

Dr. Rupert Miller, Department of Statistics, Stanford University, Stanford, California

1515-1545 BREAK

1545-1700 GENERAL SESSION II (CONTINUED) -- Ingersoll Hall, Room 122

MODELLING AND ESTIMATION OF COMPLEX SYSTEM AVAILABILITY

Dr. Donald P. Gaver, Operations Analysis Department, Naval Postgraduate School, Monterey, California

***** Friday, 21 October *****

0830-1000 TECHNICAL SESSION 8 -- Room 271 -- WEAPON SYSTEMS EVALUATION

CHAIRMAN

Agatha S. Wolman, Bethesda, Maryland

QUANTITATIVE EVALUATION OF THE M-60A1 TANK CAMOUFLAGE BY OPERATIONAL IMAGE INTERPRETERS

Edward R. Eichelman and Ronald L. Johnson, U.S. Army Mobility Equipment Research and Development Command

DESIGN ON A FULL SCALE TEST FOR U.S. ARMY HELICOPTER NAP-OF-THE-EARTH (NOE) COMMUNICATION SYSTEM

Bernard V. Ricciardi, U.S. Army Avionics Laboratory, George Hagn and Bruce Tupper, Stanford Research Institute

0830-1000 TECHNICAL SESSION 9 -- Room 322 -- METHODOLOGICAL IMPROVEMENTS

CHAIRMAN

Malcolm Taylor, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

TABLE LOOK-UP AND INTERPOLATION FOR A NORMAL RANDOM NUMBER GENERATOR

William L. Shepherd and John W. Starner, U.S. Army White Sands Missile Range

DIRECT DEGENERACY ATTAINMENT IN MARKON CHAINS

Richard M. Brugger, U.S. Army Armament Materiel Readiness Command

THE CURSE OF EXPONENTIAL DISTRIBUTION IN RELIABILITY

L. Herbach, J.A. Greenwood, S. Blumenthal, Polytechnic Institute of Brooklyn, Brooklyn, New York

1000-1030 BREAK

1030-1200 GENERAL SESSION III -- Ingersoll Hall, Room 122

***** Friday *****

CHAIRMAN OF SESSION III

Dr. Grank E. Grubbs, Program Committee Chairman, Aberdeen Proving Ground, Maryland

OPEN MEETING OF THE AMSC SUBCOMMITTEE ON PROBABILITY AND STATISTICS

Dr. Douglas B. Tang, Department of Biostatistics and Applied Mathematics Division, Biometrics and Medical Information Processing, Walter Reed Army Institute of Research, Washington, D.C.

TIME SERIES MODELLING

Dr. George E.P. Box, R.A. Fisher Professor of Statistics, Department of Statistics, University of Wisconsin, Madison Wisconsin

1200-1300 LUNCH

ANALYSIS OF UNBALANCED EXPERIMENTS

H. O. Hartley
Texas A&M University
College Station, Texas 77843

1. Introduction

The title of this talk is rather general; and I should explain, therefore, that it is really confined to a limited number of aspects of the area covered by the title. I am restricting myself to so-called "multiple factor" experiments, that is experiments in which "responses" are measured under experimental "conditions" described by specifying the "levels" for each of several "factors." The well-known "factorial experiments" represent a special case of a balanced and multiple factor experiment in which precisely one (or precisely an equal number of) experimental unit(s) is (are) used at all possible combinations of factor levels. An unbalanced experiment will have unequal numbers of units (including zero units) exposed at the possible factor-level combinations.

There are two main causes of unbalance:

(i) Experiments originally designed as balanced experiments have become unbalanced through "accidents." The best known examples are the so-called "missing value" or "missing plot" situations in which the response for a number of units entered into the experiment has been lost or has been rejected as an "outlier" generated by extraneous error-sources. Other "accidents" lead to the "censorship" or "grouping" of some or all of the responses. This means that these responses are not known "exactly" but are known to lie within certain ranges of the response and measurement scale. For other situations of unbalance described as "incomplete data" see e.g., Hartley and Hocking (1971).

(ii) The unbalanced data have not arisen from a designed experiment but are the results of an operational study involving multiple classifications of sampled units by numerous factors invariably leading to unequal representations of the "cells" (factor-level combinations) and usually involving many zero-cells.

Finally, our concept of "analysis" is here confined to the problem of estimating the parameters in a linearly additive model postulated for the data. More specifically, we shall be concerned with the so-called mixed analysis of variance model. Briefly in this model, the observed response is the sum total of a mean response plus additive effects contributed by "effect constants" of the applied levels of the "fixed factors" plus the random "effect variables" of the applied levels of the "random factors." This model is illustrated by the examples of Section 2 and mathematically defined in the Appendix.

In limiting our present objectives to the estimation of parameters, we omit the important aspect of the drawing of inferences from the data. However, we do not omit to stress that in the case of (ii) when unbalanced operational data are analyzed the drawing of inferences of a causative nature is particularly hazardous and requires the examination of potential "latent variables" (see e.g., Box (1966), Hartley (1967)) causing spurious input-response relationships.

2. Illustrative Examples of Unbalanced Data

Before turning to the mathematical details of the estimation theory, it may be helpful to illustrate the concepts of Section 1 by examples. These examples illustrate the various sources of unbalance. At the same time they recapitulate the well-known concepts of "fixed factors" and "random factors" in analysis of variance.

Example 2.1. (O. L. Davies (1956) pp. 296-297).

We quote from Davies.

> The following is an example of an experimental design of general utility in many fields. It relates to the testing of nine aluminum alloys for their resistance to corrosion in a chemical plant atmosphere. Four sites in the factory were chosen, and on each of them a plate made from each alloy was exposed for a year. The plates were then submitted to four observers, who assessed their condition visually and awarded marks to each from 0 to 10 according to the degree of resistance to attack. The observers worked independently and the plates were submitted to them in random order; in other words the observers did not assess all plates from one site at the same time. ... The aim of the experiment was to decide which, if any, of the alloys were suitable for use in the factory, and especially to select any found to be suitable on all the sites. It was also required to know whether the four observers agreed in their relative assessments.

Basically the experiment is a balanced $9 \times 4 \times 4$ factorial in which plates from 9 aluminum alloys are exposed at each of four different plant sites and these are inspected by each of four observers. The mixed ANOVA model (not spelled out by Davies) that appears to underlie his analysis is as follows:

$$y_{ios} = \mu + \alpha_i + b_o + c_s + u_{is} + e_{ios} \qquad [1]$$

where

y_{ios} = score of ith alloy on sth site tested by oth observer.

μ = mean score

α_i = differential effect constant of ith alloy (fixed factor)

b_o = effect variable of oth observer (random factor)

c_s = effect variable of sth site (random factor)

u_{is} = interaction variable of ith alloy by sth site (random factor)

e_{ios} = error.

Note that the sites are considered random variables since inferences are desired for the plant as a whole and not just for the experimental sites. It seems reasonable that a random interaction variable between sites and alloys is provided (which is rightly used as the valid error for comparing alloys) but that interactions between observers and alloys or observers and sites are regarded negligible. The above experiment is, of course, balanced and the standard analysis consequential to the above model is given by Davies. In realistic situations unbalance may easily arise through "accidents" such as certain scores getting lost or becoming invalid. We should, however, point out that the so-called "missing value analysis" is strictly speaking correct only if all factors are fixed. However, the data may be analyzed by the method given in Appendix 1.

Example 2.2. (O. H. Pfeiffer (1964)).

This is an experiment to evaluate the performance of swivel hook-type cross chain fasteners of tire chains. Again the experimental design was balanced as described by Pfeiffer. Briefly, the test comprised 8 "wheel-blocks" in the form of the 8 tires of the 4 rear dual wheels and these "blocks" were regarded as a factorial arrangement of three 2-level factors, viz. "front duals" versus "rear duals," "right duals" versus "left duals", and outside wheels versus inside wheels. Within each "block" the 3 "treatments" consisted of 3 "clusters" of three different types of hook fasteners, each cluster comprising 4 individual fasteners. The main response measured for each fastener was the log of its miles to failure.

Turning then to the factors, the type of fastener is clearly a fixed treatment factor and the individual fasteners a random repetition factor from the population of fasteners of each type but tested within a "cluster" on the wheel. The tires are also a random factor since inferences must not be restricted to the particular set of 8 tires used in the test but they have positional "main treatments" superimposed in the form of the above 2^3 factorial. Pfeiffer uses (we think conservatively) the tire × type interaction as an error which, of course, also includes any position × type interaction. This decision is proved correct since the tire × type mean square is virtually identical with the within type mean square.

In this experiment unbalance arose through accidental censorship: Certain fasteners had not failed when the experiment was terminated at 425 miles. Since the missing values are all known to exceed log 425, the customary missing value analysis (which assumes that the missing values

are a random selection from the experimental responses) is not appropriate. Likewise the analysis of the observed miles to failure as an unbalanced experiment is not appropriate as it would disregard the censored information. An appropriate analysis would be an iterative EM algorithm consisting of the following steps.

STEP (E): For each missing value compute its conditional expectation, E, given that it exceeds the value $\xi = \log 425$. This is given by

$$E = \hat{\mu} + \hat{\sigma}\left\{\frac{Z(\frac{\xi-\hat{\mu}}{\hat{\sigma}})}{Q(\frac{\xi-\hat{\mu}}{\hat{\sigma}})}\right\} \qquad [2]$$

where

$\hat{\mu}$ = iterative estimate of the cell mean for the missing value computed from the current estimates of the linear ANOVA model,

$\hat{\sigma}^2$ = iterative estimate of the within cell variance,

and $Z(\ )$, $Q(\ )$ are respectively the standard normal ordinates and tail area. The assumption of an approximate normal within cell distribution of log miles to failure requires checking.

STEP (M): Using all values of E computed by [2] along with the observed log miles to failure records, compute the customary balanced ANOVA estimates of all terms in the additive ANOVA model and return to STEP (E).

The symbol (M) of the second step stands for Maximum Likelihood estimation and the term EM algorithm was introduced by Dempster, Rubin and Laird (1977). Earlier accounts of the algorithm are given by Hartley (1958) and Hartley and Hocking (1971).

Example 2.3. (R. Bell (1963) p. 623).

> "This paper presents a typical analysis of service practice firing results and indicates the significance of these results in the Surveillance Program. An example of the evaluation of the annual service practice firings for the Honest John Rocket is presented.
>
> "934 Firings of Rocket 762MM: M31 Series, conducted for troop training and other purposes by both United States and NATO firing units have been considered. The purpose of this study was to investigate the overall accuracy performance of the M31 rocket system when fired by troop units and to establish if there is any indication of a deterioration of this accuracy performance with increasing age of the M6 series rocket motors of these M31 series rockets."

More specifically the operational data bank used in the study consisted of all firings during 6 years (1958-63) by 3 launchers (289, 386, 33) using rocket motors of varying ages (1^-, 1 to 2^-, ..., 7 to 8^-). The 6 × 3 × 8 factorial table was by necessity unbalanced with many "zero cells." Among other sources of unbalance there was a tendency of the older rocket motors to be more heavily represented in the later years. The data were acquired operationally over the years and the analysis here carried out was not originally planned.

Of the three factors both the 3 launchers and the 8 ages are fixed but there may be some question as to how the 6 years should be treated. Inferences are obviously required for the period subsequent to that covered by the data bank and there may be doubt as to whether conditions in 1958 to 1963 should be regarded as a random sample of those prevailing in future years. However, if such a proposition is accepted, the analysis of Appendix 1 could be applied to obtain estimates of the age and launcher contrasts and their interaction as well as estimates of components of variance attributable to year to year variation, the year × age, and the year × launcher interactions.

If there is some doubt about the representativeness of years 1958 to 1963 of future conditions, no useful inferences can be made unless a time series model can be formulated.

3. Relation Between Various Methods in Balanced and Unbalanced Data Analysis

As is well known the analysis of variance of balanced factorial data makes a distinction between the so-called "fixed factors" and "random factors." These concepts were introduced in Section 1 and illustrated in Section 2 by three examples. The same distinction must be made when analyzing unbalanced data. In the two-way table below we distinguish two main types of ANOVA's, namely (i) an analysis in which all factors (except the error) are fixed which is contrasted with (ii) the so-called mixed ANOVA, a situation where some factors are fixed but others are random. The so-called all random model is included in this case as one in which the only fixed parameter is the mean response. Of course (i) is also a special case of (ii), namely the case in which the only random factor is the error.

The row headings in the table are (a) balanced data and (c) unbalanced data, but an intermediate situation (b) is provided in which the data are "almost balanced" (notably missing value situations). In the body of the table we give very brief descriptions of the appropriate analysis but would amplify these as follows:

TABLE 1

Relation Between Various Methods in Balanced and Unbalanced Data Analysis

	(i) All Factors Fixed	(ii) Some Factors Fixed Some Random
(a) Balanced Data	ANOVA or regression analysis on dummy variables	ANOVA and estimation of components of variance
(b) Almost Balanced Data	Missing value formulas ANOVA = ML EST's, tests approximate	Missing value formulas, heuristic ANOVA approximate
(c) Unbalanced Data	Regression analysis on dummy variables, Exact Max. Likelihood estimation and hypothesis tests	Mixed model ANOVA components of variance estimation, Estimation of constants, Max. Likelihood Minque Present Method

(i)(a) If the random (equal variance) error is the only random factor, the data are of the form of a linear model $y = X\beta + e$ with the design matrix X consisting of 0, 1 "dummy variables." After reparameterization of β (to make X non-singular) the regression analysis is identical with the balanced data ANOVA provided we adopt the accepted hierarchy of factors main effects followed by two factor interactions, etc.

(i)(c) The same applies to the case of all factors fixed unbalanced data banks except that the reparameterization is more dependent on the adopted hierarchy in which the factors are ordered.

(i)(b) This case is separated from (i)(c) in that it is often a computational advantage to reduce the case of almost balanced data to that of balanced data by a missing value EM type algorithm.

(ii)(a) The simultaneous estimation of effect-constants and components of variance in a balanced ANOVA is well documented in the statistical literature. The (unbiased) estimation procedure may, however, lead to negative estimates of variance components for which various remedies are advocated.

(ii)(b) It should be stressed that the customary missing value estimates are M.L. estimates only for the all fixed factor models. Therefore an accurate treatment must reduce this case to (ii)(c).

(ii)(c) This is the most general situation and a computationally convenient method is described in Appendix 1 which follows. Note that all six situations (i)(a), (b), (c); (ii)(a), (b), (c) could be regarded as special cases of (ii)(c).

Before turning to a more detailed discussion of (ii)(c) in Appendix 1, we should stress that it does <u>not</u> cover unbalance through censorship and an E-algorithm should be adjoined to the M.L. estimation treatment briefly referred to in Appendix 1.

REFERENCES

Bell, R. (1963). "Statistical study of aging characteristics," Proceedings of 8th Conference on the Design of Experiments in Army Research Development and Testing, 623-647.

Box, G. E. P. (1966). "Use and abuse of regression," Technometrics, 8, 625-654.

Davies, O. L. (1956). The Design and Analysis of Industrial Experiments, Oliver & Boyd, London & Edingurgh, 1956.

Dempster, A. P., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc. Series B, 39, 1-36.

Hartley, H. O. (1958). "Maximum likelihood estimation from incomplete data," Biometrics, 14, 174-194.

Hartley, H. O. (1967). "Industrial applications of regression analysis," Transactions of the Chicago Meeting of the American Society for Quality Control, 143-148.

Hartley, H. O. and Hocking, R. R. (1971). "The analysis of incomplete data," Biometrics, 27, 783-823.

APPENDIX 1[†]

A SIMPLE 'SYNTHESIS'-BASED METHOD OF VARIANCE COMPONENT ESTIMATION

by

H. O. Hartley[*], J. N. K. Rao[+] and Lynn LaMotte[#]

1. Introduction

Two of us (HOH and JNKR) have recently had occasion (see Hartley and Rao (1977)) to consider components of variance estimation techniques in data banks arising from sample surveys. Such data banks differ from those encountered in experimental designs in that the "number of observations", n (in our case the number of elementary sampling units) is exceedingly large. We have therefore been prompted to search for computationally efficient methods for the estimation of components of variance when n is large and the algorithm here described involves a computational effort (as measured by the number of products) which is a linear function of n and this is generally regarded as computationally highly efficient. While our algorithm is new the statistical method of estimation we employ is not. In fact it represents a special case of C. R. Rao's (1971) MINQUE (with V = I). It is also identical (Communication by S. R. Searle) with a special

---

[*]H. O. Hartley, Institute of Statistics, Texas A&M University

[+]J. N. K. Rao, Carleton University, Ottawa

[#]Lynn LaMotte, Quantitative Management Science, University of Houston

[†]A shortened version of Appendices 1 and 2 will be published in Biometrics.

case of the first iterate solution of the REML equations of Corbeil and Searle (1976) whose algorithms appear to involve much larger computational efforts (proportional to n^2). The computational effort is also considerably less than that involved in the M.L. estimation by Hartley and Rao (1967) which is still fairly laborious inspite of the improvements through the W-transformation by Hemmerle and Hartley (1973).

Inspite of its computational simplicity the estimation procedure has numerous "optimality properties". Apart from being a special case of MINQUE other properties are established in Section 6 and the asymptotic consistency is proved in the Appendix under fairly general conditions. The consistency of our estimator makes it convenient as a starting point for a single M.L. cycle to obtain asymptotically fully efficient estimates.

Finally we establish simple conditions for the estimability of all variance components by our method (see Section 6). In this context we observe that with other methods (such as the Henderson 3 method (Henderson (1953)) or the Abbreviated Doolittle and square root method (see e.g. Gaylor, Lucas and Anderson (1970)) estimability depends on the subjective ordering of the components (such as with the Forward Doolittle procedure) and if the ordering is unfortunate the method may fail to yield estimates for certain components while with a different ordering (not attempted) all components may well be estimable.

2. The Mixed ANOVA Model

Employing the currently used notation we write the mixed ANOVA model in the form

$$y = X\alpha + \sum_{i=1}^{c+1} U_i b_i \tag{1}$$

where

y is an n x 1 vector of observations,

X is an n x k matrix of known coefficients,

α is a k x 1 vector of unknown constants,

U_i is an n x m_i matrix of 0, 1 coefficients,

b_i is an m_i x 1 vector of normal variables from $N(0, \sigma^2_i)$.

Specifically $U_{c+1} = I_n$ and b_{c+1} is an n-vector of "error variables". Moreover the design matrices U_i have precisely one value of 1 in each of their rows and all other coefficients 0. We denote by $m = \sum_{i=1}^{c} m_i$ the total number of random levels.

We may assume without loss of generality that

$$X'X = I \tag{2}$$

for if (2) is not satisfied we may orthogonalize X by a Gram Schmidt orthogonalization process with a consequential reparameterization of α omitting any linearly dependent columns in the Gram Schmidt process. Usually the first column of X is the column vector with all elements = $1/\sqrt{n}$. It is the objective of the method to compute estimates of the variance components σ^2_i and the vector α.

3. The Present Method

The essence of the present method is to

(α) Select c+1 quadratic forms $Q_j(y)$ in the elements of y.

(b) Use the method of synthesis (Hartley (1967), Rao (1968)) to obtain the coefficients k_{ji} in the formulas for $E(Q_j)$ in the form

$$E(Q_j) = \sum_{i=1}^{c+1} k_{ji}\sigma^2_i. \tag{3}$$

(c) Estimate σ^2_i by equating the computed Q_j to their expectations i.e. by inverting the system (3) to compute the vector $\hat{\sigma}^2$ with elements $\hat{\sigma}^2_i$

$$\underset{\sim}{\hat{\sigma}^2} = K^{-}\underset{\sim}{Q}(y) \tag{4}$$

from the vector $\underset{\sim}{Q}(y)$ with elements $Q_j(y)$ where $K = (k_{ji})$ with rank to be discussed in Section 6 and 7.

(d) Replacing any negative elements of $\hat{\sigma}^2$ by 0, with consequences to be discussed in Section 7.

We now give more details for (a), (b) and (c)

(a) The $Q_j(y)$ will be based on contrasts which do not depend on any elements of α. Accordingly we orthogonalize all U_i matrices on X and construct matrices V_i orthogonal on X as follows: Denote by u(t, i) the tth column vector of U_i and by x(r) the rth column vector of X then the columns v(t, i) of V_i are given by

$$v(t,i) = u(t,i) - \sum_{r=1}^{k} x(r)\{x'(r)u(t,i)\}$$

or

$$V_i = U_i - XX'U_i. \tag{5}$$

We now choose the c+1 quadratic forms $Q_j(y)$ as

$$Q_j(y) = y'V_jV_j'y = (V_j'y)'V_j'y \qquad j = 1, \ldots, c+1 \tag{6}$$

(b) It follows from the method of synthesis (see Hartley (1967),

J. N. K. Rao (1968) that

$$E\, Q_j(y) = \sum_{i=1}^{c+1} k_{ji}\sigma^2_i$$

with (7)

$$k_{ji} = \sum_t (V_j' u(t,i))'(V_j' u(t,i))$$

Now since $v(\tau,j)$ is orthogonal on any $x(\rho)$ (i.e. since $v'(\tau,j)x(\rho) = 0$) we can write the k_{ji} in the alternative form

$$k_{ji} = \sum_t (V_j' v(t,i))'(V_j' v(t,i))$$

$$= \sum_t \sum_\tau \{v'(\tau,j)\, v(t,i)\}^2 \qquad (8)$$

showing that $k_{ij} = k_{ji}$.

An alternative form of k_{ji} is

$$k_{ji} = \mathrm{tr}\{(V_j V_j')(V_j V_j')\}. \qquad (9)$$

We shall show in Section 6 that the symmetrical matrix $K = (k_{ji})$ will have full rank c+1 if the n x n matrices $V_i V_i'$ are not linearly dependent.

(c) We shall also show in Section 6 that the system of equations

$$\underset{\sim}{Q} = K\hat{\underset{\sim}{\sigma}}^2 \qquad (10)$$

is consistent even if the rank of K is degenerate. Solving (10) in the form

$$\hat{\underset{\sim}{\sigma}}^2 = K^- \underset{\sim}{Q} \qquad (11)$$

we shall, of course, be particularly interested in the full rank case when $K^- = K^{-1}$.

4. The Computational Load

It may be helpful to give an idea of the computational efficiency of the present method by tabulating the number of products involved in the main operations of the algorithm. To this end we first note simplified versions for the $k_{c+1,i}$: Observing that $U_{c+1} = I$ we have from (5) that $V_{c+1} = I - XX'$ and since $X'X = I$ we find that $V_{c+1}V'_{c+1} = I - XX'$ and finally from (9) that

$$k_{c+1,c+1} = \text{tr } (I - XX')(I - XX') = \text{tr } (I - XX') = n - k. \tag{12}$$

Similarly we find that

$$k_{c+1,i} = \text{tr } \{(I - XX')(V_iV'_i)\} = \text{tr } \{V_iV'_i - XX'V_iV'_i\} = \text{tr } V_iV'_i. \tag{13}$$

Further we note the form of $V'_{c+1}y$ i.e.

$$V'_{c+1}y = y - XX'y. \tag{14}$$

Defining now the adjoined matrices

$$U = (U_1 \mid \ldots \mid U_c) \qquad V = (V_1 \mid \ldots \mid V_c) \tag{15}$$

the bulk of the work consists of the formation of the elements of the symmetrical matrix $V'V = V'U = U'V$. The elements of this matrix are assembled in submatrices in accordance with the partition (15) as shown in the Schedule 1 below where it must be remembered that the range of the column index t depends on i and is $t = 1, \ldots, m_i$ and the range of $r = 1, \ldots, m_j$ so that the submatrix V'_jU_i has dimensions $m_j \times m_i$. The k_{ji} for $i \geq j = 1, \ldots, c$ are then obtained by forming the sums of squares of the elements in each submatrix in accordance with (7).

Finally, we recite the formulas for the remaining coefficients in the equations (10). The $k_{c+1,c+1}$ and $k_{c+1,i}$ are computed from (12) and

Schedule 1: Submatrices of V'U

	U_1	U_2	. . .	U_c
V_1	$v(\tau,1)'u(t,1)$	$v(\tau,1)'u(t,2)$	. . .	$v(\tau,1)'u(t,c)$
V_2		$v(\tau,2)'u(t,2)$	. . .	$v(\tau,2)'u(t,c)$
.				
.				. . . . .
.				
V_c				$v(\tau,c)'u(t,c)$

(13) respectively and the right hand sides of $Q_j(y)$ from the second form in (6) for $j = 1, \ldots, c$ while $Q_{c+1}(y)$ is given in accordance with (14) by

$$Q_{c+1}(y) = y'y - (X'y)'(X'y) \qquad (16)$$

We can now summarize the approximate number of products involved in the various operations of the algorithms.

We list the algorithms and show the associated numbers of products in (1).

1. Orthogonalization of X $(k^+(k^+ - 1)n$, where k^+ denotes the number of columns in the original matrix X)
2. Computation of $X'U_i$ for $i = 1, \ldots, c$, (0, subtotals of X)
3. Computation of $X(X'U_i)$ for $i = 1, \ldots, c$ from equation (5), (nmk)
4. Computation of $U'V = V'V$ in accordance with Schedule 1, (0 products since the elements are subtotals of the elements $v(t,i)$)
5. Computation of k_{ij} for $i,j = 1, \ldots, c$ from equation (7), $(\frac{1}{2}m(m+1))$
6. Computation of $k_{c+1,i}$ for $i = 1, \ldots, c$ from equation (13), (mn)
7. Computation of $k_{c+1,c+1}$ from equation (12), (0 products)
8. Computation of the $Q_j(y)$ for $j = 1, \ldots, c+1$ from 2nd form of equation (6) and equation (16), $((m+k+1)(n+1))$

The important point is that the number of products is only a linear function of the number of data lines n. An approximate formula for the total number of products is $n\{k^+(k^+ - 1) + (m+1)(k+1)\}$

5. A Numerical Example

A small numerical example with $n = 4$, $k^+ = 3$, $k = 2$, $c = 1$, $m_1 = 2$, $m = 2$, $m_2 = n = 4$ is shown in schedule 2 below.

Schedule 2: A Numerical Example of a Mixed Model

y	X Original		U_1		U_2				X new		V_1	
4	1	1	0	1	0	0	0	0	(1/2)	(1/2)	+(1/2)	-(1/2)
2	1	1	0	0	1	0	0	0	(1/2)	(1/2)	-(1/2)	+(1/2)
1	1	0	1	0	1	0	0	0	(1/2)	-(1/2)	0	0
2	1	0	1	0	1	0	0	0	(1/2)	-(1/2)	0	0

The orthogonolization of X (original) to X (new) follows the standard Gram Schmidt procedure and reduces the $k^+ = 3$ dependent columns to $k = 2$ columns which are orthogonal and standardized. Note that

$$x(2)_{new} = x(2)_{old} - (1/2)x(1)_{old} \text{ and}$$

$$x(3)_{old} = x(1)_{new} - x(2)_{new} \text{ must be eliminated.}$$

Using now $x(r) = x(r)_{new}$ we orthogonalize U_1 on X and compute (see (5))

$$x'(1)\ u(1,1) = +(1/2),\ x'(2)\ u(1,1) = +(1/2)$$

and hence

$$v(1,1) = u(1,1) - (1/2)x(1) - (1/2)x(2)$$

likewise

$$x'(1)\ u(2,1) = (3/2)\ x'(2)\ u(2,1) = -(1/2)$$

and hence

$$v(1/2) = u(2,1) - (3/2)x(1) + (1/2)x(2).$$

This yields the matrix V_1 in schedule 2 which has only one independent column. The elements of $V_1'U_1$ require the computation of

$$v(1,1)'\, u(1,1) = (1/2);\ v(1,1)'\, u(2,1) = v(2,1)'\, u(1,1) = -(1/2)$$

and

$$v(2,1)'\, u(2,1) = 1/2 \text{ with sum of squares of } k_{11} = 4(1/2)^2 = 1.$$

Further (equation (12)) $k_{22} = 4 - 2 = 2$ and (equation (13)) $k_{12} = k_{21} = 4(1/2)^2 + 4(0)^2 = 1$ so that the K matrix is given by $K = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$.

Finally, (equation (16))

$$Q_2(y) = 4^2 + 2^2 + 1^2 + 2^2 - (\tfrac{1}{2}\ 9)^2 - (\tfrac{1}{2}\ 3)^2 = 25 - \frac{90}{4} = 25 - 22.5 = 2.5$$

and (equation (6)) $Q_1(y) = (\tfrac{1}{2}\ 2)^2 + (\tfrac{1}{2}\ (-2))^2 = 2.$

The solution of $\underset{\sim}{Q} = K\hat{\underset{\sim}{\sigma}}^2$ therefore yields $\hat{\sigma}_2^2 = 1/2$, $\hat{\sigma}_1^2 = 1.5$.

6. Optimality Properties and the Consistency of the Equations

The estimators described in Section 3 may be seen to be "best at $\sigma^2_i = 0$, $i = 1, \ldots, c$, $\sigma^2_{c+1} = 1$" as defined by L. R. LaMotte (1973). Therefore, the consistency of equation (10), regardless of the rank of K, is established as Lemma 4 by LaMotte (1973). That the estimators defined by (11) are "best" among invariant quadratic unbiased estimators guarantees that they are admissible in that class; that is, no other invariant quadratic unbiased estimators have uniformly less variance for all $\underset{\sim}{g}$. Further, as noted by LaMotte (1973), the estimators (11) have the property that in any model for which a uniformly best estimator exists, (11) will be uniformly best. Finally, it may be seen that the "synthesis" estimators (11) are also MINQUE as in Rao (1971, Section 6) with V = I. No claim is made that this choice of

the norm has any particular merits among the rather general family of the norms covered by Minque formulas. However, it appears to be reasonable to us that in the absence of any theoretical criteria for selection of Minque norms a norm leading to simple estimators may be regarded as meritorious.

Following Section A5 in LaMotte (1973), it may be seen that the rank of K is equal to the number of linearly independent matrices among $V_i V_i'$, $i = 1, \ldots, c+1$. Thus a singular K may occur if the $U_i U_i'$ matrices are not all linearly independent or if there exists (see (5)) a linear combination of the $U_i U_i'$ matrices whose columns are contained in the linear subspace spanned by the columns of X. In the first case the singularity is caused by the design leading to the U_i matrices, while in the second the singularity is caused by confounding fixed and random effects. In either case, (10) is consistent but some linear combinations of the variance components can not then be unbiasedly estimated. We should stress however that other special cases of Minque (not necessarily invariant to α) may also deserve particular attention.

APPENDIX 2

The Asymptotic Consistency of $\hat{\sigma}^2$

In discussing the asymptotic behavior of $\hat{\sigma}^2$ it is of course necessary to specify the limiting process under which such properties are supposed to hold. Clearly it is necessary for the consistent estimation of the variances $\sigma_i^2 = \text{Var } b_i$ that the number of elements m_i in the vectors b_i all tend to ∞. For the identity matrix U_{c+1} we have $m_{c+1} = n$ the overal sample size. For the remaining m_i we assume that their limiting behavior is related to n by

$$Ln^{1-\alpha_i} \leq m_i \leq Un^{1-\alpha_i} \tag{17}$$

where $0 \leq \alpha_i < 1$ and L, U are universal constants. More specifically we assume that $\alpha_{c+1} = 0$ but $\alpha_i > 0$ for $i = 1, \ldots, c$. Generalizations to situations in which $\alpha_i = 0$ for several components are under consideration.
Denote now by

$$\nu(t, i) = \text{number of elements in } u(t, i) \text{ which are } 1 \tag{18}$$

and

$$\nu(t, i; \tau, j) = \text{number of rows in which both } u(t, i) \text{ and } u(\tau, j) \text{ have elements } 1. \tag{19}$$

Using these concepts we introduce the following conditions of 'pseudo orthogonality' of the u(t, i) vectors. We assume that

$$\ell n^{\alpha_i} \leq \nu(t, i) \leq u\, n^{\alpha_i} \tag{20}$$

(where ℓ, u are universal constants) and that

$$\nu(t, i; \tau, j) = o(\nu(t, j)) \tag{21}$$

$i \neq j$ with $i = 1, \ldots, c+1$

and $j = 1, \ldots, c$

The relationship between (17) and (20) is obvious since $\sum_{t=1}^{m_i} \nu(t, i) = n$ so that (20) implies (17) with $U = \frac{1}{\ell}$ and $L = \frac{1}{u}$ and the stronger condition (20) implies a uniform order of magnitude for all $\nu(t, i)$ in a given U_i. Since the columns of the U_i matrices are orthogonal we have $\nu(t, i; \tau, i) = 0$ for all pairs $t \neq \tau$. For columns $u(t, i)$, $u(\tau, j)$ with $i \neq j$ condition (21) is satisfied if there is an asymptotically uniform distribution of the $\nu(t, i)$ rows for which $u(t, i)$ has elements 1 over a fraction qm_j of the m_j columns of U_j where $0 < q < 1$ since the fraction of $\nu(t, i)$ which gives rise to $\nu(t, i; \tau, j)$ will be $O(q^{-1}m_j^{-1}) = O(n^{\alpha_j - 1})$ and will tend to zero.

Next we must introduce conditions on the orthogonal standardized matrix X with elements x_{sr}. Denote by $\sum_{s(t,i)} x_{sr}^2$ the sum of x_{sr}^2 over those rows for which $u(t, i)$ has a 1 element then we assume that

$$\sum_{s(t,i)} x_{sr}^2 = O(n^{\alpha_i - 1}) \tag{23}$$

Since $\sum_s x_{sr}^2 = 1$ and the number of terms in $\sum_{s(t,i)}$ is $\nu(t, i) = O(n^{\alpha_i})$ condition (23) implies that asymptotically the x_{sr}^2 have a uniform density $x_{sr}^2 = O(n^{-1})$.

Finally we place on record a consequence of conditions (18) to (23): it follows from (5) using (18), (19), (23) and Schwartz' inequality that

$$u'(t, i)\, v(\tau, j) = \begin{cases} v(t, i) + O(n^{2\alpha_i - 1}) & \text{for } t = \tau,\ i = j \\ 0 + O(n^{2\alpha_i - 1}) & \text{for } t \neq \tau,\ i = j \\ v(t, i; \tau, j) + O(n^{\alpha_i + \alpha_j - 1}) & \text{for } i \neq j \end{cases} \tag{24}$$

We now turn to the asymptotic behavior of the k_{ii} and k_{ij}. From (8), (17), (20), and (25) we have that

$$\begin{aligned} k_{ii} &= \sum_{t=1}^{m_i} \sum_{\tau=1}^{m_j} (u'(t, i)\, v(\tau, i))^2 \\ &= \sum_{t=1}^{m_i} \left\{ u'(t, i)\, v(t, i) \right\}^2 + \sum_{t \neq \tau}^{m_i} \left\{ u'(t, i)\, v(\tau, i) \right\}^2 \qquad (25) \\ &\geq \text{Const } n^{1-\alpha_i+2\alpha_i} + O(n^{2-2\alpha_i+4\alpha_i-2}) \\ &\geq C\, n^{1+\alpha_i} \qquad \text{for all } i = 1, \ldots, c+1 \end{aligned}$$

From (8), (17), (19), (21) and (24) we have for $i \neq j$; $i = 1, \ldots, c+1$; $j = 1, \ldots, c$

$$\begin{aligned} k_{ij} &= \sum_{t=1}^{m_i} \sum_{\tau=1}^{m_j} \left\{ u'(t, i)\, v(\tau, j) \right\}^2 \\ &= \sum_{t}^{m_i} \sum_{\tau}^{m_j} v(t, i; \tau, j)^2 + O(n^{\alpha_i + \alpha_j - 1}) \sum_{t}^{m_i} \sum_{\tau}^{m_j} v(t, i; \tau, j) \\ &\qquad + \sum_{t}^{m_i} \sum_{\tau}^{m_j} O(n^{2\alpha_i + 2\alpha_j - 2}) \\ &= \sum_{t}^{m_i} o(v(t, i)) \sum_{\tau}^{m_j} v(t, i; \tau, j) + O(n^{\alpha_i + \alpha_j - 1}) n \qquad (26) \\ &\quad + O(n^{2-\alpha_i-\alpha_j})\, O(n^{2\alpha_i+2\alpha_j-2}) \end{aligned}$$

$$= o(n^{1+\alpha_i}) + O(n^{\alpha_i+\alpha_j}) = o(n^{1+\alpha_i}) \tag{26}$$

since $\alpha_j < 1$. Similarly we prove by symmetry that $k_{ij} = o(n^{1+\alpha_j})$ for $i \neq j \leq c$. From (25) and (26) it is clear that for all large n the $c \times c$ matrix k_{ij} for $i, j = 1, \ldots, c$ is asymptotically diagonal with diagonal coefficients $\geq cn^{1+\alpha_i}$ while the coefficients $k_{c+1,j}$ are asymptotically equal to o(n). Moreover it is obvious from (12) that $k_{c+1,c+1} \geq Cn$. Using therefore the first c equations of $K\hat{\sigma}^2 = Q(y)$ we obtain that

$$\hat{\sigma}_i^2 = O(n^{-\alpha_i-1}) \{Q_i(y) - o(n)\hat{\sigma}^2_{c+1}\} = O(n^{-\alpha_i-1})Q_i(y) + o(n^{-\alpha_i})\hat{\sigma}^2_{c+1}$$

$$\text{for } i = 1, \ldots, c \tag{27}$$

Substituting (27) in the last equation we obtain

$$\hat{\sigma}^2_{c+1}\{cn + o(n^{1-\alpha_{min}})\} = Q_{c+1}(y) + \sum_{i=1}^{c} Q_i(y)\, o(n^{-\alpha_i}) \tag{28}$$

or

$$\hat{\sigma}^2_{c+1} = O(n^{-1})Q_{c+1}(y) + \sum_{i=1}^{c} Q_i(y)\, o(n^{-\alpha_i-1}) \tag{29}$$

Substituting (29) back in (27) we obtain

$$\hat{\sigma}_i^2 = O(n^{-\alpha_i-1})\, Q_i(y) + o(n^{-1-\alpha_i})\, Q_{c+1}(y) \tag{30}$$

Equations (29) and (30) show that $\hat{\sigma}^2$ is estimable from the $Q_i(y)$. They also show that $\hat{\sigma}^2$ is consistent provided we can show that

$$\begin{aligned} \text{Var } Q_r(y) &= o(n^{2\alpha_r+2}) \\ \text{Var } Q_{c+1}(y) &= o(n^2) \end{aligned} \qquad \text{for } r = 1, \ldots, c \tag{31}$$

since $\text{Cov}Q_i(y)Q_j(y) = O(\text{Var}Q_i(y)^{\frac{1}{2}}\ \text{Var}Q_j(y)^{\frac{1}{2}})$.

In order to prove the first result in (31) we use formulas [22], [32], [33] and [34] of J.N.K. Rao (1968) with slightly altered notation. Formula [22] gives $E\ Q_r^2(y)$ in the form

$$E(Q_r(y)^2) = 2 \sum_{i<j=1}^{c+1} \sum c_{ij}\sigma_i^2\sigma_j^2 + \sum_{i=1}^{c+1} c_{ii}\sigma_i^4 + \sum_{i=1}^{c+1} h_i\mu_{4i} \tag{32}$$

where $\mu_{4i} = E\ b_{i\ell}^4$ are the 4^{th} moments of the elements $b_{i\ell}$ of b_i. Noting that $\text{Var } Q_r(y) = E\ Q_r(y)^2 - E^2(Q_r(y))$ the leading terms of c_{ii} and c_{ij} given by J.N.K. Rao's equations [33] and [32] cancel and we are left to consider the orders of magnitude of

$$\begin{aligned} c_{ii}-2h_i &= \sum_{t<\tau=1}^{m_i} \sum \{Q_r(u(t, i) + u(\tau, i)) - Q_r(u(t, i)) - Q_r(u(\tau, i))^2 \\ &= \sum_{t<\tau=1}^{m_i} \sum \{ \sum_{s=1}^{m_r} 2(u(t, i)' v(s, r))\ (u(\tau, i)' v(s, r))\}^2 \ . \end{aligned} \tag{33}$$

Consider first the case $r = i$. We distinguish two terms when $s = t$ and $s = \tau$. For those two terms $(u(t, i)' v(s, i))\ (u(\tau, i)' v(s, i))$ is from (24) of the order of magnitude $O(n^{\alpha_i})\ O(n^{2\alpha_i-1}) = O(n^{3\alpha_i-1})$. For the remaining terms in $\sum_{s=1}^{m_r}$ the product is of the order $O(n^{4\alpha_i-2})$ but the number of terms is of the order

$O(n^{1-\alpha_i})$ so that $\{\sum_s\}^2$ is $O(n^{6\alpha_i-2})$ and hence $\dot{c}_{ii} = O(n^{2-2\alpha_i})\, O(n^{6\alpha_i-2}) = O(n^{4\alpha_i}) = o(n^{2\alpha_i+2})$ since $\alpha_i < 1$.

Consider next the case $r \neq i$ and $r \neq c + 1$. We have from (33) and (24)

$$\begin{aligned}
\dot{c}_{ii} &= \sum_{t<\tau}^{m_i} \sum \{\sum_s (\nu(t, i; s, r)\, \nu(t, i; s, r) + O(n^{2\alpha_i+2\alpha_r-2}) \\
&\quad + O(n^{\alpha_i+\alpha_r-1})\, (\nu(t, i; s, t) + \nu(\tau, i; s, r))\}^2 \\
&= \sum_{t<\tau}^{m_i} \sum \{o(\nu(s, r)) \sum_s \nu(\tau, i; s, r) + O(n^{2\alpha_i+\alpha_r-1}) \\
&\quad + O(n^{\alpha_i+\alpha_r-1})\, (\nu(t, i) + \nu(\tau, i))\}^2 \qquad (34) \\
&= \sum_{t<\tau}^{m_i} \sum \{o(n^{\alpha_r+\alpha_i}) + O(n^{2\alpha_i+\alpha_r-1})\}^2 \\
&= o(n^{2+2\alpha_r}) + o(n^{\alpha_i+2\alpha_r+1}) + O(n^{2\alpha_i+2\alpha_r}) \\
&= o(n^{2+2\alpha_r}).
\end{aligned}$$

The case $r \neq i$, $r = c + 1$ follows on the same lines as (34) except that $\alpha_r = 0$ and that $\nu(t, i; s, c+1)\, \nu(\tau, i; s, c+1) = 0$ since $u(s, r)$ has a 1 only in the s^{th} row and either $u(t, i)$ or $u(\tau, i)$ have a zero in that row. The order of magnitude of $\{\}$ will therefore be $O(n^{2\alpha_i-1})$ and $\dot{c}_{ii}$ will be $O(n^{2\alpha_i}) = o(n^2)$.

The treatment of the c_{ij} in J.N.K. Rao's formula [33] follows on similar lines to the above proof for the c_{ii} if of the two alternatives $i < j$, $j < i$ in (21) the smaller α_i, α_j is selected for majorisations.

It remains to consider the terms

$$h_i = \sum_{t=1}^{m_i} Q_r^2(u(t, i)) = \sum_{t=1}^{m_i} \{ \sum_{s=1}^{m_r} (u'(t, i)\, v(s, r))^2\}^2 \tag{35}$$

For the case r = i we have using (24)

$$h_i = \sum_{t=1}^{m_i} \{(u'(t, i)\, v(t, i))^2 + \sum_{s \neq t}^{m_i} (u'(t, i)\, v(s, i))^2\}^2$$

$$= \sum_{t=1}^{m_i} \{O(n^{2\alpha_i}) + O(n^{3\alpha_i - 1})\}^2 \tag{36}$$

$$= O(n^{1+3\alpha_i}) + O(n^{4\alpha_i}) + O(n^{5\alpha_i - 1})$$

$$= o(n^{2\alpha_i + 2}) = o(n^{2\alpha_r + 2}) \qquad \text{for } i = r \neq c + 1,$$

$$= o(n^2) \qquad \text{for } i = r = c + 1.$$

For the case i ≠ r and r ≠ c + 1

$$h_i = \sum_{t=1}^{m_i} \{ \sum_{s=1}^{m_r} (v(t, i; s, r) + O(n^{\alpha_i + \alpha_r - 1}))^2\}^2$$

$$= \sum_{t=1}^{m_i} \{ \sum_{s=1}^{m_r} o(v(s, r))v(t, i; s, r) + O(n^{\alpha_i + \alpha_r - 1}) \sum_{s} v(t, i; s, r)$$

$$+ O(n^{1-\alpha_r})\, O(n^{2\alpha_i + 2\alpha_r - 2})\}^2 \tag{37}$$

$$= \sum_{t=1}^{m_i} \{o(n^{\alpha_i + \alpha_r}) + O(n^{2\alpha_i + \alpha_r - 1})\}^2$$

$$= o(n^{\alpha_i + 2\alpha_r + 1}) + o(n^{2\alpha_i + 2\alpha_r}) + O(n^{3\alpha_i + 2\alpha_r - 1})$$

$$= o(n^{2 + 2\alpha_r}).$$

Finally for $r = c + 1$, $i \neq r$ we have

$$h_i = \sum_{t=1}^{m_i} \{ \sum_{s=1}^{n} (\nu(t, i; s, r) + O(n^{\alpha_i - 1}))^2 \}^2$$

$$= \sum_{t=1}^{m_i} \{ \sum_{s} \nu(t, i; s, r)^2 + \sum_{s} \nu(t, i; s, r)\, O(n^{\alpha_i - 1}) + O(n^{2\alpha_i - 1}) \}^2 \qquad (38)$$

Now since $\nu(t, i; s, c + 1)$ is either 0 or 1 we have that $\sum_{s} \nu(t, i; s, c + 1)^2 = \sum_{s} \nu(t, i; s, c + 1) = \nu(t, i)$ so that

$$h_i = \sum_{t=1}^{m_i} \{O(n^{\alpha_i}) + O(n^{2\alpha_i - 1})\}^2 \qquad (39)$$

$$= O(n^{1-\alpha_i})\, O(n^{2\alpha_i})$$

$$= o(n^2).$$

Since $\hat{\sigma}^2$ is unbiassed and Cov $(\hat{\sigma}^2) \to 0$ as $n \to \infty$ it follows that $\hat{\sigma}^2$ is consistent. Moreover if we replace any negative $\hat{\sigma}_i^2$ by 0 the resulting statistic say $\tilde{\sigma}_i^2$ has a smaller mean square error and hence is also consistent.

The consistent estimator $\hat{\sigma}^2$ may serve as a starting value for the iterative maximum likelihood estimation procedure described by Hemmerle and Hartley (1973). Under certain regularity conditions (not discussed here) one single cycle of the iteration will result in asymptotically efficient estimators of σ^2 and α. If the iteration is carried to convergence solutions of the ML equations are reached. If no ML cycles are performed a consistent estimator $\hat{\alpha}$ of α can be computed from the generalized least squares (ML) equations.

$$\left.\begin{aligned} \hat{\alpha} &= (X'H^{-1}X)^{-1}(X'H^{-1}y) \\ \text{where } H &= I_n + \sum_{i=1}^{c} \frac{\sigma_i^2}{\sigma_{c+1}^2} U_i U_i' \end{aligned}\right\} \qquad (40)$$

It has been shown by Henderson and Hartley (1973) that (40) can be computed directly from the $U_i U_i'$ and $X'U_i$ matrices without the inversion of the $n \times n$ matrix H using their so called W transformation. In fact the W_o matrix (their equation (19)) is essentially given by the $V_i'V_i$ matrices (see the above Schedule 1) and by the contrasts $V_i'y$ required in the computation of $Q_i(y)$.

The variance covariance matrix of $\hat{\alpha}$ can likewise be computed through the W transformation.

Acknowledgement

One of us (H.O.H.) wishes to acknowledge support from the Army Research Office.

J.N.K. Rao wishes to acknowledge support from the National Research Council of Canada.

References

Corbeil, R. R. and Searle, S. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics 15, 819-826.

Gaylor, D. W., Lucas, H. L. and Anderson, R. L. (1970). Calculation of the expected mean squares by the abbreviated Doolittle and square root methods. Biometrics, 26, 641-656.

Hartley, H. O. (1967). Expectations variances and covariances of ANOVA mean squares by 'synthesis'. Biometrics, 23, 105-114.

Hartley, H. O. (1977). Analysis of unbalanced experiments. Invited address 23rd Conference on the Design of Experiments in Army Research Development and Testing to be published in proceedings.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. Biometrika, 54, 93-108.

Hartley, H. O. and Rao, J. N. K. (1977). The estimation of non-sampling variance components in sample surveys. Presented at Symposium on Sample Surveys and Measurement, Chappel Hill, April 14-17, 1977.

Hemmerle, W. J. and Hartley, H. O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the W-transformation. Technometrics, 15, 819-831.

Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics, 9, 226-252.

LaMotte, L. R. (1973). Quadratic estimation of variance components. Biometrics, 29, 311-330.

León, L. and Senturia, J. (1976). Computation of minque variance components estimates. Technical Report #408, Department of Statistics, University of Wisconsin.

Rao, C. R. (1971). Estimation of variance and covarisnce components-minque theory. Journal of Multivariate Analysis, 1, 257-275.

Rao, J. N. K. (1968). On expectations variances and covariances of ANOVA mean squares by 'synthesis'. Biometrics, 24, 963-978.

MEASURE OF EFFECTIVENESS FOR DIVISION LEVEL MODELS

John H. Shuford and Fredrick H. Knack
Special Studies Division
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico

ABSTRACT

High level excrusions, using the Division Battle Model (DBM) or a similar game, are expected to become more important in the performance of future Cost and Operational Effectiveness Analyses (COEAs). It is therefore necessary that a good Measure of Effectiveness (MOE) for use with these games be developed. Certain MOE, such as the force exchange ratio or other ratios, have become accepted as providing good estimates of the results of high resolution, company/battalion level combat simulations. Efforts have also been made to develop analytical weighting systems for the different weapons in order to compute weighted MOE. Both of these methods have been used to analyze the outcome of DBM, a low resolution division level war game, but neigher has been entirely satisfactory. It is hoped that this paper will stimulate interest and further investigation into the analysis and interpretation of combat simulation results.

The TRADOC Systems Analysis Activity (TRASANA) has recently completed a major weapon system study, using a division level war game as one of the analysis tools. In the course of this work, the problem of finding a proper measure of effectiveness to distinguish between the competing alternatives arose. This problem, of course, is common to all studies using models or simulations, but it does take on some different aspects at division level than at company/battalion level. A broader way of stating the problem, and perhaps the better way in the long term is: How should a model or experiment be designed in order to distinguish between completing weapon systems?

Since it is not possible to do complete field testing on every proposed weapon system, the use of simulations has been an important part of the test and selection process. Now there is a growing interest in using war games, which have been used principally as training aids in the past, as analysis tools. A war game may be defined as a combat simulation that is characterized by manual interplay and takes place in a simulated combat environment. This paper describes in some detail the war game used in the TRASANA study and demonstrates the dilemma faced in attempting to apply the "accepted" measures of effectiveness to the results. It is hoped that this presentation will both identify and lead to further investigation of a problem area that is critical to the weapon system evaluation process.

The model used was Division Battle Model. It is a computer-assisted, manual war game developed by the General Research Corporation (GRC) and is designed to support studies of the performance of weapons, organizations,

and tactics employed by a division sized force. Figure 1 describes DBM schematically. The study was primarily concerned with the ground combat portion of the game, which is linked to two other GRC models: CARMONETTE, a stochastic, high resolution, company/battalion simulation, and COMANEX, an extension of classical Lanchester theory. COMANEX is both a stand alone simulation and the ground combat assessment routine in DBM.

CARMONETTE's primary activities include the movement of units, the detection of targets, and the firing of weapons. Unit resolution is variable from individual weapon system to platoons. The model is critical event sequenced with time recorded to one-ten thousandth of a minute. The spatial representation is variable but a 100 meter grid is normally used. Input to the model are detailed descriptions of the units being played, performance characteristics of the various weapon types, a set of orders for each unit, including movement and target priorities, target detection probabilities, and a detailed description of the terrain. The unit orders must be based on a predetermined scenario and on a specified tactical doctrine, either current or one to be tested. The terrain description required by grid square, includes average elevation, height of vegetation, cover and concealment. Output from a CARMONETTE run is a computer listing of every event assessed during the battle which includes the elements killed, various operational statistics, and information on engagement ranges. Various summary routines may be used to collect the data in preparation for further analysis. In preparing CARMONETTE output for use as DBM ground combat history, a sufficient number of replications of each scenario must be made to develop good estimates of battle outcome.

DBM is a game rather than a simulation. It is played on a tactical type map of scale 1:25,000 to 1:50,000 which provides sufficient detail to support the levels of unit, time, and space resolution employed. For the TRASANA study, it was more practical to resolve to the company level for the Blue reinforced division and to the battalion level for the attacking Red combined arms army, but different levels may be used depending on the gamers' purpose. Space is measured to the nearest hundred meters. Time may be measured to the nearest five minutes, but it was found that to the nearest quarter hour was generally sufficient. While the game can be played in open, semi-closed, or closed modes depending on the degree to which intelligence is considered a critical factor, it has principally been used only in the open mode. In this way, two to four hours of battle time can be gamed per working day by a player/controller team.

The manual operations of DBM consist mainly of decision making, event determination and time sequencing, while the computerized portion focuses on the determination of battle losses, tabulation and reporting of battle results, and updating of stored information. Manual play takes place over approximately four-hour increments of battle time but may be stopped sooner if the control team determines that a critical event has occurred.

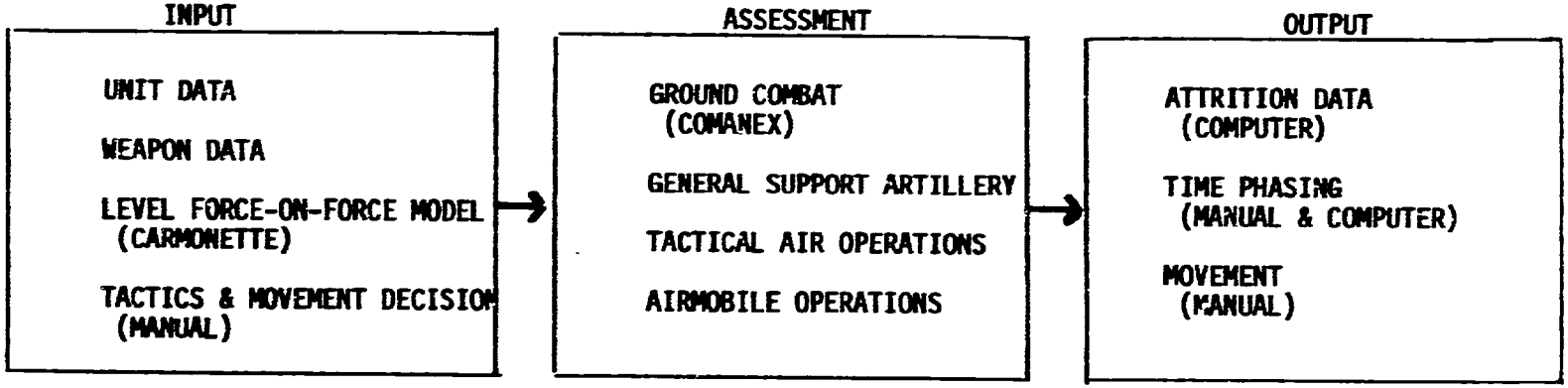


Figure 1. Diagram of Division Battle Model

At that point, computer input is prepared, describing the various combat actions that occurred during the manual phase. The computer routines then assess the casualties and provide a printout showing losses, cause of loss, and past and present unit strength. The control team makes necessary adjustments to unit locations and notifies the players of the battle outcome, after which manual play is resumed.

In order to provide the necessary background, the ground combat assessment routine must be described in some detail. The routine COMANEX solves a set of Lanchester type equations for the different weapons systems involved. These are shown for the simple case of one Blue and one Red weapon system. It may be noted that these equations reduce to the Lanchester square law for the case where all targets are acquired, that is the *Ps* approach *0*, and to the Lanchester linear law as the *Ps* approach unity, or no targets are acquired. COMANEX then treats combat situations between these two extremes of the Lanchester formulation. These equations are easily generalized to the case of several Blue and Red weapon types as is shown by the following equations:

Homogeneous Forces

$$\frac{dR}{dt} = -b(1-P_R^R)\ B \qquad (1)$$

$$\frac{dB}{dt} = -r(1-P_B^B)\ R$$

b = Rate at which one Blue weapon kills Red weapons given acquisition of at least one target.

P_B = Probability that a specific Blue target is unacquired by an individual Red firer.

B = Number of Blue weapons at time t

Similar definitions for r, P_R, and R

Heterogeneous Forces

$$\frac{dR_j}{dt} = -\sum_{i=1}^{m} b_{ij}(1-P_R^{R_j})\ B_i \qquad (2)$$

$$\frac{dB_i}{dt} = - \sum_{j=1}^{n} r_{ji}(1-P_B^{B_i}) R_j$$

b_{ij} = Rate at which one type i Blue weapon kills Red weapons of type j

B_i = Number of type i Blue weapons at time t

P_B = The same as for Homogeneous Forces

The values of the b_{ij}, P_B, and P_R are calculated by a COMANEX preprocessor from the results of each high resolution scenario. These are then used by the DBM ground combat assessment routine, the COMANEX simulator, to solve the equations and develop the results of battle groups using different but similar force structure from that used in the original CARMONETTE work. The validity of COMANEX in reproducing the results of CARMONETTE and in predicting the outcome of different scenarios has been tested both by the developer and at TRASANA and has been shown to be quite good. While these models simulate combat more or less realistically depending on our point of view, perhaps more from the point of view of a high level staff officer, devastatingly less from the aspect of an infantry private, they alone say nothing about effectiveness. In actual combat, the critical, in fact the only measure of effectiveness is mission accomplishment. Models are not as inflexible.

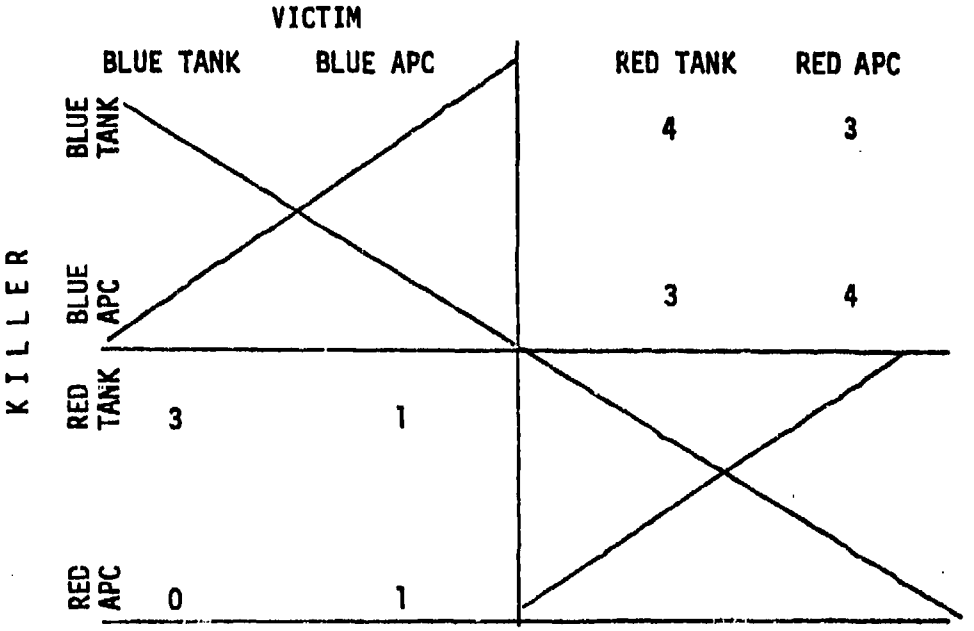
In high resolution simulations, the win or lose criteria may be difficult to define and quite arbitrary if it is done. Typically, battalion level simulations are not stopped at a logical breakpoint but are carried to extremes (e.g., 90% Red system losses) that distort both time and system losses. After making all of the necessary model runs for each weapon system, the analyst will analyze all of the data to identify a logical breakpoint. This "analysis point" is seldom driven by tactical consideration (if it were, it could be specified before hand) but rather by the necessity to find a point in the model output where all of the competing systems can be "objectively compared."

The typical numerical output from a simulation is in the form of a killer-victim scoreboard as is shown in Figure 2. These may be developed as frequently as is desired or practical during the simulation and provide a summary of the battle events.

INITIAL FORCES

BLUE		RED	
TANKS	APCs	TANKS	APCs
5	4	15	10

KILLER VICTIM SCOREBOARD



FINAL FORCES

BLUE		RED	
TANKS	APCs	TANKS	APCs
2	2	8	3

Figure 2. Example of Wargame Data

Figure 3 shows some of the traditional type measures of effectiveness used with killer-victim scoreboard data. The loss exchange ratio and force exchange ratio are often used with CARMONETTE type simulation. The total tank ratio and tank contribution are less common but have still been seen.

When one computes the value of an MOE at an analysis point, the difficulties are usually just beginning. If different values for the MOE are found (as is desired, if multiple MOE are used, it is also desired that any differences are in the same direction) some determination must be made about the significance of the differences. If a stochastic model such as CARMONETTE is being used, one can of course conduct a statistical significance test providing there is some knowledge about the distribution of the model output. If not, non-parametric statistics can be used. If, on the other hand, a deterministic battalion level model is being used, a difference of 10% is the accepted figure for significance. If no significant difference can be shown in the MOE, it is hoped that the model has provided enough "valuable insights" to come to a decision on the best (preferred) system.

When analyzing the results of a division level model, things are not as clear cut. First, it is difficult to use any of the traditional ratio type MOE because the force ratios are constantly changing with the intensity of the battle and the tactical decisions being made by the players. Analysis points can be identified as some arbitrary fraction of survivors (or losses) of the total force available and then the ratio type MOE may be used, but the problem here is that varying numbers of forces actually participate. In simulations at company/battalion level, a certain force is committed initially and fights to the conclusion, with the entire battle taking place in a time frame of approximately one-half hour or less. In contrast to this, a division game may require a period of one to four days of combat time, while the intensity varies not only with time, but also with space along the division front. The numbers of engaging forces change as a result of both combat attrition and the tactical decisions made, such as commitment of the reserve or withdrawal of a unit to another position.

The strong point of the division game, however, is that tactical stopping points can be easily identified prior to the start of the game. For the TRASANA study the end of game criteria was simply mission accomplishment by Red or Blue. The game was stopped with Red accomplished his mission by penetrating the Blue rear boundary or when Blue accomplished his mission by causing Red to break off the attack and go on the defensive. It was fortunate in the study that there were three distinct outcomes for our three leading candidates. With one candidate Blue lost; with the second, he prevented a penetration, but at a cost of an entire division. However, with the third candidate Blue not only prevented the breakthrough but had the capability to mount a strong counter-attack. Even with results that diverse, a quantitative MOE is required if only to have something to use with cost comparisons.

TRADITIONAL MOE

Loss Exchange Ratio (LER)

$$LER = \frac{\text{Number of Red Systems Lost}}{\text{Number of Blue Systems Lost}}$$

Force Exchange Ratio (FER)

$$FER = \frac{\text{Number of Red Systems Lost/Initial Number of Red Systems}}{\text{Number of Blue Systems Lost/Initial Number of Blue Systems}}$$

$$= \frac{\text{Loss Exchange Ratio}}{\text{Engaging Force Ratio}}$$

Total Tank Ratio (TTR)

$$TTR = \frac{\text{Red Tanks Killed}}{\text{Blue Tanks Killed}}$$

Tank Contribution (TC)

$$TC = \frac{\text{Red Systems Killed by Blue Tanks}}{\text{Blue Tanks Killed}}$$

Figure 3

Figure 4 shows an example of the Force Exchange Ratio calculated for each alternative for the previous manual interval at various times during the game. Comment is unnecessary on the difficulty of using this as an MOE.

Figure 5 shows the Loss Exchange Ratios for the same case. Here the curves have been smoothed by taking cumulative values throughout the course of the battle, but the differences lie only in the relative positions of the curves and are still difficult to interpret. The arrows show points of equal Red losses.

Simulations have long been used as test beds for weapon systems; in contrast, war games have traditionally been used as training aids. It is becoming recognized that the games, particularly high level ones, have a legitimate use in the analysis process. In fact, TRASANA and the Combined Arms Center at Fort Leavenworth are devoting considerable joint effort toward improving existing games and developing new ones for use in both training and analysis. Use of the game does, however, present some problems in experiment design and data interpretation that have not been fully explored.

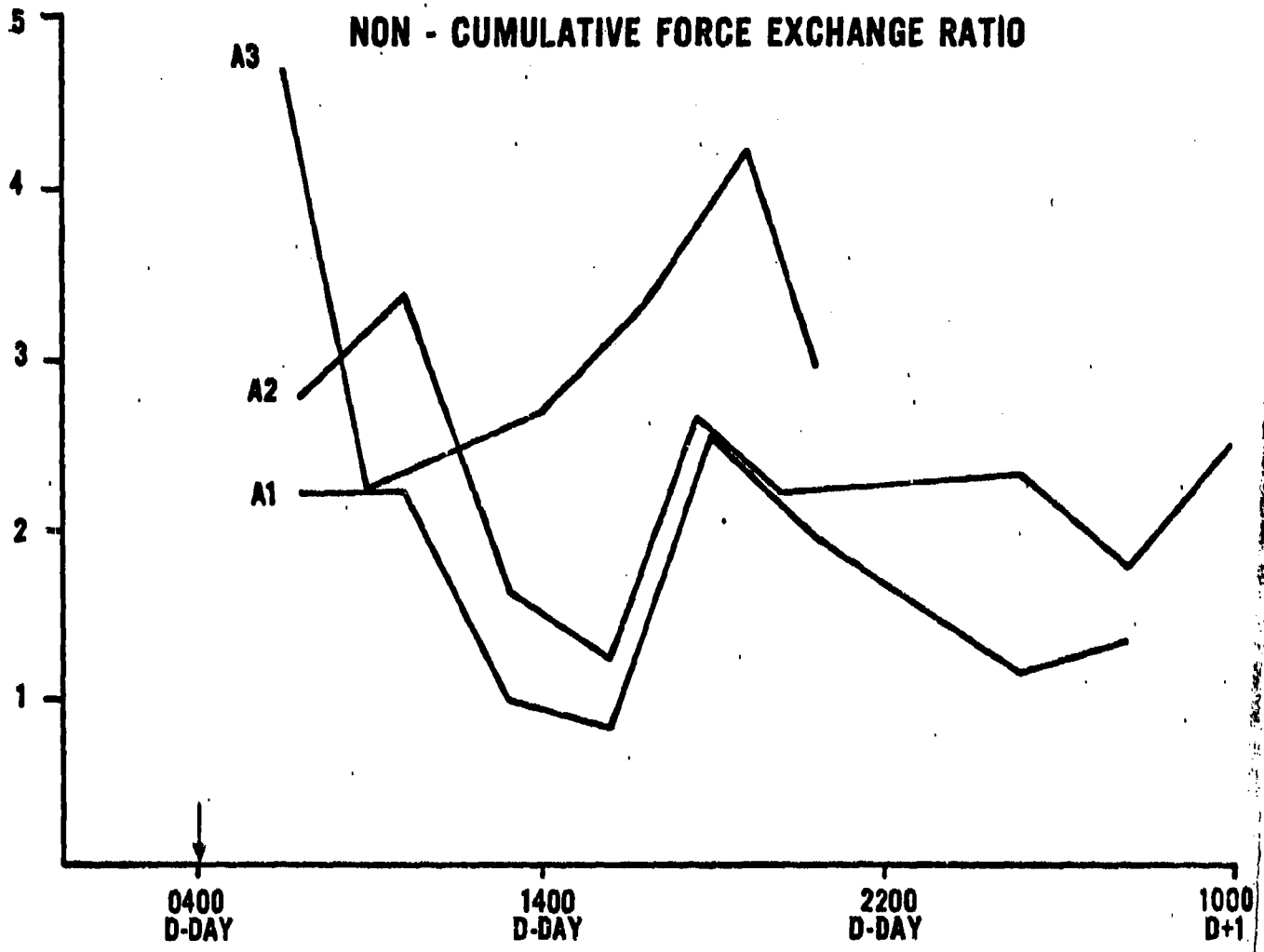

NON - CUMULATIVE FORCE EXCHANGE RATIO
5
4
3
2
1
A3
A2
A1
0400
D-DAY
1400
D-DAY
2200
D-DAY
1000
D+1

Figure 4

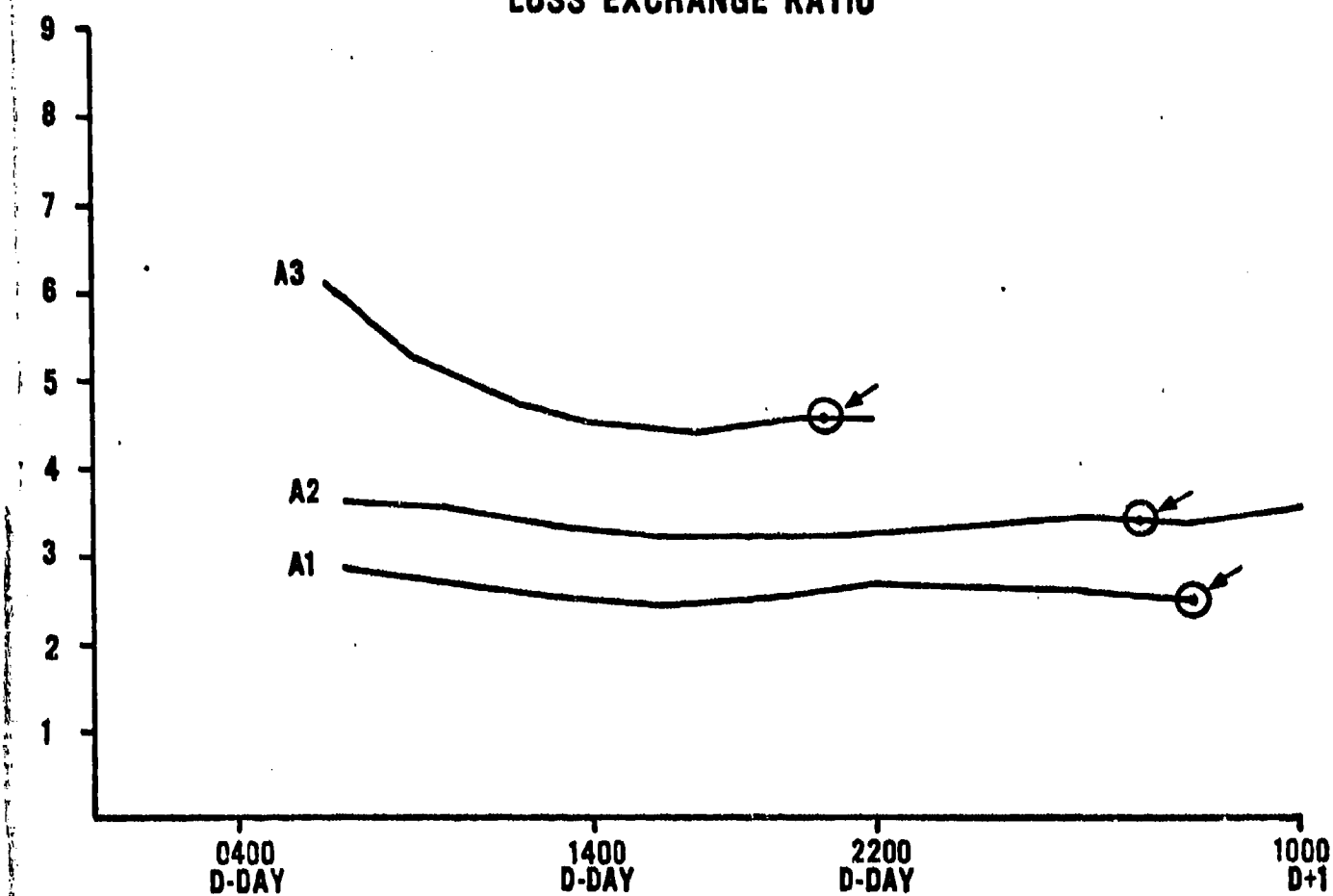

LOSS EXCHANGE RATIO
9
8
7
6
5
4
3
2
1
A3
A2
A1
0400
D-DAY
1400
D-DAY
2200
D-DAY
1000
D+1

Figure 5

ANALYSIS OF RATIO DATA FROM FIELD EXPERIMENTATION

Brian Barr
US Army Combat Developments Experimentation Command
Fort Ord, California

ABSTRACT. Measures of effectiveness which result from taking the ratio of two dependent variables are difficult to analyze. The problem becomes further complicated when the data come from field experimentation where the data is rarely "clean".

Examples of the type of data involved are presented along with the reasons why the data cannot be analyzed using standard techniques. The analysis approach of looking at the numerator and denominator separately is discussed along with the reasons why this technique cannot be universally applied to ratio data.

I. INTRODUCTION. The Combat Developments Experimentation Command (CDEC) conducts field experiments for the U.S. Army. These experiments quite often take the form of instrumented force-on-force field tests in which one tactical unit engages another in a relatively free play environment. The instrumentation permits the collection of detailed data on the engagement sequences as they occur. Normally, four or five independent variables are controlled, but the number of uncontrolled or nuisance variables can be almost infinite.

Examples of the types of measures of effectiveness that have been used in previous experiments include the ratio of red kills to blue kills, the ratio of detections to engagements, the ratio of targets exposed to detections, and the ratio of ammunition expended to hits or kills. One ratio in particular that has appeared repeatedly is the casualty exchange ratio, the ratio of red kills to blue kills. (Many arguments can be presented for and against using this as a measure of effectiveness. Without getting into that topic, it should suffice to say that this MOE has appeared before and will probably continue to be used.)

II. THE PROBLEM. The problems with analyzing the casualty exchange ratio from field experimentation data start before the calculation of the MOE. The first problem is that the sample size is usually severely limited by practical constraints (field experiments are extremely expensive). Time and cost constraints quite often overshadow statistical considerations and the analyst must do the best with what he is given. The sample size is further complicated because up to 25 percent of the trials may be invalidated due to operational problems or instrumentation failures. When these trials cannot be rerun the result is unequal sample sizes. The sample sizes may also be unbalanced by the nature of the MOE. The sample size of the ratio of targets to detections, for example, is dictated by the number of detection opportunities which randomly appear during the field trial.

A typical field experiment design might look like this:

	A1	A2	A3
B1	6	6	6
B2	6	6	6

The independent variables A and B have three and two levels respectively. Two more variables may be nested equally in the cells and an undetermined number of nuisance variables may appear during execution. (In PARFOX VII, for example, with 54 trials, nine variables could be shown to influence the dependent variable.) These nuisance variables normally result in great variability of the data within each cell.

The two elements of the ratio MOE are rarely if ever independent of one another. The number of red players who have been killed obviously influences the number of blue players who will be killed. Also the distribution of the number of kills on either side is usually skewed in one direction and often truncated by an arbitrary end of trial criteria. Thus, the distribution is rarely normal.

III. PAST TRIALS. CDEC has been relatively successful in analyzing casualty exchange ratio data by using analysis of covariance techniques; however this has only been possible because the basic statistical question of how to test hypotheses on ratio data has been avoided. Instead of analyzing the ratio, the numerator and the denominator have been analyzed independently, then conclusions have been drawn from the results of these two analyses.

The approach taken so far has followed the following logic:

$$\frac{R1}{B1} \lesssim \frac{R2}{B2}$$

If R1 is greater than R2 and if B1 is smaller than B2, then the ratio R1/B1 must be greater than the ratio R2/B2. Likewise, if R1 equals R2, and B1 is smaller than B2, then R1/B1 is greater than R2/B2; or if B1 equals B2, and R1 is greater than R2, then R1/B1 is greater than R2/B2. This logic doesn't appear to bother anyone until the case where R1 is greater than R2 and B1 is greater than B2 (so far CDEC has not had this appear, but is would seem to be just a matter of time). Look at the possibilities:

(a) $\frac{4}{12} = \frac{1}{3}$

(b) $\frac{4}{12} > \frac{1}{10}$

(c) $\frac{4}{12} < \frac{3}{6}$

In each case, the statistical testing on the separate variables tells us the same thing (R1 is greater than R2 and B1 is greater than B2); but the ratios are equal, greater, and smaller respectively.

An additional consideration that will not be discussed but should be mentioned is the case where we have:

$$\frac{1}{2} = \frac{10}{20}$$

The ratios are equal, but obviously the battles are not identical since the casualties on both sides vary by a factor of 10.

IV. SUMMARY. Every indication points to the fact that ratio type measures of effectiveness will continue to appear in field experimentation. Literature searches have failed to reveal acceptable solutions to the analysis of ratio data, and eventually the case will arise where the separation of numerator and denominator will no longer be adequate. Further work needs to be conducted in this area, both to strengthen Army field experimentation and to benefit the whole statistical community.

PHYSIOLOGICAL AND PERCEPTUAL ADAPTATION TO SUSTAINED AND MAXIMAL WORK IN YOUNG WOMEN

D. Kowal[1], D. Horstman[1], and L. Vaughan[2]
Exercise Physiology Division, US Army Research Institute of Environmental Medicine, Natick, MA[1]
Department of Physical Education, Wellesley College, Wellesley, MA[2]

In order to better understand differences in physical work performance between men and women, a study was carried out to determine if a.) women perceive physical effort differently than men; b.) does previous activity experience influence the perception of effort; and c.) how does acute and chronic training affect the perception of effort and ability for prolonged work in women. Preliminary analysis suggests that perceived exertion in women is influenced by activity history and self concept prior to participation in aerobic training. The perceptual measures displayed a substantial interaction depending upon self concept/prior activity and group affiliation of these women. Psychological estimates of physical self concept improved for the previous low activity training group but not for the previous high activity training group when compared to controls.

Background:

Presently, about 5% of the workforce of the US Army is comprised of women, the highest percentage in peacetime history. This figure will increase substantially within the next few years with a projected contingency of 50,000 women soldiers. The role of the Army's women has also undergone drastic change; whereas previously confined to less physically demanding tasks (such as clerical work), all Military Occupation Specialties are presently available to women, with the exception of combat arms. With the prospect of increasing numbers of women serving in a greater variety of work roles, our interests have focused on the performance of prolonged physical work by women. Sustained performance of physical work is governed by two distinct factors: (a) one's capacity for work and (b) one's willingness to endure hard physical work. Capacity is objective in nature and dependent to a large extent upon genetic traits, but can be modified by other influences (primarily physiological), such as training, diet, and environment (1,2,3,4). Willingness to endure is more complex and subjective in nature, and probably governed by psychosocial factors (5,6,7,8).

Given this situation, the question is obvious: Are there physiological or perceptual differences between men and women that may obviate the latter from performing sustained heavy work. Currently available research provides little information. However, observations in our laboratory suggest that, when asked to perform tests which require a maximum voluntary contraction, women tend to score less than could be predicted on the basis of physiological indices, e.g., lean body mass. It has also been reported that women possess approximately half the

arm and shoulder strength of men, 3/4 the leg strengths, and 3/4 the aerobic capacity of the average man (9). Further we recognize that perception of work is related to experience. However, because society has often considered women incapable or it unfeminine, many women have not experienced strenuous physical work.

This study was designed to evaluate the following questions:

1. Do women perceive work differently than men and are the physiological and psychological factors related to work capacity the same for both groups?

2. How does prior experience influence the perception of effort and the capacity for sustained work performance?

3. Do women who have had high activity experience differ from those with low activity history in their response to training?

Progress:

Seventy-five women volunteers ages 18-22 served as subjects. They were assigned to one of 5 groups: Low previous activity, experimental (N = 14) and control (N = 15), high previous activity, experimental (N = 15) and control (15) and an intercollegiate athlete (high fitness) group (N = 15). The following measurements were made during the first and last week of the program. Anthropometric measurements were made of height and weight. Body composition was determined by measuring skin fold thickness with a Harpenden caliper at four anatomical sites: triceps, biceps, subscapular, and suprailiac. An interrupted treadmill test for maximal aerobic power ($\dot{V}O_2$ max) was performed following the procedure of Taylor (10). During the last minute of each run, the expired gas was collected into vinyl Douglas bags and analyzed for oxygen and CO_2 content. Subjects were monitored electrocardiographically during all runs. $\dot{V}O_2$ max was determined when the oxygen uptake did not increase with an increase in work load. At the end of each run the subject rated her perceptual response during the workload using Borg's report of perceived exertion (RPE). The RPE is a ratio scale from 6-20 with verbal labels: 6 = very, very light to 20 = very, very hard. The treadmill test for aerobic fitness was performed during the first week of study (Pre-training), a week later (Acute) and following the 12-week training program (Post-training). These replications

were performed to assess changes that may have occurred in both physiological and perceptual responses to maximal work, and the aerobic training program. The training program consisted of 12 weeks during which the women ran for progressively longer periods of time at a faster pace. Each week a 30 minute test run was performed to assess improvement in stamina and endurance. During the pre- and post-testing sessions the subjects were asked to complete a battery of cognitive and behavioral self-evaluation questionnaires designed to assess their attitude toward exercise, expectations of their physical capacity and performance.

Anthropometric measures are summarized in Table 1. The findings suggest that women engaged in an aerobic training program can expect to lose body fat but gain some weight even though they are maintaining high energy expenditures. This is attributed to the increase in caloric intake reported by the members of the training groups. Table 2 summarizes the physiological and perceptual responses to initial, acute and post training maximal exercise. The anticipated improvement in aerobic fitness is evident with improvement in $\dot{V}O_2$ max increasing 8% for the high activity group and approximately 15% for the low activity group. It is difficult to equate the perceptions of effort (RPE) reported because of the different workloads involved at the end of the training program. The other measures of aerobic fitness, ventilation (V_E max), maximum heart rate (HR_{max}) and maximum workload (speed/grade) also showed the anticipated improvement as a result of training.

Table 3 describes the physiological and perceptual responses to a 20 minute endurance run at 70% of $\dot{V}O_2$ max. While the first two endurance runs were based on initial $\dot{V}O_2$ max values, the post-training 70% workload was calculated based on the subjects post-training $\dot{V}O_2$ max; i.e. absolute workload was increased from 8-15% for the groups. It can be seen that the perceptual responses to the same workload (pre-acute) were quite different. This finding suggests that exposure and activity experience alone may play an important role in understanding work performance in women even if no training is involved.

Data analysis of these physiological and perceptual measures across replications of the maximal performance and endurance tests are in progress. It can be seen in Table 4 that psychological measures of attitude toward activity, physical self-estimation, hidden shapes, motor satisfaction, perceived control of the environment (lack of control) and physical self concept did not demonstrate

substantial differences between the high and low activity groups. However, it is noteworthy that many of these measures were apparently different from the college norm population scores. This could be expected in light of the activity experience of the latter group.

In general the preliminary data analysis indicates that perceptual responses are intricately involved in the development of physical work capacity in women. Comparison of differences in peripheral responses between women and men will be reported subsequently. The population studied appears to be rather unique and superior to the college norms making psychological comparisons difficult; however, additional analysis is in progress.

LITERATURE CITED

1. Klissouris, V. J. Appl. Physiol. 29:358.

2. Pollack, M. In: Exercise and Sports Sciences Review, J. H. Wilmore (ed) N.Y., Academic Press, 1970.

3. Horstman, D. H. In: Ergogenic Aids and Muscular Performance, W. P. Morgan (ed) N.Y., Academic Press, 1972.

4. Astrand, P. and K. Rodahl. In: Textbook of Work Physiology, NY, McGraw, 1970.

5. Hilgard, E. R. Am. Psychologist, 24:103, 1969.

6. Sternback, R. A. In: Pain, A Psychophysiological Analysis. N.Y., Academic Press, 1968.

7. Buss, A. H. and N. W. Portnoy. J. Personality and Soc. Psych. 6:106, 1967.

8. Morgan, W. P. Med. Sci. in Sports 5:97, 1973.

9. Vogel, J. A., Ramos, M. U. and Patton, J. F. Med.Sci. in Sports 9:85, 1977.

10. Taylor, Buskirk et al. J. Appl. Physiol. 8:73, 1955.

TABLE 1. Anthropometric Characteristics of Women with Different Activity Patterns Before and After an Endurance Training Program.

		Experience			
		Low Activity		High Activity	
		Pre -Training	Post-Training	Pre -Training	Post-Training
Height (cm)	E	166.7 ± 1.4	166.8 ± 1.4	166.1 ± 1.3	166.2 ± 1.3
	C	162.9 ± 2.4	163.4 ± 2.6	166.1 ± 2.0	165.9 ± 2.1
Weight (Kg)	E	60.9 ± 1.6	61.4 ± 1.6	57.8 ± 1.6	58.9 ± 1.5
	C	57.7 ± 2.2	57.6 ± 2.2	59.2 ± 1.7	60.2 ± 2.1
Body Fat (%)	E	18.8 ± 0.9	18.3 ± 0.9	14.9 ± 1.0	14.7 ± 1.2
	C	16.1 ± 1.1	16.0 ± 1.1	16.5 ± 0.6	17.2 ± 0.9

Values represent means ± S.E. E = experimental (aerobic training)
C = control (non-training)

TABLE 2. Physiological and Perceptual Responses to Acute and Chronic Exposure to Maximal Treadmill Performance

		Experience					
		Low Activity			High Activity		
		Pre	Acute	Post Training	Pre	Acute	Post Training
$\dot{V}O_2$ max (L/min)	E	2.38 ± .07	2.38 ± .07	2.72 ± .07	2.48 ± .09	2.51 ± .07	2.77 ± .08
	C	2.43 ± .09		2.55 ± .09	2.58 ± .08		2.73 ± .11
$\dot{V}O_2$ max (ml/kg per min)	E	36.9 ± .08	37.9 ± .4	42.3 ± 1.1	43.9 ± 1.2	44.2 ± 1.0	47.8 ± 1.1
	C	38.8 ± 1.2		40.6 ± 1.1	45.6 ± 1.1		46.6 ± 0.9
Max HR	E	194 ± 1	192 ± 1	187 ± 2	189 ± 1	187 ± 2	187 ± 2
	C	193 ± 2		193 ± 2	191 ± 2		190 ± 2
Max V_E	E	91.6 ± 3.5	91.5 ± 4.1	96.6 ± 4.3	91.5 ± 4.0	95.9 ± 3.5	102.7 ± 3.4
	C	90.1 ± 3.7		102.6 ± 3.0	89.5 ± 4.8		99.1 ± 5.9
RPE max	E	17.2 ± .3	15.7 ± 0.6	17.9 ± 0.4	17.6 ± .4	17.5 ± 0.4	18.2 ± 0.4
	C	16.9 ± .7		17.7 ± 0.7	17.2 ± .4		16.8 ± 0.5
Max work load	E	5.3/6.1	5.4/7.7	6.4/9.9	6.0/7.0	5.7/8.6	6.5/9.7
	C	5.4/6.1		5.7/7.6	5.7/7.5		6.0/8.8

Values are mean ± S.E.

E = experimental (aerobic training)
C = control (non-training)

TABLE 3. Physiological and Perceptual Responses to a 20 Minute Endurance Run at 70% Max during the Last Minute of Performance

		Low Activity Experience			High Activity		
		Pre	Acute	Post	Pre	Acute	Post-Training
State Anxiety	E	5.9 ± 0.5	5.3 ± 0.4	4.8 ± 0.1	5.8 ± 0.5	5.3 ± 0.4	5.1 ± 0.2
	C	5.8 ± 0.8		4.7 ± 0.1	5.1 ± 0.4		5.0 ± 0.1
RPE	E	12.6 ± 0.6	11.3 ± 0.6	10.9 ± 6.7	11.7 ± 0.4	11.1 ± 0.4	10.8 ± 0.4
	C	13.3 ± 0.8		12.8 ± 6	11.5 ± 0.7		11.0 ± 0.6
Heart	E	172 ± 4	162 ± 4	171 ± 4	163 ± 4	154 ± 3	161 ± 4
	C	176 ± 5		182 ± 3	167 ± 3		174 ± 5

Values represent means ± S.E.

E = experimental (aerobic training)
C = control (non-training)

TABLE 4. Performance Expectations and Self Evaluation of Physical Abilities in Women of Different Activity Patterns Before and After an Endurance Training Program

Variable	N		Low Activity Experience Pre	Low Activity Experience Post-Training	N		High Activity Experience Pre	High Activity Experience Post-Training	College* Norms
Attitude Toward Physical Activity	22	E	31.75 ± 6.4	33.36 ± 6.7	18	E	38.78 ± 4.5	36.11 ± 11.0	32.5
	10	C	31.90 ± 7.9	34.5 ± 10.2	10	C	36.85 ± 4.3	37.29 ± 12.4	
Physical Self-Estimation	22	E	17.70 ± 5.5	19.13 ± 6.0	18	E	22.33 ± 5.9	23.78 ± 7.8	18.7
	10	C	19.10 ± 6.9	19.18 ± 8.4	10	C	21.64 ± 5.4	23.57 ± 8.5	
Hidden Shapes	22	E	32.5 ± 4.5	33.41 ± 2.6	18	E	32.16 ± 4.7	34.22 ± 2.6	15.37
	10	C	30.6 ± 6.8	35.1 ± 1.4	10	C	32.35 ± 3.4	32.7 ± 4.7	
Motor Satisfaction Scale	22	E	163.3 ± 25.7	175.8 ± 28.3	18	E	167.8 ± 31.1	195.9 ± 27.8	143
	10	C	163.0 ± 24.1	171.8 ± 30.9	10	C	162.1 ± 36.6	184.4 ± 32.1	
Locus of Control I-E	22	E	9.8 ± 4.1	9.7 ± 3.9	18	E	9.24 ± 4.1	8.05 ± 4.4	9.62
	10	C	10.5 ± 3.3	10.3 ± 4.7	10	C	8.5 ± 2.4	10.0 ± 3.7	
Physical Self Concept	22	E	62.7 ± 9.2	63.4 ± 9.4	18	E	69.22 ± 7.3	71.8 ± 7.2	
	10	C	64.0 ± 10.2	61.4 ± 10.8	10	C	67.14 ± 7.9	70.2 ± 10.7	

Values Represent (Means ± S.D.)

E = experimental (aerobic training)

C = control (non-training)

*Represent combined male and female values

THEORY OF LEAST CHI-SQUARE FOR POLYNOMIALS: IMPLICATION FOR DESIGN OF EXPERIMENTS

Richard L. Moore*
US Army Armament Research and Development Command
System Evaluation Office
Dover, NJ 07801

ABSTRACT. This paper extends the least Chi-Square theory (which was previously developed[1] for fitting data to non-linear functions of the parameters) to fitting polynomial functions of an independent variable. The underlying concept is that a Chi-Square is minimized. This Chi-Square is the ratio of the sum of the square of the residuals to the variance of the instrumental error plus the sum of the ratio of squares of an appropriate number of autocorrelation coefficients (with delay times which are integral increments of the interval between observations) to their variances.

The normal equations are extensions of, and reduce to, the ordinary least squares when the autocorrelation coefficients are zero. Iterative solution is required since the sum of squares of residuals and the autocorrelation coefficients depend on the values of the parameters. Two different approaches for the iterative solution have been programmed for a commercial programmable calculator. Typical results will be presented.

Effective use of this theory requires measurement of instrumental errors, and if appropriate, randomization of the order in which the independent variable(s) are varied.

The use of the theory is expected to give a set of values of the parameters which are "more probable" than those determined by ordinary least squares. It is expected to be "robust" to outliers and give an estimate of the probability that a particular outlier came from the same population as the other observations.

I. INTRODUCTION. The aim of the investigations which led to this paper was to find a better method to estimate the parameters in mathematical models of physical phenomena. Several assumptions are inherent in such a problem: Two of them were essential in our considerations:

*Based partially on work done in Logistics Executive Development Course, USA Logistics Management Center, Ft. Lee, VA.

First: The mathematical model or models under test are completely specified by a priori knowledge; only the parameters are unknown.*

Second: The errors are assumed to be measurement errors, and independent means are available (and have been used) to determine the precision of the measurement devices whose variance is given as σ_e^2. These measurement errors are assumed to be independent, and thus to form a random sequence.

Because of the first assumption, we do not permit ourselves to use the established statistical curve-fitting procedure of generalized least squares in which the variance-covariance matrix is transformed to a diagonal matrix. The procedure is rejected because in effect it changes the mathematical model to a different model, in which "periodic" terms are added to account for the observed values of the autocorrelation of the errors.

Because of the second assumption, we must provide a test as to whether, in fact, the errors remaining after the parameters have been estimated are consistent with a random generation of errors with a variance of σ_e^2, and if at the same time the autocorrelations observed are consistent with a random sequence of errors.

The last criteria is essential from an experimental point of view since, try as he may, the experimenter may not have succeeded in eliminating all sources of bias. To help him determine whether he has done so, many tests of the residuals are available[3,4]. However, these tests are essentially go/no go, and offer no method to improve the estimate of the parameters by reducing the autocorrelation.

Our object is to provide a data reduction method which will give a single test to answer the question: What is the probability that the set of residuals corresponding to a given set of parameters arise by a random sequence from a population with variance σ_e^2. Given this probability, can the probability be increased by a change in the parameters?

*Most (if not all) basic theories of physics can be derived from the least-square principle. This principle was stated by Gauss in 1828, and has recently been confirmed by Moore[2]. Because of this fact, it would be inappropriate to add additional terms to the physical theory merely to reduce the autocorrelation.

II. CONSIDERATION OF CRITERIA. In considering what statistical criteria could or should be used for our purpose, several well-known criteria such as "run" probability, error normality, etc. as considered by Anscombe[4], were proposed but were rejected either because a given test was not expressible easily in terms of the residuals and thus in terms of the parameters, or it was not directly applicable to the question of interest.

Evidently some form of chi-square tests would be desirable in view of the well-known fact that the sum of squares of the residuals follow a chi square distribution. The variance, covariance matrix (V_c^{-1}) was considered as a candidate by using the following, (where (III...) is a column vector amd (III...)′ is its transpose) (see Aitken[5]).

$$(III...)' V_c^{-1} (III...) = n \sum_{ij} \sigma_{ij}^2 \qquad (1)$$

The expected value of this expression is just $n\sigma^2$. Because σ_{ij}^2 where $i \neq j$ can be either negative or positive and because of the tendency for alternating positive and negative values in some cases, this expression was found to be unsatisfactory for a chi square test.

The next criteria which could be used is $(III...)' (V_c^{-1})^2 (III...)$ which equals $n (1 + \Sigma r_j)^2 \sigma^4$ whose expected value is $n (1 + 2 \Sigma r_{ij} + 2 \Sigma r_i r_j + \Sigma r_j^2) \sigma^4$.

If one should expand this square, on the assumption of r_i being not correlated with r_j (consistent with our second assumption) one might expect the sum of the cross product terms to vanish leaving only the sum of the squares. If this is the case, then the sum of the squares criteria (an alternate which follows) should be a more sensitive criteria.

A third alternative is the combination of (a) the "F" test of the variance of the residuals where the measurement variance σ_e^2 is the standard against which the sample sum of squares is compared, and (b) the Box-Pearce test of the sum of the squares of the autocorrelation coefficients divided by the individual variance v_j (Box and Pearce[6]).

The chi-square formed by combining these two tests is a single test of the joint probability of a given value of the sum of the squares and the corresponding values of the autocorrelation coefficients arising by chance from a particular set of estimates of the parameters.

The mathematical process to find the parameters which maximize the probability that both the "F" test and the autocorrelation test are satisfied will be called the "least chi-square method." Its derivation follows.

III. LEAST CHI PROCEDURE. In this derivation we will follow the procedure and most of the notation of Aitken [5] for generalized least squares:

Let the representation of the vector of data:

$$u = \{u(x_1),\ u(x_2),\ \ldots\ u(x_n)\} \tag{2}$$

by the vector:

$$y = \{y(x_1),\ y(x_2),\ \ldots\ y(x_n)\} \tag{3}$$

be linear in terms of a set of assumed functions

$$p_1(x),\ p_2(x),\ \ldots,\ p_{k+1}(x). \tag{4}$$

These functions are restricted only by the condition that they must be linearly independent over the n values of x.

If we let P be the matrix of these functions, the ith row of P is the row vector.

$$p_i = [p_1(x_i),\ p_2(x_i),\ \ldots\ p_{k+1}(x_i)]. \tag{5}$$

In this event, p is of the order of n x (k+1).

Let θ^* denote a column vector of k+1 coefficients independent of x, such that

$$\theta^* = \{\theta_1^*,\ \theta_2^*,\ \theta_3^*,\ \ldots\ \theta_{k+1}^*\} \tag{6}$$

(The asterisk symbol * will be used to indicate an estimate of the indicated symbol where convenient. However, it will not be used on complex expressions involving χ_T^2, σ^2, and r because of typographical difficulties). By definition then the vector y is $P\theta^*$ and we let the vector d be:

$$d = u - y = u - P\theta^*. \tag{7}$$

If χ_T^2 is defined in the first way considered, i.e., $(d)'(d)/\sigma_e^2$ plus the covariance normalized to σ_e^2, it is

$$\sigma_e^{-2}\, d'd + \sigma_e^{-2}\, \{[d'\, (V_c^{-1})^2 d] - (d'd)^4\} \tag{8}$$

and let

$$(V_c^{-1})^2 = [d\, (I + \sum_{j=1}^{s} V_j^{-1})\, d']^2 \tag{9}$$

In this expression V_j^{-1} is defined as follows:

$$V_1^{-1} = \begin{bmatrix} 0\,1\,0\,\ldots\,0 \\ 0\,0\,1\,0\,\ldots\,0 \\ 0\,0\,0\,1\,0\,\ldots\,0 \\ \ldots \end{bmatrix};\quad V_2^{-1} = \begin{bmatrix} 0\,0\,1\,0\,\ldots\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0 \end{bmatrix}$$

$$V_j^{-1} = \begin{bmatrix} 0\,0\,\ldots\,0\,1\,\ldots \\ 0\,0\,0\,\ldots\ldots\,0\,1\,\ldots \\ \ldots \end{bmatrix} \tag{10}$$

In these, the subscript "j" indicates a unit value in each of the ith rows and (i + j)th column. Thus equation (8) becomes

$$\chi_T^2 = \sigma_e^{-2}\, \{d'd + \sigma_e^{-2} d'[d'(2\sum_{j=1}^{s} V_j^{-1} + \sum_{j=1}^{s} V_j^{-2})d]d\} \tag{11}$$

The partials of equation (11) with respect to θ_r are clearly a complex expression, when compared to the method which follows this discussion, so that further analysis is not presented.

If the chi-square is taken as the final alternative, and V_j is the variance of r_j^2 then:

$$\chi_T^2 = \sigma_e^{-2}\, d'd + \sum_{j=1}^{s} r_j^2 / V_j \tag{12}$$

Following the generalized principle of least squares, the partials of $\chi_T^2/2$ are

$$\frac{\partial \chi_T^2/2}{\partial \theta^*} = \sigma_e^{-2} \{(P' \Gamma P\theta^*) - P' \Gamma u\} = 0 \tag{13}$$

where Γ, α_j are defined in terms of the unit vector I and the factors d, r_j, and V_j^{-1} as follows. Let

$$\alpha_j = \frac{2 r_j V_j^{-1}}{(d)'(d)/\sigma_e^2 - 2\sum_{i=1}^{s} (r_i)^2 V_i^{-1}}$$

thus

$$\Gamma = I + \sum_{i=1}^{s} \alpha_i r_i V_i^{-1} \tag{14}$$

Solving for θ^*, we find

$$\theta^* = [P' \Gamma P]^{-1} P' \Gamma u. \tag{15}$$

Since Γ depends on the values of r_j and d'd, which again depends on θ^*, the values of θ^* must be determined iteratively, with each iteration being used to determine r_j and d'd until the values converge.

IV. LEAST CHI SQUARE FOR FUNCTIONS WHICH ARE NOT LINEAR IN THE PARAMETERS. In the previous paper [1] the expression for a new estimate of the parameters has been derived for "the least chi square" procedure. That derivation will be understood by the present notation as follows:

$$\text{Let } y_i^* = y(x_1, \theta^*), y(x_2, \theta^*) \ldots y(x_n, \theta^*) \tag{16}$$

and let $u_i^* = u_i - y_i^*$.

Define the matrix P* as the matrix whose ith row is

$$\frac{\partial y_i^*}{\partial \theta_1}, \frac{\partial y_i^*}{\partial \theta_2} \cdots \frac{\partial y_i^*}{\partial \theta_r}. \tag{17}$$

Let $\{d^*\} = P^* [\delta\theta^*] - u^*$. (18)

From this it is clear that d^*, P^*, $\delta\theta^*$, and u^* may be substituted for d, P, θ^*, and u in the formula for θ^* so that

$$[\delta\theta^*] = [P^{*\prime} \Gamma P^*]^{-1} P^{*\prime} \Gamma u^*. \quad (19)$$

V. EXPLICIT EXPRESSION FOR POLYNOMIAL LEAST-CHI SQUARE.

Equation (11) can be explicitely expressed in terms of x_i, u_i, and α_i if $p_i(x)$ are polynomials. For computing purposes this may be desirable since the various "moments" can be evaluated from the data (u_i) and from the values of the independent variable x_i in several ways.

To calculate the matrix elements explicitely, let the value of Γ be $I + 2 \Sigma \alpha_j V_j^{-1}$ as in (15). From this expression the matrix elements of the equation

$$P' \Gamma u = P' \Gamma P \theta^*$$

are calculated and the results are given in Figure (1). (Note that in Fig (1) y_i is used as the vector of the observed data instead of u_i as previously.) The matrix terms include the ordinary least square terms plus the added terms as may be seen by inspection of each term. The added terms can be distinguished from the ordinary terms by the fact that each of the added terms are proportional to α_p. The calculation of the "moments" can be done in a variety of ways. Assuming that x_i are equally spaced integers, two different approaches have been used to program a Texas Instrument programmable computer (SR-52). These were:

(a) calculation and storage of all the "moments," calculation of x_T^2 and α_p from assumed values of θ^* followed by calculation of the matrix elements, and concluding with a new estimate of θ^* by a standard ordinary least square program routine such as the Texas Instrument "Trend Analysis Program." This program calculates the new values of θ_i^* by the usual techniques of solution of simultaneous linear equations.

(b) A second way is to calculate the residuals d from an initial estimate of θ_i^*. From them calculate $(d)'(d)$ and $(d)' V_j^{-1} (d)$. From these two, x_T^2 and α_p are calculated; followed by the matrix elements of the trend analysis program and then the values of θ^*.

VI. EXAMPLES. The first to be discussed uses the data on "national paperboard production per quarter" given by Butler, Kanesh, and Platt[7]. This case illustrates the situation where serial correlation due to seasonal effects is present, and offers a comparison between ordinary least squares, and least chi square. The second case uses data on the gross national product (8), in a case where the "eyeball test" indicates that a linear least squares is not adequate. The purpose of the study of this case is to provide a case where a priori one would not expect a good fit.

In all cases, 30 data points were used. The variance of the autocorrelation squared was taken as approximately $(n-4)^{-1}$, and the expected value for $\sqrt{2\chi_T^2}$ was assumed to be $[2(n+s-q)]^{\frac{1}{2}}$ where n is 30, s is 3, and q is 2. The validity of this formula as compared with alternates such as one where the degrees of freedom are n+s-2q is not important for these cases.

Table 1 shows the results of the calculation. For each case, as designated under the "DATA SOURCE" column, values were estimated for the variance of the measurement error. Under "initial" and "final" columns are given the estimates of θ_0, θ_1, χ_1^2, χ_2^2 and χ_T^2. Using the final values of $2\chi_T^2$, an estimate of its deviation (Δ) in multiples of the standard deviation from the expected value $E(\sqrt{2\chi_T^2})$ is obtained.

The first case of Gross National Product (fig. 2) used a straight line fitted by eye to the data. The second case used ordinary least squares as the initial estimate. The ordinary least squares gave the same final estimate of the parameters after one iteration as did the initial "eye ball" fit did after two iterations. (There was no change between the second and the third.) The amount of calculation would be somewhat less with the ordinary least squares as the initial point. The eyeball fit was used to check the ability of the program to converge when given an initial condition which was not the "best" estimate.

The third case of the GNP used a value of the estimate of the measurement variance of the GNP as four times that initially estimated. The same initial "eyeball" estimate was used as before and a rapid iteration to nearly the same final values of the parameter resulted. The large value of χ_1^2 nearly always dominated the value of χ_2^2.

The initial estimate on the "Paperboard Production" was taken from the result given by the authors using many more data points. This estimate was: θ_0 is 3671.8 and θ_1 is 74.12. A change of variables was made for convenience as follows:

$$y' = .2y - 760 \qquad (20)$$

Using this value as "normalized" production the initial and final values are given in table 1. In terms of the original parameters the estimates of θ_0, and θ_1 are 3712.5 and 72.25 respectively for case A. For case B they are 3715.0 and 72.15.

In case A, the initial estimate of the parameters was changed by the iteration so that the least squares error became smaller and the sum of the squares of the autocorrelation coefficients became larger. The final parameters of case A were used as the initial estimate for case B, but the estimated variance of measurement was increased by a factor of ten. The iteration procedure produced a change in the final value such that the sum of the squares (χ_1^2) was slightly increased, but the autocorrelation decreased. This is the only case studied where "Δ" is less than one standard error.

The reasons for the large values of Δ, follow for each case: For the GNP cases, the linear model is obviously insufficient to fit the data. Making allowance for a larger estimate of the measurement error does not compensate for the correlation of the residuals. We conclude: the GNP case does not satisfactorily fit a linear model as assumed.

For the Paperboard Production cases the "measurement" variance is larger than 100 but probably less than 1,000. Because there has been no attempt in the present study to adjust for "seasonal" fluctuations which may be real, the "seasonal" fluctuation then represents an additional (and correlated) error in each quarterly estimate. Further analysis will be done for this case when a computer of larger storage capacity than the one used in this study is available.

To investigate the possibility that outlier rejection would be assisted by this technique the 25th data point was chosen by Monte-Carlo techniques and a $-3\sigma_e$ deviation from the original fitted line was introduced. Two cases were calculated using this set of data: in the first case ordinary least squares was used to initiate the calculations; in the second case an initial estimate of the values of the parameters near to the fitted line of the unmodified set of data was used.

The final parameter estimates agreed in both cases and the values of both χ_1^2 and χ_2^2 greatly increased. The result is that "Δ" became greater than 2.3 as compared to the previous result of 0.68. (The variance of $\sqrt{2\chi_T^2}$ is unity.)

Thus we find this test sensitive to a single outlier and indicates that further study should be done of this technique.

Questions such as the relation to the ARMA technique (9) have not yet been investigated.

VII. SUMMARY. It was observed that the fit criteria, χ_T^2, was improved in each case from the ordinary least square value by the iteration procedure. In this process the chi square of the autocorrelation coefficients was always reduced from that which occurred at minimum variance of the errors at the expense of permitting a slight increase in the variance of the errors.

Based on this result and on the theory of the tests, least chi square gives an improved estimate of the parameters as compared to ordinary least squares.

The convergence was rapid. The number of iterations required to converge was approximately three. It is yet to be demonstrated that an "eyeball" initial fit might reduce the number of iterations required but it is believed likely.

When performing experiments involving measurements, the measurement error should be independently observed so that data will be available to apply the least-chi square test if appropriate.

VIII. ACKNOWLEDGMENT. Professor G. E. P. Box and Dr. J. R. Moore each made significant suggestions to clarify the historical background and goal of this paper.

IX. REFERENCES.

1. Moore, R. L., Proc. 1975 Army Numerical Analysis Conference, ARO Report 75-3.

2. Moore, R. L., Found. of Phys. 7, 129-135, 1977 and to be published.

3. J. R. MacDonald, Rev. Mod. Phys. 41, 316, 1969.

4. Anscombe, F. J., and J. W. Turkey, Technometrics, 5, 141, 1963.

5. Aitken, A. C., Proc. Roy. Soc. Edinb. A 55, 42-47 (1934), and R. L. Plackett, Principles of Regression Analysis, Osford, Claredon Press, 1960.

6. Box, G. E. P. and Pierce, D. A., J. Amer. Stat. Assoc. 64, 1509, 1970.

7. Butler, W. F., Kavesh, R. A., and Platt, R. B., Methods and Techniques of Business Forecasting, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.

8. U.S. Dept of Commerce, The National Income and Product Account of the United States, 1929-74, Statistical Tables, Supt of Documents, GPO, Washington, and U.S. Dept of Commerce, Survey of Current Business, 56, 1, July 1976.

9. Box, G. E. P. and Jenkins, G. M., Time Series Analysis; Forecasting and Control, San Francisco, Holden-Day 1970.

Table 1

Summary of Calculations

INITIAL

Data Source		σ_e^2	θ_0^*	θ_1	χ_1^2	χ_2^2	χ_T^2	No. of Iterations
Gross National Product	(1)	23.31	50.0	6.3	1329.02	42.25	1371.27	3
ORD LS	(2)	23.31	57.03	6.65	—	—	.370	0
	(3)	93.28	50.0	6.3	332.36	42.38	374.74	1
Paperboard Production	(A)	100	-25.64	14.82	367.64	.0514	388.16	2
	(B)	1000	-17.50	14.45	36.277	.3604	36.637	3
Same with outlier								
ORD LS	(A)	1000	-12.37	13.78	46.7292	5.829	52.5584	1
	(B)	1000	-17.47	14.45	48.6209	4.86	53.484	2

FINAL

Data Source		θ_0^*	θ_1^*	χ_1^2	χ_2^2	χ_T^2	$\sqrt{2\chi_T^2}$	$E(\sqrt{2\chi_T^2})$	Δ
Gross National Product	(1)	56.89	6.66	1130.64	38.22	1168.86	48.350	7.87	40.48
ORD LS	(2)	56.89	6.66						
	(3)	56.5	6.68	282.8	38.22	321.02	25.34	7.87	17.47
Paperboard Production	(A)	-17.50	14.45	362.77	.364	363.14	26.95	7.87	19.08
	(B)	-17.01	14.43	36.28	.2907	36.57	8.55	7.87	0.68
Same with outlier									
ORD LS	(A)	-12.84	13.87	46.729	5.309	52.038	10.20	7.87	2.33
	(B)	-12.83	13.86	46.768	5.8030	52.57	10.254	7.87	2.384

$$\sum_{1}^{n} y_i + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (y_i + y_{i+p}) = \theta_0 (n + \sum_{p=1}^{s} 2(n-p)\alpha_p) +$$

$$\sum_{i=1}^{n} x_i y_i + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (x_{i+p} y_i + x_i y_{i+p}) = \theta_0 \left(\begin{array}{l} \text{SAME AS COEFFICIENT} \\ \text{OF } \theta_1 \text{ IN ROW 1} \end{array} \right) +$$

$$\sum_{i=1}^{n} x_i^2 y_i + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (x_{i+p}^2 y_i + x_i^2 y_{i+p}) = \theta_0 \left(\begin{array}{l} \text{SAME AS COEFFICIENT} \\ \text{OF } \theta_2 \text{ IN ROW 1} \end{array} \right) +$$

FIG.1A LEFT SIDE OF REGRESSION EQUATION

$$\theta_1\left(\sum_{i=1}^{n} x_i + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (x_i + x_{i+p})\right) + \theta_2\left(\sum_{i=1}^{n} x_i^2 + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (x_i^2 + x_{i+p}^2)\right) + \cdots$$

$$\theta_1\left(\sum_{1}^{n} x_i^2 + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} 2\, x_i\, x_{i+p}\right) + \theta_2\left(\sum_{i=1}^{n} x_i^3 + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (x_i\, x_{i+p}^2 + x_{i+p}\, x_i^2)\right) + \cdots$$

$$\theta_1(\text{SAME AS COEFFICIENT OF } \theta_2 \text{ IN ROW 2}) + \theta_2\left(\sum x_i^4 + \sum_{p=1}^{s} \alpha_p \sum_{i=1}^{n-p} (2 x_i^2\, x_{i+p}^2)\right) + \cdots$$

FIG.1B RIGHT SIDE OF REGRESSION EQUATION

FIGURE 2

GROSS NATIONAL PRODUCT
(BY QUARTER, JAN 68 TO JAN 76)

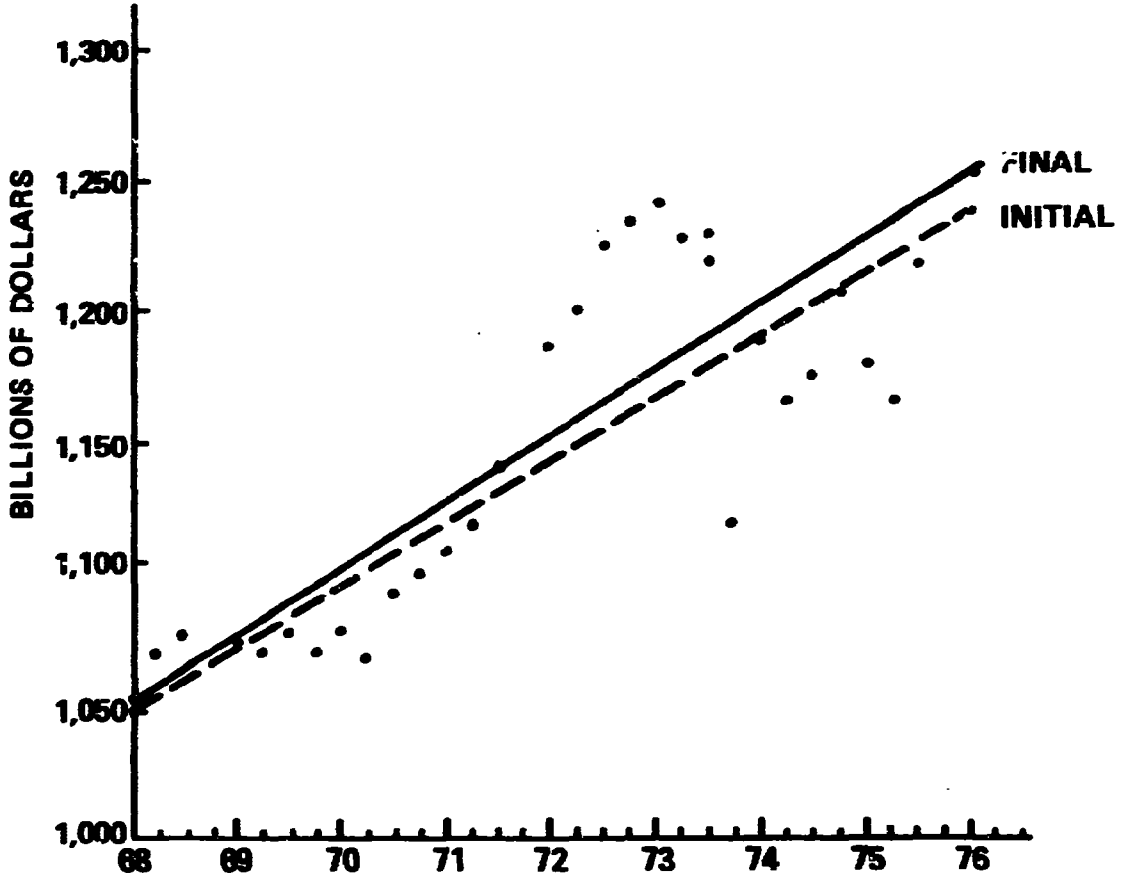
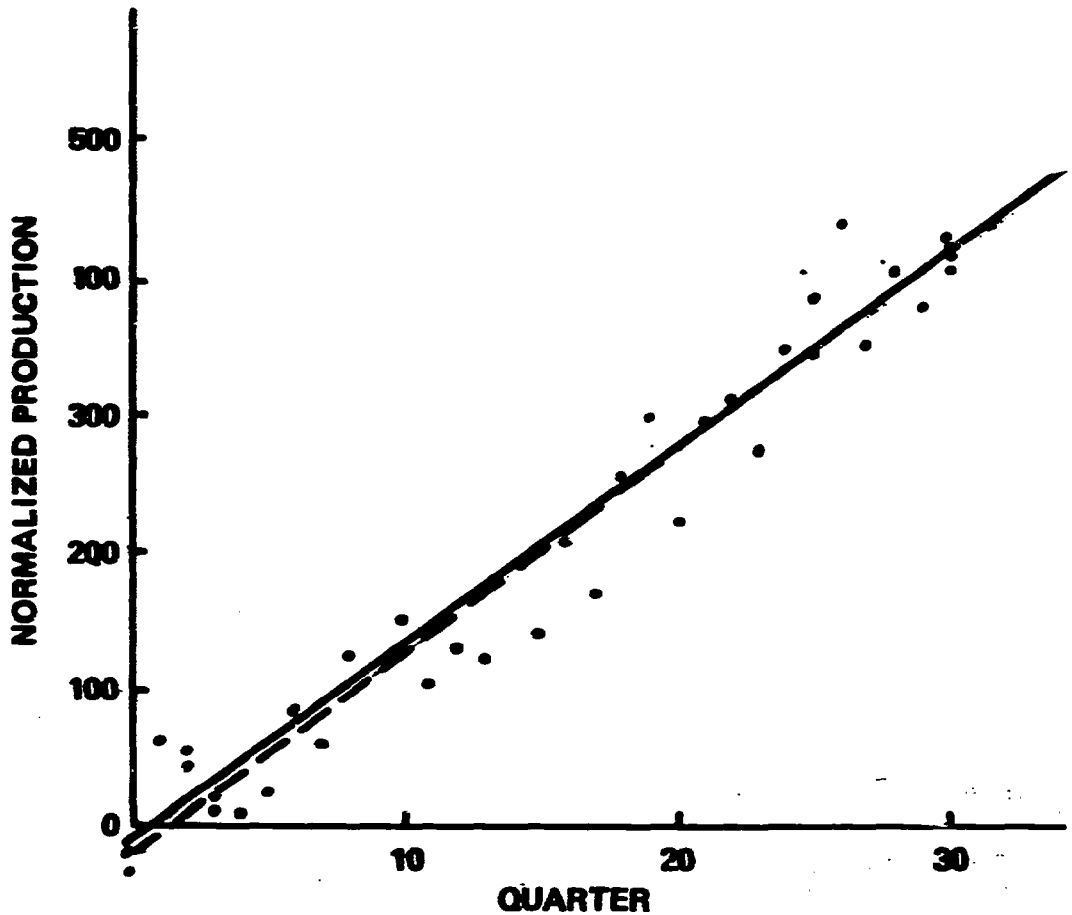


FIGURE 3

PAPERBOARD PRODUCTION
(30 QUARTERS BEGINNING JAN 60)



SIMPLIFIED CONSTRUCTION OF BASIS FUNCTIONS FOR POLYNOMIAL SPLINES

J. J. Heimbold
MARK Resources, Inc., Marina del Rey, California

A simple, straightforward procedure is presented for generating polynomials over a set of contiguous intervals. The polynomials can be constructed to be continuous or to have an arbitrary number of derivatives continuous across the interval boundaries (knots). The constructed functions are ordinary polynomial splines of given degree with any specified number of derivatives continuous across the boundaries.

A minimum mean-square error criterion in fitting the spline polynomials to a set of data points requires solving a set of linear equations. In actual applications it is efficient to express the polynomial splines as a set of basis functions, which simplifies the solution of the linear equations. A set of spline basis functions is presented which does simplify the solution to the minimum mean-square fit. The functions are created in such a way that many pairs of basis functions are mutually orthogonal. In addition they are ordered in a way that results in a banded matrix in the set of linear equations. Both of these properties lead to a numerically simple solution and a reasonably small amount of computer storage.

MOTIVATION

The need to construct splines grew out of a requirement to obtain trajectory estimates from noisy radar data. It was known that some of the trajectories could be modeled by fourth to sixth degree polynomials over short time intervals, and it was desired that the range, velocity and sometimes the acceleration or higher-order range derivatives be continuous across interval boundaries.

As a result of the need for the trajectory estimates, a technique was developed for constructing spline basis functions for polynomials of arbitrary degree with an arbitrary number of constrained derivatives at the knots.

The motivation for deriving the spline basis functions came from the need to quickly implement a spline program. A survey of the spline literature found it to be either limited to second or third degree polynomials, or to be unreadable without a specialized background.

CONSTRUCTION OF POLYNOMIAL SPLINES

It can be shown that a polynomial spline of degree D in $C^{P-1}(x_1,\infty)$ over the set of strictly increasing knots $\{x_1,x_2,\ldots,x_N\}$ can be written as

$$Q(x) = \sum_{j=0}^{D} r_{1j}(x-x_1)_+^j + \sum_{j=P}^{D} r_{2j}(x-x_2)_+^j + \ldots + \sum_{j=P}^{D} r_{N-1,j}(x-x_{N-1})_+^j$$

where

$$(x-x_i)_+ \triangleq \begin{cases} x-x_i & x \geq x_i \\ 0 & x < x_i \end{cases} .$$

Expressing the splines in this form yields a concise mathematical formulation of the splines. The first summation term is a polynomial of degree D on (x_1,∞) and is in $C^{P-1}(x_1,\infty)$. The rest of the summation terms are in $C^{P-1}(-\infty,\infty)$, and hence Q(x) is in $C^{P-1}(x_1,\infty)$.

The functions $(x-x_i)_+^j$ are basis functions for Q(x), and are not necessarily mutually orthogonal for any two basis functions. Consequently, a direct

computation of a minimum mean-square fit of Q(x) to noisy data will require matrix storage and inversion for a square matrix with dimension (N-1)(D+1-P)+P. A change of basis functions can reduce the matrix storage requirements if the basis functions are chosen such that many pairs of the functions are mutually orthogonal.

SPLINE BASIS FUNCTIONS

The basis functions which lead to a banded matrix are

$$ {}_jB_i(x) = \begin{cases} \displaystyle\sum_{k=1}^{P+1} {}_jC_{ik}(x-x_{k+k-1})_+^{P+j-1} & x_i \le x < x_{i+P+1} \\ 0 & \text{otherwise} \end{cases} $$

where

$$ {}_jC_{ik}(x_{i+P+1}-x_{i+k-1})^{j-1} = \frac{\displaystyle\prod_{\substack{1\le m\le P+1 \\ m\neq k}} (x_{i+P+1}-x_{i+m-1})}{\displaystyle\prod_{\substack{1\le m\le P+1 \\ m\neq k}} (x_{i+m-1}-x_{i+k-1})} , $$

for

$j=1,2,\ldots,D+1-P$,

$k=1,2,\ldots,P+1$, and

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

These basis functions span all terms of the form $(x-x_i)_+^j$; j=P,...,D; i=1,...,N-1. The other terms in the polynomial Q(x), viz., $(x-x_1)^j$, j=0,1,...,P-1, are spanned by creating P knots $x_{-(P-1)},\ldots,x_{-1},x_0$ which are strictly increasing with $x_0 < x_1$. Then the set of basis functions $\{{}_1B_i(x)\}$, i=-P+1,-P+2,...,0,1, is a set of P+1 linearly independent polynomials of degree P on the interval $[x_1,\infty]$, and hence

$$\sum_{j=0}^{P} r_{ij}\,(x-x_1)_+^j$$

can be formed as a linear combination of these basis functions.

These basis functions have the property that

$${}_kB_i(x)\,{}_\ell B_j(x) = 0 \text{ for } |i-j| \geq P+1 .$$

They can be ordered as:

$${}_1B_{-P+1},\ {}_1B_{-P+2},\ \cdots,\ {}_1B_0 ,$$

$${}_1B_1,\ {}_2B_1,\ \cdots,\ {}_{D+1-P}B_1 ,$$

$$\vdots$$

$${}_1B_{N-1},\ {}_2B_{N-1},\ \cdots,\ {}_{D+1-P}B_{N-1} .$$

With this ordering, the matrix of dot products of the basis functions is banded with bandwidth (P+1)(D-P+1), P=0,1,...,D.

VALT PARAMETER IDENTIFICATION FLIGHT TEST

Robert L. Tomaine*, Wayne H. Bryant,** and Ward F. Hodge**

NASA-Langley Research Center
Hampton, Virginia 23665

ABSTRACT

The Langley Research Center is engaged in a research program to develop the technology to maximize the capability of helicopter operation in confined areas. The program, VALT (VTOL Approach and Landing Technology), uses an integrated approach involving the helicopter, avionics system, control system, displays, and the pilot. An important task in the study is to develop an accurate model of the helicopter system for flight control design and simulation studies. A flight test designed utilizing the VALT approach profile was performed at the NASA Wallops Island test facility to obtain data for verifying existing mathematical models through use of parameter identification techniques. Briefly, parameter identification as applied to flight vehicles consists of identifying the aerodynamic coefficients of the vehicle equations of motion utilizing the measured vehicle states and accelerations resulting from measured control inputs. Theoretically, these coefficients can be determined very accurately; however, in actual applications many problems and limitations are encountered. In addition, the research vehicle used (CH-47) and the VALT flight regime introduced problems specific to this application. The unique facilities utilized to minimize these problems for the CH-47 parameter identification flight test included the CH-47 fly-by-wire control system and on-board computer, the Wallops Test Center radar tracking system, the Langley Research Center mobile research Aircraft Ground Station (RAGS) and Piloted Aircraft Data System (PADS), and the CH-47 Sperry flight director display.

Data runs were performed to include test points along the entire VALT approach trajectory, including straight and level flight, straight descending and ascending flight, and spiral descents. Complete data sets were measured at 40 sps on PCM recorders and stored on board to include attitudes, velocities, angular rates, linear accelerations, pilot stick positions, actuator positions, SAS positions, rotor RPM, and other pertinent information.

In addition to describing the details of this flight program, preliminary results of parameter identification processing utilizing advanced statistical methods are presented.

---

*Structures Laboratory, US Army Research and Technology Laboratories (AVRADCOM)
**National Aeronautics and Space Administration, Langley Research Center

INTRODUCTION

V A L T is an acronym for VTOL Approach and Landing Technology (Ref. 1). It is a comprehensive program including flight management, guidance and control, and display technology with the ultimate goal of the development of avionics technology for optimum VTOL short haul transportation in the 1980's time regime. One important task of the VALT program is to develop an accurate model of the VALT research vehicle, which is required for guidance and control system design. This paper is concerned with the approach taken to determine this model.

The method of obtaining an accurate model of the VALT research vehicle is verification of prior developed analytical models by processing selected flight maneuver data with advanced parameter identification algorithms.

The VALT flight regime consists of cruise, transition and hover flight conditions. Anticipated VALT trajectories include straight and level flight, straight ascending and descending flight, and spiral descending flight. A comprehensive flight test program was conducted at the NASA Wallops Flight Test Center to obtain data for all of the flight conditions anticipated for the VALT trajectories.

Parameter identification of flight vehicles consists of disturbing the test aircraft with a known control input to produce a response in the vehicle states which are measured as a function of time (see Fig. 1). Given a form, the vehicle model (plant) and the measured states, the algorithms compute the coefficients (stability and control derivatives) of the model. The equation set governing the identification process is as follows:

$$\dot{\bar{X}} = A(p)\bar{X} + B(p)\bar{U}$$

where $\bar{X}$ refers to the vehicle state vector, $\bar{U}$ is the control input vector, and A and B are the stability and control matrices which compose the assumed plant. The plant is the equations of motion of the vehicle.

The general identification problem is complicated by the presence of two primary error sources. First of all, the measurements of the states contain noise due to the vehicle vibration, instrument limitations, and data processing. This results in the equation $\bar{Z} = \bar{X} + \bar{V}$, where the measurement vector $\bar{Z}$ is a combination of the actual state vector $\bar{X}$ and a measurement noise vector $\bar{V}$. In addition, some of the response of the vehicle may be due to external disturbances such as wind gusts, and the assumed model may not be representative of the actual vehicle. These error sources in combination are referred to as process noise. Therefore, the problem is to determine the components of the A and B matrices of the assumed plant in the presence of both measurement and process noise.

In practice, several specific problems occur in parameter identification; and in this study additional problems associated with the VALT flight regime and the VALT research vehicle are encountered. The general problems include

the presence of winds, which as discussed earlier introduces process noise. Additional problems result from the form of the vehicle equations of motion (plant) chosen to represent the vehicle. These equations are linear 6 degree-of-freedom small perturbation equations chosen for compatibility with control system design procedures and limitations on existing parameter identification algorithms. The equations require obtaining an accurate and steady vehicle trim and linear response in the vehicle state variables.

The VALT flight regime introduces the problem of determining accurate vehicle velocity measurements at low airspeeds where conventional pitot-static instruments are useless. The vehicle itself introduces furtner difficulty in that it has unstable modes and its rotors introduce high frequency noise in the measurement system and the possibility of rotor/fuselage coupling. Lastly, flight testing introduces the need to evaluate on board and at the test location the accuracy and quality of the data being acquired. The next section will discuss how the test was designed to minimize the aforementioned problems, and to obtain an accurate and appropriate data set. To provide a background for discussing how these problems were handled and the testing approach taken in these flight tests, the facilities utilized are discribed first.

DESCRIPTION OF FLIGHT TEST FACILITIES

The parameter identification flight tests were carried out at NASA's Wallops Flight Center, Wallops Island, Virginia in March of 1977. The Wallops facilities crucial to these flights were the Aeronautical Radar Research Complex (ARRC radar); the Transponder Data System (TDS); wind data measurement equipment, including a wind measurement tower and weather balloons; and the Research Aircraft Ground System (RAGS). The current VALT research vehicle is a Boeing-Vertol CH-47 transport helicopter from NASA's Langley Research Center. Each of these systems are briefly described below.

The ARRC radar facility consists of an FPS-16 radar used in conjunction with a laser tracking radar to provide vehicle position data accurate to one foot. This information is processed by a minicomputer within the facility to provide highly accurate data in a Cartesian coordinate system aligned with the runway chosen for each day's flights. The data is then telemetered to the vehicle using the Transponder Data System (TDS). The TDS is a data link that uses the time between radar pulses to send pulse position modulated (ppm) digital data to and from the vehicle on the same frequencies as the ground radar (uplink) and the airborne transponder (downlink). The data transmission rate is one ten-bit digital word on both uplink and downlink per pulse of the radar, and for these tests was configured to give approximately 34 complete position updates per second to the on-board digital computer.

The ARRC radar facility was also used to track weather balloons released at regular time intervals to obtain wind velocity and direction information at 100-foot intervals from 200 feet to 2,500 feet. A 100-foot weather data tower was used to obtain low altitude and surface wind data.

The RAGS is a mobile station with a telemetry link to the aircraft measurement system as well as magnetic tape playback equipment. It provides the capability for both real-time data display of selected parameters as well as a post-flight quick look capability at all of the measured parameters.

As previously mentioned, the research vehicle is a Boeing-Vertol CH-47 tandem rotor transport helicopter equipped with a fly-by-wire control system. The cockpit has both a standard mechanical control stick arrangement (the safety pilot) as well as an electrical stick (the research pilot). The mechanical control arrangement controls the position of the vehicle's actuators. The electrical stick serves as input to the computing system, which can manipulate the signals in a variety of ways through programming of the Sperry 1819A digital flight computer. Outputs from the Sperry 1819A are converted to analog signals used as inputs to electrohydraulic actuators. The outputs from these actuators are then used to drive the standard mechanical control stick arrangement through a clutch arrangement, which allows rapid disconnection of the computing system in the event a potentially dangerous control input to the vehicle is generated.

The Sperry 1819A flight computer is a general purpose, fixed-point 18-bit stored program integer machine with 16,384 words of ferrite core memory for program and data storage. This computer communicates through a variety of interfaces to the research pilot's control sticks, motion sensors, the control system actuators, the transponder data system, and its own control panels, which allow data examination and modification.

Measurement, recording, and telemetering of spatial, control, and discrete variables is handled by the Piloted Aircraft Data System (PADS), a pulse code modulated (pcm) recording system. Sensor outputs are first routed through buffer amplifiers and then sent to the computing system, the on-board recording system, and to the telemetry system.

TESTING APPROACH

The first category of flight testing problems; accurate knowledge of the winds, precise aircraft trim, and low-speed air data measurements were handled through the combined use of the ARRC radar facility, the TDS, the on-board digital computer, and an electromechanical flight director. As described previously, wind data were obtained at periodic time intervals by releasing and radar tracking a weather balloon. The subsequent reduction of the radar track provided wind velocity and direction at regular altitude intervals.

Accurate low-speed air data measurements were obtained through processing of radar derived position data (telemetered to the vehicle using the TDS) with on-board acceleration measurements in a complementary filter implemented in the Sperry 1819A digital computer to obtain an estimate of ground speed. To this ground speed estimate, the current wind velocity was

added so that when flying directly into the wind an accurate estimate of airspeed was obtained. This airspeed determination system was used for all flights and covered the range from hover to 80 knots.

Precise trim conditions were established by using an electromechanical flight director, driven by the flight computer, indicating deviation from desired trim. Figure 2 is a photograph of the research pilot's cockpit and shows, in addition to the flight director, other standard aircraft instruments. Starting at the top on the left-hand side, is an airspeed indicator, a torque meter, and a flight-altitude indicator. At the right, starting at the top is an altimeter, a vertical-speed indicator, and a magnetic compass. The CRT shown just below center is used for display evaluation, but was not used in these flights. The flight director horizontal pointer was used to indicate error from desired airspeed; the vertical bar, error from desired sideslip; the doughnut (at the left side), error from desired descent rate; and the localizer (at the bottom), error from desired rate of turn. The pilot's task to obtain precise trim was to simultaneously center the four flight director pointers. To accomplish this task, the pilot first would obtain an approximate trim using the standard aircraft instruments, and then focus his attention to centering the four flight director pointers. Gains and damping for each flight director pointer were individually selectable through the entry of appropriate constants in the Sperry 1819A computer. This feature allowed the flight director to be "tuned" to the pilot to obtain the most satisfactory overall performance.

The second category of problems were all handled through the combination of control input design, its implementation in the on-board digital computer, and the electrically-driven control surface actuators. The basic control input design was carried out under contract to NASA's Langley Research Center by Systems Control, Inc. of Palo Alto, California. These designs were based on exciting the Stability Augmentation System-on closed-loop modes of an analytic model of the CH-47, and consisted of a high and low frequency sinusoid. Figure 3 represents a typical control input generated by the flight computer for the pitch axis and shows the two components of the designed control input. To strike a balance between adequate modal excitation and the linearity constraints on the vehicle response imposed by the small perturbation model used in the parameter identification sequence, scaling was provided in the computer implementation of the automatic control inputs. Repeatability and accurate knowledge of the control input was inherent in the digital computer implementation. To account for a known speed instability at higher airspeeds, a longitudinal stabilization input (also implemented in the digital computer) was added to the programmed input to maintain the resultant vehicle response within the small perturbation equations' linearity constraints.

Two major systems were primarily used to address the third category of problems, real-time data evaluation. The Piloted Aircraft Data System (PADS) on board the vehicle was used to both record a wide selection of measurements on magnetic tape and also telemeter a subset of these measurements to the Research Aircraft Ground Station (RAGS) for subsequent real-time display on multi-channel chart recorders. In the RAGS, transparent overlays of the expected measurements, prepared earlier using the CH-47 analytic model (Ref. 2), were then compared with the real-time data for use

in evaluating the success of a particular run. This information was then relayed to the research project engineer on board the helicopter for his use in determining the next flight test point. After each flight, the on-board tape was used in the RAGS to create additional stripchart recording of measurements that proved useful in planning subsequent flights.

TEST POINT SEQUENCE

Figure 4 is a pictorial of NASA's Wallops Flight Center which illustrates the systems used by this series of flight tests. Each of these systems has been described earlier. This figure is useful in understanding the sequence of events in obtaining flight test points.

Since the airspeed estimator required the vehicle to be flown into the wind for all test points, the test sequence naturally divided into a downwind leg and an upwind leg. On the downwind leg, a weather balloon is released and tracked by the radar to obtain the requisite wind data. This data is then relayed via radio to the research project engineer on board the helicopter, who then decides what test points will be flown, and establishes the constant wind velocity to be entered into the digital computer for airspeed estimator calculation.

On the upwind leg, the research project engineer first selects the test point (based on wind magnitude), then provides the computer operator with his reference number and the desired magnitude (in per cent) of the computer-generated control input. When the computer operator enters these values, the appropriate trim values are obtained from a look-up table stored in the computer, and trim error signals are sent to the electromechanical flight director. The research pilot then obtains the desired trim using conventional aircraft instruments to obtain an approximate trim and the flight director to obtain a more precise trim. When an accurate trim is obtained, the evaluation pilot's electric stick inputs are disconnected by the computer and programmed control inputs are substituted. At the end of an individual data run (approximately 20 seconds), the computer system is disengaged from the basic vehicle and the safety pilot regains control of the helicopter setting up for the next data run. While these activities are underway, a comparison of the real-time data collected with the appropriate analytic model overlay provides valuable insight into the success of the test point. The results of this evaluation are then relayed to the research project engineer on board to aid in his selection of the next data point. Typically, several data points were obtained during each upwind leg, and wind information was updated during each downwind leg.

PRELIMINARY RESULTS

The post-flight processing consists of converting the PADS data tapes to engineering units and selecting the best data sets for each flight

condition. Selection is based upon attained trims and state variable responses. For data from helicopters, better identification results have been obtained from filtered flight data measurements. For this study, the data has been filtered by a zero-phase-shift Graham digital filter (Ref. 3) with cutoff and termination frequencies chosen above any expected rigid body modes and below frequencies associated with the rotor system. This step reduces the noise content of the measured state variables appreciably and provides only rigid body vehicle responses. The data is further processed using a Kalman filter/estimator based upon the aircraft kinematic equations. The Kalman filter estimates and removes the measurement biases, and provides estimates of the vehicle states based on measured attitudes, rates, and accelerations.

After data reduction and prefiltering, the data sets are ready for parameter identification processing. This data will ultimately be processed using two differing advanced algorithms capable of handling both measurement and process noise, and the results will be compared. An Extended Kalman Filter algorithm (Refs. 4 and 5) will be used by USARTL personnel to identify six degree-of-freedom stability and control derivatives, and a maximum likelihood algorithm (Ref. 6) will be used by NASA personnel; and selected data sets will be processed under contract to SCI (Vt.), who will also use a maximum likelihood approach.

Some preliminary results are presently available from the Extended Kalman Filter algorithm and the major derivative values identified are compared with existing analytical values in figure 5. The majority of the identified derivatives agree very well with the analytically-predicted values. These results are encouraging since the responses generated by the analytical values produced responses very close to those measured in flight. Figure 6 shows the eigenvalues (characteristic roots) for both the identified and analytical derivatives. Good agreement between analytical and identified results are shown with all the basic vehicle modes represented, including the expected unstable Dutch roll mode and speed instability. The results presented are preliminary, and many data sets remain to be processed. Final acceptance of the derivatives will be based upon a combination of tests; including comparison with analytical values and expected values based on engineering judgment, responses generated by identified derivatives (regeneration), responses generated by identified derivatives for data not used in the identification process (simulation), derivative uncertainties and convergence characteristics, and comparison of eigenvalues (roots) computed using identified derivatives with analytical results and engineering judgment.

CONCLUDING REMARKS

A specialized flight test was designed and implemented to provide data acceptable for parameter identification for an unstable rotorcraft operating in the presence of winds at flight conditions from hover through transition

to cruise. General problems in parameter identification, flight testing, and problems specific to this flight test were considered; and a unique test procedure utilizing existing facilities was performed. Preliminary data processing has resulted in identified parameters which agree well with existing analytical results.

REFERENCES

1. Kelly, J. R.; Niessen, F. R.; Thibodeaux, J. J.; Yenni, K. R.; and Garren, J. F., Jr., "Flight Investigation of Manual and Automatic VTOL Decelerating Instrument Approaches and Landings," TND-7524, 1974, NASA.

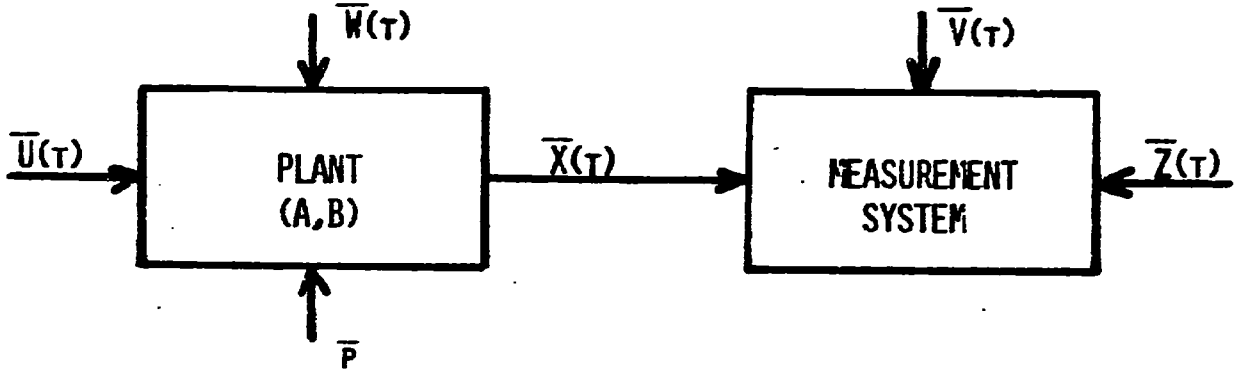
2. Ostroff, Aaron J.; Downing, David R.; and Rood, William J., "A Technique Using a Nonlinear Helicopter Model for Determining Trims and Derivatives," TND-8159, 1976, NASA.

3. Graham, R. J., "Determination and Analysis of Numerical Smoothing Weights," TR R-179, 1963, NASA.

4. Molusis, J. A., "Analytical Study to Define a Helicopter Stability Derivative Extraction Method," CR-132371, 1973, NASA.

5. Tomaine, Robert L., "Flight Data Identification of Six Degree-of-Freedom Stability and Control Derivatives of a Large "Crane" Type Helicopter," TMX-73958, 1976, NASA.

6. Hodge, W. F. and Bryant, W. H., "Monte Carlo Analysis of Inaccuracies in Estimated Aircraft Parameters Caused by Unmodeled Flight Instrumentation Errors," TND-7712, 1975, NASA.



$$\dot{\overline{X}} = A(p)\overline{X} + B(p)\overline{U} \qquad \overline{Z} = \overline{X} + \overline{V}$$

GIVEN: INPUT VECTOR $\overline{U}$ WITH: MEASUREMENT NOISE $\overline{V}$

MEASUREMENT VECTOR $\overline{Z}$ PROCESS NOISE $\overline{W}$

FIND: COMPONENTS OF $A(p)$, $B(p)$

Figure 1. Parameter Identification Schematic

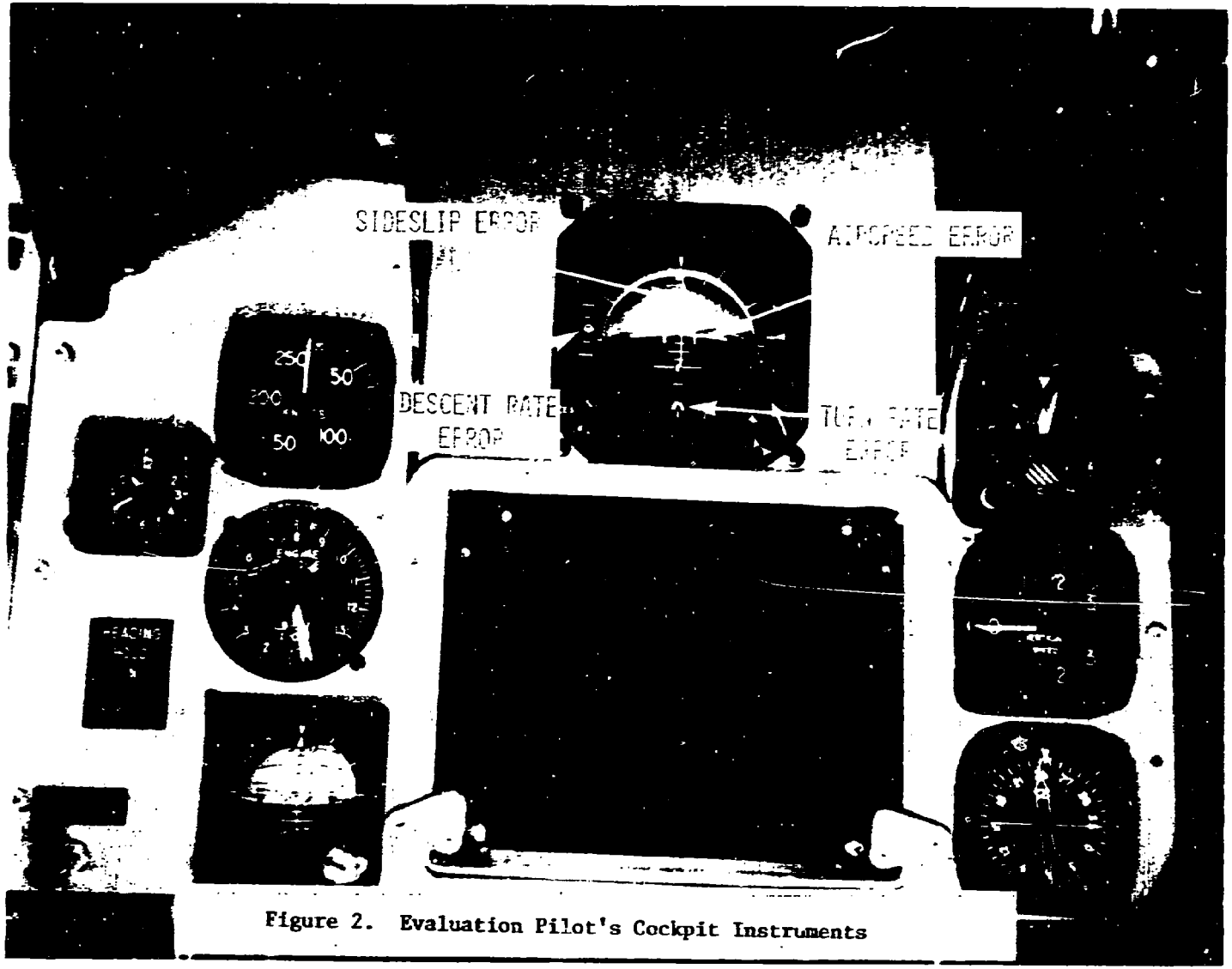


Figure 2. Evaluation Pilot's Cockpit Instruments

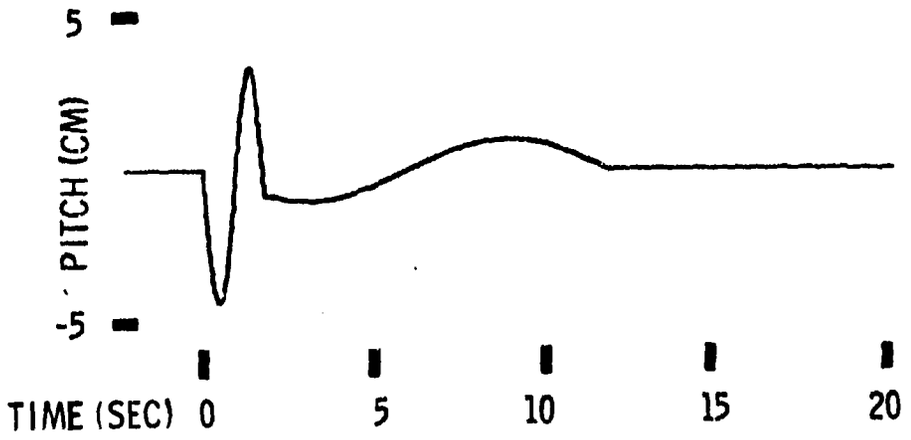


Figure 3. Typical Computer Generated Input

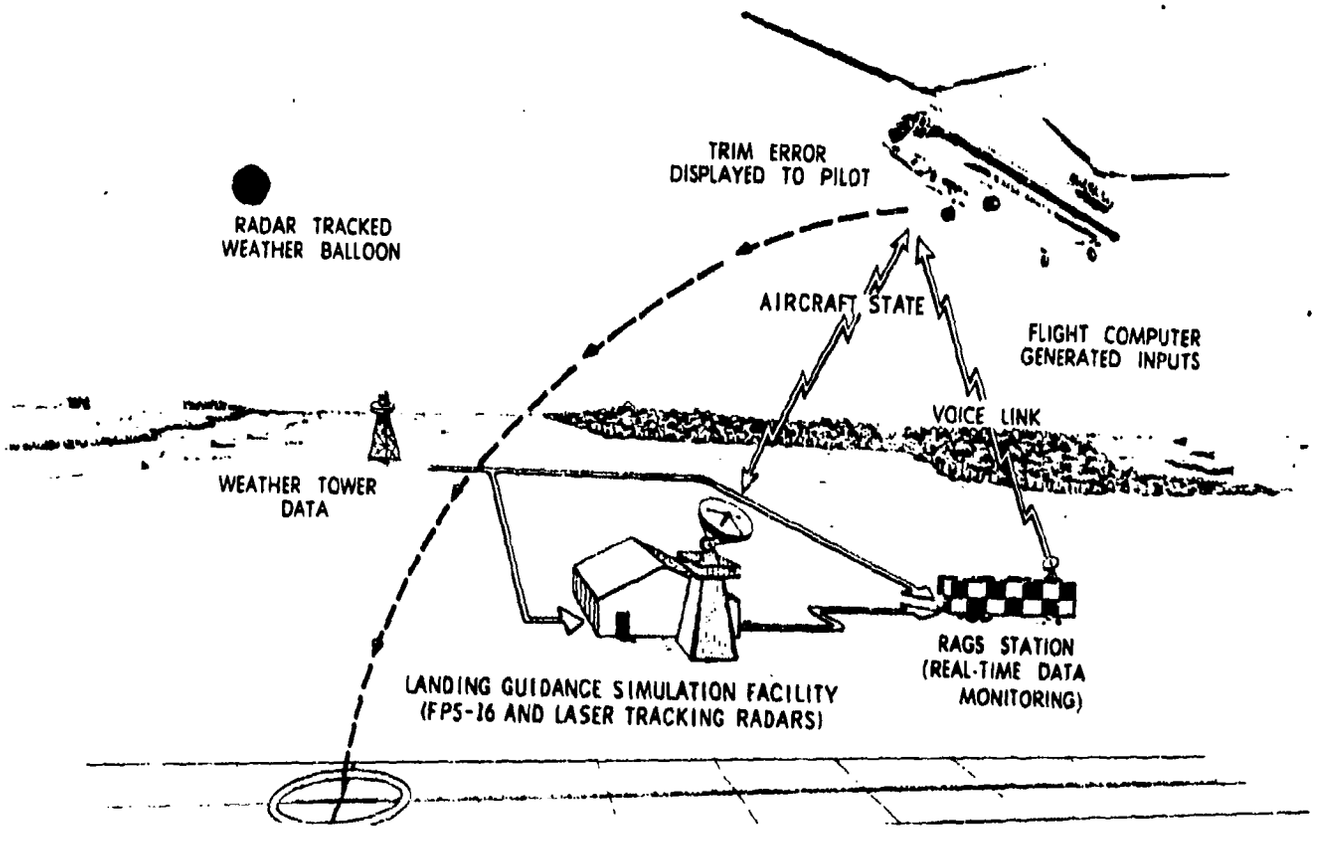


Figure 4. Stability and Control Derivative Flight Test Facilities

PRELIMINARY RESULTS

DERIVATIVE	ANALYTICAL VALUE	IDENTIFIED VALUE
X_u	-.0204	-.0208
M_u	-.0042	-.0046
Y_v	-.07	-.054
L_v	-.0055	-.0050
N_v	-.00009	-.00009
Z_w	-.551	-.213
M_w	.0176	.016
L_p	-.818	-.834
M_q	-1.68	-1.76
N_R	-.0398	-.0398

Figure 5. Derivative Comparison

PRELIMINARY RESULTS

60 KNOT LEVEL FLIGHT

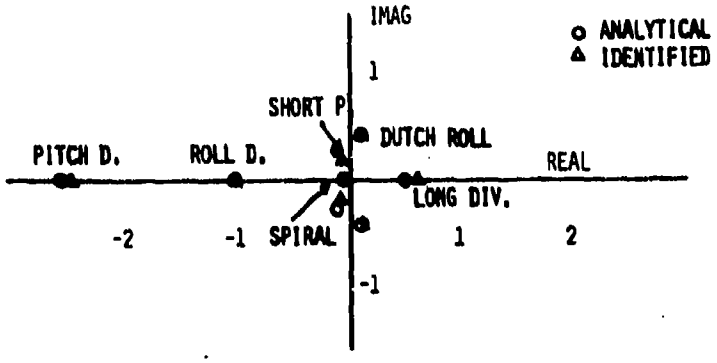


Figure 6. Eigenvalue Comparison

EXPERIMENTAL DESIGNS FOR SENSITIVITY EXPERIMENTS OF COMPUTER SIMULATION MODELS

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

ABSTRACT. Large stochastic computer simulation models usually have a large number of input variables. After model development and/or before the model is used for production runs or used in a particular study, sensitivity testing of input variables is usually required. Because of the size of the model and the intended future use of the model, the list of input variables desired to be tested is invariably long. Also, because of the absence of a priori information on the interaction of input variables, the experimental design for the sensitivity experiment must provide for the testing of main effects and first-order interactions. The application of fractional factorial designs in sensitivity testing is illustrated and their shortcomings for sensitivity testing of large computer simulation models is discussed.

1. INTRODUCTION

The US Army Concepts Analysis Agency (USACAA) is a staff support agency under the Deputy Chief of Staff for Operations and Plans (DCSOPS). The agency's mission is to conduct mid- and long-range force concept studies to establish the framework and guidance for development of doctrine, organizations, and materiel requirements for Army forces. Agency studies and analyses support Department of the Army planning and programing and provide the basis for materiel acquisition. The Agency develops, within resource constraints, the most effective force structure and weapon and/or system mix. The primary tool for the performance of the studies is computer simulation models. After computer simulation model development and/or before a model is used in a particular study, sensitivity testing of input variables is usually required. That is, if no a priori knowledge exists concerning model sensitivity, an investigation must be made of the sensitivity of selected output variables to changes in input variables. This is necessary in order to evaluate model performance or to assess the ability of models satisfying specific study requirements.

The models range from high resolution, low (division) level, to low resolution, high (theater) level models. A commonality, however, of all the models is their size and complexity. All of the simulation models are large and very complex. The number of input variables is in the hundreds and the number of input data is in the thousands.

2. PROBLEM DESCRIPTION

Statisticians at CAA are within a service support Directorate. They provide experimental design and statistical analysis support to all study Directorates within the agency. Analysts who are study team members and who have responsibility for model sensitivity testing of a particular model come to the statisticians with experimental design problems.

Invariably, the list of input variables which are desired to be investigated is in the order of 50 to 100 variables. One case involved 350 variables. Naturally, time constraints never permit a thorough investigation of all variables on the original "laundry list" of variables. Because no a priori information exists, the minimum objective of the sensitivity testing is to test and estimate main effects and first-order interactions.

The list of candidate variables for testing are those suspect of being significant. The small subset of input variables ultimately tested are those most strongly suspected of being highly significant. That is, the variables eventually tested are anticipated and expected to significantly influence model output. Past experience in model sensitivity testing has shown that, in general, most input variables ultimately tested are, in fact, significant. Moreover, most of the first-order interactions are also

significant. With study and hindsight, it is generally agreed that this is consistent with reality. That is, the simulation model does adequately portray the real world which does, in fact, consist of many interacting parts or components.

3. EXAMPLE PROBLEM

A recent experimental design problem involved a sensor model. It had been decided that three levels would be investigated for each input factor considered. The pessimistic estimate of the number of model runs was 100, and the optimistic estimate was 250 runs. All input factors under consideration were completely crossed. Therefore, a factorial experiment in a completely randomized design was appropriate for the computer simulation model sensitivity experiment. Two designs were ultimately developed, one requiring approximately 100 runs and the other requiring approximately 250 runs.

A $(1/9) \times 3^7$ fractional factorial experiment requiring 243 model runs was designed using $I = ABCDE = CD^2EF^2G^2$ as the defining contrast. The design, plan 9.7.9 in Connor and Zelen (1959), permits estimation of the 7 main effects and the $\binom{7}{2} = 21$ first-order interaction effects. The ANOVA table is given below.

Table 1. ANOVA for the (1/9) x 3^7 Design

Source	DF
7 main effects	14
21 first-order interactions	84
residual	144
total	242

A smaller fractional factorial experiment was then designed such that its design points were a subset of the design points of the above seven factor experiment. This was accomplished by using the aliases from the (1/9) x 3^7 fractional factorial to determine the five factors having a full design within the 243 design points. The factors were B,C,D,E, and G. Then, using I = BCDEG as the defining contrast gave a (1/3) x 3^5 fractional factorial experiment requiring 81 model runs. The ANOVA table for the five factor experiment is given below.

Table 2. ANOVA for the (1/3) x 3^5 Design

Source	DF
5 main effects	10
10 first-order interactions	40
residual	30
total	80

The design points of both designs were provided to the analysts responsible for exercising the sensor model. The 81 factor level combinations of the (1/3) x 3^5 design were run first. Experiment execution proceeded smoothly and the remaining 162 runs for the seven factor fractional factorial were also run. The analysis of Table 1 was performed on each of a number of output variables selected during the design phase of the simulation model sensitivity experiment.

4. CONCLUSIONS

Sensitivity experiments of large complex computer simulation models involve a large number of input factors. The number of input factors normally involved far exceeds the number of factors involved in past field and laboratory experiments. A priori information concerning interactions among the input factors almost never exists. Minimum experimental objectives are, therefore, that the design permits the estimation of main and first-order interaction effects. Input factors selected for testing are those suspected of being highly significant. Past experience has shown that most main effects and first-order interaction effects are, in fact, statistically significant.

Fractional factorials for 2^n and 3^n designs, developed by Finney (1945) and (1946) and available in Cochran and Cox (1957), Connor and Zelen (1957) and (1959), and Davies (1960) do provide

designs which may be applied to sensitivity testing of computer simulation models. However, the largest 2^n design in Connor and Zelen (1957) which yields estimable first-order interactions is for 15 factors. The design has 256 design points. The largest 3^n design in Connor and Zelen (1959) which gives estimable first-order interactions is for 10 factors and it has 243 design points. The large number of computer simulation model runs required by fractional factorial designs do not normally permit assessment of the number of input factors desired when performing sensitivity experiments of large computer simulation models. Designs containing less design points than fractional factorial designs but permit the testing of main and first-order interaction effects are needed. Tabulations and catalogs of designs and/or computer software for generation of the designs are also needed. Analysis methodology as well as fast and efficient software for performing the statistical analysis dictated by the designs are naturally required.

5. REFERENCES

1. Cochran, W. G. and Cox, G. M., Experimental Designs, John Wiley and Sons Inc., New York, 1957.

2. Connor, W. S. and Zelen, M., Fractional Factorial Experiment Designs for Factors at Two Levels, NBS Applied Mathematics Series 48, US Government Printing Office, Washington, D.C., 1957.

3. Connor, W. S. and Zelen, M., Fractional Factorial Experiment Designs for Factors at Three Levels, NBS Applied Mathematics Series 54, US Government Printing Office, Washington, D.C., 1959.

4. Davies, O. L. (Ed.), Design and Analysis of Industrial Experiments, Hafner Publishing Company, New York, 1960.

5. Finney, D. J., The Fractional Replication of Factorial Arrangements, Annals of Eugenics, Vol. 12, pp 291-301, 1945.

6. Finney, D. J., Recent Developments in the Design of Field Experiments-III, Journal of Agricultural Science, Vol. 36, pp 184-191, 1946.

ON VALIDATING MISSILE SIMULATIONS: FIELD DATA ANALYSIS AND TIME-SERIES TECHNIQUES*†

Naim A. Kheir
School of Science & Engineering
The University of Alabama
in Huntsville
Huntsville, Alabama 35807

Donald Sutherlin
Aeroballistics Directorate
U. S. Army Missile Research
& Development Command
Redstone Arsenal, Alabama 35809

Abstract

The research reported here focused on an ARM/CM field data analysis, and models' fitting using time-series techniques. The immediate objective is to build, for the field data, an adequate model that fits a noise signal corrupting a deterministic one. The data happened to be seasonal and nonstationary. The ultimate goal, however, is to use the generated model in updating an all-digital computer simulation model, and be able to use simulation-data and field-data in validating the model. Few computer programs have been developed to help in the data analysis, the fitting and checking the adequacy of selected models. The fitted model is of the integrated autoregressive moving-average type.

---

*This research was supported by the U. S. Army Research Office under Contract DAAG 29-76-D-0100/D.0 534.

†Copies of a detailed report with the same title can be obtained from the authors.

SELECTED REFERENCES

A. ON TIME SERIES

(1) Brubacher, S. R. and Wilson, G. T., "Interpolating Time Series with Application to the Estimation of Holiday Effects on Electricity Demand," Applied Statistics, Vol. 25, No. 2, pp. 107, 1976.

(2) Yamamoto, T., "Asymptotic Mean Square Prediction Error for an Autoregressive Model with Estimated Coefficients," Applied Statistics, Vol. 25, No. 2, pp. 123, 1976.

(3) Anderson, T. W.,"The Statistical Analysis of Time Series," New York, John Wiley, 1971.

(4) Bloomfield, P., "On the Error of Prediction of a Time Series," Biometrika, Vol. 59, pp. 501, 1972.

(5) Tiao, G. C., "Asymptotic Behavior of Temporal Aggregates of Time Series," Biometrika, Vol. 59, 3, pp. 525, 1972.

(6) Box, G. E. P., and Tiao, G. C., "A Change in Level of a non-stationary Time Series," Biometrika, Vol. 52, pp. 181, 1965.

(7) Stralkowski, C. M., CeVor, R. E. and Wu, S. M., "Charts for the Interpretation and Estimation of the Second Order Moving Average and Mixed First Order Autoregressive Moving Average Models," Technometrics, Vol. 16, No , pp. 275, May 1974.

(8) Box, G. E. P., and Hill, W. J., "Correcting Inhomogeneity of Variance with Power Transformation Weighting," Technometrics, Vol. 16, No. 3, pp. 385, Aug. 1974.

(9) Melsa, J. L., "Identification of Autoregressive Processes Corrupted by Colored Noise," IEEE Conference on Decision and Controls, New Orleans, Dec., 7-9, 1977.

(10) IEEE Transactions on Automatic Control, Special Issue on System Identification and Time-Series Analysis, Volume AC-19, No. 6., December 1974.

(11) Cleveland, W. S., "Fitting Time Series Models for Prediction," Technometrics, Vol. 13, No. 4, pp. 713, November 1971.

(12) _____, "The Inverse Autocorrelations of a Time Series and Their Applications," Technometrics, Vol. 14, No. 2, pp. 277, May 1972.

(13) Hannan, E. J., "The Estimation of Mixed Moving Average Autoregressive Systems," Biometrika, Vol. 56, No. 3, pp. 579, 1969.

(14) _____, "The Identification of vector mixed autoregressive-moving average systems," Biometrika, Vol. 56, No. 3, pp. 223, 1969.

(15) Anscombe, F. J., and Tukey, J. W., "The Examination and Analysis of Residuals," Technometrics, Vol. 5, pp. 141, 1963.

(16) Stoica, P. and Soderstrom, T., "A Method for the Identification of Linear Systems Using the Generalized Least Squares Principle," IEEE Trans. AC-22, No. 4, pp. 634, August 1977.

(17) Bendat, J. S. and Piersol, A. G., "Random Data: Analysis and Measurement Procedures," Wiley-Interscience, 1971.

(18) Box, G. P. and Jenkins, G. M., "TimeSeries Analysis, forecasting and Control," Holden-Day, San Francisco, 1970.

(19) The International Mathematical and Statistical Libraries, Inc., (IMSL, Vol. 1 and 2), Edition 5, 1975.

(20) Enochson, L. D. and Otnes, R. K., "Programming and Analysis for Digital Time Series Data," The Shock and Vibration Information Center, U. S. Department of Defense, 1969.

(21) Box, C. E. P. and Jenkins, G. M., "Statistical Models for Prediction and Control," Technical Reports # 72, 77, 79, 94, 95, 99, 103, 104, 116, 121 and 122, Department of Statistics, University of Wisconsin, Madison, Wisconsin, 1967.

(22) Singer, A., "ARM/CM, Field Test Instrumentation," Harry Diamond Laboratory, U. S. Army ARM/CM Meeting, July 19-20, 1977, Maryland.

(23) Caines, P. E., and Chan, C. W., "Feedback between Stationary Stochastic Processes," IEEE AC-20, No. 4, August 1975, pp. 498-508.

(24) Caines, P. E., and Chan, C. W., "Estimation, Identification and Feedback," University of Toronto, Control Science and Eng. Report No. 7510, May 1975.

(25) Akaike, H., "Some problems in the application of the cross spectral method," in Advanced Seminar on Spectral Analysis of Time Series, B. Harris, Editor, New York, Wiley, 1967.

B. ON VALIDATION

(26) Kheir, N. A., "A Validation Procedure for Missile-Systems Simulation Models," The Proceedings of the Seventh Annual Pittsburgh Conference on Modeling and Simulation, April 26 - 28, 1976, pp. 534-539.

(27) Kheir, N. A. and Holmes, W. M., "On Validating Simulation Models of Missile Systems," to appear in SIMULATION (January 1978).

(28) Nolan, R., "Verification/Validation of Computer Simulation Models," Proceedings 1972 Summer Computer Simulation Conference, Vol. II, p. 1254.

(29) Driscoll, T. R. and Stoc dale, R. C., "Validation Methodology of a Digital 6-DOF Tactical Missile Simulation Model," Proceedings 1975 Summer Computer Simulation Conference, p. 1217.

(30) Fishman, G. S., "Problems in the Statistical Analysis of Simulation Experiments," Communications of the ACM, Vol. 10, No. 2, February 1967, p. 94.

(31) Richards, F. M., et. al., "A Validation Procedure for Discrete Digital Simulation Models," Proceedings 1973 Summer Computer Simulation Conference, p. 1145.

(32) Senge, P. M., "Some Issues in Evaluating the Validity of Social System Models," Proceedings 1973 Summer Computer Simulation Conference, p. 1176.

(33) Van Horn, R. L., "Validation of Simulation Results," Management Science, Vol. 17, No. 5, January 1971, p. 247.

(34) Nicholson, G. E. and Sewell, W. E., "Review of Model Validation Seminar," Proceedings U. S. Army Operations Research Symposium, May 1970, p. 73.

(35) Schrank, W. E. and Holt, C. C., "Critique of: Verification of Computer Simulation Models," Management Science, Vol. 14, No. 2, October 1967.

(36) Biggs, A. G., and Cawthorne, A. R., "Bloodhound Missile Evaluation," Journal of the Royal Aeronautical Society, September 1962, pp. 571-587.

(37) Wigan, M. R., "The Fitting Calibration and Validation of Simulation Models," SIMULATION, May 1972, pp. 188-192.

(38) Holmes, W. M., et. al., "Modular Missile Simulation Model," Technical Report RG 75-21, December 1974, U. S. Army Missile Command, Redstone Arsenal, Alabama.

(39) Cyert, R. M., "A Description and Evaluation of Some Firm Simulations," Proceedings IBM Scientific Computing Symposium, Simulation Models and Gaming, IBM Processing Division, White Plains, New York, 1966.

(40) Wright, R., "Validating Dynamic Models," Proceedings 1972 Summer Computer Simulation Conference, Vol. II, pp. 1286.

(41) Schlesinger, S., "Developing Standard Procedures for Simulation Validation and Verification," Proceedings 1974 Summer Computer Simulation Conference, Vol. I, p. 927.

(42) Naylor, T. and Finter, J. M., "Verification of Computer Simulation Models," Management Science, Vol. 14, No. 1, October 1967.

STATISTICAL VALIDATION OF GUIDED PROJECTILE/MISSILE SIMULATION MODELS

Harold L. Pastrick
Guidance and Control Directorate
Technology Laboratory
US Army Missile Research and Development Command
Redstone Arsenal, Alabama 35809

ABSTRACT. This paper discusses the statistical analysis which is proposed for aiding in the validation of several Laser Designator/Weapon System Simulation models. The primary objective is to provide a means for insuring that simulation responses to input signals match hardware responses under similar driving conditions to some "goodness-of-fit" criteria. The method involves generating several statistics on the point by point differences between the "true" data and the simulation data. These statistics include subinterval mean errors, confidence bounds for those errors, Theil's Inequality Coefficient, and the cumulative mean error.

I. BACKGROUND. Simulations of guided projectile and missile systems are used for a variety of purposes including flight stability analyses, trajectory studies, and lethality predictions. The computer simulation of these systems in many ways predicts the results that may be obtained only by actual flight tests or enhances analyses already generated by flight data. The potential for significant cost savings by using simulations vis-a-vis flight tests creates a firm case for making many program judgements based on simulation data with the understanding that they are truly representative of the real world. The general skepticism that program managers and decision makers previously placed on simulation data is slowly being replaced by their belief in simulation results given that a quantitative match, to some level of confidence, can be established between hardware and simulation models.

Recently a computer program entitled, "Laser Designator/Weapon System Simulation" (LDWSS) was generated to enable program managers for COPPERHEAD, HELLFIRE, and Ground Laser Designators as well as Army policy makers to judge alternatives among those systems. A significant objective in the LDWSS chronology is to validate the projectile/missile characteristics modeled in the software. The approach is being directed toward generating simulation responses under specified input conditions that match some level of goodness-of-fit to the actual hardware. LDWSS is the product of an evolution of simulations of semiactive laser guided missiles which had been developed by US Army Missile Research and Development Command (MIRADCOM) Technology Laboratory. Modeling formats and computer executive structures which had been proven in prior missile simulations were used as the base from which LDWSS was built [1-4].

The one-on-one engagement scenario employed a fixed foreground false target and a randomly selected background or overspill target. The randomly selected distance between the tank target and the background false target was based upon a statistical representation of this parameter obtained for certain observation posts in a digitized terrain model. All energy returns were subjected to appropriate geometric and atmospheric attenuation to determine the reflected energy received at the seeker. Utilizing seeker false target rejection logic to select the return to be tracked (tank or false target), the selected track point for each pulse was used as input to the appropriate dynamic model of seeker and delivery system. An overview of the organization of LDWSS and associated data relative to simulation elements has several features. The executive structure is designed to preserve a great deal of the internal system operation information which is generated during the calculation of hit probabilities [5].

II. DATA BASE. A simulation was developed to generate a set of meaningful statistics which aided in the validation of several of the models used in LDWSS [6]. The models examined included HELLFIRE and COPPERHEAD components. In general, model validation was accomplished by comparing "real" data with that generated by the appropriate LDWSS subroutines (under identical input conditions). The real data came from either field experiments or the hardware-in-the-loop simulation. In any case, two sets of data were generated. They are referred to as actual (real world data) and simulation data (LDWSS). Figure 1 is a sample plot of these data. In actuality, both curves are generated from digital simulations of an actuator. The outputs shown are time response curves to a step function input. Figure 2 is a plot of point by point difference (actual - simulation) between the two curves. The more closely the two curves are alike, the smaller the residuals. These residuals form the basis for the statistical analysis programs.

The derivation of the statistics used for verification has already been covered in detail [7, 8] and will only be reviewed briefly here. The time series shown in Figures 3 and 4 are analysed on a subinterval and a cumulative basis, respectively.

Subinterval Statistics

a) Mean residuals between the real and simulated data are defined as:

$$\bar{\epsilon}_j = \sum_{r=1}^{n} \left[A_{n'(j-1)+k} - S_{n'(j-1)+k} \right] \cdot \frac{1}{n'} \qquad (1)$$

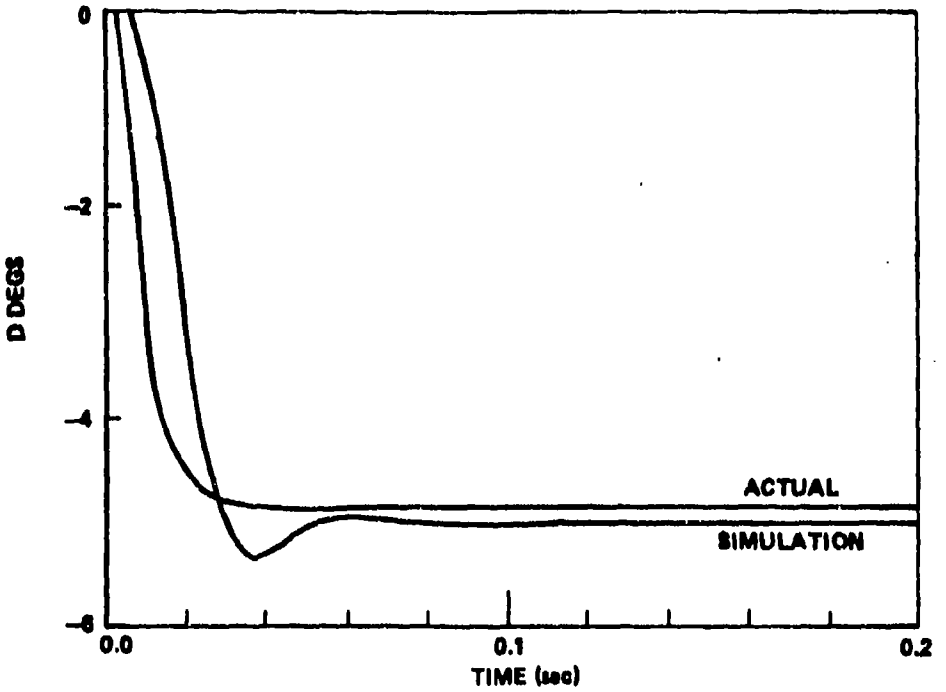


Figure 1. Test data.

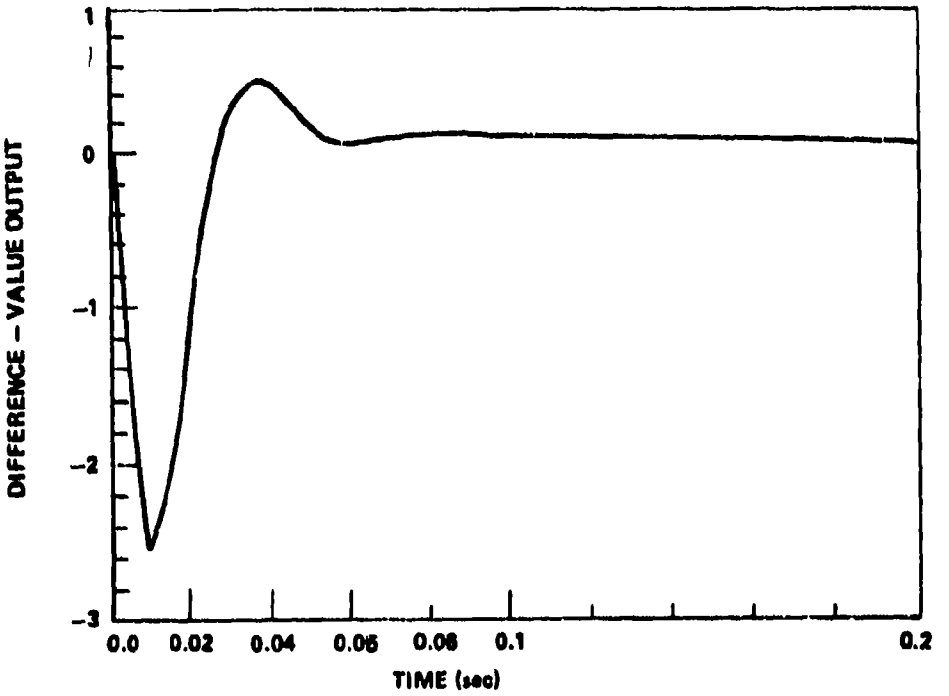


Figure 2. Difference plot (actual - simulation).

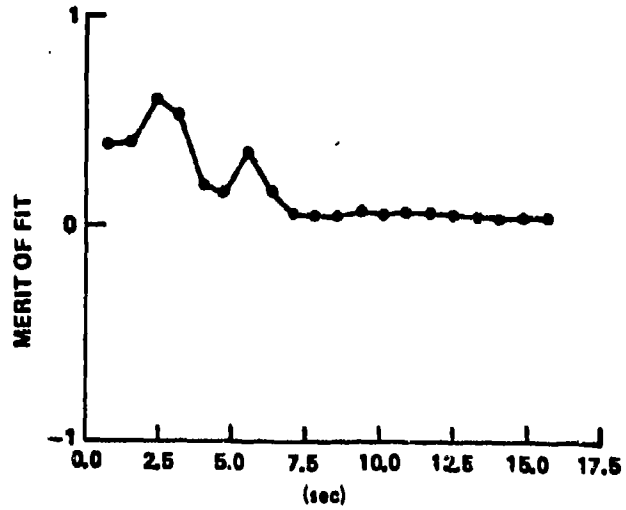


Figure 3. Subinterval TIC.

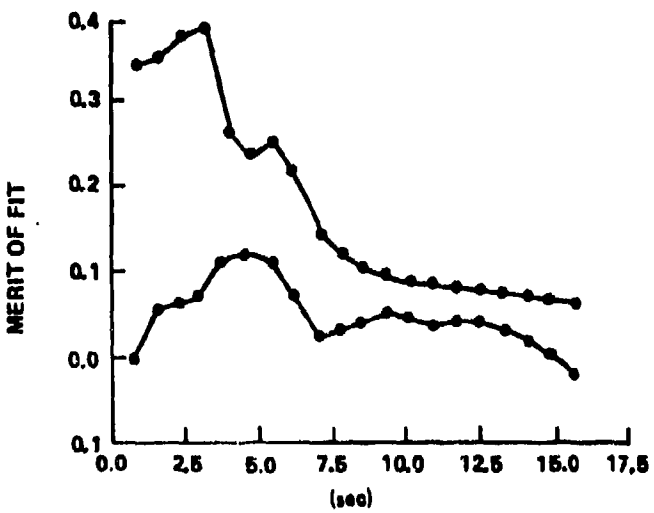


Figure 4. Cumulative mean residual and TIC (real versus hardware model).

where A_i is the ith sample from the real data, S_i is the ith sample from the simulated data, j is the subinterval counter, and n' is the number of points on the subinterval.

b) Confidence bounds on $\bar{\epsilon}_j$ are given by:

$$LB = \bar{\epsilon}_j - \theta \sqrt{\sigma_{\bar{\epsilon}_j}^2/n'} \tag{2}$$

$$UB = \bar{\epsilon}_j + \theta \sqrt{\sigma_{\bar{\epsilon}_j}^2/n'} \tag{3}$$

where $\sigma_{\bar{\epsilon}_j}^2$ is the variance of $\bar{\epsilon}_j$ on the jth interval corrected for correlation effects and $100 \cdot (1 - 1/\theta^2)$ is the percent confidence desired.

c) Theil's Inequality Coefficients (TIC)

$$U = \frac{\sqrt{\frac{1}{n'}\sum_{k=1}^{n'}\left(A_{n'(j-1)+k} - S_{n'(j-1)+k}\right)^2}}{\sqrt{\frac{1}{n'}\sum_{k=1}^{n'}A_{n'(j-1)+k}^2} + \sqrt{\frac{1}{n'}\sum_{k=1}^{n'}P_{n'(j-1)+k}^2}} \tag{4}$$

d) Theils Coefficient of Unequal Central Tendency

$$U^M = \left[\frac{\bar{S} - \bar{A}}{NUM}\right]^2 \tag{5}$$

where $\bar{S}$ is the mean of the S_i on interval j, $\bar{A}$ is the mean of the A_i on interval j, and NUM is the numerator given in Equation (4).

e) Theil's Coefficient of Unequal Variation

$$U^S = \left[\frac{\sigma_A - \sigma_S}{NUM}\right]^2 \tag{6}$$

where σ_A is the sample standard deviation of the A_i on interval j and σ_S is the sample standard deviation of the S_i on interval j.

f) Theil's Coefficient of Imperfect Covariation

$$U^C = \left[\frac{\sqrt{2(1 - r)\ \sigma_S \sigma_A}}{NUM}\right]^2 \tag{7}$$

where r is the correlation coefficient between the A_i and S_i on interval j.

In addition to the subinterval statistics, two cumulative statistics are also computed. These are cumulative mean residual and cumulative TIC.

III. DESCRIPTION OF EXPERIMENT. The statistics package was run on three sets of data:

a) Real versus hardware simulation.

b) Real versus LDWSS.

c) Hardware simulation versus LDWSS.

Each time series consisted of 2020 data points with a delta time of 0.0078125 sec. The series was divided into 20 intervals each containing 100 data points. The percent confidence requested for the mean residual was 95%. Each run produced a tabular output of the statistics as well as several plots.

The hardware chosen for the experiment is the actuator which is shown in its most complete form. That is, the model in Figure 5 represents the best information available for the actuator. It was subsequently reduced to the model shown in Figure 6 for use in the LDWSS simulation program. The objectives were to determine whether the complete model, referred to as the "hardware simulation" was well represented by the reduced model, referred to as the "LDWSS model" and whether either or both were high fidelity models of the hardware test data, referred to as "real data." The real data were obtained from flight recordings of the output of the actuator as a response to input commands. Consequently, the input-output command and response time series history is an accurate portrayal of the transfer function characteristics of the actuator in Figure 7.

An example of real data compared to simulated hardware data is shown in Figure 8. It is a plot of the mean residuals and confidence bounds (shown as vertical lines). From this plot, it can be seen that the means of the real and simulated data agree rather well with small mismatches on Intervals 8 and 9, where the mean residuals are -0.23 and -0.32, respectively. Considering the range of values for the original data, these residuals are quite small. Remember the ideal case is zero mean residual with a small confidence bound. Subintervals 8 and 9 correspond to the time period immediately following guidance initiate. This is the point in the flight of the missile where the reflected laser energy starts to contribute to the guidance loop. The small degradation in the last couple of subintervals is due to the fact that the missile is in terminal guidance where an acceleration in rate changes is common.

Figure 3 is a graph of the subinterval, the TIC for the same two sets of data, i.e., real and hardware model time series. A TIC of zero indicates equality between the two series, which in turn indicates that

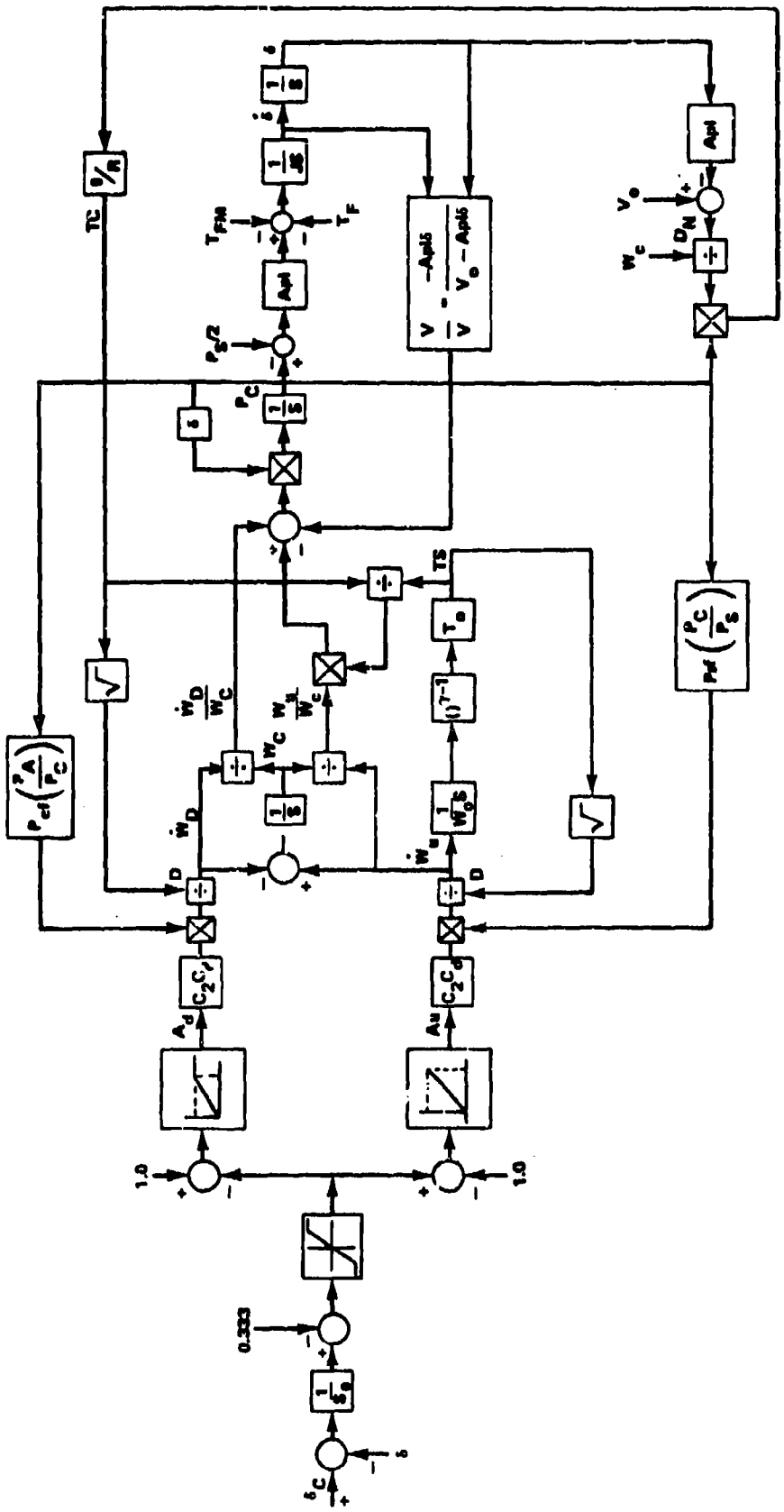


Figure 5. Block diagram of actuator module hardware model.

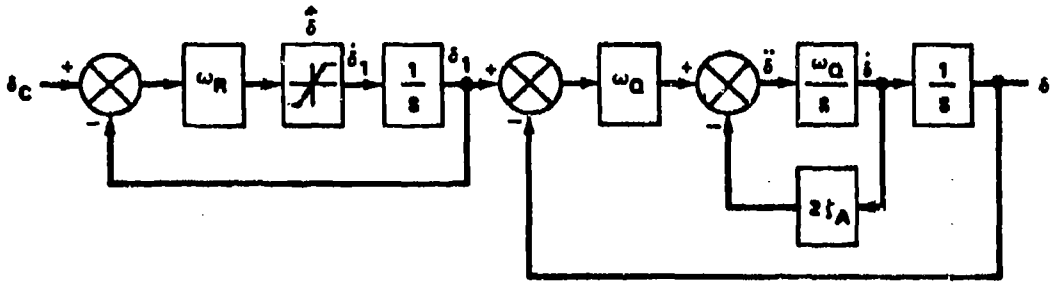


Figure 6. Actuator model block diagram.

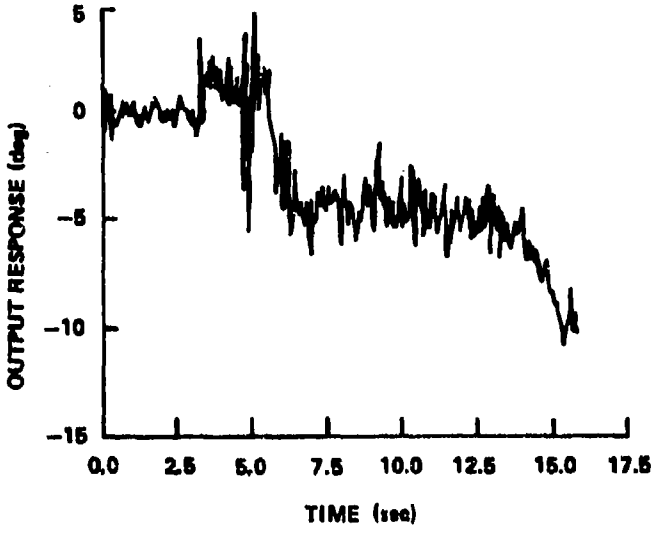


Figure 7. Real flight data digitized.

a perfect model had been hypothesized for the actuator. By the eighth subinterval, the value is reduced to approximately 0.05 and it remains small thereafter. The data in Figure 4 are a more detailed view of the same data and include the cumulative mean residual for comparison with the cumulative TIC.

IV. CONCLUSIONS. The preceding statistics represent a small subset of those available to the analyst for validating dynamic systems. Many (Bibliography) agree that these can supply useful and meaningful information for validation purposes. However, there are some who feel that

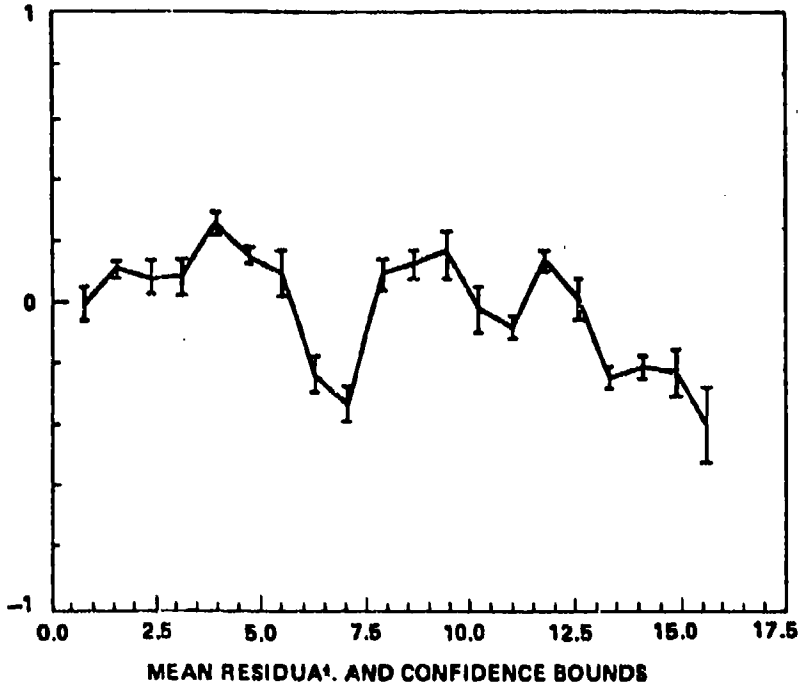


Figure 8. Real versus hardware simulation residuals and confidence bounds.

special techniques must be used to analyze nonstationary systems and the straightforward statistical quantities (as those discussed in this summary) are questionable in the cases where the models being analyzed produce highly nonstationary data. Work is underway using variations of these techniques as well as spectral techniques to circumvent the problem. Early results appear promising.

REFERENCES

1. Lewis, C. L., Taylor, D. S., and Lee, A. W., Jr., User's Guide for a Monte Carlo Point Target Terminal Homing Simulation Program, US Army Missile Command, Redstone Arsenal, Alabama, 20 May 1974, Technical Report RG-74-37 (UNCLASSIFIED).

2. Lewis, C. L., Hooker, W. R., Lee, A. W., Jr., and Harrison, J. S., THAD T-7 Missile Monte Carlo Terminal Homing Simulation Utilizing ALS, Digital/Linear and TV Seekers, US Army Missile Command, Redstone Arsenal, Alabama, 1 July 1976, Technical Report RG-7T-2 (UNCLASSIFIED).

3. 6-DOF Simulation Program of HELLFIRE Engineering Development Missile, Rockwell International Corporation, Columbus, Ohio, 30 September 1976, (UNCLASSIFIED).

4. Pastrick, H. L. and Hollman, H. C., Analysis and Digital Simulation Models for CLGP: Martin Marietta Aerospace Design, US Army Missile Command, Redstone Arsenal, Alabama, 18 December 1974, Technical Report RG-75-29 (UNCLASSIFIED).

5. Leonard, J. P., Yates, R. E., and Lewis, C. L., Laser Designator/Weapon System Simulation, Phase I Report, US Army Missile Research and Development Command, Redstone Arsenal, Alabama, 31 March 1977, Technical Report TG-77-8 (CONFIDENTIAL).

6. Mango, John, Statistics Generation Program, Science Applications, Inc., Huntsville, Alabama, 5 August 1977, Final Report, Contract DAAH40-77-C-0187 (UNCLASSIFIED).

7. Kheir, Naim A., "A Validation Procedure for Missile Systems Simulation-Models," US Army Missile Research and Development Command, Redstone Arsenal, Alabama, Summer 1977, Internal Research Paper (UNCLASSIFIED).

8. Naylor and Finger, "Verification of Computer Simulation Models," Management Science, Vol. 14, No. 2, October 1967 (UNCLASSIFIED).

BIBLIOGRAPHY

Fishman, George S., Concepts and Methods In Discrete Event Digital Simulation, New York: Wiley-Interscience, 1973.

Gordon, Geoffrey, System Simulation, Edglewood Cliffs, New Jersey: Prentice-Hall, 1969.

Nolen, Richard L., "Verification/Validation of Computer Simulation Models," Proceedings 1972 Summer Computer Simulation Conference.

Schlesinger, Stuart, et al., Developing Standard Procedures for Simulation Validation and Verification, The Aerospace Corporation, El Segundo, California.

Shrank and Holt, "Critique of Verification of Computer Simulation Models," Management Science, Vol. 14, No. 2, October 1967.

Shannon, Robert E., Systems Simulation, Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

Wright, Richard D., "Validating Dynamic Models: An Evaluation of Tests of Predictive Power," Proceedings 1972 Summer Computer Simulation Conference.

ANALYSIS OF VARIANCE OF MULTIVARIABLE FLIGHT TEST DATA

A CALL FOR ASSISTANCE

James S. Hayden
US Army Aviation Engineering Flight Activity
Edwards Air Force Base, California 93523

INTRODUCTION: The flight test community is frequently called upon to define changes in performance resulting from a change in configuration of an aircraft. Even with extreme attention to control of test condition state variables, the problem of duplication of conditions is an order of magnitude more difficult than in a laboratory environment. Further complicating the problem is the fact that depending on the flight regime, up to three non-linear or eleven linearized independent variables are involved. Measurement errors may be present in each of the independent variables. Published methods for analysis of variance are inadequate to treat this problem. A brief case history of determination of the change in hovering performance due to a rotor system change is presented to illustrate the problem. Measurement accuracies, test techniques and analysis methods are discussed to highlight the problem and suggest areas where discussion of statistical analysis techniques would be most useful.

HELICOPTER PERFORMANCE TEST TECHNIQUES: Pre-test preparation includes calibration of most performance instrumentation data sensors and indicators to N.B.S. secondary reference standards. Wherever possible "end to end" calibrations are performed on complete measurement subsystems after installation in the test vehicle. Certain systems such as engine torquemeters and instrumented rotor shafts are of necessity calibrated by contractors. Prior to testing, the fully instrumented helicopter is subjected to multiple precision weighings to accurately determine weight, center of gravity location, and to calibrate fuel cells. Strict inventory control of useful load items such as ballast, armament load, parachutes, oxygen equipment, individual crew composition and pre/post flight fuel mass are kept on a flight by flight basis. Recalibrations and re-weighings are performed periodically during the test program.

The vast majority of precision performance data is gathered under stabilized conditions. Using the great outdoors as your laboratory has esthetic advantages but your ability to carefully control the environment is quite limited. Smooth air is essential for all tests and steady winds not exceeding three knots are required for hover performance tests. Wind is not as critical for tests performed at altitude but caution must be exercised to avoid mountain waves which may seem smooth as glass while the air mass is rising and falling sinusoidally in a pattern relatively stationary with respect to the ground. Errors equivalent to rates of climb of $\pm$ 500 ft/min are not uncommon in these atmospheric formations.

It should be clear that the ability of the test pilot to stabilize the aircraft with a minimum of control motions and to hold this condition for the required data recording time period is of primary importance. On many tests the state variables are also controlled to hold certain non-dimensional variables (to be discussed) constant for a series of data points. This process is itself quite involved and requires the flight test engineer to calculate a target altitude and rotor RPM for the next data point based on cockpit observed values of airspeed, altitude, air temperature and fuel used. The calculations are quite involved and require use of charts, a programmable calculator or a telemetry down link with voice up link. Errors which may accumulate in the various steps (engineer reading of cockpit indicators, calculations, and pilot setting of conditions using cockpit indicators) are reduced significantly by the use of telemetry. The key point is that the accuracy of establishing desired flight conditions is limited.

PERFORMANCE DATA PARAMETER MEASUREMENT ACCURACY: As has been pointed out, flight testing is not conducted in a laboratory environment. Test instrumentation is exposed to a host of alien environmental factors; vibration, temperature extremes, shock, dirt, etc. The flight test engineer quickly recognizes that brochure accuracies are unrealistic in practical application. Experience has shown that the following accuracies can be achieved with reasonable attention to detail.

FLIGHT TEST PERFORMANCE DATA MEASUREMENT ACCURACY

PARAMETER	SYSTEMATIC/ERROR	POINT ERROR (σ)
Gross Weight	30 Lb	15 Lb
Engine Torque	1.5%	1%
Calibrated Airspeed	0.5 KT	0.3 KT (>100KCAS)
Rotor Speed		0.1%
Air Temperature		0.5°C
Pressure Altitude		20 Ft

The impact of these uncertainties on the analysis of hovering data will be discussed in more detail later.

HELICOPTER PERFORMANCE MODELS: The versatility of the helicopter expressed in its ability to fly literally in any direction presents an extremely complex performance analysis statement. For the purpose of describing the subject statistical analysis challenge, we will restrict the discussion to two important flight regimes; hover and cruise.

Non dimensional methods are commonly used in helicopter performance analysis. The parameters of interest, for our restricted discussion, are power coefficient (CP), thrust (or weight), coefficient (CT), advance ratio (μ), and advancing tip mach number (M).

NON DIMENSIONAL PERFORMANCE PARAMETERS

$$CP = \frac{Q}{\rho o\ (\delta/\theta)\ \pi R^5 \Omega^2}$$

$$CT = \frac{W}{\rho o\ (\delta/\theta)\ \pi R^4 \Omega^2}$$

$$\mu = \frac{KTAS \times 1.68781}{\Omega R}$$

$$M = (1 + \mu)\ \frac{\Omega R}{1116.45\ \sqrt{\theta}}$$

WHERE:

CONSTANTS

ρo = S.L. Std Atmospheric density, slug/ft^3.

R = Rotor radius , ft.

π, 1.68781, 1116.45 = CONSTANTS

MEASURED PARAMETERS.

Q = Total delivered torque at rotor speed, Lb-ft.

Ω = Rotor rotational speed, rad/sec.

δ = Ambient atmospheric pressure/S.L. std ambient pressure, Dim.

θ = Ambient atmoshperic absolute temperature/S.L. std. Ambient absolute temperature, Dim.

W = Aircraft gross weight , lb.

KTAS = True airspeed, Kt.

Hover power required, in simple terms, may be considered to be composed of induced power (energy required to produce lift) and profile power (energy raquired to overcome rotational drag of the blades). Rotational drag is composed of a base drag, a component of drag due to lift, an additional drag due to compressibility and in some cases an additional drag due to stall. In coefficient form, a model which has proven effective is:

HOVER POWER REQUIRED MODEL

$$CP = A + B\ CT^{3/2} + C\ CT^3 + f\ (CT, M)$$

This is the equation form which will be used with the specific example to be discussed.

Forward flight power required includes additional components; parasite power (energy required to overcome airframe drag), additional profile power due to forward speed (μ), and stall and compressibility power which is a function of μ, CT, and M. A typical forward flight power required model is:

FORWARD FLIGHT POWER REQUIRED MODEL

$$CP = A\ (1 + 3\mu^2) + D\left(\frac{CT^2}{\mu} + E\mu + F\mu^3 + f\ (CT,\mu,M)\right)$$

The functional relationship indicated for stall and compressibility power understates the complexity of the phenomena. The onset of stall is usually defined for a specific aircraft as a unique relationship between CT and μ. This unique relationship, or boundary, is however a function of both the drag configuration (i.e., rocket pods, doors open, etc.) and the rotor tip mach number. The onset of compressibility effects is usually defined as unique relationship between CT and M, however, this boundry is also a function of μ.

The gross trends of these additional power components are illustrated in Figures 1 and 2.

Now that you have been introduced to the complexity of our forward flight problem, lets turn our attention to the simple example problem to be used to illustrate our challenge - determination of the change in hovering performance due to a rotor system change.

COMPARATIVE HOVERING PERFORMANCE TESTS: The United States Army Engineering Flight Activity conducted comparative tests of two types of rotor blade installed on an AH-1R helicopter. The tests were conducted at field elevations from approximately 2,300 ft to 10,000 ft over a span of approximately three months. The comparison of out of ground effect hovering performance was only one of the many objectives of the test and is the only subject which will be discussed here.

All tests were flown on the same aircraft with the same engine and with the same basic instrumentation. Data were obtained with each blade type, "Back to Back", at each of the three test sites.

Data were obtained by stabilizing the helicopter in hover at a skid height of 100 $\pm$ 2 ft for a period of not less than 20 sec. Data were recorded continuously for a period of approximately 10 sec at a sample rate of 100 samples/sec. The data records were then edited from time history strip charts. Acceptable data points were then edited to the most stabilized 6 sec of record. The edited record was then used to calculate the non-dimensional parameters based on actual data every tenth of a second. The calculated non-dimensional parameters were then averaged over the period. This leads to the first question to be posed in this clinical discussion:

1. "Should data be averaged as measure or after calculation?"

The data gathered during these tests is presented graphically in Figures 3 and 4.

DATA ANALYSIS: The edited averaged data points were analyzed by performing a multiple linear regression of the hover power required in the form:

$$CP = A + B\,CT^3 + C\,CT^{3/2} + DM$$

Results of the regressions are summarized as follows:

MULTIPLE LINEAR REGRESSION DATA

$$CP + A + B\,CT^3 + C\,CT^{3/2} + DM$$

	WITH MACH NUMBER		NO MACH NUMBER	
BLADE	A	B	A	B
n	82	58	82	58
A	-3.755-8	-3.189-8	-1.380-7	-7.457-8
B	3.673+2	2.430+1	-1.879+2	-2.040+2
C	9.458-1	1.120+0	1.376+0	1.319+0
D	1.250-4	6.431-5	0	0
R^2	9.859-1	9.843-1	9.832-1	9.839-1
s	8.965-6	1.085-5	9.760-6	1.099-5

The nominal performance design point for the AH-1R is 9,000 lb gross weight at 4,000 ft, 35°C or a thrust coefficient (CT) of 55.34×10^{-4} and a tip mach number (M) of 0.6465. Evaluation of the polynomials yields the following power coefficients for the two blade sets.

$$CP(A) = 53.24 \times 10^{-5}$$
$$CP(B) = 50.67 \times 10^{-5}$$

If the problem being addressed was linear with a single independent or even multivariate, the analysis of variance would be straight forward. Recall the description of the functional relationships given in AMCP 706-110 where X values can be measured exactly (FI) and where errors may be present in the X measurement (FII).

Now recall the possible point errors of the present example, as implied by instrumentation accuracies. The vectors representing the individual effect of 1σ data errors on the non-dimensional coefficients are illustrated in Figure 5 as are the maximum possible 1σ measurement errors and clearly illustrate that we are confronted with an FII situation.

This leads to the concluding question of this clinical presentation:

2. "What procedures are recommended for calculating a specified difference in average performance, with a chosen degree of confidence with a multivariable FII relationship?"

SUMMARY OF QUESTIONS:

1. "Should data be averaged as measured, or after calculation?"

2. "What procedures are recommended for calculating a specified difference in average performance, with a chosen degree of confidence with a multivariable FII relationship?"

REFERENCES:

1. "Airworthiness and Flight Characteristics, Improved Main Rotor Blade on the YAH-1R", Yamakawa Et, Al, USAAEFA Project No 76-08.

2. AMCP 706-110, DEC 1969.

FIGURE 1. COMPRESSIBILITY EFFECTS ON GENERALIZED LEVEL FLIGHT PERFORMANCE

C_T = CONSTANT

$M_{DD} = \frac{N/\sqrt{\theta}R}{9.5493a_o}(1 + FVR\sqrt{C_T/2})$

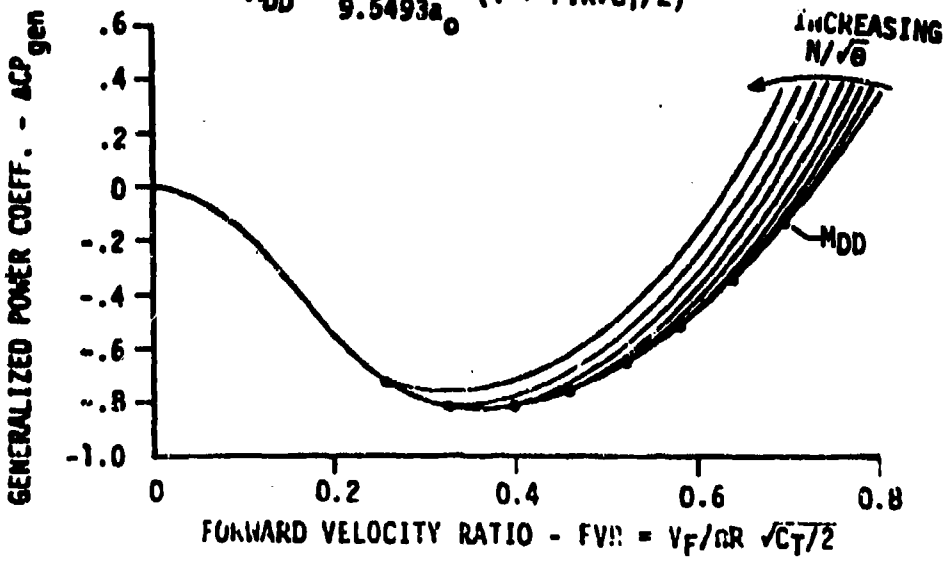


FIGURE 2. RETREATING BLADE STALL EFFECTS ON GENERALIZED LEVEL FLIGHT PERFORMANCE

N/√θ = CONSTANT

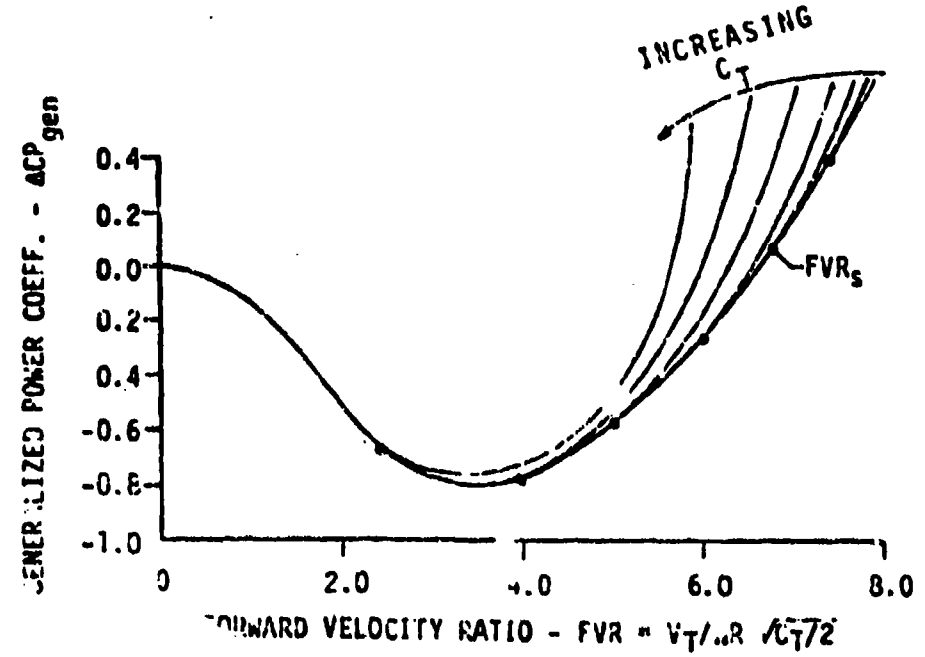
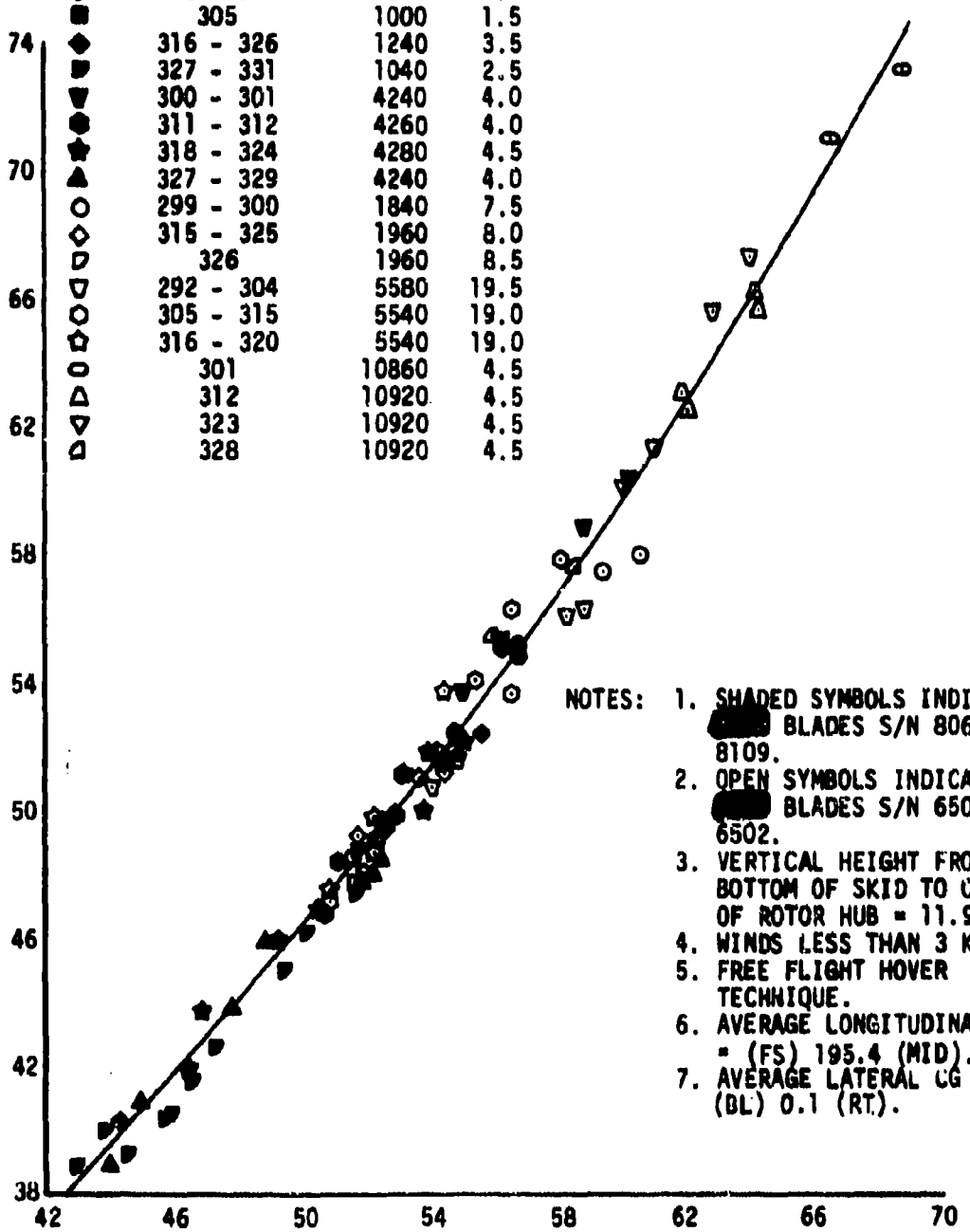


FIGURE 3
OUT-OF-GROUND EFFECT NONDIMENSIONAL HOVERING PERFORMANCE
YAH-1R USA S/N 70-15936
ENGINE T53-L-703 S/N LE15124Z
SKID HEIGHT = 100 FEET

SYM	REFERRED ROTOR SPEED RANGE (RPM)	DENSITY ALTITUDE (FEET)	OAT (°C)
●	303 - 304	1300	4.0
■	305	1000	1.5
◆	316 - 326	1240	3.5
▬	327 - 331	1040	2.5
▼	300 - 301	4240	4.0
⬢	311 - 312	4260	4.0
★	318 - 324	4280	4.5
▲	327 - 329	4240	4.0
○	299 - 300	1840	7.5
◇	315 - 325	1960	8.0
▷	326	1960	8.5
▽	292 - 304	5580	19.5
⬡	305 - 315	5540	19.0
☆	316 - 320	5540	19.0
◇	301	10860	4.5
△	312	10920	4.5
▽	323	10920	4.5
▱	328	10920	4.5



NOTES:
1. SHADED SYMBOLS INDICATE [redacted] BLADES S/N 8063 AND 8109.
2. OPEN SYMBOLS INDICATE [redacted] BLADES S/N 6500 AND 6502.
3. VERTICAL HEIGHT FROM BOTTOM OF SKID TO CENTER OF ROTOR HUB = 11.9 FEET.
4. WINDS LESS THAN 3 KNOTS.
5. FREE FLIGHT HOVER TECHNIQUE.
6. AVERAGE LONGITUDINAL CG = (FS) 195.4 (MID).
7. AVERAGE LATERAL CG = (BL) 0.1 (RT).

FIGURE 4
OUT-OF-GROUND EFFECT NONDIMENSIONAL HOVERING PERFORMANCE
YAH-1R USA S/N 70-15936
ENGINE T53-L-703 S/N LE15124Z
SKID HEIGHT = 100 FEET

SYM	REFERRED ROTOR SPEED RANGE (RPM)	DENSITY ALTITUDE (FEET)	OAT (°C)
○	300 - 301	1620	6.0
□	315	1640	6.0
◇	316 - 325	1660	6.5
▽	326 - 329	1560	5.4
○	314	4740	12.0
⬠	319 - 324	4800	12.5
◇	304	10700	1.5
△	311 - 313	10580	0.5
▽	318 - 325	10700	2.0
△	328 - 336	10520	-0.5

NOTES:
1. [redacted] BLADES S/N 1005 & 1009.
2. VERTICAL HEIGHT FROM BOTTOM OF SKID TO CENTER OF ROTOR HUB = 11.9 FEET.
3. WINDS LESS THAN 3 KNOTS.
4. FREE FLIGHT HOVER TECHNIQUE.
5. AVERAGE LONGITUDINAL C.G. = (FS) 195.1 (MID).
6. AVERAGE LATERAL C.G. = (BL) 0.1 (RT).

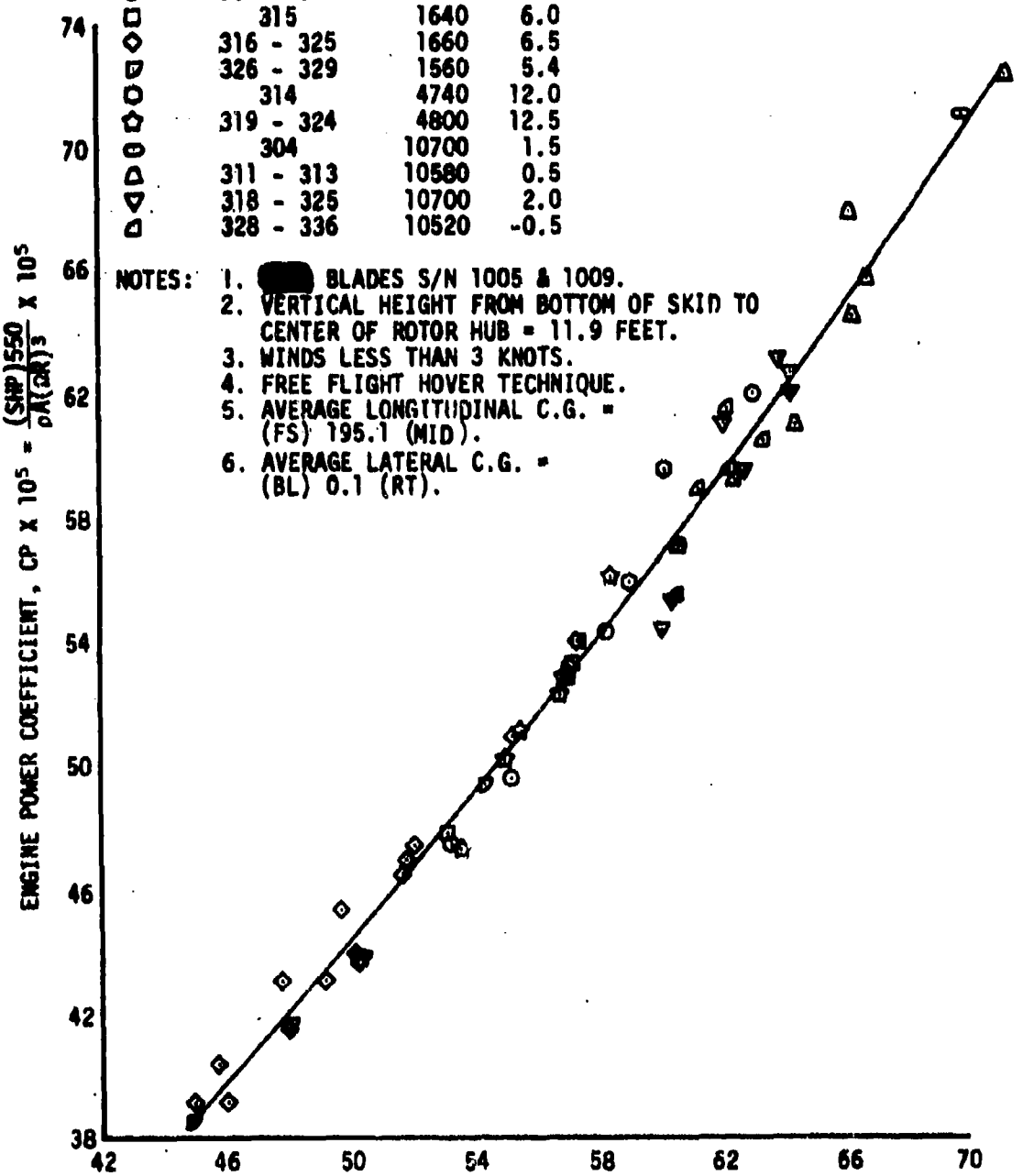
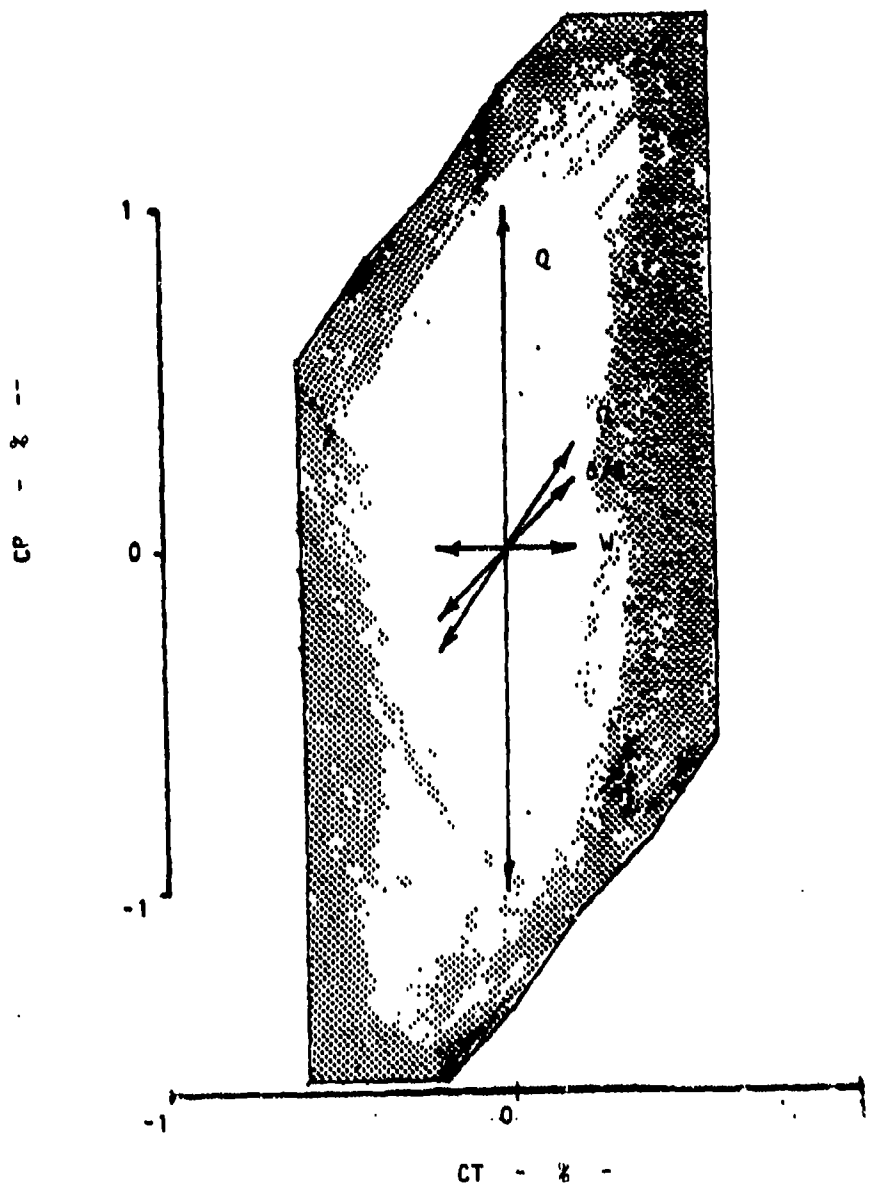


FIGURE 5.

RELATIVE EFFECTS OF POINT ERROR ON HOVER NON DIMENSIONAL COEFFICIENTS



ANALYSIS OF VARIANCE: SELECTION OF A MODEL AND SUMMARY STATISTICS

Frederick Steinheiser, Jr. & Kenneth I. Epstein

Army Research Institute for the Behavioral and Social Sciences

Alexandria, VA 22333

SUMMARY

Three models can be used to perform ANOVA: fixed, random, or mixed. The choice of a model is determined by the sampling plan of the treatments, e.g., if sampling was exhaustive, then no generalization beyond such sampled levels is allowable. Two summary statistics may also be computed: the F-ratio (to test the hypothesis of an effect due to a given treatment), and an index of the magnitude of experimental effect (also called the proportion of variance accounted for by a given treatment effect). This paper examines the relationship between ANOVA models, summary statistics, and the inferences that can be drawn from them. Data from a completely crossed repeated measures experiment are presented, to show how some inferences about effects can change as a function of the model selected and the summary statistics which are then computed.

Introduction

The topics of this paper are models for the analysis of variance (fixed, random, or mixed ANOVA models), and the subsequent summary statistics (F-ratio, quasi-F-ratio, and magnitude of treatment effect) which may be computed following the ANOVA. ANOVA is a useful method for assessing the statistical significance of treatment effects. But the significance of an effect is a function of two decisions. First is the selection of a model and an appropriate sampling plan for elements within each of the treatment factors. Second is the choice of summary statistics which indicate the extent of significance achieved. In this paper, comparisons will be made between models, and between summary statistics. Specific issues will be clarified concerning the interpretation of results when various models and summary statistics are used on the same set of data.

Selection of an ANOVA Model

In the fixed-effects model, the levels of the independent variables are assumed to have been exhaustively sampled. No generalization beyond those levels sampled is intended, or theoretically permissible. The random effects model assumes that the selected treatment variables have been randomly selected from a very large population of such variables. Generalization of results from the random sample to the population is allowed. The mixed model allows both fixed and random factors to be studied in the same experiment, with the results for each factor to be interpreted according to that factor's sampling plan.

The choice of a model has an impact upon the probability of obtaining the observations under the null hypothesis for each treatment

(factor). Behavioral research is particularly vulnerable to the choice of a model, because often the investigator can use only a limited sample of the possible number of stimuli (items, drug doses, etc.). Furthermore, because of the difficulty in creating comparable sets of stimuli, the same stimulus set may, by necessity, be given to all subjects.

As a simple hypothetical experiment (adapted from Clark, 1973), suppose that two classes of stimuli, nouns and verbs, are individually shown to subjects. We want to see if it takes the same time to identify each word as a member of the correct part-of-speech class. This simple hypothesis will be shown to have interesting implications for both experimental design and statistical analysis.

First of all, fixed sets of nouns and verbs which are matched on relevant parameters, such as number of letters and frequency of occurrence, should be prepared. If we want to be able to generalize to the full domain of nouns and verbs, each subject should receive a different random sample of words from the two lists. However, it is impossible to match the words on all relevant variables. It is also practically impossible to use a different random sample of words for each subject.

Consider, then, the following experimental design, in which "s" subjects are each presented "w" different nouns and verbs:

TABLE 1. Assignment of Subjects and Parts of Speech.

Subject:	Part of Speech P_1 (nouns)	P_p (verbs)
S_1	$w_1 \cdots w_{w/2}$	$w_{w/2+1} \cdots w_w$
⋮		
S_s		

In order to compare the adequacy of the several possible F ratios for testing the difference in response time to the two "treatment" (part of speech) conditions, the following tables of expected mean squares will be helpful:

TABLE 2. EMS Assuming Parts of Speech is a Fixed Factor, and Subjects and Words are Random.

Source	EMS
P (Part of speech)	$\sigma_e^2 + sw\sigma_p^2 + s\sigma_{w(p)}^2 + w\sigma_{pxs}^2 + \sigma_{sxw(p)}^2$
W(P) (Words within part of speech)	$\sigma_e^2 + s\sigma_{w(p)}^2 + \sigma_{sxw(p)}^2$
S (Subjects)	$\sigma_e^2 + pw\sigma_s^2 + \sigma_{sxw(p)}^2$
P x S	$\sigma_e^2 + w\sigma_{pxs}^2 + \sigma_{sxw(p)}^2$
S x W(P)	$\sigma_e^2 + \sigma_{sxw(p)}^2$

TABLE 3. EMS Assuming Parts of Speech and Words are Fixed, and Subjects are Random.

Source	EMS
P	$\sigma_e^2 + sw\sigma_p^2 + w\sigma_{pxs}^2$
W(P)	$\sigma_e^2 + s\sigma_{w(p)}^2 + \sigma_{sxw(p)}^2$
S	$\sigma_e^2 + pw\sigma_s^2$
P x S	$\sigma_e^2 + w\sigma_{pxs}^2$
S x W(P)	$\sigma_e^2 + \sigma_{sxw(p)}^2$

If we choose to test the significance of the Parts of Speech treatment, the appropriate F-ratio for the model illustrated in Table 2 is: $F_1 = MS_p/MS_{pxs}$. The only term in the numerator that is not in the denominator is $sw\sigma_p^2$. However, if this same F-ratio is used with the model in Table 3 (applicable when generalization is desired to all nouns and verbs), then this F-ratio will contain two terms that are not in the denominator: $s\sigma_{w(p)}^2$ and $sw\sigma_p^2$. And, using alternative error terms in the parts-of speech fixed, words random model (Table 2) also leads to the same problem. For example, if we test the parts of speech effect against the words within parts of speech effect, we obtain $F_2 = MS_p/MS_{w(p)}$. In this case, EMS_p exceeds $EMS_{w(p)}$ by the amount of $w\sigma_{pxs}^2 + ws\sigma_p^2$. Therefore,

this F_2 ratio would also be significant when the true contribution of σ_p^2 due to parts of speech (treatments) is really zero. In summary, both F_1 and F_2 could be significant when $\sigma_p^2 = 0$, provided that σ_w^2 and σ_{pxs}^2 exceed zero.

A possible solution to this dilemma is to take the "quasi-F" ratio, or F', which equals $(MS_p + MS_{sw(p)})/(MS_{pxs} + MS_{w(p)})$. Now the only term in the numerator which is not in the denominator is σ_p^2. However, F' is only approximately distributed as F, although the error involved is not large, provided that adjustments are made to the degrees of freedom.

Another, more conservative solution is minimum F', which assumes that $MS_{sxw(p)}$ is zero. A more detailed discussion of this problem may be found in Clark (1973).

A series of Monte Carlo computer simulations (Forster & Dickinson, 1976) explored the relationship between all of the above F-ratios and and the resulting type I error rates. Generally, F_1 and F_2 alone produced unacceptably high error rates, whereas F' and min F' were more conservative, as can be seen in Table 3.

TABLE 3. Type I Error Rates as a Function of Variation in MS_{sxp} and $MS_{w(p)}$. (500 observations per situation, alpha = .05, p = 2, q = 5, r = 9)

Source of Variance Manipulated	$s.d._1$	$s.d._2$	F_1	F_2	min F'	F'
Neither	0	0	.044	.046	.010	.026
$MS_{w(p)}$	5	0	.228	.052	.038	.044
	10	0	.484	.070	.060	.060
	15	0	.586	.056	.048	.052
	20	0	.724	.050	.048	.048
MS_{sxp}	0	5	.042	.146	.024	.036
	0	10	.064	.388	.042	.042
	0	15	.036	.520	.032	.034
	0	20	.042	.586	.038	.042

Both	5	5	.124	.096	.034	.042
	10	10	.190	.090	.040	.040
	15	15	.220	.138	.056	.064
	20	20	.208	.118	.048	.048

As can be seen in Table 4, increasing the number of items and subjects tends to decrease F_1 type I error for the fixed effects model, where only subjects are random. Min F' and F' continue to have lower error rates.

TABLE 4. Type I Error Rates as a Function of the Numbers of Subjects and Items. (300 observations per situation, $s.d._1 = s.d._2 = 20$, and alpha = .05.)

Number of Subjects	Number of Items	F_1	F_2	minF'	F'
10	5	.240	.070	.040	.040
10	20	.090	.290	.053	.053
20	5	.307	.077	.067	.067
20	20	.193	.217	.060	.060

The "Magnitude of Effect" as a Summary Statistic

The F ratio indicates the level of statistical significance that can be attributed to a particular treatment. The degree of statistical significance is a joint function of the "true" strength of that factor, the error variability (which reflects the degree of experimental control), and the sample size (i.e., number of subjects tested). As sample size increases, there is increasing power to reject a false null hypothesis. Thus, in conducting large scale experiments with hundreds of subjects, the large "n" may be necessary in order to detect a weak "signal" buried in a background of "noisy" data. But the large n may also lead to spuriously significant F-ratios which are actually statistical artifacts.

One index for assessing the significance of effects is the "magnitude of effect," also sometimes referred to as the "proportion of variance accounted for." It is interesting to note that relatively few research papers have included this index, compared to the ubiquitous F-ratio.

Basically, the magnitude of effect (m.e.) measures the degree of association between the independent variable(s) and the dependent variable(s). In the simplest case for ANOVA having fixed factors, none of which are repeated, the m.e. formula is:

magnitude of effect = $(SS_{effect} - (df_{effect} \times MS_{error}))/(SS_{total} + MS_{error})$.

Rules for deriving m.e. indices are provided by Dodd & Schultz (1973), along with tables for representiave ANOVA designs.

The concern of the present paper is with the interpretation of these summary statistics, since both F and m.e. can be computed from the same set of data. It is clear that as the statistical significance for a given effect increases--i.e., the p(observation|null) decreases--the magnitude for that effect also increases. But it is also possible that an F-ratio may be highly statistically significant, yet the m.e. for that effect could account for only some very small proportion of the overall variance. The results from an experiment summarized in the following section show that when statistical significance ($p < .001$) was achieved by several treatments, the m.e. for these treatments ranged from 1% to 23%.

A Study of Marksmanship

Consider the following experiment which was conducted for the U.S. Army Military Police School at Fort McClellan, Alabama. Each of 237 students shot a total of 240 handgun rounds from eight different position-distance combinations. There were three repetitions of 80 shots each, at stationary silhouette targets. Within each repetition, five shots were taken, the weapon was reloaded, and five more shots were fired in the adjacent test lane. (Each subject had previously passed a training course with a score of at least 35 hits out of 50 shots.) In the test, 160 trials (2 repetitions) were taken on Thursdays, the third

was taken on Fridays. The completely crossed design was therefore: A x B x C x D, or 237 x 2 x 8 x 3, or subjects x lanes x tables x repetitions.

Table 5 highlights the results of the ANOVA from this experiment. The first column of F-ratios assumes a mixed model, with B,C,D as fixed factors. The second column of F-ratios assumes that only Tables was a fixed factor. The third F-ratio column assumes that all four factors were randomly sampled from their respective populations. The point is rather obvious: different ANOVA models produce different F-ratios for null hypothesis rejection, given the same set of data.

TABLE 5. Changes in F-Ratios as a Function of ANOVA Model

Source	d.f.[1]	M.S.	F^2	F^3	F^4
A (Subjects)	236	12.80		3.93****	2.54****
B (Lanes)	1	7.70	7.33****	5.96**	2.26
C (Tables)	7	732.71	385.64****	79.11****	79.11***
D (Repetitions)	2	34.75	14.18****	12.55****	4.71**

****:$p < .001$ ***:$p < .01$ **:$p < .025$ *:$p < .05$

1. d.f. for F-ratios were obtained using the Satterthwaite approximation.
2. A random; B, C, D fixed effects.
3. A, B, D random, C fixed.
4. A, B, C, D all random effects.

The problem of interpreting the F-ratios now needs to be addressed. Is there, for example, a significant effect due to lanes or to repetitions? If these effects are assumed to be fixed, the answer is yes; if they are assumed to be random, the answer for lanes is no, and for repetitions the level of statistical significance has greatly decreased.

We offer the suggestion that the choice of the ANOVA model (and ultimately the level of significance reached) lies in the eye of the beholder--the scientist himself. From a sponsor's persepctive, it may well be that only those conditions which are studied in the experiment are of interest. If many lanes, repetitions, or even tables are never to be studied or added to his testing program, then those factors would never be sampled from a larger population of such factors. However, one might argue from a scientific point of view that many additional lanes, repetitions, and firing positions could have been tested. That is, we happen to have chosen only three repetitions, two lanes per subject, and eight different distance-position combinations. Thus, the sponsor-practitioner wishes information that is specific to his particular test. In contrast, the scientific "purist" may perceive this one test or experiment as merely one of many different kinds which could have been conducted by him for the sponsor. Hence, the choice of model indeed influences the significance levels obtained.

The power of the F-ratio to reject a false null hypothesis is a function of (1) the "true" strength of the particular factor, and (2) the sample size. Although a large sample size may help to detect a weak signal in a noisy background, the result of using such a large sample can lead to increasingly significant F-ratios, with little, if any concomittant increase in the m.e. It is to this latter summary statistic that we now turn our attention, in the analysis of the same set of marksmanship data.

The m.e. results are shown in Table 6, where it may be seen that the largest effect, other than random error, was due to the "Tables" factor, which captured a 23% share of the total score variability.

The effect due to Persons, reflecting individual differences among the students, reached nearly 10%. Several interaction terms, in which Tables was a factor, accounted for about 6% to 7%.

TABLE 6. Changes in Magnitude of Effect Index as a Function of ANOVA Model.

	Proportion of Total Variance, Assuming:		
Source	A Random, B,C,D Fixed	A,B,D Random, C Fixed	A,B,C,D Random
A (Subjects)	.0852	.1027	.1030
B (Lanes)	.0004	.0006	.0005
C (Tables)	.1643	.2454	.2631
D (Repetitions)	.0027	.0041	.0042

Note that the effect due to Repetitions in Table 5 was statistically significant, whereas according to Table 6, Repetitions contributed an effect worth only about .4%. The reason for this apparent discrepancy between the two summary statistics is due to the large number of subjects, which in turn produced a large number of degrees of freedom. This allows small F-ratios to more readily achieve statistical significance. Thus, the values for m.s. in Table 6 act as a check upon the significance levels listed in Table 5. Therefore, the effect due to Repetitions reveals a slight, but probably inconsequential learning effect. A similar line of reasoning holds for the interpretation of the Scores variable in Tables 5 and 6.

Summary and Conclusions

In actual experimental testing situations, it may not be easy to determine whether a given treatment should be classified as a fixed or as a random effect. For example, in the experiment outlined, the Scores,

Repetitions, and Tables factors could be considered as either fixed or as random. Recall that Tables had eight levels, representing the eight specific position-distance combinations that comprise the marksmanship test. Since there are theoretically an infinite number of distance-position combinations, Tables could be interpreted as a sampling of eight from this much larger population. Since an experimenter is often interested in generalizing his results beyond the specific treatment levels to a larger set of "real-world" circumstances, a random effects assignment to Tables could easily be justified. Furthermore, the probability of falsely rejecting a true null hypothesis is less when a treatment is considered to be random as opposed to fixed.

In summary, the wise use of an ANOVA model involves the following points: (1) determination of fixed vs. random factors, (2)computation of complete sets of summary statistics, (3) interpretation of the statistics.

References

Clark, H.H. The language-as-fixed-effect-fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 335-359.

Dodd, D.H., & Schultz, R.F. Computational procedures for estimating magnitude of effect for some analysis of variance designs. Psychological Bulletin, 1973, 79, 392-395.

Forster, K.I., & Dickinson, R.G. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F' and min F'. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 135-142.

EXPERIMENTAL DESIGN FOR TESTING EFFECT OF INGESTING CRUDE FIBER ON PLASMA ZINC LEVELS IN HUMAN VOLUNTEERS

Walter D. Foster, AFIP, and Barbara F. Harland, FDA
Washington, D.C. 20306

ABSTRACT. The benefits of ingesting dietary fiber may be offset by a possible depression of plasma zinc levels. An experiment was designed to detect a loss of 10ug/100ml in plasma zinc (if it existed) at the .01 significance level with a power of .95. Variance estimates were deduced from serum (not plasma) distributions in the literature and restructured to offer between (and within) subject variance components. According to the non-central F-distribution, these design parameters required 14 volunteers to finish the experiment, each with three plasma determinations before treatment and three more at the end. Treatment consisted of daily ingestion of bran muffins and bread containing 2.7 grams of crude fiber for a period of 14 weeks. A similar group of controls ingested this diet without added fiber.

I. INTRODUCTION AND OBJECTIVES. For at least 20 years, the scientific literature has noted the general health benefits that might accrue from the ingestion of crude fiber, with specific emphasis on crude fiber's potential for reducing the incidence and severity of atherosclerosis. The popular literature of recent years has reiterated this theme. Thus, a growing proportion of the reading public is actively altering diets to include more crude fiber. The manufacturers of bread and breakfast foods have instituted advertising programs to sell newly developed, high fiber products.

What is not well known is the possibility of detrimental effects from increased crude fiber, specifically the excretion of zinc and other minerals from the body. This problem has been acknowledged in the medical literature only recently and has been slow to reach the popular literature and the advertising media.

The Food and Drug Administration bears the responsibility for monitoring (and regulating, if necessary,) the production and sale of food. To augment the information currently available, FDA asked for experimentation specifically designed to measure the decrease (if any) in zinc and other minerals in blood plasma as a result of the daily ingestion of 2.7 grams of crude fiber in addition to self-selected diet.

It is the objective of this report to describe in detail the design and suggested analysis for this experiment and to document the experimental protocol selected.

The hypothetical time trend shown in Figure 1 formed the basis of the plan. Measurements of plasma levels were to be obtained before treatment. Treatment was defined to be the daily ingestion of 2.7 grams of crude fiber derived from unprocessed bran, incorporated into muffins, date bread, and brownies. After a transition period of several weeks to allow serum levels to reach a new

equilibrium, further plasma determinations were to be obtained.

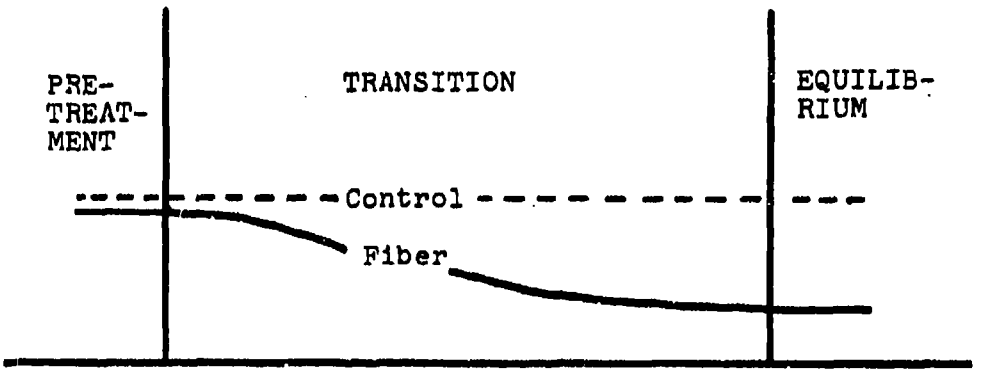


FIGURE 1. Hypothetical time trend of plasma zinc

The specific questions were: 1. duration of transition period; 2. number of subjects in the treatment group; 3. number of subjects in the control group on the same regimen but without bran; and 4. number of plasma measurements in the pre-treatment and equilibrium periods.

II. ESTIMATION OF SAMPLE SIZES. Neither our own experience nor the literature was helpful in answering objective #1: length of transition period. Our solution was arbitrary--12 weeks, a most conservative estimate to allow for complete transition. Two weeks were allotted for the pre-treatment baseline testing; two weeks were added for the equilibrium period, making a total of 16.

Objectives 2-4, how many subjects and how many periods, were approached simultaneously. The paradigm below shows the detailed experimental design and suggested analysis of variance, but does not specify how many subjects and how many periods.

		DIET	
		Pre-Treatment	Equilibrium
Treatment Group	SUBJECTS	Pds.	Pds.
Treatment Group	A	- - -	- - -
	B	- - -	- - -
	.		
	S	- - -	- - -
Control Group	a	- - -	- - -
	b	- - -	- - -
	.		
	.		
	s	- - -	- - -

A.V.
GROUPS
DIETS
GxD
SUBJECTS IN G
SxD
PERIODS IN D
GxP
SxP

It was convenient to consider the treatment group alone as an approach to suggesting the number of subjects and periods.

		DIETS: Pre-Treatment	DIETS: Equilibrium
Treat-	A	- - -	- - -
ment	B	- - -	- - -
Group	·		
	·		
	S	- - -	- - -

A. V.	EXPECTED MEAN SQUARE
SUBJECTS	$\sigma^2_{SP} + dp\sigma^2_S$
DIETS	$\sigma^2_{SP} + s\sigma^2_P + p\sigma^2_{SP} + ps\sigma^2_D$
SxD	$\sigma^2_{SP} + p\sigma^2_{SD}$
PERIODS IN D	$\sigma^2_{SP} + s\sigma^2_P$
SxP	σ^2_{SP}

The problem was to secure estimates of those variance components to be used to test the effect of Diets and to determine s and p. Design criteria were defined as follows: require that a difference in plasma level due to diet of as much as 10 ug/100ml be statistically significant at the .01 level with the power of the test set at .95. In terms of the non-central F-distribution, we have

$$\text{Non-Central} \quad F : \quad \phi^2 = \frac{\sum_1^k (\mu_i - \bar{\mu})^2/k}{\text{EMS/Sample Size}}$$

We set $k = 2$

$$\phi^2 = 3^2 \quad \text{for } \alpha = .01,\ \beta = .05,\ \delta = 10$$

$$\therefore \quad \phi^2 = \frac{\delta^2/2}{\frac{\sigma^2_{SD}}{s} + \frac{\sigma^2_P}{p} + \frac{\sigma^2_{SP}}{sp}}$$

or, solving for p,

$$p = \frac{\sigma^2_{SP}/s + \sigma^2_P}{\delta^2/2\phi^2 - \sigma^2_{SD}/s} \qquad \cdots \qquad (1)$$

Measurements of serum (not plasma) levels repeated in time for subjects on a steady-state but self-selected diet were available from Pekarek (72), but not in analysis of variance format. An approximate reconstruction of Pekarek's data in AV form is shown below.

	d.f.	MS	EMS
Subjects	98	1834	$\sigma^2_{SP} + p\sigma^2_S$
Periods in S	728	81	$\sigma^2_{SP} + \sigma^2_P$

If we assume that $p \doteq 827/99 = 8.35$, then $\sigma^2_{SP} + \sigma^2_P \doteq 81$ and $p\sigma^2_S - \sigma^2_P \doteq 1753$. These estimates were not out of line with those reconstructed similarly from other investigators, Davies (69), Pecoud (75), Halstead (74), and Nichols (76). However, there was a problem in changing scale from serum values to the expected equivalent in plasma levels.

A currently used conversion from serum to plasma means is a simple percentage drop: ΔX = plasma - serum = $X_P - X_S$ $\doteq X_S/1.16 - X_S = -.14X_S$.
A plot of s(X) vs $\bar{X}$ using both serum and plasma reports revealed the consistent relation: $\Delta s = \Delta X/5$ so that
$\Delta s \doteq -.0275X_S \doteq -2.8$ for typical serum levels of 100 ug/100ml. In terms of variances, the estimates become $\sigma^2_{SP} + \sigma^2_P \doteq 38.4$ (Plasma). Equation (1) requires estimates of σ^2_{SP}, σ^2_{SD}, and σ^2_P; thus far, the literature has yielded only $\sigma^2_{SP} + \sigma^2_P \doteq 38.4$. Table 1 contains values of s and p for a variety of relationships between σ^2_{SP}, σ^2_{SD}, and σ^2_P in an effort to "box in" a portion of hypersurface represented with the hope that impracticable values of s and p would be accompanied by unlikely values of the variances. Clearly a considerable degree of guessing was involved when the values of s = 14 and p = 3 were chosen from the center of Table 1. Thus, 14 subjects who would finish the experiment was a minimum requirement. A similar number was recommended for the control group with the emphasis on a greater number in the treatment group if absolute balance was not possible.

III. ALLOCATION OF SUBJECTS. Allocation of the 34 persons who answered the request for volunteers was based on a balance of height, weight, sex, level of physical activity, and a measure of body fat. The physical factors were combined to give an index number Y as follows:

$$Y = 2(t - T) + (w - W)/3, \text{ where}$$

t = triceps skinfold, mm; T = median skinfold for that age,sex;
w = weight, pounds; W = median weight for height,sex, and frame from the Metropolitan Life tables.

The index numbers Y were found to be reasonably related to a somewhat similar index constructed by Lamphier in Nichols(76). After

TABLE 1. Values of s, p to meet design criteria for various component ratios under the restraint, $\sigma_P^2 + \sigma_{SP}^2 = 38.4$

σ_{SP}^2/σ_P^2	$\sigma_{SD}^2/\sigma_{SP}^2$ = 1/3		2/3		1		3/2		2		3	
	s	p	s	p	s	p	s	p	s	p	s	p
1	20	3.9			20	4.4					20	7.5
	30	3.7			30	4.0					80	4.0
	40	3.6			40	3.3					200	3.7
3			8	4.2	12	4.0	18	3.6				
			12	3.1	16	3.0	22	3.0				
			16	2.6	20	2.7	26	2.8				
4	6	3.3	8	3.8	10	5.0	14	4.4	20	3.7	20	10
	10	2.4	12	2.7	14	3.0	18	3.1	25	2.9	25	5.0
	14	2.1	16	2.2	18	2.3	22	2.6	30	2.5	35	3.0
5					8	7.0						
					12	3.1						
					16	2.4						
9	4	4.7			8	6.6			16	4.9	22	6.5
	6	2.6			12	2.5			20	2.6	26	3.3
	8	2.0			16	1.8			24	2.0	30	2.4

ranking the subjects by their index number Y and according to their level of physical activity, adjacent subjects were allotted to groups at random. Neither the subjects nor the technicians who made the plasma determinations knew the group allocations; every precaution possible was employed to make it truly a blind experiment.

REFERENCES

Determination of serum zinc concentrations in normal adult subjects by atomic absorption spectrophotometry. Robert S. Pekarek, William Bisel, Peter Bartelloni, Karen Bostian. Am J Clin Path 57:506.1972

Measurements of plasma zinc. I. J. T. Davies, M. Musa, and T.L. Dormandy. J Clin Path 21:359. 1969

Effect of foodstuffs on the absorption of zinc sulfate. A. Pecoud, P. Donzel, & J. L Schelling. Clin Pharmacol & Therapeutics:17: 469. 1975

A conspectus of research on zinc requirements of man. J. A. Halstead, J. C. Smith, Jr., & M. I. Irwin. J of Nutrition 104:#3, 345 March 1974.

Independence of serum lipid levels and dietary habits: The Tecumseh Study. Allen B. Nichols, Catherine Ravenscroft; Donald E. Lamphiear; Leon D. Ostrander. JAMA 236: #17, 1948. Oct 1976.

Analysis of variance. Henry Scheffe. John Wiley & Sons. New York 1959.

FIELD VERIFICATION OF RADIATION CHARACTERISTICS OF RADARS

JL HARRIS
Aeroballistics Analysis Branch
Aeroballistics Directorate
Technology Laboratory
US Army Missile Research and Development Command
Redstone Arsenal, Alabama

ABSTRACT. This paper deals specifically with work done to determine from field test data, the radiation patterns of the radars of the Improved HAWK system. It does not attempt to treat the subject in general. The problem of data analysis is the underlying subject of this paper. Many problems were encountered when doing the analysis which would yield a radiation pattern. These are discussed. Some results are presented and conclusions are drawn. The conclusions deal with measures which will make the job of data analysis easier and quicker, and should apply generally.

I. INTRODUCTION. In 1975, from July to November, field tests were conducted with the radars of the Improved HAWK system. The tests were conducted at Naval Weapons Center (NWC), China Lake, CA. The tests were motivated by the Anti-Radiation Missile problem (ARM). The primary objectives and findings of the tests are not the subject of this paper. During the tests, data was collected from which the transmit patterns of the primary radars could be determined. Pattern data had been made available by the system prime contractor. This was data taken on a radar range, in a receive rather than transmit mode, and in a free space environment, to whatever extent this latter was achieveable. It was felt that the data taken under field test conditions should be processed to yield the patterns of the antennas in a transmit mode, in a natural environment (if China Lake can be judged natural), with multipath present. It was also felt that the data could be processed in such a way that it would provide a check point for a multipath model which had been developed. For these reasons an effort was started to develop the radar patterns from the data which had been collected.

II. DATA COLLECTED. Figure 1 shows the geometry of the tests and test set-up. An RF sensor was mounted in the gondola of a hot air balloon. The balloon was then permitted to rise to various altitudes and as the radar of interest was allowed to rotate with its main beam at a fixed elevation, the output of the RF sensor was recorded. Thus, the geometry of the radar relative to the balloon borne RF sensor was widely variable,

0 to 360 degrees of azimuth and from almost zero elevation up to about 60 degrees (the mechanical limit). Also mounted in the gondola of the ballon, and boresighted with the RF sensor, was an IR seeker, a television camera, and a riflescope. The riflescope allowed the operator to point the seeker cluster toward the radar. The TV camera provided a visual record of where the seeker cluster was pointed. The IR seeker provided a quantitative history of where the seeker cluster had been pointed because an IR source was provided at the radar and the seeker was gimballed and free to track the IR source. The IR seeker gimbal angles provide a record of the pointing error of the seeker cluster.

The quantities which were recorded are the intensity output, the two (right-left and up-down) direction finding outputs, and a status indicator from the RF sensor; the two gimbal angles from the IR sensor; and north marks from the radars. Also, the geometric data to relate the balloon position to the radar position was recorded. This was named the "Call Out" data because of the way it was collected and recorded. A person was stationed with a sextant and he kept sighting on the balloon and calling-out the balloon azimuth and elevation. Someone would write it down in the log with time of occurrence. The balloon operator would observe range lines painted on the ground and call out the range and someone would write it down. There was also data from an altimeter to be called out and written down. This handwritten log was the only source of the"Call Out" data. The other data was recorded on FM tape, copies of which were furnished to MICOM for use in data analysis. Copies of the log were also furnished. Some of the FM tapes were digitized and copies of these were furnished to MICOM.

III. ANALYSIS APPROACH. The problem with analysis was not so much a problem of approach as of retreat. As soon as some of the digitized tapes were available at MICOM, people began to be solicted to "do something" with the data. One young man started to do something with the data and found that some of the digital tapes could not be read at all, the rest were digitized at only 20 samples per second, and that there were chronic tape reading problems with the computer system which he had chosen to use. Being a very capable and many faceted individual, he quickly found something else to do and has been busy ever since. So it went, for about a year. Then the author was solicited to "do something" with the data, and got stuck with it. To abbreviate the story, the data tapes were digitized by the Test and Evaluation Directorate of the Research, Development and Engineering Laboratory of MICOM. The digitization rate was 100 times per second, and tapes were generated which were compatible with the CDC 6600 computer system which was chosen for the analysis.

No big problems have been encountered with this part of the effort, just communications.

Only carefully selected portions of the FM tapes have been digitized because of the large amount of data which exists. For a segment of interest, chosen with the aid of the test conductors' log, the digitized tapes provide the outputs of both the sensors and the radar north marks, as a function of time. The test conductor's log is used to make a table of balloon elevation angle, azimuth angle, and range as a function of the same time base. These are entered into a computer program which reads the tape, and then calculates the relative geometry which existed for every time recorded on the digital tape. To represent the radar intensity pattern as a function of the relative geometry, angular space was divided into cells which were 1 degree of elevation and 3 degrees of azimuth. All samples occuring in a particular cell were then averaged and a standard deviation calculated. The quantities processed were the intensity output (which indicates radar pattern) and the direction finding outputs of the RF seeker. The latter provide information about the multipath situation. The number of samples which occurred in each cell was also recorded. For a particular radar, data from several different days of testing were lumped together if the RF conditions were the same.

IV. STATUS. The only analysis which has yet been done is that just described. No time series approach or spectral analysis approach has been attempted. The most complete set of results is for the low altitude search radar. Much less data was recorded for the illumination radar. No analysis has yet been done with data from the high altitude search radar.

V. RESULTS.

A. Radiation Pattern. Figure II shows a three dimensional plot of the intensity data from the RF sensor, for the low altitude search radar. In this figure 0. relative azimuth means that the radar beam is pointed in azimuth toward the balloon. Negative relative azimuth is to the right. The elevation is balloon elevation angle above the radar beam. Note that the intensity scale is not provided here. It can be seen that the most power is with the main beam pointed toward the balloon and that power decreases with balloon elevation. Data for main beam on the balloon is not shown here and was not taken in this test. Other places where no data is shown are at high balloon elevations where none was recorded, and at a few orientations where there was insufficient received power at the RF sensor. Figure III shows a representation of data from the contractors tests. The intensity scale is again unspecified, and is different from previous figure. The thing which seems worthy of note here is that the intensity levels in some regions are approximately the same, but the patterns measured by the contractor show much steeper gradients. Indeed, the plot is full of spikes. There

is higher intensity in a quite narrow region at zero relative azimuth for all elevation angles shown. Within about ten degrees to either side of this region the intensity drops abruptly down to a region which is approximately 180 degrees total width. In this region, the intensity spikes seem randomly scattered and their height decreases roughly linearly as the edge of the region is approached. Another striking difference between the two plots is the shape variation with azimuth at a particular elevation. The field test data is high at zero azimuth, drops for a few degrees to each side of zero, then rises again and drops again. Some behavior of this nature can be seen at all elevation angles. The contractors data shows this sort of tendency only at approximately 40 degrees elevation angle, and in regions approximately 90° to either side of zero azimuth. Still another difference is that the field test data shows intensity to decrease consistently with balloon elevation, but the contractors data does not change much with elevation angle, except at the 40 degrees elevation angle just discussed. The contractors data was based upon a single set of measurements and no averaging was done. Consultation with people who are experts in the field has revealed that there may be a good deal of randomness in the structure of a radiation pattern determined from a single set of measurements. In other words, if the measurement set were to be repeated by the contractor, the radiation pattern would not be duplicated, but should have the same general characteristics. If several measurement sets were averaged together, then the resulting pattern should be much more similar to the pattern determined by averaging field test results, as I have done. This argument would lead to a conclusion that the examination of the field test data on a scan by scan basis should reveal a rapidly changing intensity history as the various radiation spikes are oriented toward the balloon. The field test data has been inspected on a scan by scan basis and the intensity variation within a scan does not appear to be of this rapidly changing nature. In fact, many of the scans have the same characteristics as the plot of the averaged data. Figure IV shows three scans of this data. It is thought that the data recording process (the RF sensor, telemetry process, and stripchart recorder) do not introduce enough filtering to prevent response to intensity spikes. But effort is being put forth to determine whether or not this is true.

B. Multipath Model. The multipath model validation effort will now be discussed. The model hypothesizes that multipath is produced by a diffuse type of reflection of the main beam radiation of this radar. For some radars it might be necessary to include other high level lobes also. It must be emphasized that diffuse rather than specular radiation is assumed. The main lobe is assumed to "paint" a swath of ground as illustrated in Figure V. This area then becomes a distributed radiator. The model calculates the area and centroid of the swath and using empirically derived

data taken at NWC in a previous test, calculates the power radiated from the swath of ground. Assuming this power effectively originates from the area centroid of the swath, the centroid of the direct path and the multipath radiation combined can be calculated. Obviously this centroid will be at some point displaced from the radar on a line toward the centroid of the swath of area. The model would then predict that the sensed emitter location would revolve around the actual radar location at the rotation rate of the radar. If the RF sensors were directly overhead, the sensed emitter location would be on a circle and the azimuth and elevation components of the angular error would be equal. In the general case the elevation angular error is smaller because the circle appears to be elliptical when viewed at an angle. This multipath error model has not been extensively validated. One objective of analysis of the field test data is to validate the model, or to discover its short comings. Figure VI shows idealized error plots for this low altitude search radar, at a particular balloon elevation. When the radar main beam is 90 degrees to the right of the balloon, the azimuth channel error should be a maximum value and to the right, while the elevation channel error should be zero. When the main beam is pointed toward the balloon (or away from it), the elevation channel error should be a maximum, and down (or up) and the azimuth channel error should be zero. Figure VII shows a three dimensional plot of the azimuth channel error from the field test. It is to be noted that at a particular elevation angle the error behaves in the same manner as the idealized error of Figure VI. Figure VIII shows the elevation channel error, where again the behavior is as the model would predict, in a qualitative sense.

The preceding figures have demonstrated that multipath seems to originate by diffuse scattering of radiation from the radar main beam on the ground, because this assumption seems to describe what was observed in the field test. The comparison is qualitative, however. The multipath model has not been exercised to see to what degree it will reproduce the field test results. To do this, a good representation of the radar pattern is needed. At this point it is not clear what to use. The pattern from the field test data contains an intensity contribution from the multipath, and there are no measurements which are free of multipath, except the contractors measurement. These look a good bit different from the pattern derived from the field test, and it is felt that the difference cannot be attributed to the multipath power alone. Also, these would be very difficult to represent.

The next step toward validation of this model is likely to be the calculation of the multipath intensity contribution for each angular cell where field test data was collected, using the model as is. This intensity can then be subtracted from the intensity measured in this field test and the difference taken as the radar contribution. The multipath model can then be used to produce error data for all geometrics of the field test, and this compared to the error data from the field test. An iterative process could be used to refine the model.

V. PROBLEMS. There exists the problem of the radiation pattern being different from that measured by the contractor. On one hand, there is the opinion that if the contractors facility does not yield the same results as field tests, then it's no good at all. The other extreme of opinion is that the agreement is as close as should be expected.

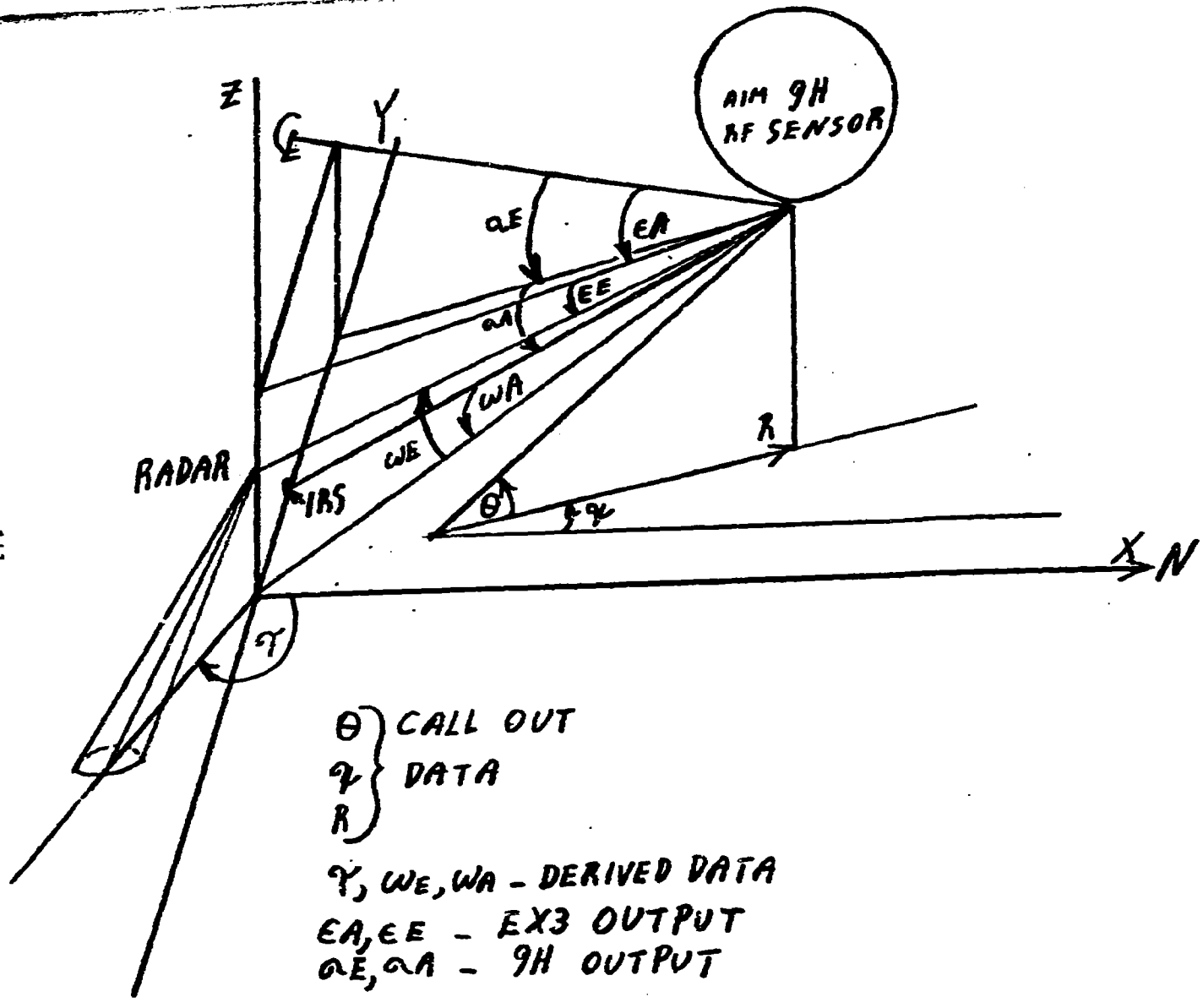
The number of samples which have been averaged to find mean intensity, and mean angular error components, is variable. Near the lower and upper extremes of balloon elevation, fewer samples were taken. This contributes to the raggedness of the estimates in those regions. In regions where the received intensity was low there are also fewer samples. In this case, there is a double contribution to the raggedness of the estimates because the RF sensor noise becomes more important at low signal. But, all the available data has been used.

VII. CONCLUSIONS.

A. A data reduction/analysis plan should be prepared prior to the test.

B. Where exchange of magnetic tapes is contemplated, it would be very good to verify compatibility with a pre-test sample.

C. The person or persons who will ultimately end up doing the analysis should be intimately involved in test planning, determination of data requirements, and perhaps the conduct of the test. At a minimum, he should observe some typical portions of testing.



θ } CALL OUT
φ } DATA
R

γ, ω_E, ω_A - DERIVED DATA

εA, εE - EX3 OUTPUT

αE, αA - 9H OUTPUT

FIGURE 1 GEOMETRY

INTENSITY FROM FIELD TEST

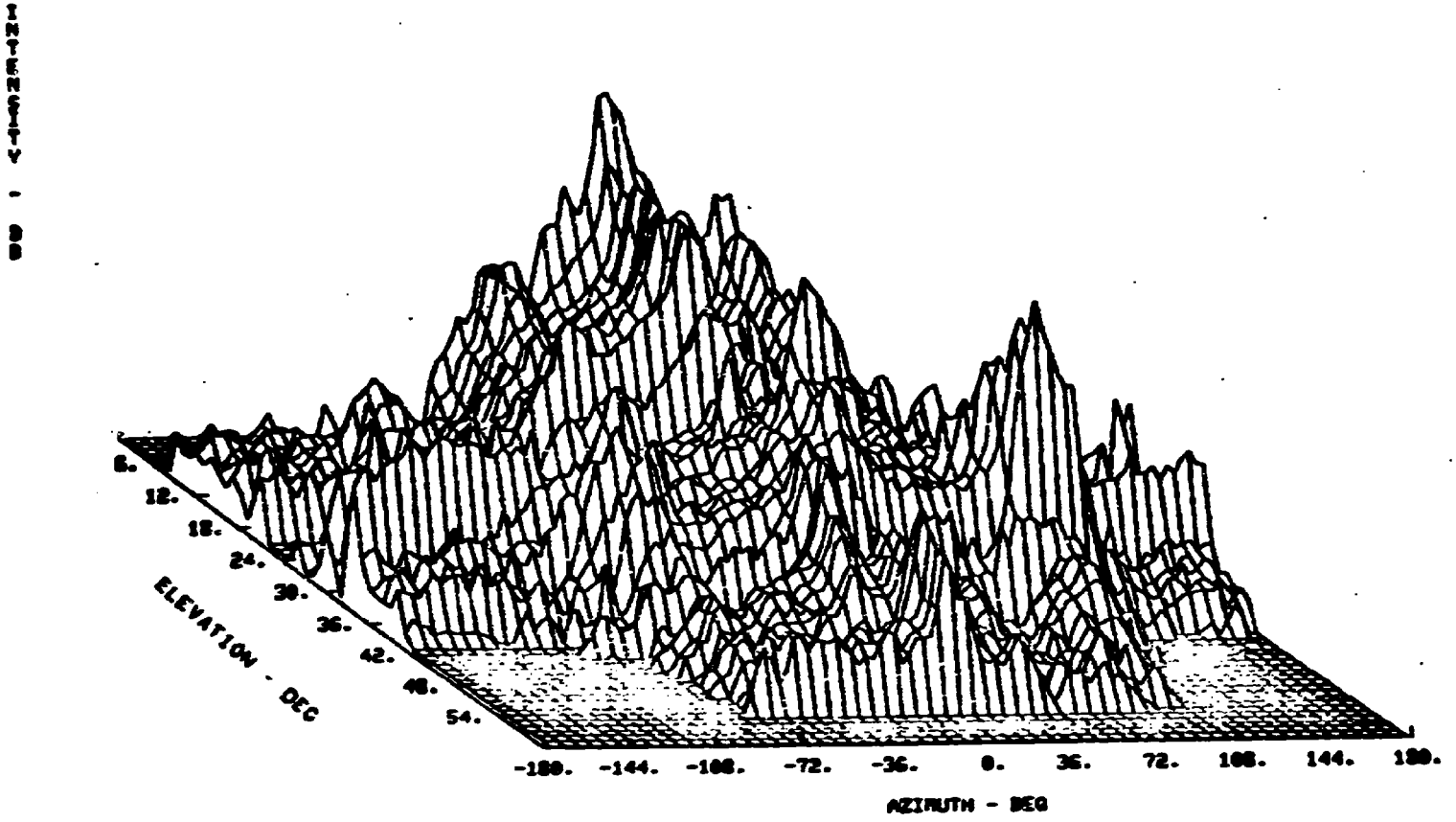


FIGURE 2

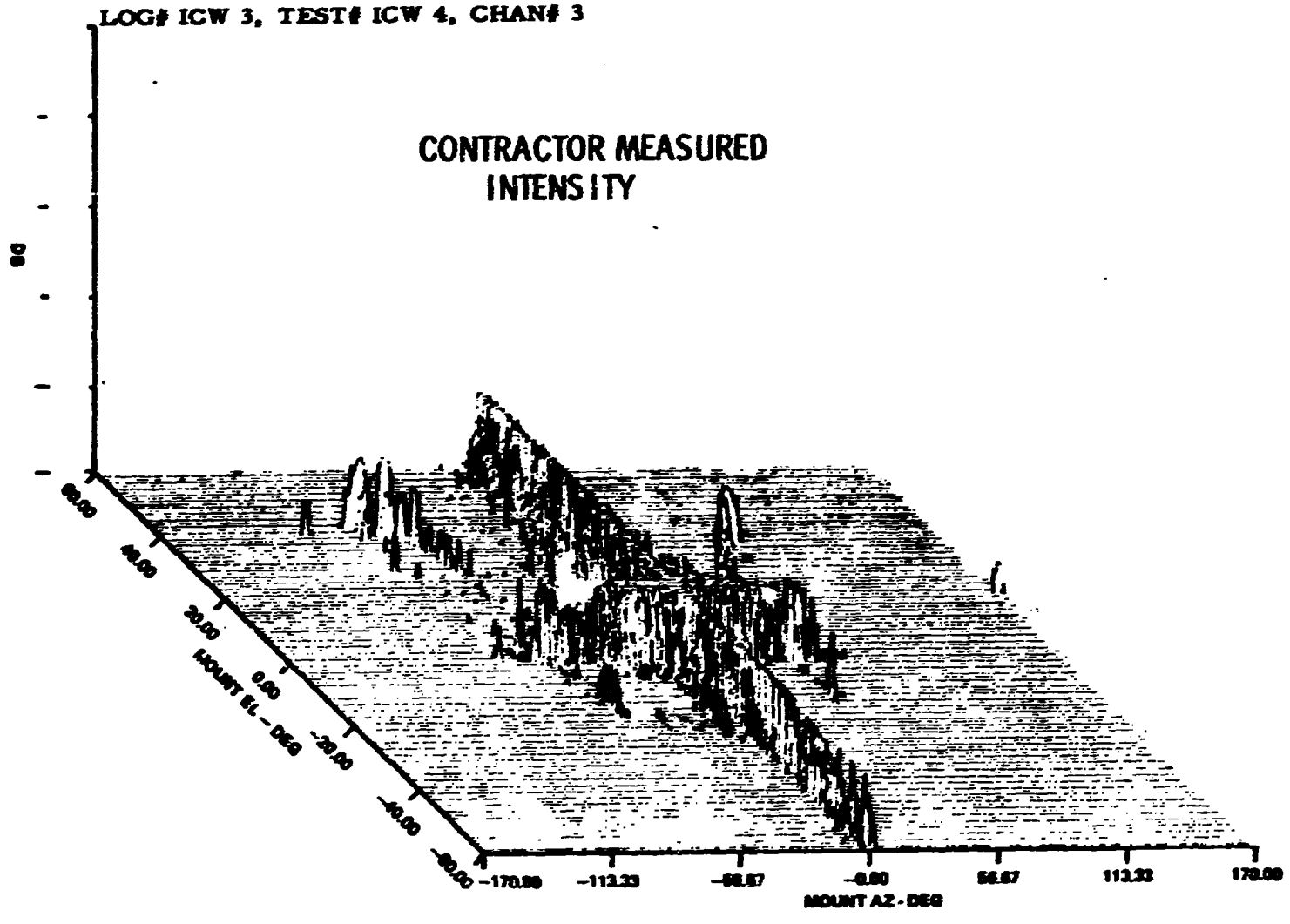
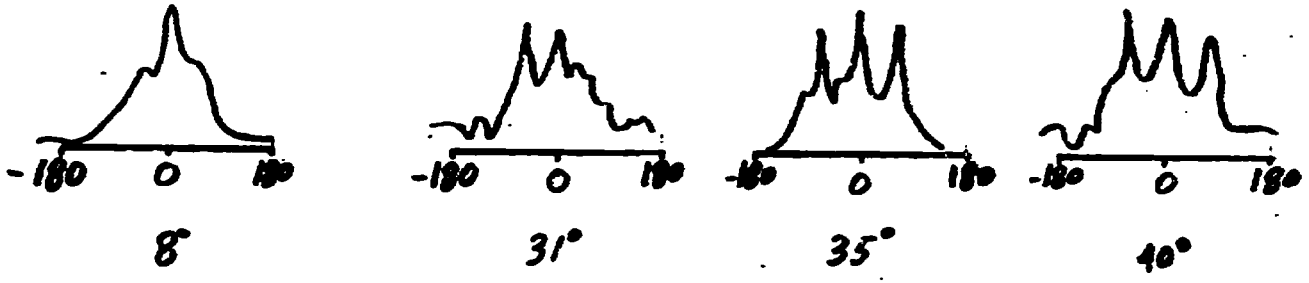


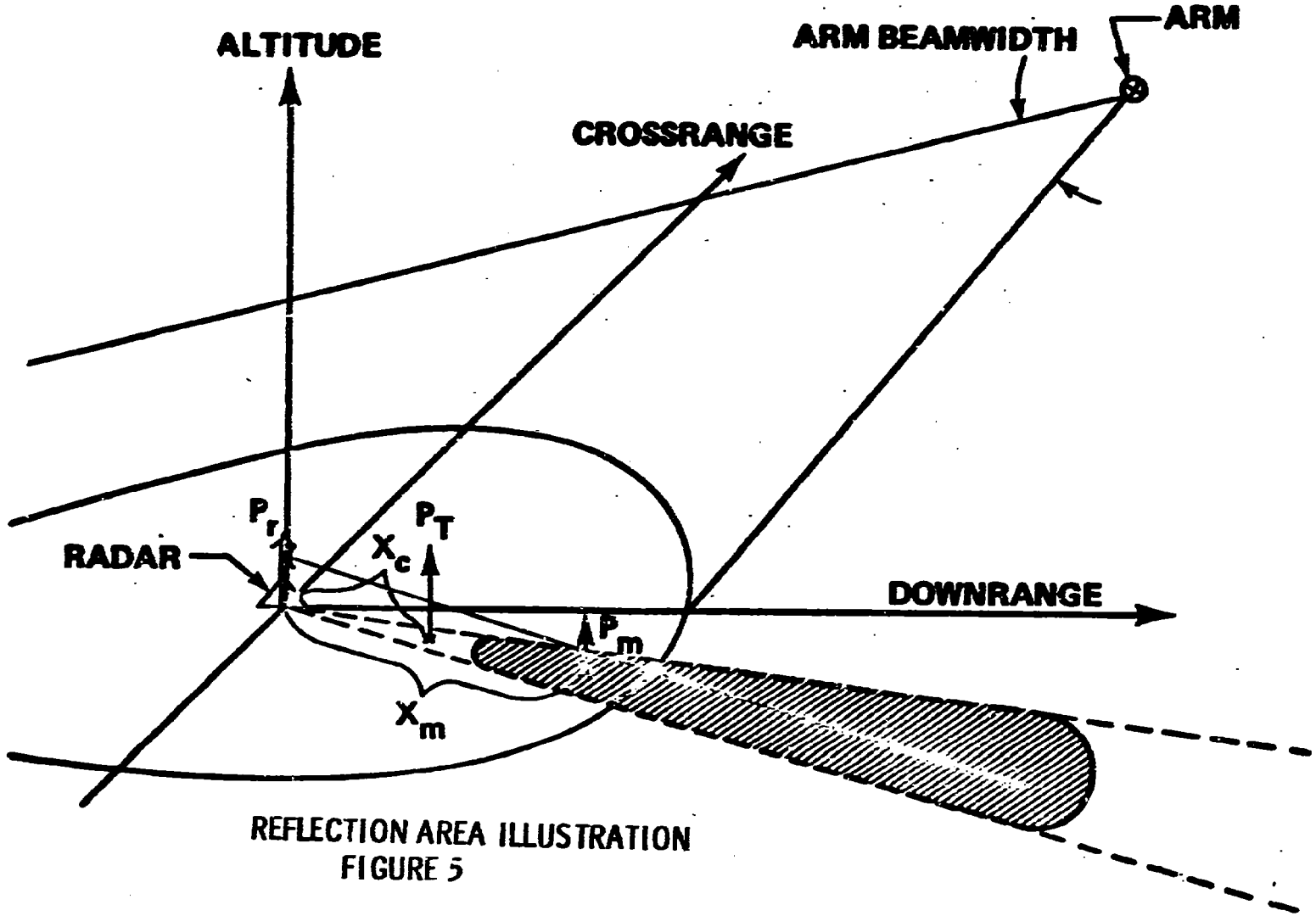
FIGURE 3 . ICWAR TRANS. PRIMARY POL. 3-D CONTOUR @ MID-FREQ.

SINGLE SCAN INTENSITY DATA



BALLOON ELEVATION ANGLE

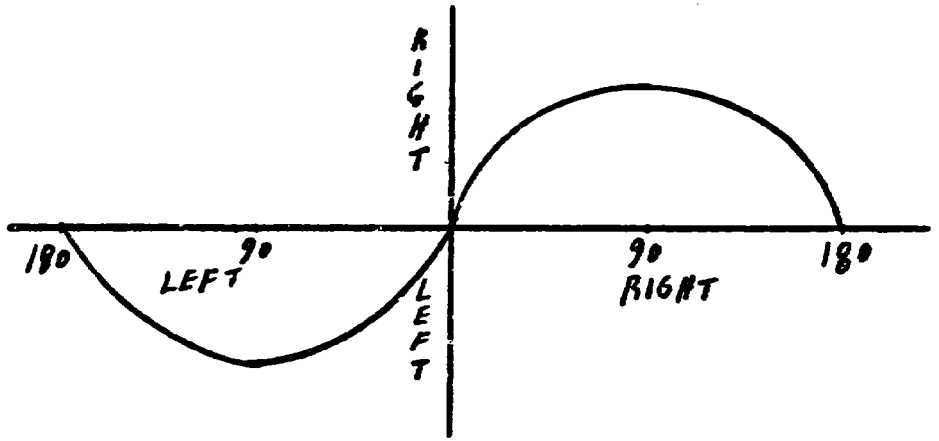
FIGURE 4

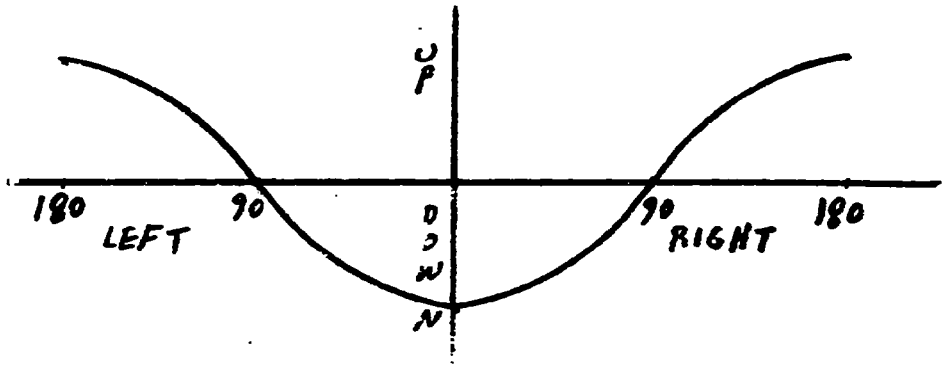


REFLECTION AREA ILLUSTRATION
FIGURE 5

IDEALIZED ERROR PLOTS

AZIMUTH ERROR





ELEVATION ERROR

FIGURE 6

AZIMUTH PLANE ERROR

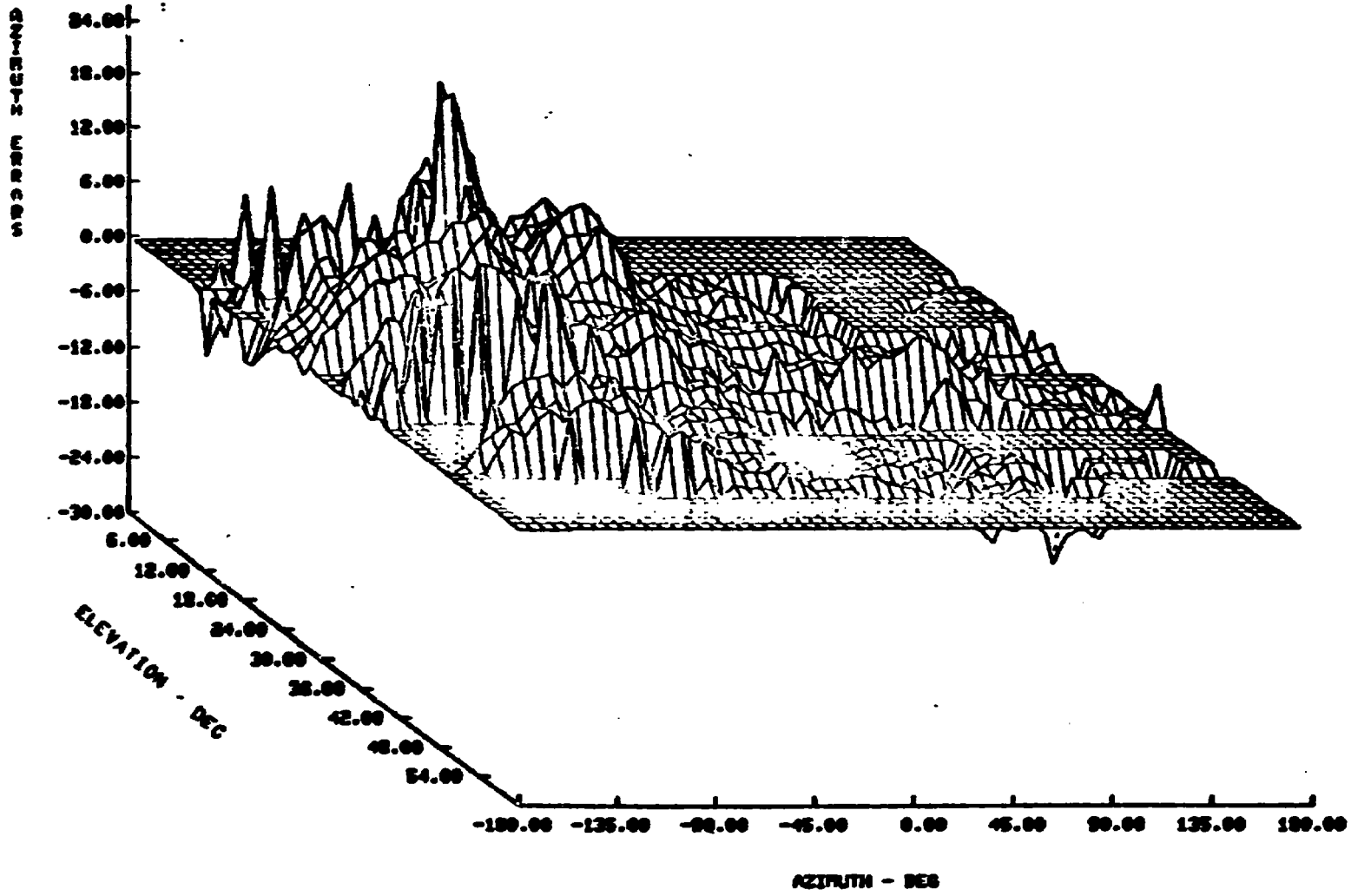


FIGURE 7

ELEVATION PLANE ERROR

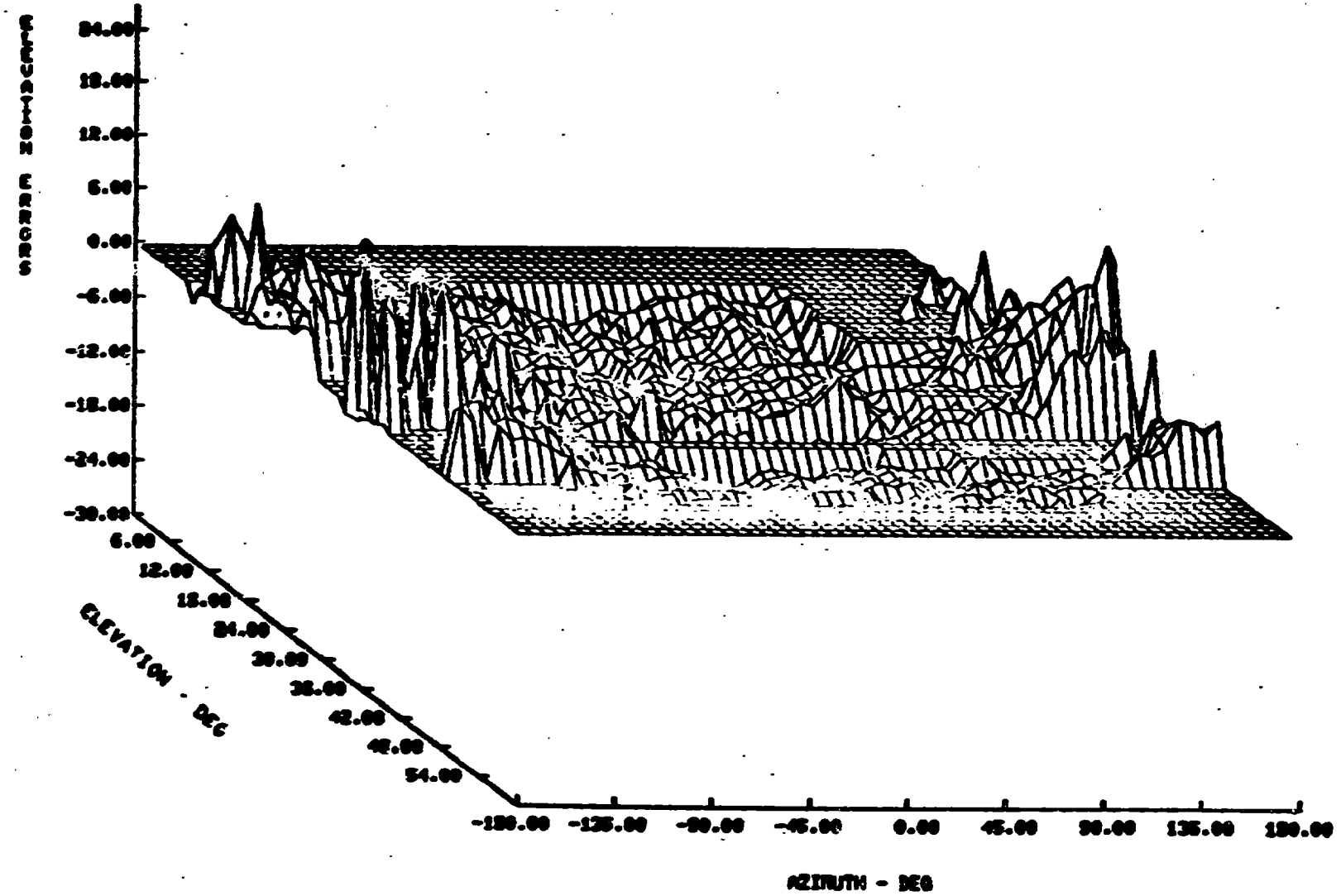


FIGURE 8

CONSTRUCTION OF CONFIDENCE LIMITS IN A NONLINEAR REGRESSION

C. MAXSON GREENLAND
LYNN H. DAVIS
SYSTEMS ASSESSMENT OFFICE
Chemical Systems Laboratory
Aberdeen Proving Ground, Maryland

ABSTRACT. This problem was presented in a clinical session at the Twenty-Third Conference on the Design of Experiments. It arises from the need to assess the uncertainties associated with calibration curves which have been fitted to observed data. The discussion includes a particular nonlinear model for the curve, the regression procedures, and several attempted methods for calculating $100(1-\alpha)\%$ confidence limits for the curve. A detailed description is given of an approach outlined by panel members to whom the problem was presented. Finally, a complete example is given, including graphical representation of a portion of a $100(1-\alpha)\%$ confidence region in the parameter space, and a description of the computer work necessary to obtain numerical results.

1. BACKGROUND. Sensitive electronic analyzers which are now in use are capable of measuring very low concentrations (on the order of 1-15 nanograms per milliliter) of chemical substances in solution. The uncertainties inherent in the development of calibration curves for this type of equipment assume great importance in quantitative analyses of highly toxic materials. At a given significance level, α, a properly-constructed confidence band for a calibration curve is the basis for obtaining interval estimates of concentration x (the independent variable) for an observed value y (the dependent variable) of the analyzer output. An interval of particular interest is determined by the intersection of the upper confidence limit curve and the Y-axis. This point, y_c, is called the decision limit since an observed instrument response of less than or equal magnitude has a non-negligible probability of having been produced by a zero X-value. The X-value, x_D, corresponding to y_c and determined by the lower confidence limit curve is called the detection limit, the lowest value of X which can be distinguished from zero. Hence, for concentration measurements at a significance level of α, x_D

is the lowest concentration which can be detected, and y_c is the lowest reading which distinguishes between the presence or absence of a chemical substance. These relationships, which have been discussed by Hubaux and Vos (ref. 1), are illustrated for a hypothetical nonlinear curve in Figure 1.

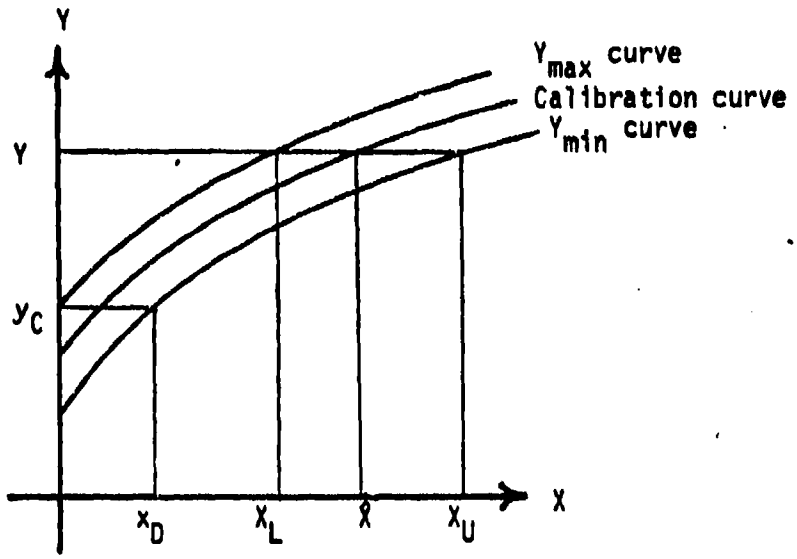


Figure 1. Calibration curve and confidence band; minimum (X_L), maximum (X_U) and regression value ($\hat{X}$) corresponding to chart reading (Y); decision limit (y_C) and detection limit (x_D).

II. REGRESSION PROCEDURES. The calibration data for analyzer instruments used in a recent Chemical Systems Laboratory study demonstrated a configuration similar to Figure 1, where the abscissa represents the concentration in nanograms/milliliter (1 ng = 10^{-9} grams), and the ordinate represents observed chart readings.

Because of time constraints, the first calibration curves were developed by means of linear interpolation between points. Later, when more time became available, a model of the form $y = a + b\ln x$ was examined; it had the approximate configuration of the data plot and was linear in the parameters a and b, but y decreases without bound as x approaches zero. In order to translate the curve to the left so that the point (0,10) falls

reasonably close to the curve, the following modified equation was tried:

$$y = a + b\ln(x+k)$$

The unknown parameters were obtained as follows:

1) The parameter k was estimated; then $z=\ln(x+k)$ transformed the model into $y=a+bz$, which is linear in a and b.

2) The three parameters (p=3) of the regression curve were determined by the method of least squares.

3) The value of k was varied in increments of 0.1, and new fits were calculated by means of an HP25 handheld electronic calculator until a maximum value of the correlation coefficient was obtained.

4) The equation which produced the greatest value for the correlation coefficient was the model selected for the calibration curve.

A representative example of eight data pairs (n=8) resulted in the following regression equation:

$$y = -38.405 + 41.167 \ln (x+3.2)$$

and a correlation coefficient of 0.9996. The standard error of the estimate is:

$$S_E = \frac{1}{n-p}\sqrt{\sum(Y_i - a - b\ln Z_i)^2}$$

$$= 0.9131$$

Then, following a procedure described by Natrella (ref. 2), $100(1-\alpha)\%$ confidence intervals were calculated for the inverse function

$$X = \exp\left[\frac{1}{b}(\overline{Y}'-a)\right]-k$$

where $\overline{Y}'$ is the average of n' chart readings. The equation for the interval, which was also programed on the HP25, is

$$X' = \overline{X} + \frac{b(\overline{Y}'-\overline{Y})}{C} \pm \frac{t_{1-\alpha/2}\, s_Y}{C}\sqrt{\frac{(\overline{Y}'-\overline{Y})^2}{S_{xx}} + \left(\frac{1}{n}+\frac{1}{n'}\right)C}$$

where $C = b^2 - (t_{1-\alpha/2})^2 S_b^2$

$\overline{X}$ = mean of the observed X-values

$\overline{Y}$ = mean of the observed Y-values

$S_{xx} = \sum X_i^2 - (\sum X_i)^2/n$

s_Y = standard error of the estimate of Y

n = number of calibration observations

n' = number of new observations of Y

s_b = standard error of the estimate of b

Although this appears to be a somewhat more refined approach than successive linear interpolations, several theoretical objectives occur:

1) There is no physical reason to assume an underlying logarithmic relationship between concentration and the electrical output of the analyzer.

2) The equation cannot be transformed to one which is linear in the parameter k.

3) The size of the increments applied to k was arbitrarily chosen.

4) The correlation coefficient is a questionable criterion of selection of the parameter values.

These considerations led to a search for improved procedures.

In this instance, the analyzer operates on the principle of light absorption. The intensity of light transmitted through a sample of the solution is inversely proportional to concentration and affects the output of a photocell, which causes the deflection of a continuously-recording pen. The process of radiation absorption is described by the Beer-Lambert Law:

$$I = I_0 e^{-klx}, \text{ where } \begin{cases} I_0 = \text{intensity of light before transmission} \\ I = \text{intensity of transmitted light} \\ k = \text{absorption coefficient} \\ l = \text{length of light path through solution} \end{cases}$$

Assuming a simple linear relationship between intensity of transmitted light and instrument reading leads to the following:

$$\begin{aligned} y &= \alpha + bI \\ &= \alpha + bI_0 e^{-klx} \\ &= \alpha + \beta\gamma^x \end{aligned}$$

Note: The symbol α for the parameter should not be confused with the symbol α for the statistical significance level.

At Chemical Systems Laboratory (CSL) there is available an International Mathematics and Statistics Library (IMSL) subroutine (ref. 3) which estimates α, β and γ for this function, calculates the standard error of the estimate (S_E), and determines the variance-covariance (VCV) matrix for α, β and γ. Only partial details of this proprietary procedure are available, but an estimate of γ is determined iteratively to a specified accuracy using a Fibonacci technique. Then α and β are determined by the method of least squares. When this example was run on the UNIVAC 1108 computer at CSL, ten iterations of the subroutine gave the regression equation

$$y = 92.394 - 81.868\ (0.88352)^X$$

and $S_E = 0.3167$, which is approximately one-third the value of S_E obtained for the logarithmic model.

A method described by Snedecor and Cochran (ref. 4), based on a Taylor's series expansion of the function, is particularly satisfactory if good initial estimates of the parameters are available. Consider y as a function of γ:

$$y = f(\gamma) = \alpha + \beta\gamma^X, \text{ where } \gamma \in [k,1] \text{ and } 0<k<1.$$

If f is continuous on $[k,1]$ and differentiable on $(k,1)$, and if $r_1 \in [k,1]$, then for each $\gamma \in (k,1)$

$$f(\gamma) = f(r_1) + (\gamma - r_1)f'(r_0), \text{ where } \gamma < r_0 < r_1$$

If r_1 is chosen very close to γ, the following approximation holds:

$$f(\gamma) \approx f(r_1) + (\gamma - r_1)f'(r_1)$$

Therefore,

$$\begin{aligned} y &\approx \alpha + \beta r_1^X + (\gamma - r_1)x r_1^{X-1} \\ &= \alpha X_0 + \beta X_1 + \lambda X_2 \end{aligned}$$

where $X_0 = 1$, $X_1 = r_1^X$, $X_2 = x r_1^{X-1}$, and $\lambda = \beta(\gamma - r_1)$. If the above equation were exact, it would be possible to obtain estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$ of the coefficients α, β and λ; $\hat{\gamma}$ could then be calculated. The truncation of the Taylor's series introduces an error; hence, the calculated values are estimates a, b and c of the estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$, respectively. It is then possible to use procedures applicable to multiple linear regression to fit the model

$$Y = aX_0 + bX_1 + cX_2$$

where $c = b\,(r_2 - r_1)$

If X is the matrix of observed values of the X_i, i.e.

$$X = \begin{pmatrix} 1 & r_1^{x_1} & \cdots & x_1 r_1^{x_1-1} \\ 1 & r_1^{x_2} & \cdots & x_2 x_1^{x_2-1} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 1 & r_1^{x_n} & \cdots & x_n r_1^{x_n-1} \end{pmatrix}, \quad A = \begin{pmatrix} a \\ b \\ \\ c \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix},$$

and X' is the transpose of X, then the matrix equation X'XA = X'Y can be solved for A:

$$A = (X'X)^{-1} X'Y$$

Then, if a, b and $r_2 = r_1 + \frac{c}{b}$ are substituted into the original equation for α, β and γ, respectively, S_E can be calculated. The procedure is repeated until S_E reaches a minimum value. In this example the final equation is

$$y = 91.263 - 80.916\ (.88012)^x$$

and $S_E = 0.2581$, a further reduction in the standard error of about 19%. This two-step approach was used to fit all the calibration curves in the CSL study.

Since the calibration curves are used to obtain concentration values from chart readings, the inverse function

$$x = \frac{\ln(y-\hat{\alpha}) - \ln\hat{\beta}}{\ln\hat{\gamma}}$$

$$= \frac{\ln\left[(y-\hat{\alpha})/\hat{\beta}\right]}{\ln\hat{\gamma}}$$

provides the necessary transformation.

III. CONFIDENCE REGIONS

A. Preliminary Calculations

The inverse matrix of Gauss multipliers, $(c_{i,j}) = (X'X)^{-1}$, was used to calculate the standard error of each coefficient:

$$s_a = S_E \sqrt{c_{11}} = 0.7185$$

$$s_b = S_E \sqrt{c_{22}} = 0.6528$$

$$s_r = \frac{S_E c}{b} \sqrt{\frac{c_{33}}{c^2} + \frac{c_{22}}{b^2} + \frac{2c_{23}}{bc}}$$

$$= S_E \sqrt{c_{33}}/b = 0.00213$$

Then $100(1-\alpha)\%$ confidence limits for each estimated parameter are given by

$$a \pm t_{\alpha,n-p} s_a$$

$$b \pm t_{\alpha,n-p} s_b$$

$$r \pm t_{\alpha,n-p} s_r$$

A simultaneous $100(1-\alpha)\%$ confidence band is required, i.e., a confidence band which will contain the calibration curve $100(1-\alpha)\%$ of the time. Breiman (ref. 5) has shown that individual $100(1-\alpha)\%$ confidence intervals for k parameters, form $100(1-k\alpha)\%$ simultaneous confidence intervals. Hence, a 95% simultaneous confidence region for the three parameters α, β and γ represents 98.3% individual intervals. Then $t_{0.017,5} = 3.5$, and

$$\alpha \in [88.748, 93.778]$$

$$\beta \in [-83.201, -78.631]$$

$$\gamma \in [0.87266, 0.88758]$$

By selecting combinations of the parametric values within these intervals which give maximum and minimum values for y, a 95% simultaneous confidence region for y was calculated. The procedure is relatively crude and leads to fairly wide intervals. The detection limit is approximately 1 ng/ml, and the interval estimates become wider at the higher concentrations. A method is needed to calculate improved (more restricted) confidence regions, if possible.

B. Suggested Procedure

The following method for determining a $100(1-\alpha)\%$ simultaneous confidence band for the calibration curve $y = \alpha + \beta\gamma^x$ developed from suggestions of the panel members to whom this problem was presented at the Twenty-Third Design of Experiments Conference in Monterey, CA. A new model, $y = 91.269 - 80.921(.88014)^x$, based on 24 calibration points,

fitted as previously described, is introduced here.

In the preceding section 100(1-α)%, individual confidence intervals were calculated for each parameter. A set of 3 linearly independent unit vectors, A, B, Γ can be considered as an orthornormal basis for a vector space P called the parameter space. Every point of P can be written as a linear combination of A, B and Γ, i.e., as a 3-tuple of real numbers (α, β, γ). All points of P whose components lie within the separate confidence intervals determine a rectangular parallelopiped in P. An extension of a method described by Draper and Smith (ref. 6) permits the further restriction of the points (α, β, γ) to a subset which represents an approximately 100(1-α)% simultaneous confidence region C for the three parameters.

All points must satisfy the equation

$$S(\alpha, \beta, \gamma) = S(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \left[1 + \frac{p}{n-p} F(p, n-p, 1-\alpha)\right]$$

where $S(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}\hat{\gamma}^{Xi})^2$

and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ is the point in P whose components are the parameter values of the calibration curve. Hence, the right side of the equation is a real-valued number, S, which is a function of (i) the sum of the squared residuals of the calibration curve, (ii) the number of parameters to be estimated, p (here p=3), (iii) the number of data pairs, n (here n=24) and (iv) the confidence level, 1-α (here α=0.05). Expanding the equation,

$$\begin{aligned} S(\alpha, \beta, \gamma) &= \sum_{i=1}^{n} (Y_i - \alpha - \beta\gamma^{Xi})^2 \\ &= \sum_{i=1}^{n} \left[(Y_i - \alpha) - \beta\gamma^{Xi}\right]^2 \\ &= \sum_{i=1}^{n} \gamma^{2Xi} \cdot \beta^2 - 2\sum_{i=1}^{n} (Y_i - \alpha)\gamma^{Xi} \cdot \beta + \sum_{i=1}^{n} (Y_i - \alpha)^2 \\ &= A\beta^2 - 2B\beta + C \\ &= S \end{aligned}$$

Since $A\beta^2 - 2B\beta + (C-S) = 0$,

$$\beta = \frac{-(-2B) \pm \sqrt{4B^2 - 4A(C-S)}}{2A}$$

$$= \frac{B \pm \sqrt{B^2 - A(C\ S)}}{A}$$

where $A = \sum_{i=1}^{n} \alpha^{2xi}$, $B = \sum_{i=1}^{n} (Y_i - \alpha)\ \gamma^{xi}$, and $C = \sum_{i=1}^{n} (Y_i - \alpha)^2$.

Since for this example it is simpler to calculate the β-values from α and γ, the 3-tuples will be denoted by (α,γ,β) to conform to the usual coordinate convention (x,y,z) in three-dimensional drawings. A visualization of the parameter surface is achieved by use of the fact that every point (α,γ,β) of *P* which lies on the surface or interior to it has real-valued conponents. As α was held constant the two real β values were calculated for successively incremented γ values; the process was repeated for successive α increments. To obtain sufficiently small initial values and sufficiently large final values for α and γ to bracket the entire surface it was necessary to widen the 98.3% individual confidence intervals by about 25%. In this example, the 98.3% confidence intervals for α and γ are [90.031, 92.508] and [0.87647, 0.88382], respectively; the intervals [89.800, 92.909] and [0.87544, 0.88454] are sufficiently large to include the α and γ values which apply to C. Increment sizes of 0.01 and 0.0001 for α and γ, respectively, require approximately 3 minutes central processor time to produce approximately 11,500 points of C. Figure 2 illustrates the mapping procedures, using a hypothetical sphere as an example.

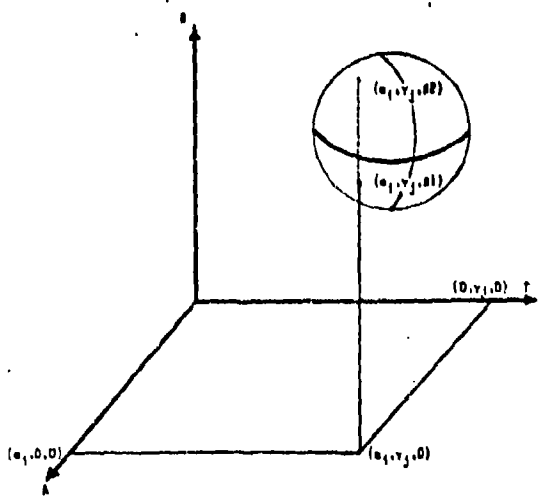


Figure 2. Illustration of the method used to determine the points on the surface of a solid figure.

A package of FORTRAN callable subroutines, the Perspective Plotting System, is available at CSL for producing drawings of perspective views of three-dimensional objects. Appropriately ordered arrays of data values are necessary as input to this graphics package. The array of approximately 11,500 β-values for the region C, which was created as described from the (α, γ) pairs, contains values between -82.311 and -79.620; hence, C lies in the negative β half-space. Attempts to achieve graphical representations of C using the calculated array were only partially successful because of limitations of the Perspective Plotting System. For example, the coordinates for an acceptable "observer's position" must be selected, and plots of closed solid figures are not now possible. However, a perspective drawing of approximately one-half of the region was produced by means of a two-step transformation on the coordinates. First, subtraction of the centroid coordinates from each point $(\alpha, \gamma, \beta 2)$ translated the upper portion of C to the vicinity of the origin in P. Second, a transformation matrix applied to the translated points rotated the figure in such a way that the vector $[\alpha_{max}-\hat{\alpha}, \gamma_{max}-\hat{\gamma}, \beta_{min}-\hat{\beta}]$ is rotated into the AΓ-plane. The net effect is approximately a one-to-one linear mapping of the points $(\alpha, \gamma, \beta 2)$ onto points $(\alpha', \gamma', \beta')$, where $\beta' \geq 0$.

A computer graphics drawing of this object was produced by a Tektronix 4051 Graphic System using approximately 2700 points. The final version shown in Figure 3 was produced by means of a CalComp Pen Plotter. The true scales of the A and B directions have a ratio of approximately 1:1. The ratio of the true scale of A to that of Γ is approximately 100:1. Despite the unavoidable distortion of scale in the drawing, interesting geometrical characteristics of the region are apparent. The figure appears to have symmetries with respect to certain axes and planes. Alternating ridges and grooves encircle C in the Γ-direction. Analysis of the mathematical properties of the function which defines C has not been completed.

UPPER PORTION OF THE 95% CONFIDENCE REGION C IN THE PARAMETER SPACE P

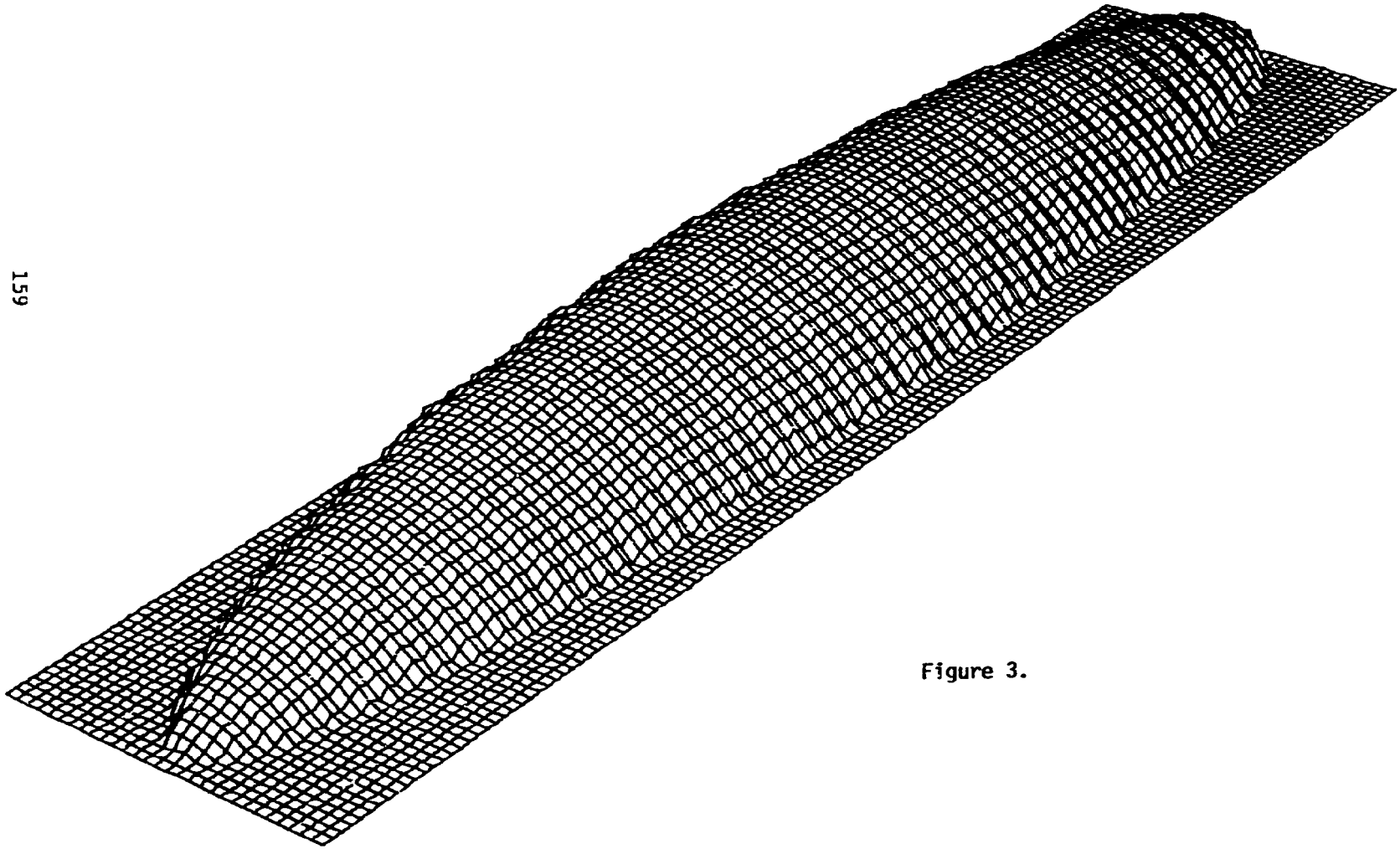


Figure 3.

For each concentration x, the calibration curve $y(x) = \alpha + \beta\gamma^x$ lies within a $100(1-\alpha)\%$ confidence band determined by the $100(1-\alpha)\%$ confidence region C in P. To determine the curves which define the band, it is necessary to calculate, for each x, the maximum and minimum values of y for all values of the parameters in the region. It is possible to eliminate from consideration all points (α,γ,β) in the interior of C for the following reasons. The directional derivatives of y in the coordinate direction are

$$D_\alpha y = 1$$
$$D_\beta y = \gamma^x$$
$$D_\gamma y = \beta x \gamma^{x-1}$$

For all x and for all parameter values obtained here (note $\gamma \neq 0$) these derivatives are defined. At no point $(\alpha_0, \gamma_0, \beta_0)$ is it true that $D_\alpha y = D_\beta y = D_\gamma y = 0$. Since it is necessary that the three partial derivatives equal zero simultaneously for an extreme value of the function to exist at a point, it follows that extreme values of y on the closed region must occur on the boundary C.

For each x from 0 to 15.1 ng/ml (in increments of 0.01) the value of y was computed for approximately 11,500 points of C. The maximum and minimum y values for each x represent the $100(1-\alpha)\%$ confidence limits for the instrument response. At a 95% confidence level the decision limit for this curve is 10.74 divisions and the detection limit is 0.08 ng/ml; at the higher concentration levels (about 14.5 ng/ml), the interval represents an uncertainty of approximately ± 0.3 ng/ml. The table gives the maximum and minimum chart readings for concentrations from 0 to 15 ng/ml. Figure 4 is a graph of these points to illustrate the calibration curve confidence band.

TABLE

ANALYZER CHART READINGS

Concentration, X (ng/ml)	Y_{min} (divisions)	Y_{max} (divisions)	$\hat{Y}$ (divisions)
0	9.95	10.74	10.35
1	19.78	20.31	20.05
2	28.33	28.84	28.58
3	35.80	36.39	36.10
4	42.39	43.03	42.71
5	48.20	48.86	48.53
6	53.32	53.98	53.65
7	57.85	58.47	58.16
8	61.84	62.42	62.13
9	65.35	65.89	65.62
10	68.43	68.97	68.70
11	71.12	71.69	71.40
12	73.47	74.10	73.78
13	75.51	76.24	75.88
14	77.30	78.15	77 72
15	78.87	79.03	79.35

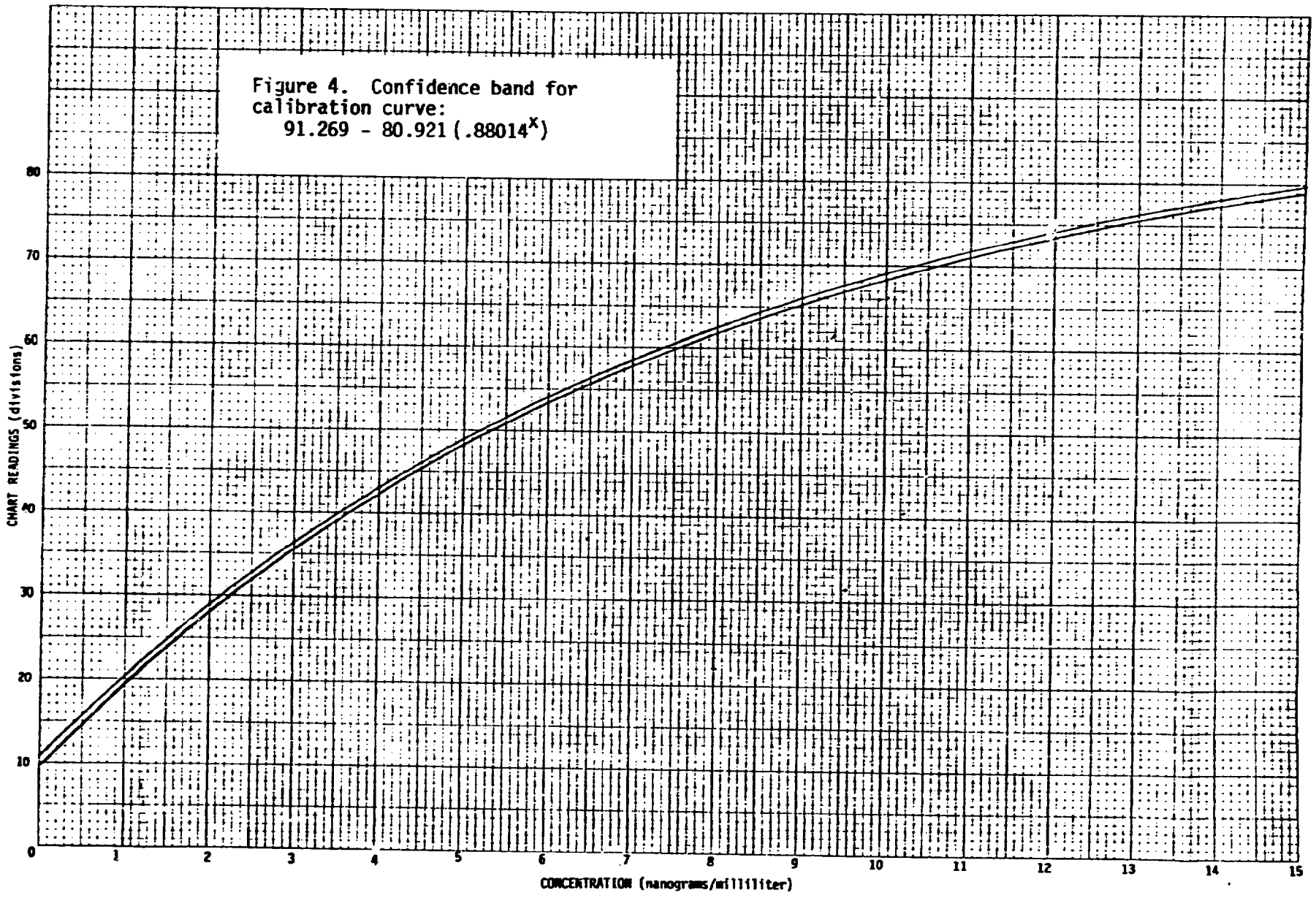


Figure 4. Confidence band for calibration curve:
91.269 - 80.921 (.88014^X)
CHART READINGS (divisions)
0
10
20
30
40
50
60
70
80
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
CONCENTRATION (nanograms/milliliter)

C. Future Investigations

The work described here has involved fitting the model $y = \alpha + \beta\gamma^{x}$ by least squares and the development of a numerical method of determining a $100(1-\alpha)\%$ confidence band for the curve. Continuing investigations will include (i) extension, if possible, of the methods to additional nonlinear models which are of importance in testing and other experimental work, (ii) analytic investigation of the functions, and (iii) development of a computer program to permit a more complete visualization of closed surfaces of the type encountered in this study.

REFERENCES

1. A. Hubaux and G. Vos, Anal. Chem., 42, 849 (1970).
2. M. Natrella, US Dept. of Commerce SP300, Vol 1, 204 (1969).
3. International Mathematical and Statistical Libraries, Inc., GNB Bldg., 7500 Bellaire Blvd., Houston, TX 77036
4. G. W. Snedecor and W. G. Cochran, Statistical Methods, Iowa State University Press, 1967.
5. L. Breiman, Statistics: With a View Toward Applications, Houghton Mifflin Co., 1966.
6. N. R. Draper and H. Smith, Applied Regression Analysis, John Wiley & Sons, Inc., 1966.

COMPUTING THE DEFINITE INTEGRAL $\int_0^\infty e^{-(px^2 + qx + r)}dx$ ON A PROGRAMMABLE CALCULATOR

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

ABSTRACT. When a reliability function is expressed by the exponential of a quadratic form, computation of mean life or mean time to failure requires evaluation of the definite integral

$$\theta = \int_0^\infty e^{-(px^2 + qx + r)}dx.$$

A transformation of variables is effected by completing the square. This allows θ to be expressed rather simply in terms of the complementary error function of the new variable. The latter can be evaluated by either of two well-known infinite series.

In using these series and, indeed, in selecting which of the two should be employed in a given case, certain difficulties are met with and there are some pitfalls to be avoided. A reasonably economical solution to the problems encountered is found.

I. THE PROBLEM. Recently, in conducting a software reliability analysis, employment of the modified Schick-Wolverton model was indicated [5]. This gives rise to the following equation:

$$\text{MTTF} = \theta = \int_0^\infty e^{-(ac^2x + \frac{c^2}{4}x^2)}dx, \tag{1}$$

a and c^2 being constants obtained by observation. Solution is by "completing the square". Thus

$$\theta = \int_0^\infty e^{-c^2\left(-a^2 + a^2 + ax + \frac{x^2}{4}\right)}dx$$

$$= e^{a^2c^2}\int_0^\infty e^{-c^2\left(a + \frac{x}{2}\right)^2}dx.$$

Let $t = ca + \frac{c}{2}x$. Then $dt = \frac{c}{2}dx$. Note that when $x = 0$, $t = ca$, whence

$$\theta = \frac{2}{c} e^{a^2c^2} \int_{t=ca}^{t=\infty} e^{-t^2}\, dt. \qquad (2)$$

In passing, observe that the absence of a constant term in the first exponent entails no loss of generality.

The error function and its complement are defined:

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt \text{ and}$$

$$\operatorname{erfc} z = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2}\, dt = 1 - \operatorname{erf} z.$$

Thus, setting $z = ca$,

$$\theta = \frac{\sqrt{\pi}}{c} e^{z^2} \operatorname{erfc} z. \qquad (3)$$

II. THE ASYMPTOTIC SERIES FOR erfc z. For large values of z, a useful asymptotic expansion is [2]

$$\sqrt{\pi}\, e^{z^2} \operatorname{erfc} z \sim \frac{1}{z}\left[1 - \tfrac{1}{2} z^{-2} + \frac{1\cdot 3}{2^2} z^{-4} - \frac{1\cdot 3\cdot 5}{2^3} z^{-6} + \cdots\right]. \qquad (4)$$

The general term is $T_n = \dfrac{(-1)^n \left[1\cdot 3\cdot 5\cdots(2n-1)\right]}{2^n z^{2n+1}}$ and

the recurrence ratio $T_n = \dfrac{\frac{1}{2} - n}{z^2} T_{n-1}$.

It is easy to see that the smallest term will occur when $0 < \frac{1}{2} - n + z^2 \leq 1$, the series diverging after that point. Using this inequality to identify the smallest term, and truncating the series immediately thereafter, results in a (nearly) minimum error. The worst case occurs when $z^2 = n - \frac{1}{2}$, n being the integer subscript of the smallest term. Some values of the relative error in this sum, together with the relative value of the smallest term, are tabulated for illustration:

TABLE 1

z^2	$\frac{\lvert\epsilon\rvert}{c\theta}$	$\frac{\lvert T_n\rvert}{c\theta}$	z^2	$\frac{\lvert\epsilon\rvert}{c\theta}$	$\frac{\lvert T_n\rvert}{c\theta}$
0.5	1.0000	1.5251	7.5	4.3261 E-4	8.3823 E-4
1.5	0.23446	0.41149	8.5	1.5741 E-4	3.0609 E-4
2.5	0.075564	0.13867	9.5	5.7399 E-5	1.1193 E-4
3.5	2.6047 E-2	4.8859 E-2	10.5	2.0964 E-5	4.0976 E-5
4.5	9.2158 E-3	1.7514 E-2	11.5	7.6657 E-6	1.5012 E-5
5.5	3.3030 E-3	6.3317 E-3	12.5	2.8056 E-6	5.5034 E-6
6.5	1.1926 E-3	2.3002 E-3	13.5	1.0276 E-6	2.0185 E-6

III. A SERIES FOR erf z. For small values of z, the infinite series

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} e^{-z^2} \sum_{n=0}^{\infty} \frac{2^n z^{2n+1}}{1\cdot 3\cdot 5\cdots(2n+1)} \tag{5}$$

is employed [4]. Although the series converges for all finite values of z, it is of little practical use when z is large. Convergence is then very slow -- hundreds, even thousands of terms being required -- and an unacceptably high number of significant digits are lost when the subtraction $\operatorname{erfc} z = 1 - \operatorname{erf} z$ is performed, even though the computations be done in multiple precision.

Recalling that $x\Gamma(x) = \Gamma(x+1)$ and that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we have

$\Gamma\left(m + \frac{1}{2}\right) = \frac{1}{2}\cdot\frac{3}{2}\cdot\frac{5}{2}\cdots\left(m - \frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)$, where m is an integer. This can be rewritten

$$\Gamma\left(m + \frac{1}{2}\right) = \frac{1\cdot 3\cdot 5\cdots(2m-1)\sqrt{\pi}}{2^m}$$

Setting $m = n + 1$, we can write immediately

$$\operatorname{erf} z = e^{-z^2} \sum_{n=0}^{\infty} \frac{z^{2n+1}}{\Gamma\left(n + \frac{3}{2}\right)} \tag{6}$$

The wanted function, of course, is

$$\theta = \frac{\sqrt{\pi}}{c} e^{z^2} \operatorname{erfc} z = \frac{\sqrt{\pi}}{c} e^{z^2} (1 - \operatorname{erf} z)$$

$$= \frac{\sqrt{\pi}}{c} \left\{ e^{z^2} - \sum_{n=0}^{\infty} \frac{z^{2n+1}}{\Gamma\left[n + \frac{3}{2}\right]} \right\}$$

$$= \frac{\sqrt{\pi}}{c} \left\{ \sum_{n=0}^{\infty} \frac{z^{2n}}{\Gamma(n+1)} - \sum_{n=0}^{\infty} \frac{z^{2n+1}}{\Gamma\left[n + \frac{3}{2}\right]} \right\} \qquad (7)$$

This last form not only points up a problem -- that e^{z^2} must be computed to the same precision as $\sum_n \frac{z^{2n+1}}{\Gamma\left[n + \frac{3}{2}\right]}$ -- but suggests the answer:

The two parts can be summed using the same subroutine, varying only the first term and the first value of the summing index. This advantage (programming simplicity) was decisive in the choice of series for erf z, even though one is known which converges slightly faster [3].

It is interesting to note that a simple change of summing index produces the elegant form

$$\theta = \frac{\sqrt{\pi}}{c} \sum_{r=0}^{\infty} \frac{(-z)^r}{\Gamma\left[1 + \frac{r}{2}\right]} \qquad (8)$$

An estimate of the number of significant digits lost by subtraction is given by

$$\log_{10} \frac{e^{z^2}}{e^{z^2} (1 - \operatorname{erf} z)} = - \log_{10} \operatorname{erfc} z.$$

Some values of $- \log_{10} \operatorname{erfc} z$ are tabulated:

TABLE 2

z	$-\log_{10}\operatorname{erfc} z$	z	$-\log_{10}\operatorname{erfc} z$	z	$-\log_{10}\operatorname{erfc} z$
1.5	1.470	2.4	3.162	3.3	5.515
1.6	1.626	2.5	3.390	3.4	5.818
1.7	1.790	2.6	3.627	3.5	6.129
1.8	1.962	2.7	3.872	3.6	6.449
1.9	2.142	2.8	4.125	3.7	6.777
2.0	2.330	2.9	4.386	3.8	7.113
2.1	2.526	3.0	4.656	3.9	7.459
2.2	2.730	3.1	4.934	4.0	7.812
2.3	2.942	3.2	5.220	4.1	8.174

Table 2 does not take into account the effect of round-off error in the individual terms.

It can be seen at once that, as z increases, significant digits are lost at an accelerating rate. An actual single-precision program on a 13-digit calculator produced the following result:

TABLE 3

argument range (value of z)	significant digits	number of terms in sum
0.83 to 1.42	10	13 to 20
1.43 to 2.01	9	19 to 26
2.02 to 2.51	8	25 to 33
2.52 to 2.93	7	31 to 38
2.94 to 3.30	6	37 to 43
3.31 to 3.63	5	42 to 48
3.64 to 3.94	4	46 to 53
3.95 to 4.22	3	51 to 57
4.23 to 4.48	2	55 to 61
4.49 to 4.73	1	59 to 65
4.74 to ∞	noise only	63 or more

IV. WHICH SERIES TO USE? To point up the problem which remains, let us assume there is a requirement to compute e^x to six significant digits on a machine which computes e^x with a maximum relative error of 10^{-9}. For values of the argument up to about 2.33 ($z^2 = 5.43$), the second series (see eq. 7) can be used, and for values above 3.68 ($z^2 = 13.54$), the asymptotic series (see eq. 4) can be used if truncated after the smallest term. But what is to be done when the argument falls "in-between"? (i.e., when $2.33 < z < 3.68$?)

The answer, surprisingly enough, lies in the asymptotic series itself. Asymptotic series of this type* have a most interesting and useful property: Provided that the truncated series consists of at least two terms (i.e., $n \geq 1$), and further provided that the series is terminated immediately after the smallest term, the approximation ALWAYS is improved by halving the last term. Performing this operation and tabulating (see Table 4), it is seen that the improvement, though quite noticeable, is not yet enough to solve the problem.

TABLE 4

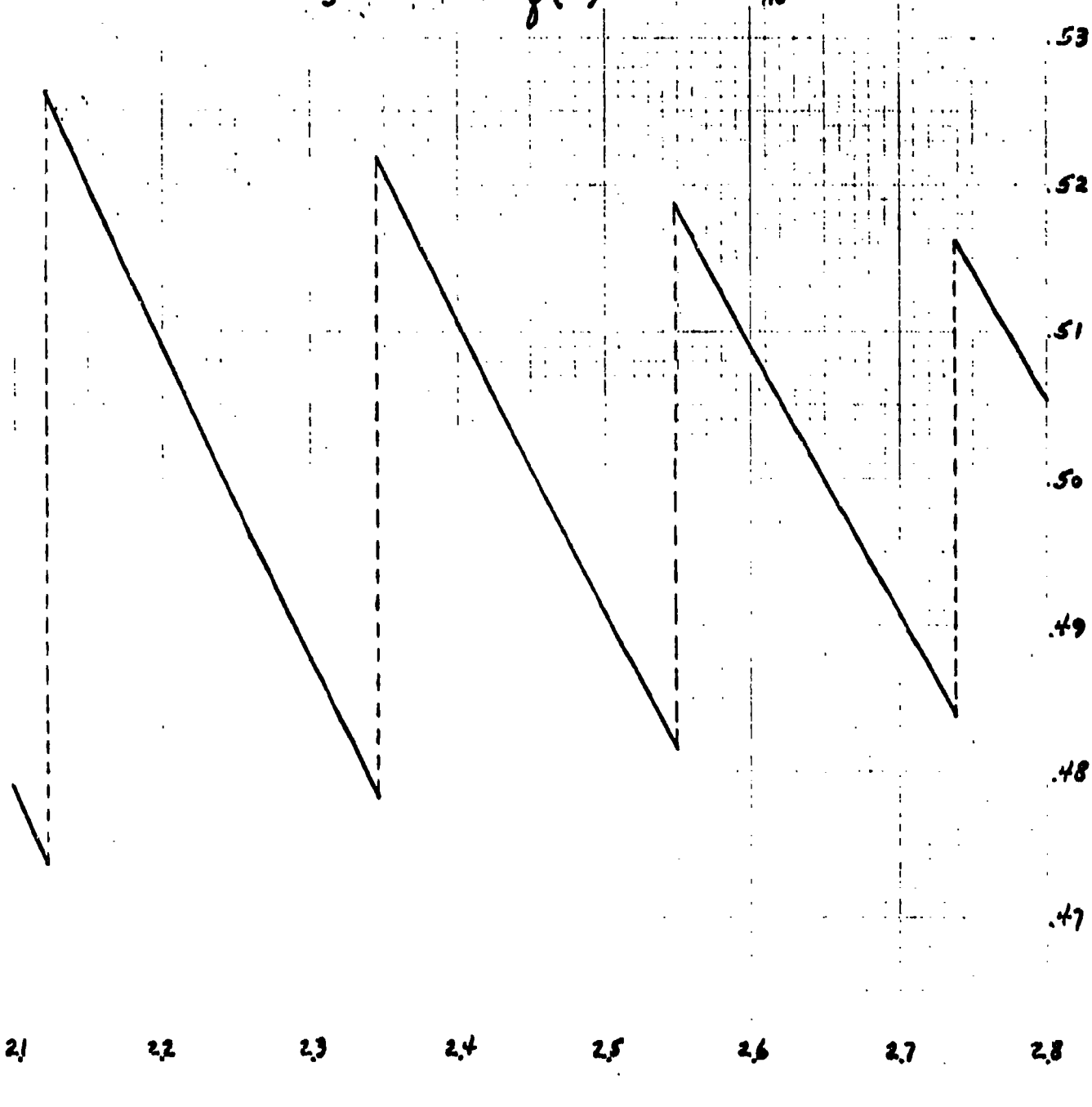
z^2	$\frac{\lvert\varepsilon\rvert - \lvert\frac{1}{2} T_n\rvert}{c\theta}$	z^2	$\frac{\lvert\varepsilon\rvert - \lvert\frac{1}{2} T_n\rvert}{c\theta}$	z^2	$\frac{\lvert\varepsilon\rvert - \lvert\frac{1}{2} T_n\rvert}{c\theta}$
0.5	0.23743	5.5	1.3721 E-4	10.5	4.7599 E-7
1.5	0.028718	6.5	4.2497 E-5	11.5	1.5957 E-7
2.5	6.2314 E-3	7.5	1.3495 E-5	12.5	5.3917 E-8
3.5	1.6177 E-3	8.5	4.3663 E-6	13.5	1.8340 E-8
4.5	4.5884 E-4	9.5	1.4333 E-6	14.5	6.2743 E-9

It is both interesting and informative to compute and plot the ratio $\frac{\varepsilon}{T_n}$. (See Figure 1.) Since an alternating series always "overshoots", the last term used and the error in the partial sum will be of the same sign, and their ratio will be positive definite. The function

$$f(z) = \frac{\varepsilon}{T_n}$$

*i.e., with simple terms. Should -- say -- Bernouilli's numbers appear, the adjective "useful" may no longer be applicable, due to increased programming difficulties.

Fig. 1. $f(z) = \epsilon / T_m$



Note: $\xi = z^2 + \frac{1}{2}$

is a "saw-tooth", having two values at those points where $z^2 + \frac{1}{2}$ is an integer. (There are two equal "smallest" terms, and it is arbitrary whether one or both are used.) It is obvious that the sum of the two values is unity.

After applying the half-term correction, the remaining error can be stated as $\varepsilon - \frac{1}{2} T_n$, of course. Using a similarly-formed ratio, we define

$$g(z) = \frac{\varepsilon - \frac{1}{2} T_n}{T_n} = f(z) - \frac{1}{2}. \qquad (9)$$

Let us tabulate, not g(z), but its reciprocal, at the points where $z^2 + \frac{1}{2} = n$ (i.e., an integer), using the greater of the two values

TABLE 5

z^2	n	$\frac{1}{g(z)} = \frac{T_n}{\varepsilon - \frac{1}{2} T_n}$	z^2	n	$\frac{1}{g(z)} = \frac{T_n}{\varepsilon - \frac{1}{2} T_n}$
0.5	1	6.4234 5156	7.5	8	62.1131 7520
1.5	2	14.3283 6175	8.5	9	70.1017 6038
2.5	3	22.2527 7233	9.5	10	78.0924 2300
3.5	4	30.2036 5992	10.5	11	86.0846 4634
4.5	5	38.1700 2696	11.5	12	94.0780 7107
5.5	6	46.1457 3873	12.5	13	102.0724 3985
6.5	7	54.1274 3674	13.5	14	110.0675 6367

By inspection, we can approximate $\frac{1}{g(z)}$ at these end-points reasonably well by the function

$$\frac{1}{g^*(z)} = 8n - 2 + \frac{1}{n+1} \text{ from which}$$

$$g^*(z) = \frac{n+1}{8n^2 + 6n - 1}.$$

The right-hand end of the ramp is then estimated by

$$-g^*\left(\sqrt{z^2 + 1}\right) = \frac{-(n+2)}{8n^2 + 22n + 13}$$

V. APPROXIMATING THE RAMP FUNCTION. We can improve both the notation and the accuracy as follows. Let

$$\xi = z^2 + \tfrac{1}{2} = n + x. \tag{10}$$

The integer part of ξ is represented by n, the decimal part by x. ξ is a continuous variable, n a discrete one. The general form of the approximating function is taken to be

$$\frac{1}{g^*(\xi)} = 8\xi - 2 + \frac{1}{\xi + \alpha}$$

which upon development yields

$$g^*(\xi) = \frac{\xi + \alpha}{8\xi^2 + (8\alpha - 2)\,\xi + (1 - 2\alpha)} \tag{11}$$

A little investigation reveals that in the region of interest ($z > 2$), a near-optimum formula is given by assigning the value $\alpha = \frac{7}{8}$. Thus

$$g^*(\xi) = \frac{\xi + \frac{7}{8}}{8\xi^2 + 5\xi - \frac{3}{4}} \tag{12}$$

As a fortunate happenstance, the denominator is factorable, allowing the expression to be reduced to partial fractions.

$$g^*(\xi) = \frac{1}{7}\left[\frac{8}{8\xi - 1} - \frac{1}{8\xi + 6}\right] \tag{13}$$

An extremely close approximation to the ramp is given by

$$\left[1 - 2x + \frac{x(1 - x)}{\xi}\right] g^*(\xi).$$

Adopting the notation $\sum_{i=0}^{n} T_i$ for the finite series truncated after the smallest term, we find

$$c0 = \sum_{i=0}^{n} T_i - T_n\left[\frac{1}{2} + \left(1 - 2x + \frac{x - x^2}{\xi}\right) g^*(\xi)\right] \tag{14}$$

Some worst-case results are given in Table 6, below.

TABLE 6

Residual error, η, in $c\theta$ from
"corrected" asymptotic series for erfc z

z	η	$\frac{\eta}{T_n}$
2.00	-1.255331 E-7	-9.7940 E-6
2.07	2.103996 E-7	22.3722 E-6
2.17	1.429149 E-7	-24.3138 E-6
2.30	-0.407994 E-7	13.1637 E-6
2.39	-0.279393 E-7	-14.2805 E-6
2.50	8.7552 E-9	8.0323 E-6
2.59	6.0404 E-9	-9.0574 E-6
2.70	-2.0204 E-9	5.6537 E-6
2.78	-1.3582 E-9	-6.0688 E-6
2.88	0.4902 E-9	3.9942 E-6
2.95	3.409 E-10	-4.278 E-6
3.04	-1.248 E-10	2.772 E-6
3.12	-0.833 E-10	-3.105 E-6
3.20	0.328 E-10	2.081 E-6
3.27	0.229 E-10	-2.331 E-6
3.36	-9 E-12	1.69 E-6
3.42	-6 E-12	-1.80 E-6
3.50	3 E-12	1.35 E-6
3.57	2 E-12	-1.47 E-6
3.64	-1 E-12	0.88 E-6

It is found that employment of the "corrective" term extends the use of the asymptotic series down to an argument of z = 1.99, thereby overlapping the useful range of the other series and providing a solution to

the six-place problem posed in Section IV. In fact, if "break points" of $z = 2.1$ and $z = 4.1$ are chosen*, the relative error throughout the whole spectrum probably does not exceed 3.5×10^{-7}. A program written for a thirteen-digit calculator, with break points at $z = 2.5$ and $z = 4.4$ (summing the first eleven terms thereafter), produces a value of $c\theta$ which errs no more than one in the eighth decimal place.

VI. INCREASING THE ACCURACY. In the remote event that additional accuracy is required, two avenues of approach offer themselves.

a. The calculations can be performed in double- (or triple-) precision. This will extend the useful range of the argument when employing the series for erf z. This procedure is NOT recommended, since it will increase the running time by many orders of magnitude.

b. The accuracy of the "corrective" term can be improved, thereby extending downward still further the use of the asymptotic series for erfc z. Since we will be operating in a region where the asymptotic series contains very few terms anyway, it is unlikely that running time will be too adversely affected. In pursuit of our goal, two steps are taken.

1. The degree of the rational expression for $g^*(\xi)$ is increased. It is found to be

$$g^*(\xi) = \frac{\xi^2 + \alpha\xi + \beta}{8\xi^3 + (8\alpha - 2)\,\xi^2 + (1 - 2\alpha + 8\beta)\,\xi + \left[\alpha - 2\beta - \frac{3}{4}\right]} \tag{15}$$

Selecting $\alpha = 1.2$ and $\beta = 1.05$ results in

$$g^*(\xi) = \frac{\xi^2 + 1.2\xi + 1.05}{8\xi^3 + 7.6\xi^2 + 7\xi - 1.65} \tag{16}$$

2. More terms are added to the ramp function. Thus

*When $z > 4.1$, merely sum the first eleven terms ($i = 0, 1, 2, \cdots, 10$) of the asymptotic series.

$$c\theta = \sum_{i=o}^{n} T_i - T_n \left[\frac{1}{2} + \left(1 - 2x + \frac{x - x^2}{\xi} - \frac{(1 - 2x)(x - x^2)}{4\xi^2} + \frac{(x - x^2)^2}{2\xi^3}\right) g^*(\xi)\right] \qquad (17)$$

Using these refinements, with break points at 2.34 and 4.77, reduces the maximum error on a 13-digit calculator to less than $1.7 \times E^{-9}$. Attempts to further reduce the maximum error will prove to be tedious and somewhat unrewarding, since the "smallest" term in the asymptotic series becomes too large to lend itself to the process.

BIBLIOGRAPHY

[1] Abramowitz, M. and Stegun, I. A., eds., "Handbook of Mathematical Functions", 1965, Dover Publications, Inc., New York. See sec. 7.1

[2] Fike, C. T., "Computer Evaluation of Mathematical Functions", 1968, Prentice-Hall, Inc., Englewood Cliffs, N. J. See esp. Ch. 11

[3] Peirce, B. O., "A Short Table of Integrals", 4th ed., 1967, Ginn and Co., New York. See #808, 809, 810, 811, 812, 874

[4] Smith, J. M., "Scientific Analysis on the Pocket Calculator", 1975, John Wiley & Sons, New York. See sec. 4.4

[5] Sukert, A. N., "A Software Reliability Modeling Study", 1976, Rome Air Development Center, Griffiss A. F. Base, New York. See sec. 2.4

[6] Von Alven, W. H., ed., "Reliability Engineering", 1964, Prentice-Hall, Inc., Englewood Cliffs, N. J. See sec. 3.3

A FRESHMAN ERROR CAN BE FATAL

OR

I'M NOT SO SURE ABOUT BEING 95 PERCENT SURE

NORMAN L. WYKOFF
US ARMY JEFFERSON PROVING GROUND
MADISON, IN 47250

ABSTRACT.

The testing of artillery ammunition involves the use of control rounds to measure the "day-to-day" variations caused by different tubes, recoils and weather conditions. The control rounds are assembled from components that have been tested (separately and in combination) in sufficient quantity to establish the performance characteristics of control components and complete rounds.

The difficulty comes when a component is nearly depleted and must be replaced. Unless the match is perfect, the performance characteristics of the control will shift. The accepted technique thus far has been to check the match or mismatch using a 95 percent confidence interval for the means of rounds with the old (n=20) and new (n=20) component. Obviously, this criteria does little to assure the integrity of the control and thus can jeopardize troops in the field.

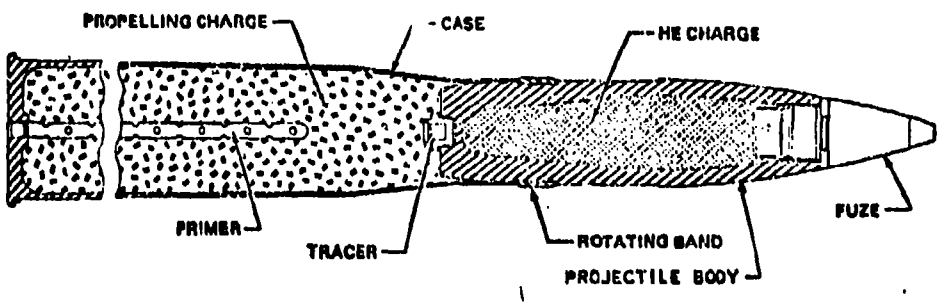
The problem is two-fold: (1) what is an optimal technique, considering both cost and control integrity; and (2) how can we eliminate the idea that use of a 95 percent confidence interval means you are almost certain to make a good choice.

1. INTRODUCTION:

Part of the mission of US Army Jefferson Proving Ground is to ballistically test large caliber ammunition. Statistically, the process is not overly complicated, but there are many factors that vary, independently and dependently, that keep the process from being a simple one.

A round of ammunition is a complex machine. There are many components that must do their particular job in exactly the right way, at exactly the right time for the complete round to behave properly.

FIGURE 1



In this example there are eight major components plus the whole assemblage to be ballistically tested. That is, performance parameters such as velocity, chamber pressure, target accuracy, range accuracy and/or functioning must be evaluated for each component when the round is fired from the appropriate weapon.

There are many different factors that can affect a parameter such as the velocity of the round. For example, tube wear, recoil system, give of the earth under the weapon, size of the projectile, type of rotating band, burning rate of the propellant, and of course the amount of propellant will each have an effect. Some of these factors cannot readily be measured and may, in fact change from trial to trial. The obvious way to estimate the total of all these extraneous effects is to use a control round. A control round with a long history of performance including many very carefully monitored firings can be used to estimate the trial-to-trial or day-to-day variation as it is usually called. In brief, if the control rounds have a mean velocity that is 20 foot-seconds lower than normal in a trial, we assume that the sum total of all those effects yields a 20 foot-second decrease in the velocity of the test rounds also. Therefore we add 20 foot-seconds to the observed test velocities to "correct" them to standard conditions.

In order to reduce the number of interactions, a test component is tested against the control component by loading each into rounds that are "identical" except for the component being tested. In this way we can measure the change in performance of the test component from the control component.

By now you can see the dependence on the performance integrity of the control round for a critical parameter such as velocity. It is exactly this dependence that creates my concern in this present problem. Before I describe the problem more fully, let me emphasize that obtaining the long history on the control rounds is expensive in time, money and material.

II. The Problem: Because of the variety of uses of the control round, one component may be nearly depleted long before the others. It only makes sense then, because of economics to substitute a new lot of the component, rather than restart the whole process.

Suppose the component in question is the projectile, it obviously has an effect on the velocity. Incidentally, we will not considar the propellant since the substitution process is different for the propellant. The question now is, what is the best procedure to use in substituting a new projectile lot?

Figure 2 shows the description of the accepted practice.

FIGURE 2

8. FIRING WITH SUBSTITUTE COMPONENTS. The purpose of these firings is to determine the effect of changing a selected component in the master or reference established values. When any change of a component in the reference round is required, the following steps are taken:

a. From engineering judgment and past data decide whether the change is likely to affect the velocity or pressure level of the round.

b. If a change in velocity or pressure is expected, fire 20 rounds from the check tube with the old component and 20 with the new, keeping all other components the same.

c. If the firing in b above is not statistically different (significance level of 5%), accept the new component.

d. If the firing in b above shows a significant difference, fire 20 additional rounds with the new component and 20 with the old in each of two tubes with not less than 90 percent life remaining (total: 80 rounds). If this firing also shows a significant difference, discard the new component and select a second replacement component. Repeat the test procedure in b until a satisfactory replacement component is obtained.

e. For multicharge systems, conduct the firings under b at zones at which ballistic differences would be at a maximum. If the difference is significant at that charge or charges, follow the procedure of d, above.

f. Before testing a substitute lot, evaluate the performance of the existing calibration rounds (para 3.5) and submit the evaluation to ARMCOM.

It leaves quite a bit to the imagination of the reader doesn't it. Although, perhaps not too much. The underlying assumption in the process is that the continuous parameters (velocity in this case) have a normal distribution with μ and σ unknown and estimable for a given trial only by the results of that trial. There are a few possible interpretations for the meaning of the statement above but the one that seems to have been used by those who have the task of interpreting it is to use the 2 sample t-test (2 sided). That is, the test is based on:

FIGURE 3

$$(1)\quad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

or if $n_1 = n_2 = n$

$$(2)\quad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

The acceptance region for the statistic (2) is:

$$(3)\quad \sqrt{\frac{s_1^2 + s_2^2}{n}}\; t_{\alpha/2,\, 2n-2} < \bar{x}_1 - \bar{x}_2 < \sqrt{\frac{s_1^2 + s_2^2}{n}}\; t_{\alpha/2,\, 2n-2}$$

This is exactly what you saw in that freshman statistics course a few years ago. However, hopefully you saw more. You understood that the so called 95 percent confidence interval given in (3) is an interval big enough to contain the difference of the sample means (given these values of S_1 and S_2) 95 percent of the time if the two samples actually come from the same population and that you didn't fall prey to the freshman fallacy of believing that if $\bar{x}_1 - \bar{x}_2$ fell in this interval you were 95 percent sure that μ_1 and μ_2 were actually the same. If you made this mistake you probably never did reconcile the implication that the larger 99 percent interval made you even more certain that the match was good. Of course, we don't make such errors. Perhaps if the phrase "confidance interval" wasn't used others wouldn't either. I wish we could change this to a "95 percent location interval".

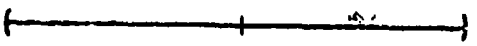
The problem is hopefully now clear. The process is good for the seller, but not for the buyer and I represent the buyer. To say it a different way, this is the classic case where α (the probability of rejecting a test lot that is an exact match) is fixed and β (the probability that a poorly matched lot is accepted) varies and for some reason that I prefer not to put in print, we choose α to be small.

In the following example, the numbers are realistic although they do not represent actual data. Suppose $\bar{x}_1$, S_1, n_1, and $\bar{x}_2$, S_2, n_2 represent the old and new sample means, standard deviations, and numbers. Suppose further than $n_1 = n_2 = 20$, $\bar{x}_1 = 5050$, $S_1 = S_2 = 26.6$ (the maximum allowable value for acceptance tests for this round) and $\bar{x}_2 = 5067$. The acceptance region is shown in Figure 4 below.

By now some are asking, why not use the location interval based on the first sample and see if $\bar{x}_2$ falls inside?

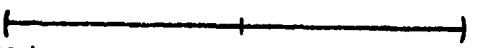
FIGURE 4

95% Location Interval Based on 2 Sample Test



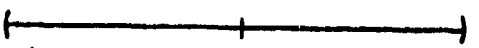
-17.03	0	17.03	d
5032.97	5050	5067.03	$\bar{x}_2$

95% Location Interval Based on Sample 1



-12.45	0	12.45	d
5037.55	5050	5062.45	$\bar{x}_2$

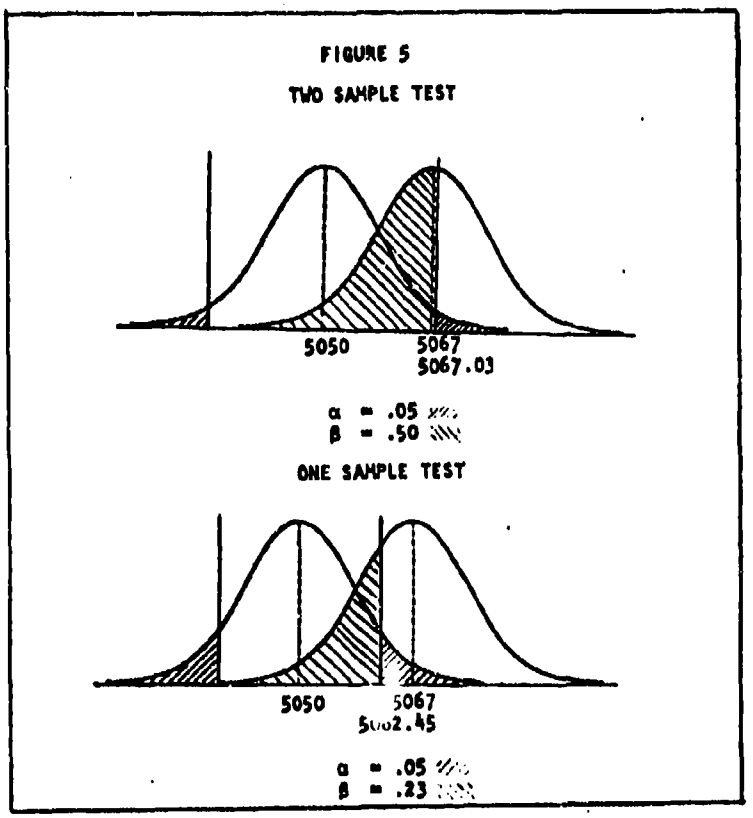
95% Location Interval Based on Sample 2



-12.45	0	12.45	d
5054.55	5067	5079.45	$\bar{x}_1$

It seems to be a good idea, it does shorten the interval with no increase in sample size or cost. I believe, however, that this has not been used because it doesn't make use of S_2 (or S_1, if you center the interval about $\bar{x}_2$) and it is necessary to have samples taken from each of the populations and therefore both means and standard deviations are available.

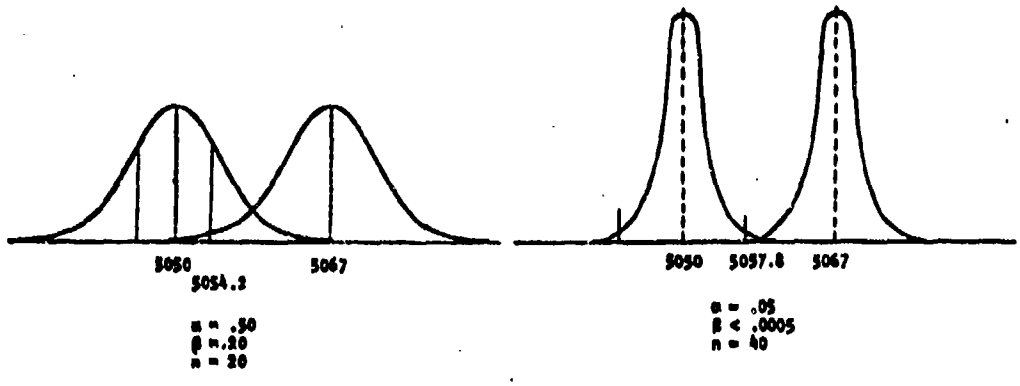
Looking at this another way for the numbers already given.



Quite a change in β from the two sample to the one sample technique. Better, but not good enough for me. I still represent the buyer and I am very insistent that the product "prove" to me that it is good. In other words, I want to have a high degree of confidence that the product is good, the same thing that freshman student thinks he has. I am not content to reject or not reject the claim that they are the same. You see, I must think of that poor GI who must use this ammunition to defend himself. If we are lax and let the velocity level change too much, he just might miss the tank that is bearing down on him.

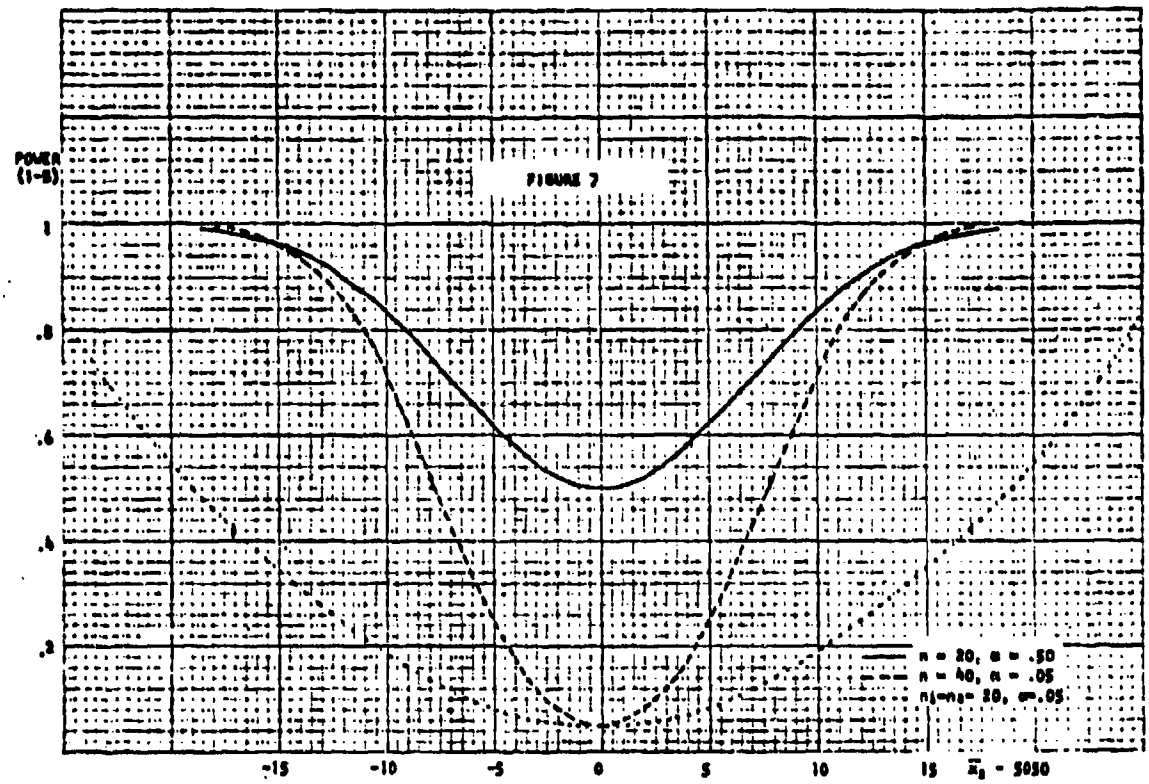
There seem to be two things for me to do. First, to convince the people involved that the one sample technique is preferrable in terms of the power of the test, even though it doesn't use all the sample information, and second, to take the best steps to decrease β even more. The two alternatives for decreasing β are to increase the sample size or increase α. Both techniques dramatically increase the cost in my application, but it is difficult to predict the exact amount. My choice is to increase α. This will result in rejecting more lots of good components, but I will be more content to accept a lot that passes the more severe test. Increasing the sample size will increase the confidence in the decision but at a greater cost in each lot considered.

FIGURE 6



In Figure 6 we have a comparison of the distribution for $\alpha = .50$, $n = 20$ and $\alpha = .05$ with $n = 40$ for $\bar{X}_1 = 5050$ and $\bar{X}_2 = 5067$. I like the tight cut-offs on the first one and the separation on the second. However, Figure 6 only tells part of the story. The question I originally thought I would pose for the panel is: Which is better, increase α or n?

But after I drew the OC curves for $n = 20$ with $\alpha = .50$, $n = 40$ with $\alpha = .05$ and the currently used 2 sample test, I have changed the question to: How much should I increase α?


POWER (1-β)
FIGURE 7
1
.8
.6
.4
.2
-15
-10
-5
0
5
10
15
$\bar{x}_0$ - 5050
n = 20, α = .50
n = 40, α = .05
$n_1=n_2$= 20, α=.05

LASER VELOCIMETER DATA INTERPRETATION BY HISTOGRAM AND SPECTRAL ANALYSIS

Warren H. Young, Jr.,* James F. Meyers,** and Danny R. Hoad*
Structures Laboratory, U.S. Army Research and Technology Laboratories (AVRADCOM)
NASA Langley Research Center
Hampton, Virginia 23665

ABSTRACT. A laser velocimeter has been used to survey turbulent, unsteady flows. Data have been analyzed in histogram form. The time-averaged flow field has been found from ensemble averages. By assuming stationary flow, the standard deviation and excess give the RMS unsteadiness in the flow and the statistical uncertainty in the mean and standard deviation.

The calculation of part of the unsteady flow field has been attempted by a Monte Carlo method. Partial success in explaining bimodal and skewed histograms has been achieved. This approach has been limited by the necessity of constructing a hypothetical flow field and the inability to define a mathematically unique solution.

The definition of power spectra has been achieved for single components of velocity. Autocorrelation has been chosen to construct the power spectrum because of the random sample time. Measurements of velocity are available only when seed particles pass through the sample volume. This is a random event with a Poisson distribution so that the usual time series analyses are precluded.

Theory has been developed for cross-correlation and cross-spectral analyses for two velocity components. However, methods for analysis of nonstationary flow have not yet been explored.

I. INTRODUCTION. The reduction and interpretation of data acquired by laser velocimetry in large wind tunnels has illustrated several unique aspects of the data analysis. The distinctive characteristics of the laser velocimeter that contribute to the need for new data interpretation techniques are primarily the ability to calculate errors prior to the test, the acquisition of discrete, digital measurements, and the randomness of the time between measurements. The purpose of this paper is to illustrate several techniques that have been developed specifically for handling laser velocimeter measurements, to outline the limitations of the present techniques, and to anticipate opportunities and problems that lie in the immediate future.

In order to define the source of the unique aspects of laser velocimetry, the apparatus is briefly described. This description is sufficient to explain the interaction between the error analysis and the histogram moments. Monte Carlo methods extend the usefulness of the histogram as an interpretative tool.

---

*Structures Laboratory, USARTL (AVRADCOM)
**NASA Langley Research Center

The second part of the paper deals with the analysis of the time dependence of the flow. The capabilities of time analysis are linked both to the manner in which laser velocimeter measurement times relate to the time scales of the flow and to the method of analysis of the data. The most general method in use, power spectra, is described in detail. Differences between laser velocimeter and traditional frequency analyses are identified. The basic requirements of conditional sampling are outlined, and several future needs are identified.

II. APPARATUS

Example tests: The laser velocimeter has been used in large wind tunnels at Langley Research Center to measure flow velocities about aerodynamic models such as wings. Two such test setups are shown in figure 1 (Ref. 1) and figure 2. These particular models are wings at very high angles of attack (about 19.5°). The two tests used flow velocities of 170 m/sec and 50 m/sec, respectively. In both cases, measurements were taken of the two components of velocity which lie in a plane perpendicular to the wing span. This plane cut the center of the span of the wing. Thus, from figures 1 and 2, it can be seen that the velocity measurements were made perpendicular to the laser beams.

Laser velocimeter operation: In order to measure two components of velocity, three separate laser beams were used (Ref. 1). These beams intersected at the center span of the wing. The beams were 0.3 mm in diameter so the volume of intersection (called the sample volume) was about 0.3 mm in diameter and 1 cm in length. Seed particles that pass through this sample volume scatter laser light back through the optics system to photomultiplier tubes. The two photomultiplier-tube outputs are separately checked for consistency and strength. A signal of sufficient quality will allow the measurement of one or both velocity components to be measured for these particular tests with a bias error between -1.33 percent and +0.91 percent and a $\pm$0.47 percent random uncertainty.

Example data: The example tests required the analysis of the several million velocity measurements acquired at several hundred points in the velocity field about an airfoil. Figure 3 shows a section of the wing at the center span and the directions, labeled U_L and V_L, in which the two components are measured. The tail of each arrow in figures 4 and 5 represents a measurement point. At each measurement point several hundred (up to 4096) individual velocity measurements were made in a period that varied from 10 seconds to several minutes.

III. DATA INTERPRETATION BY HISTOGRAM

Histogram moments: The most elementary, compact means of presenting laser velocimeter data is by means of a histogram of ensembles of each component of velocity. Figure 6 shows four pairs of histograms measured at four points above the airfoil. The ordinate, C_i, is the percentage of measurements that lie between $U_i - \frac{\Delta U}{2}$ and $U_i + \frac{\Delta U}{2}$. In this case, ΔU was 2.56 m/sec.

The histogram shape approximates a probability density function, P(U). Thus, C_i is approximately $P(U_i)\Delta U$. The mean of the probability density function is equal to the time-averaged velocity, that is

$$U_a = \lim_{T \to \infty} \frac{1}{T} \int_0^T U(t)dt = \int_{-\infty}^{\infty} U\, P(U)\, dU$$

The histogram mean approximates the time-averaged velocity under the following assumptions:

1. The true velocity probability density function is independent of time (i.e. stationary in time).

2. The laser velocimeter is equally likely to measure all velocities, or else any velocity bias (Ref. 2) has been removed before the formation of the histogram. Therefore, this source of error, as well as particle tracking errors, will be ignored in this discussion.

3. The number of velocity measurements, D, is large. The statistical uncertainty in the mean for a 95 percent confidence limit is given by:

$$\text{Uncertainty in } U_a = \pm \frac{2\sigma}{\sqrt{D}}$$

where σ^2 is the variance of the histogram.

In a similar manner, the standard deviation, σ, is identified as the root-mean-square value of $U(t) - U_a$. The statistical uncertainty in σ for a 95 percent confidence limit is given by (Ref. 3):

$$\text{Uncertainty in } \sigma = \sigma \sqrt{\frac{2}{D}} \left(1 + \frac{E}{2}\right)^{\frac{1}{2}}$$

where E is the excess (or kurtosis - 3) of the histogram. The uncertainty in σ is usually an order of magnitude larger than the random error in the individual velocity measurement. (The random error in individual velocity measurements was discussed in section II). For example, if 6 is 3.50 m/sec and the excess is zero and if 2000 measurements were made, then the uncertainty is 0.11 m/sec.

Histogram interpretation: The histograms are extremely useful in the physical interpretation of the flow field. Figure 7 shows contours of constant resultant mean velocities, and figure 8 shows contours of constant resultant standard deviation. The aerodynamic interpretation of these contours was hindered by the complete lack of any time history or frequency information in the histograms.

For example, the cause of the locus of maximums of standard deviations (shown in figure 8 by a dotted line) may be caused by (1) high levels of random turbulence, (2) a moving continuous vortex sheet, (3) by a series of discrete vortices that move down the airfoil, or (4) any combination of the above. In order to distinguish between these possibilities, a Monte Carlo simulation of the histograms was used.

Monte Carlo simulation: The first step in the Monte Carlo simulation is the creation of a flow model. A vortex model, shown in figure 9, was hypothesized. By adjusting the physical parameters (such as vortex strength and height above the airfoil), calculating the velocities caused by the vortex model, and simulating the laser velocimeter measuring process, a series of simulated histograms were generated. Figure 10 shows a comparison, above the 15 percent chord of the airfoil, of the actual measured pairs of histograms along with the simulated histograms for each component. The Monte Carlo method has qualitatively reproduced the measured bimodal histograms. The simulation reproduces the high velocities in both components at point "a." Although points "b" through "d" have low measured velocities, the simulation does not show the lower velocities before points "d" and "e." Point "f" shows good simulation of the bimodal histogram. For points "g" through "j" a gradual shift to a low mean velocity is reasonably simulated. Although this and other Monte Carlo type simulations were considered to be successful, the hypothetical flow model cannot be accepted with complete assurance because other models could yield the same result. The Monte Carlo method cannot define a unique time variation of velocity. This is one of several severe defects in the present method of histogram interpretation.

Limitations on interpretation: The value of the histograms is augmented by one of the distinctive characteristics of the laser velocimeter. The laser velocimeter is an unusual measurement tool in that the errors, both random and bias, of individual measurements are calculable, and therefore known, before the experiment begins. Since several hundred, or even several thousand, individual velocity measurements are available to calculate each value of mean velocity and standard deviation, the statistical uncertainty of these two histogram moments are also calculable. However, it has not been possible to fully utilize the advantages of precalculable errors. For example, the uncertainty in the higher moments of the histogram, skew and excess have not been derived, and their physical significance, apart from indicating large deviation from a Gaussian shape, is not readily interpretable. Also, it has not been possible to assign any quantitative degree of certainty to the histogram shape. Thus, no numerical measure for the goodness-of-fit of the Monte Carlo simulation and the measured histogram has been found.

The splitting of histograms into a steady (mean) part and unsteady part is a familiar process. Although each of the two peaks of a bimodal histogram represents a flow state, there is no analysis available to separate the unsteadiness in each state so that the two states may be analyzed separately. The goal of such an analysis should be a means of defining the set of time histories that may reasonably have yielded the histogram.

Extension of histogram applicability: The difficulties in histogram interpretation will be compounded when the laser velocimeter is structured to

simultaneously measure both velocity components so that pairs of components, (U_i, V_i) are recorded. The type of histograms that might result from this process is shown in figure 11. Although much more information is available in the two-dimensional histograms, the type of analysis needed to fully utilize this information is not available.

The need for histogram analysis will not disappear with the newer laser-velocimeter data-acquisition systems that record time of measurement. The histogram analysis requires at least an order of magnitude fewer velocity measurements and measurement rates than frequency spectrum representations. Conditional sampling techniques yield many histograms. Also, the histogram analysis will continue to be used for online confirmation of the data validity and online selection. An optimization of data-acquisition cost may eventually consist of histogram representation at most points in the velocity field and selective use of temporal or frequency analyses.

IV. TIMED VELOCITY MEASUREMENTS. In order to analyze velocity data by time-based methods, it is necessary to record the time lapse, or interarrival time, between successive velocity measurements. The task of measuring interarrival times has been performed by a clock with three ranges. For interarrival times between 0.1 μsec and 6.55 msec, the clock has a resolution of 0.1 μsec. Using automatic ranging, the clock measures up to 0.655 sec with a resolution of 1 μsec and up to 0.655 sec with 10 μsec resolution. The typical time scales for large-scale wind-tunnel power spectrum measurements are shown in Table I.

TABLE I.- TYPICAL TIME SCALES SUITABLE FOR POWER SPECTRA

Maximum resolution of the interarrival clock	0.1 μsec
Instrument reset time between measurements	0.4 μsec
Time required for one velocity to be measured	2 μsec
Residence time of a particle in sample volume	2 to 20 μsec
Average particle arrival time (T/D)	0.5 to 2 msec
Maximum interarrival time measurable by clock	0.655 sec
Measurement period, T	2 to 100 sec

Since existing instrumentation is capable of making a velocity measurement about every 2.4 μsec (this depends on particle velocity, Bragg cell frequency, and fringe spacing), the limiting factor on data rate is the rate at which particles pass through the sample volume. Since each particle must pass through 10 fringe planes (spaced, in the second test, 26.5 μm apart) in order to register a velocity measurement, and since the planes are moving in the measurement direction at a speed governed by the Bragg cell frequency (e.g., 132 m/sec), the time required for one velocity measurement is the reset time plus (10 x fringe spacing)/(Bragg velocity + measured particle velocity). The average particle interarrival time depends on the average flow velocity, the diameter of the sample volume (e.g., 0.314 mm), and the density of particles of measurable size. Although the minimum time between measurements will vary from 2.4 μsec for various test conditions, it is unlikely to be a restrictive factor in the data analysis. This can be demonstrated by a comparison to the average data rate.

The arrival of particles through the sample volume has been found to approximate a Poisson distribution in time (Ref. 4). This distribution takes the form

$$p(n;\Delta T) = \frac{(\lambda\Delta T)^n e^{-\lambda\Delta T}}{n!} \quad n = 0, 1, 2 \ldots$$

where λ is the mean particle arrival rate. In an experimental test case, where λ was 317 measurements per second, it was found that the limitations imposed by the system, minimum interarrival time of 2.4μsec and the maximum interarrival time of .655 second, pose no limitations on the measurement of interarrival times, figure 12.

V. POWER SPECTRA. The best developed method of presentation of the time dependence in unsteady flows is power spectra. The most accurate calculation method that has been found to use the laser velocimetry measurements for power spectra is an indirect method. The first step is the calculation of a weighted estimate of the autocovariance. In order to apply a fast Fourier transform to obtain the power spectra, the autocovariance is extended to form an even function. This method has been selected over Fourier series methods and periodgram methods entirely on a trial and error basis (Ref. 4). Its superiority has not been established analytically, and there is little understanding of the reasons for the smaller errors that result from the autocovariance approach.

Formation of the autocovariance: The autocovariance estimate $C(k\Delta\tau)$ for $k = 1 \ldots K$ is based on a minimum lag time $\Delta\tau$. The value of $\Delta\tau$ must be greater than the resolution of the interarrival clock. However, much larger values are required to avoid excessive errors. Of course, $K\Delta\tau$ must not exceed the measurement period, T.

These limitations on the choice of $\Delta\tau$ and K are the same as they are in determining the autocovariance function of a uniformally sampled data set from a continuous signal. That is, the frequency resolution is determined by $\Delta f = 1/2K\Delta\tau$ where the maximum (possible lag) value of $K\Delta\tau$ is the total measurement time, T, and the maximum frequency is limited by $f_{max} = 1/2\Delta\tau$ where $\Delta\tau$ is the minimum time between samples. For the system under consideration, the minimum possible $\Delta\tau$ is 2.4μsec and the maximum value of $K\Delta\tau$ is related to the average arrival rate by $K\Delta\tau = 4096/\lambda$ where the value 4096 is the maximum number of measurements that can be stored in the memory buffer and λ is the mean data sample rate (mean particle arrival rate). However in a practical random sampling situation, the choice of K and $\Delta\tau$ should be made with regard to the following: 1) desired frequency resolution and resulting variability error, and 2) the value of $\Delta\tau$ must be chosen so that K is fixed at 512 due to data processing limitations. If the chosen value of $\Delta\tau$ is found to be undesirable, the fact that the data is sampled randomly in time allows another choice of $\Delta\tau$ to be made without repeating the experiment.

To obtain the autocovariance, the mean velocity is subtracted from each measurement to give velocity values U_i, $i = 1 \ldots D$. The time delay between two velocity values U_i and U_ℓ has been recorded and is denoted as $t_i - t_\ell$.

The lag product, $A_u(k)$ is defined as

$$A_u(k) = \sum_{i=1}^{D} \sum_{\ell=i}^{D} U_i U_\ell S[(t_i - t_\ell - (k - \tfrac{1}{2})\Delta\tau)((k + \tfrac{1}{2})\Delta\tau - t_i + t_\ell)]$$

where $S(x) = 1$ for $x \geq 0$

$= 0$ for $x < 0$

Thus only those velocities whose interarrival times, $t_i - t_\ell$, lie between $(k - .5)\Delta\tau$ and $(k + .5)\Delta\tau$ contribute to the lag product.

For efficient computation, each possible pair of velocities is examined; the lag time ratio, $\frac{t_i - t_\ell}{\Delta\tau}$, is calculated; and no action is taken if for that pair the time lag ratio is not less than K. Otherwise the product $U_i U_\ell$ is added to the k^{th} location (where k is the integer nearest the lag time ratio) of the array $A_u(k)$ and k^{th} location in array H(k) is incremented by one. The accumulations of $A_u(k)$ and H(k) found for several separate periods (e.g., lots of 4096 velocity measurements) of measurement may be summed. The autocovariance is then formed as

$$C(k\Delta t) = \frac{A_u(k)}{H(k)} \text{ for } k = 1 \ldots K$$

This has been shown to be an unbiased estimate (Ref. 5) if the true mean velocity has been subtracted from the data. The variance of $C(k\Delta t)$ is, under very restrictive assumptions,

$$\sigma_k^2 = \frac{\sigma^4 + C^2(k\Delta\tau)}{H(k)}$$

where σ^2 is the variance of the velocity data. No error estimate is available for spectra that do not describe a stationary broadband Gaussian process. The value of C(0) may be calculated by the above scheme or more simply as

$$C(0) = \frac{1}{D} \sum_{i=1}^{D} U_i^2 = \sigma^2$$

Power spectra results: The power spectrum is the Fourier transform of the autocovariance. Figure 13 shows results obtained by calculating a 512 bin autocovariance array, and then defining an additional 512 bins so that an even function is formed. This allows a fast Fourier transform using Bartlett (triangular) window and a frequency resolution of $\Delta f = 1/2K\Delta\tau$ to be used.

Figure 13(c) is a power spectrum of the V_L velocity component of the circled point in figure 7. About 40,000 measurements, taken at an average data rate of about 400 measurements per second, were used. The minimum lag time, $\Delta\tau$, for the autocovariance was 0.5 msec and 512 values of lag time were calculated. These data were taken in 10 lots with about 4000 measurements in each.

Figure 13(a) shows the distribution of number of lag products, $H(\Delta t)$. The histogram bin size is 0.5 msec. The autocovariance is shown in figure 13(b). The spectrum is displayed in figure 13(c) up to the maximum calculated frequency of 1 kHz. The reason for the negative values that occur above 280 Hz is not known. The negative values persisted for the recalculation of the power spectrum for a doubled (1 msec) minimum lag time (Fig. 13(d)). Since negative power is undefined, the source of the anomalous behavior above 280 Hz must be an error. At this particular data rate, for this 2 Hz frequency resolution, and for 40,000 measurements, a satisfactory spectrum has apparently been calculated up to 280 Hz although attempts to calculate this spectrum with only 4000 measurements gave very inconsistent curves. There is, unfortunately, no general analysis of the error in the spectra, no theory that explains how the researcher may compensate for a low data rate by increasing the number of measurements, and especially no means of calculating the effects of the known random measurement errors in velocity and time.

Frequency limitation on spectra: Work is proceeding, on an experimental basis, on the maximum frequency limitation. Because of the random interarrival times of the velocity measurements, the Nyquist criterion does not apply to the average data rate. Theoretically, a maximum frequency far exceeding the Nyquist criterion could be achieved. In simulations (Ref. 4) and experiments, attempts to exceed twice the Nyquist criterion frequency have led to erratic spectra. A method of predicting the data rate, number of samples, and error bounds necessary to achieve a desired frequency limit and accuracy is needed.

Although satisfactory power spectra have been obtained by the indirect method using an autocovariance estimate, it is possible that an improved technique could be devised.

VI. ANALYSIS OF PERIODIC PHENOMENA. As research into the fluid mechanics of turbulent flow has progressed, more patterns have been discerned in flow fields traditionally considered to be random variations in velocity (Ref. 6). The aerodynamicist must also analyze patterned or organized flow even in the presence of random or broadband velocity fluctuations (Ref. 7). Flows such as those beneath a helicopter rotor (Ref. 8) are difficult to break into three categories as suggested in reference 9: mean flow, organized or patterned or repetitive velocities, and random velocity fluctuations.

The spectrum in figure 14 is taken from a heuristic test in a water tunnel. An oscillating vane imposed a discrete frequency oscillation on the axial flow velocity. The spectra alone is sufficient to relate the velocity response to the vane oscillation to the magnitude of the random turbulence and confirm the absence of higher harmonics.

The spectrum is not sufficient to define the phase angle between the flow and vane oscillation. This information can be obtained by a conditional sampling technique. The preferred method is to record, at the time of each velocity measurement, the vane angle. A plot of velocity versus vane angle will reveal the phase, the waveform of the response, and the variance of the response at each vane angle.

A similar technique has been applied to a rotor tip vortex flow (ref. 10) and will be applied to an oscillating airfoil test now under construction. The airfoil is large enough (66 cm in chord) to contain a small laser velocimeter. Figures 15 and 16 show the airfoil mounted in Langley Transonic Dynamics Tunnel. As the airfoil oscillates, the flow velocity with respect to the airfoil-fixed coordinates and roughly parallel to surface will be measured. This velocity is expected to be highly nonsinusoidal due to presence of shock waves. Therefore, the primary data reduction technique will be conditional sampling based on the airfoil angle. Other techniques could be devised (such as basing the condition on the leading edge pressure drop as zero time and plotting against time). Each experimental setup will contain unique conditional sampling techniques but each will share three common demands on the data acquisition and reduction process.

1. Each velocity will be linked to a condition of some measurement.

2. Velocities must be sorted into bins on the basis of the condition measurement.

3. Each bin must be analyzed by the techniques developed for histograms.

The use of current and planned methods of data analysis of laser velocimeter measurements will allow the aerodynamicist to investigate flow fields that have not been amenable to probe investigation. Such complex flow fields as helicopter rotor wakes, separated and recirculating flow on airfoils, and transition from laminar to turbulent flow are especially likely candidates for laser velocimeter measurements. Because of the continuing need for evermore penetrating analysis of experimental data and because of random phenomena that occur in each of these flows, a need will arise for the reduction of laser velocimeter data by such techniques as temporal and spatial cross-correlation of two velocities, time history reconstruction, moving block analyses, and pattern recognition.

The potentialities and difficulties of the more refined data-analysis techniques will become more apparent as deeper understanding of conditional sampling and power spectrum technique is gained. However, these two techniques are clearly not the end point of laser velocimeter data analysis.

VII. CONCLUSIONS. The rapid development of the laser velocimeter as a routine tool for flow measurement in large wind tunnels has given rise to new demands for data interpretation and analysis. The distinctive characteristics of laser-velocimeter systems with respect to the traditional flow measurement systems are primarily the following:

1. Rapid acquisition of thousands of individual, digital velocity measurements is possible at data rates limited, at present, only by the capacity of the wind-tunnel seed-particle injection process.

2. The seed velocity measurement errors are not only small but they are predictable before the experiment is begun.

3. The time between measurements is a random variable.

These distinctive characteristics offer opportunity for more precise control of errors and more efficient and more complete analysis of time-dependent data and real-time data analysis. To achieve these advantages much work remains to be done on both the existing methods of analysis, such as histogram displays and power spectra, and on methods now being developed such as two-component histograms, conditional sampling, and cross-correlation.

Histograms are the most efficient means of data interpretation because of much lower requirements for the amount of data and data rate. Better means of quantizing histogram shape and errors in shape are needed, especially for bimodal and other highly non-Gaussian shapes. Even a shape classification to guide the present cumbersome Monte Carlo methods would be helpful. The use of histograms in the conditional sampling process will compound the need for these analytical tools.

Also urgently needed are better error analyses for power spectra. Any optimization of the calculation process would be very useful.

The problems that will be presented by the untried data reduction techniques are not as clearly defined as for the histogram and power spectra. These problems may include the reconstruction of time histories from data with random interarrival times, error analysis of cross-correlations using noncoincident (in time) measurements of the two velocity components, and sufficiency conditions for moving block analyses.

REFERENCES

[1] W. H. Young, Jr.; J. F. Meyers; and T. E. Hepner, "Laser Velocimeter Systems Analysis Applied to a Flow Survey Above a Stalled Wing," NASA TN D-8408 (1977).

[2] W. G. Tiederman; D. K. McLaughlin; and M. M. Reischman, "Individual Realization Laser Doppler Technique Applied to Turbulent Channel Flow," Proc. of Symposium on Turbulence in Liquids (Rolla, Missouri) (1973).

[3] G. Udny Yule; and M. G. Kendall, "An Introduction to the Theory of Statistics," Charles Griffin & Co., Ltd. (1940).

[4] W. T. Mayo, Jr.; M. T. Shay; and S. Riter, "Development of New Digital Data Processing Techniques for Turbulence Measurements with a Laser Velocimeter," AEDC TR-7453 (Aug. 1974).

[5] W. T. Mayo, Jr.; M. T. Shay; and S. Riter, "Digital Estimation of Turbulence Power Spectra from Burst Counters," Second International Workshop on Laser Velocimetry - Proc., H. D. Thompson and W. H. Stevenson, eds., Purdue University, (March 1974).

[6] Anatol Roshko, " Structure of Turbulent Shear Flows: A New Look," AIAA Paper 76-78 (Jan. 1976).

[7] F. O. Carta, "Analysis of Oscillatory Pressure Data Including Dynamic Stall Effects," NASA CR-2394 (1974).

[8] J. C. Biggers; and K. L. Orloff, "Laser Velocimeter Measurements of the Helicopter Rotor-Induced Flow Field," J. of the American Helicopter Society, vol. 20, no. 1, (Jan. 1975), pp. 2-10.

[9] D. P. Telionis, " Unsteady Boundary Layers, Separated and Attached," Symposium on Unsteady Aerodynamics - Proc.. AGARD, Ottawa, Canada (Sept. 1977, pp. 16.1-18.

[10] F. K. Owen, "Measurement of Unsteady Vortex Flow Fields," to be published AIAA Sixteenth Aerospace Sciences Meeting, Huntsville, Ala. (Jan. 1978).

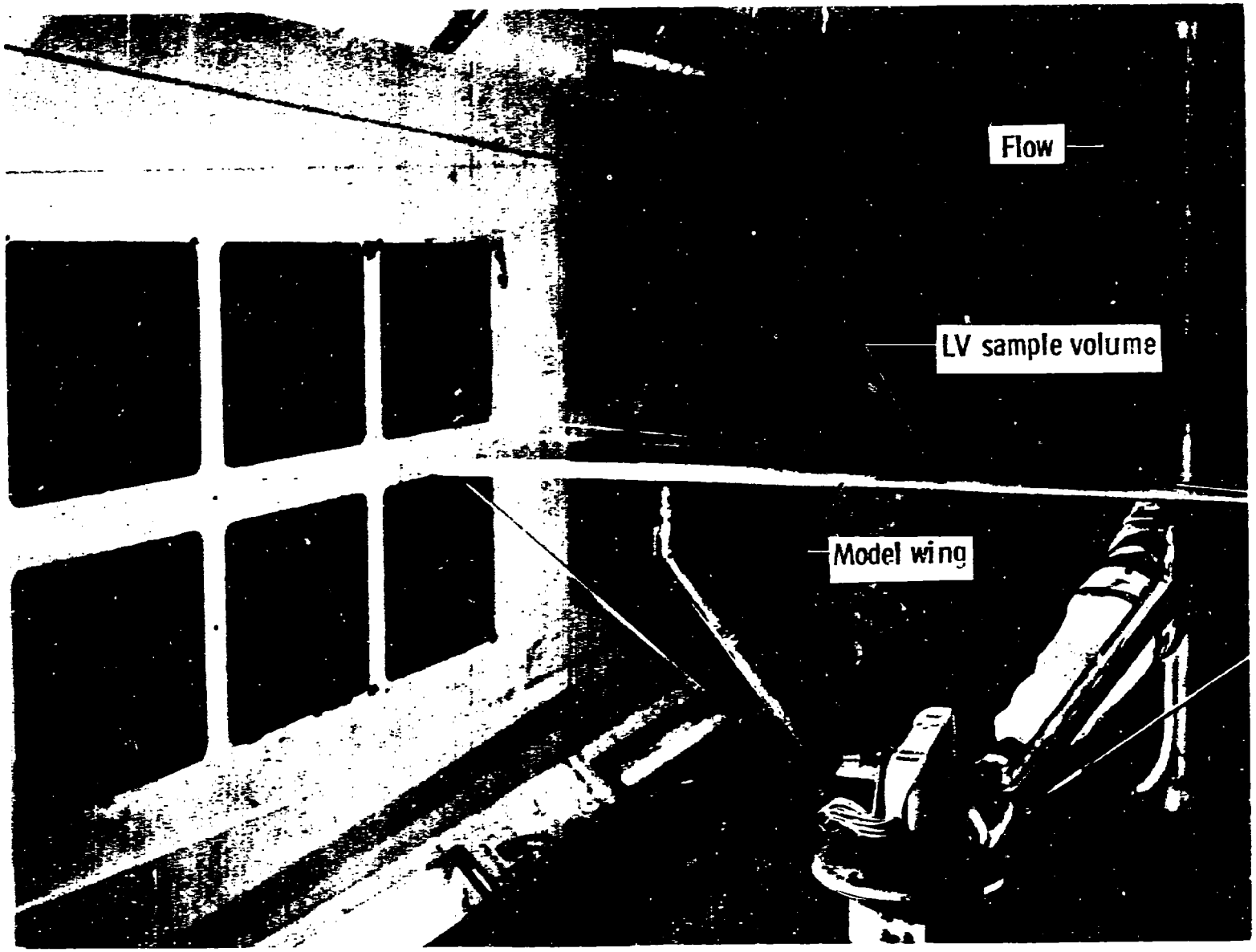


Figure 1.- Laser velocimeter measurements in the Langley high-speed 7- by 10-foot wind tunnel.

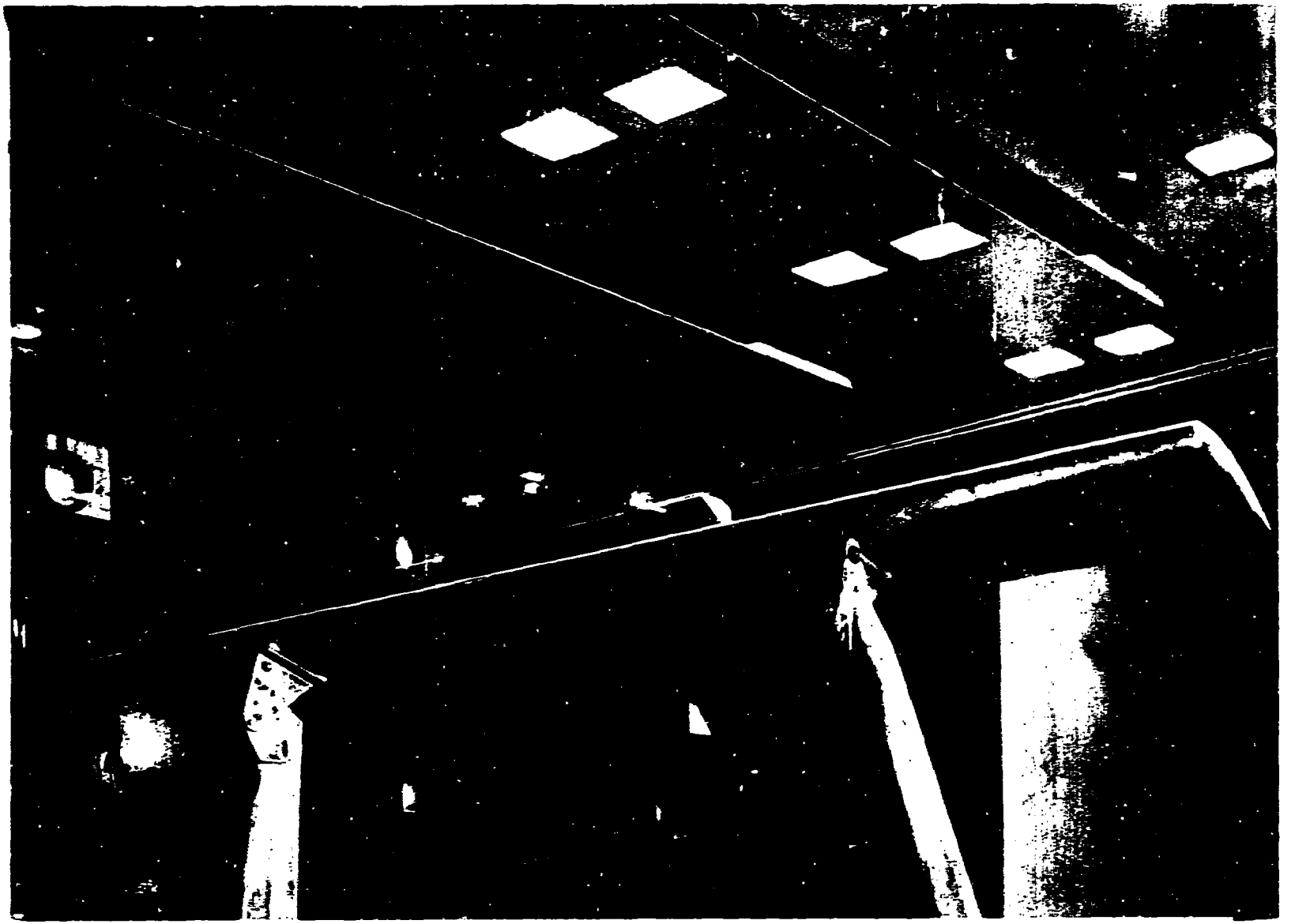


Figure 2.- Laser velocimeter measurements in the Langley V/STOL wind tunnel.

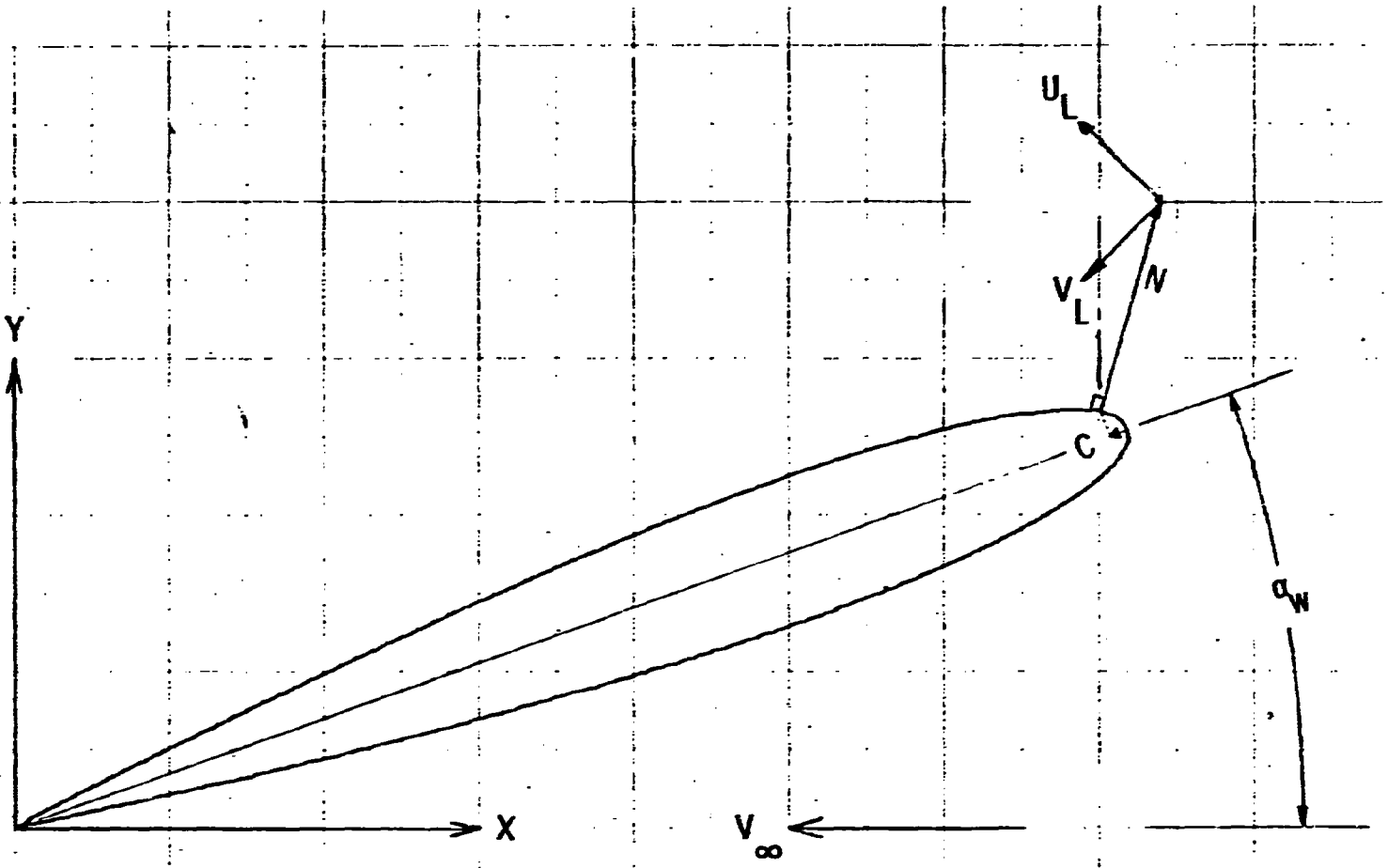


Figure 3.- Wing and laser velocimeter coordinate systems.

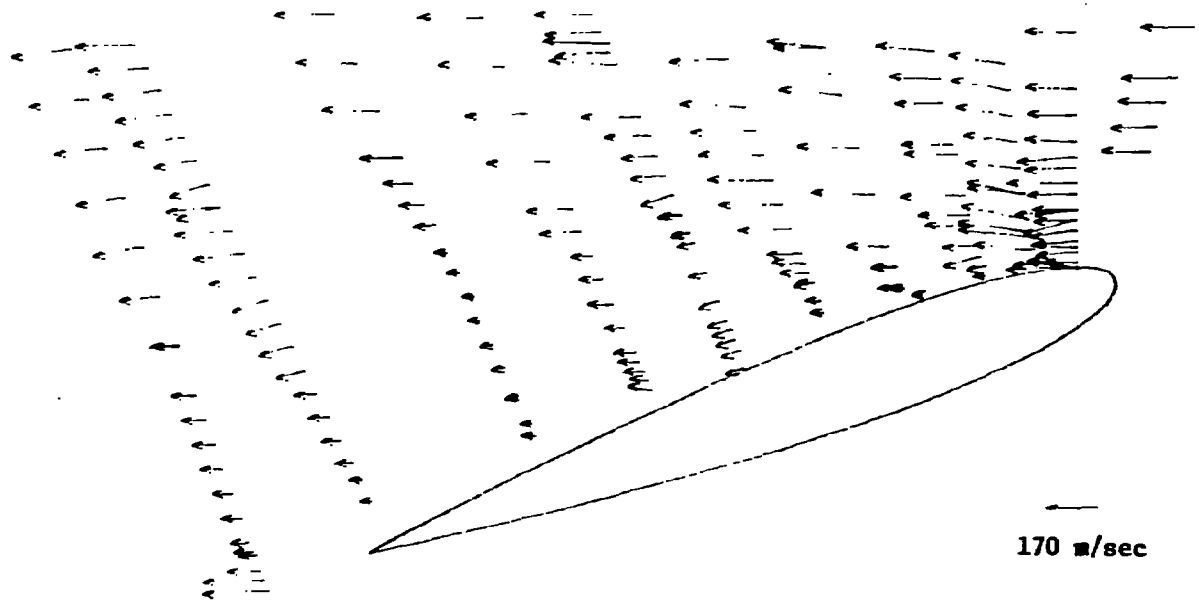


Figure 4.- Mean airspeed with all velocities shown in the prime direction.

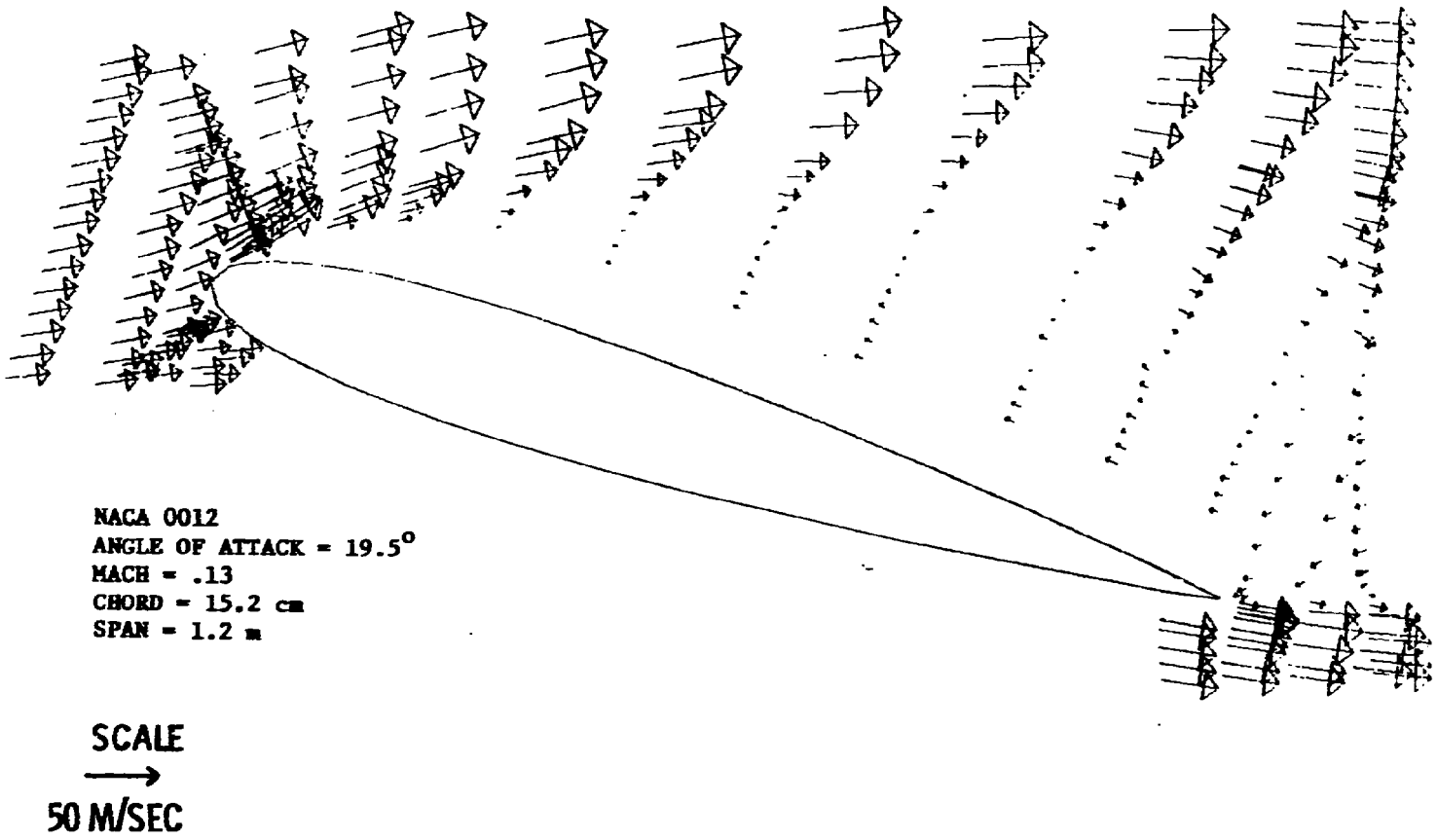


Figure 5.- Mean velocities above a stalled wing.

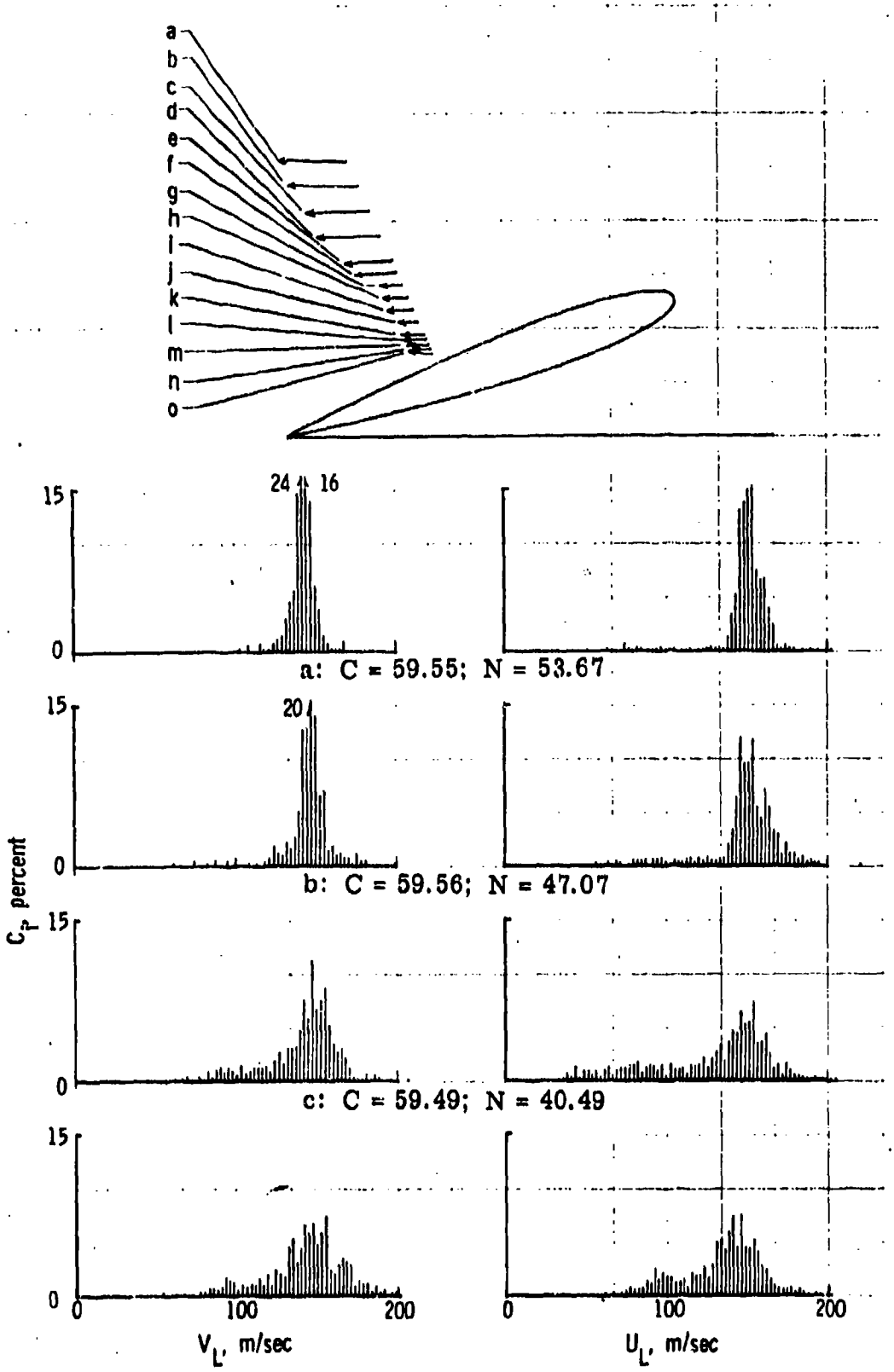


Figure 6. - Histograms above transducer 15.

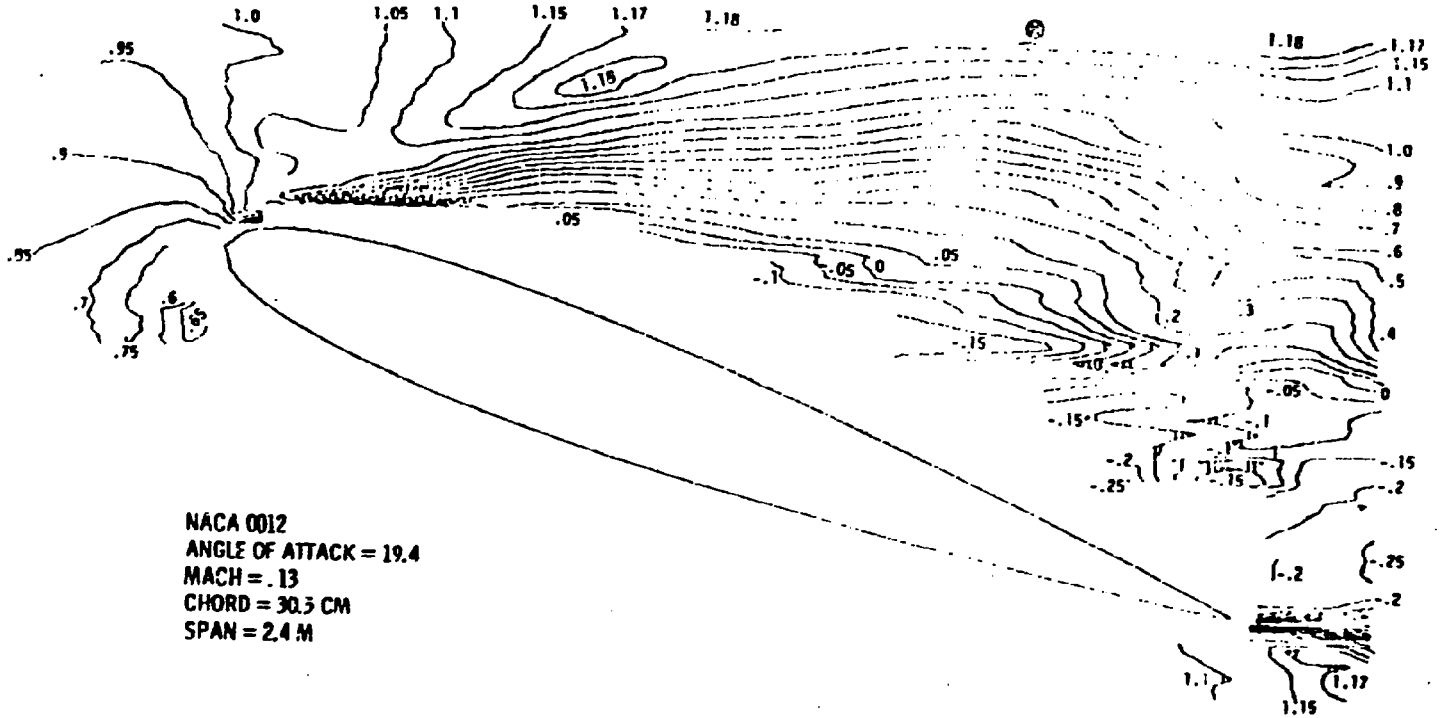


Figure 7.- Contours of constant resultant mean velocity normalized by the freestream velocity.

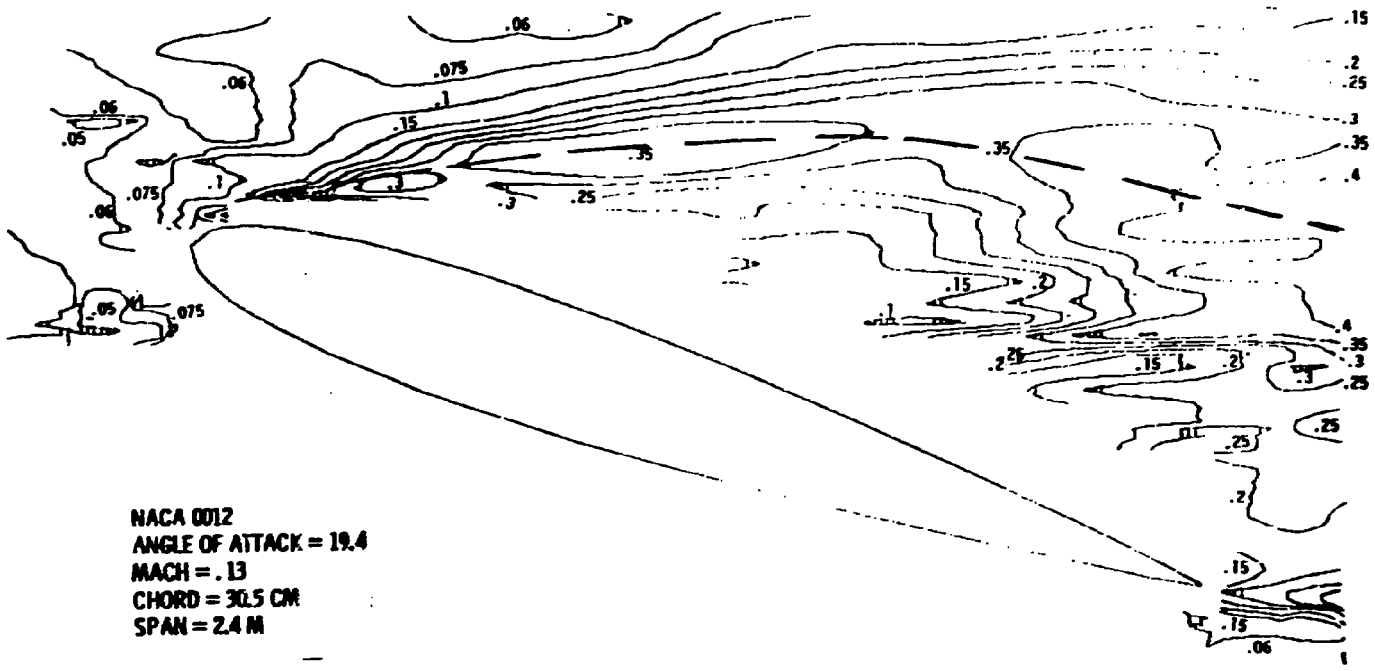


Figure 8.- Contours of constant values of standard deviation of velocity normalized to freestream velocity.

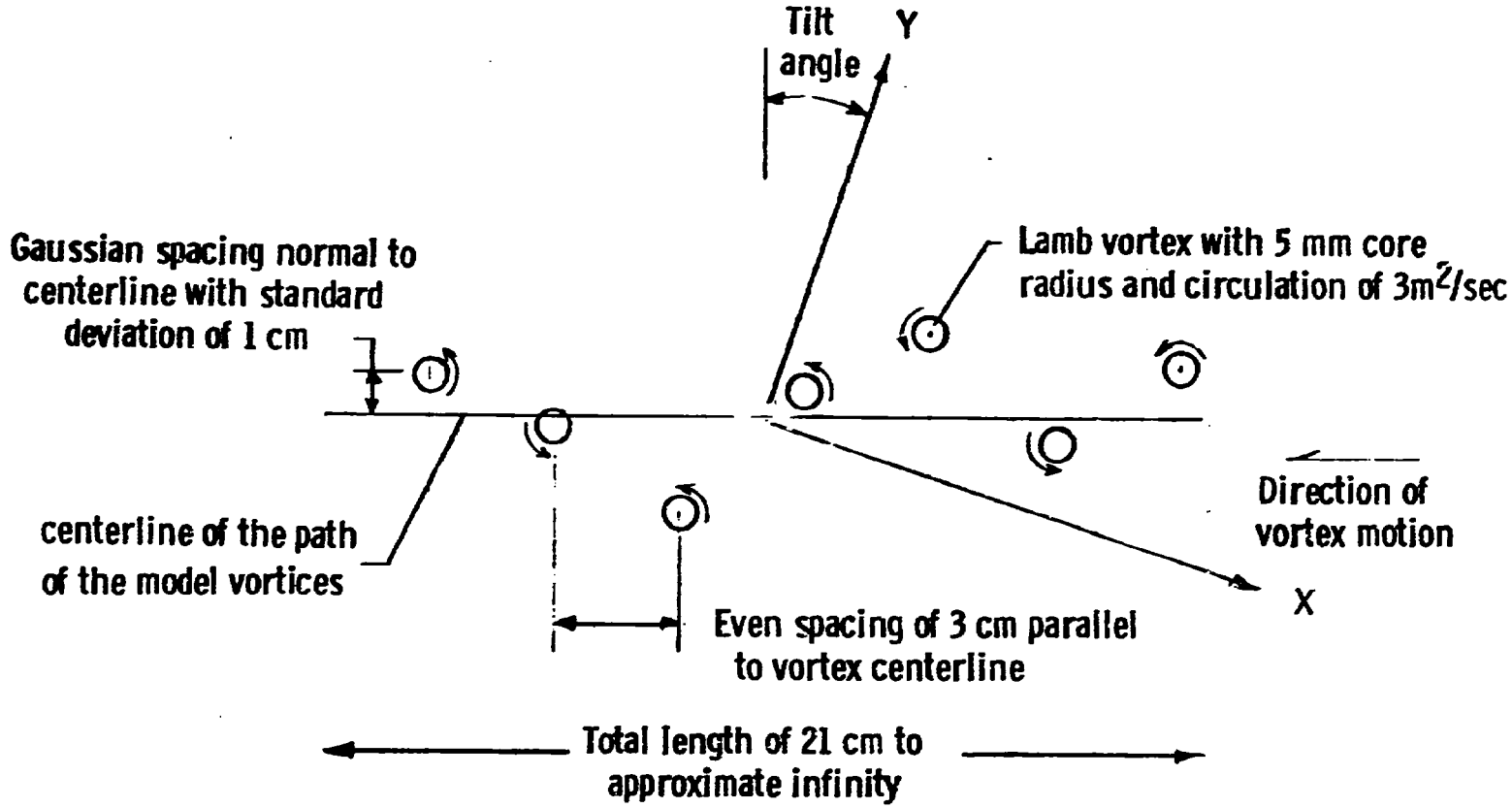


Figure 9.- Model of vortex stream.

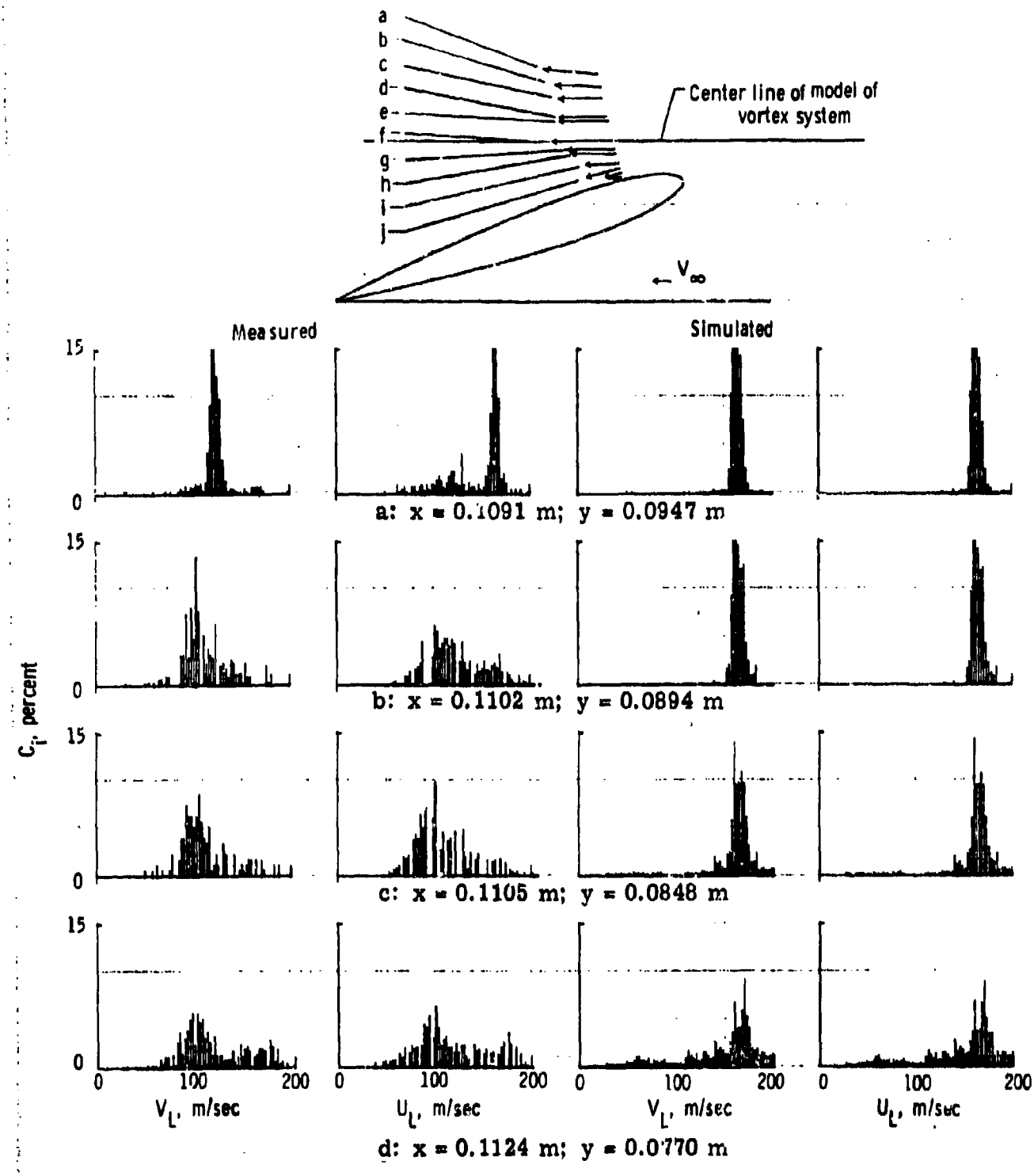


Figure 10.- Measured (left 2 columns) and simulated (right 2 columns) histograms above transducer 7.

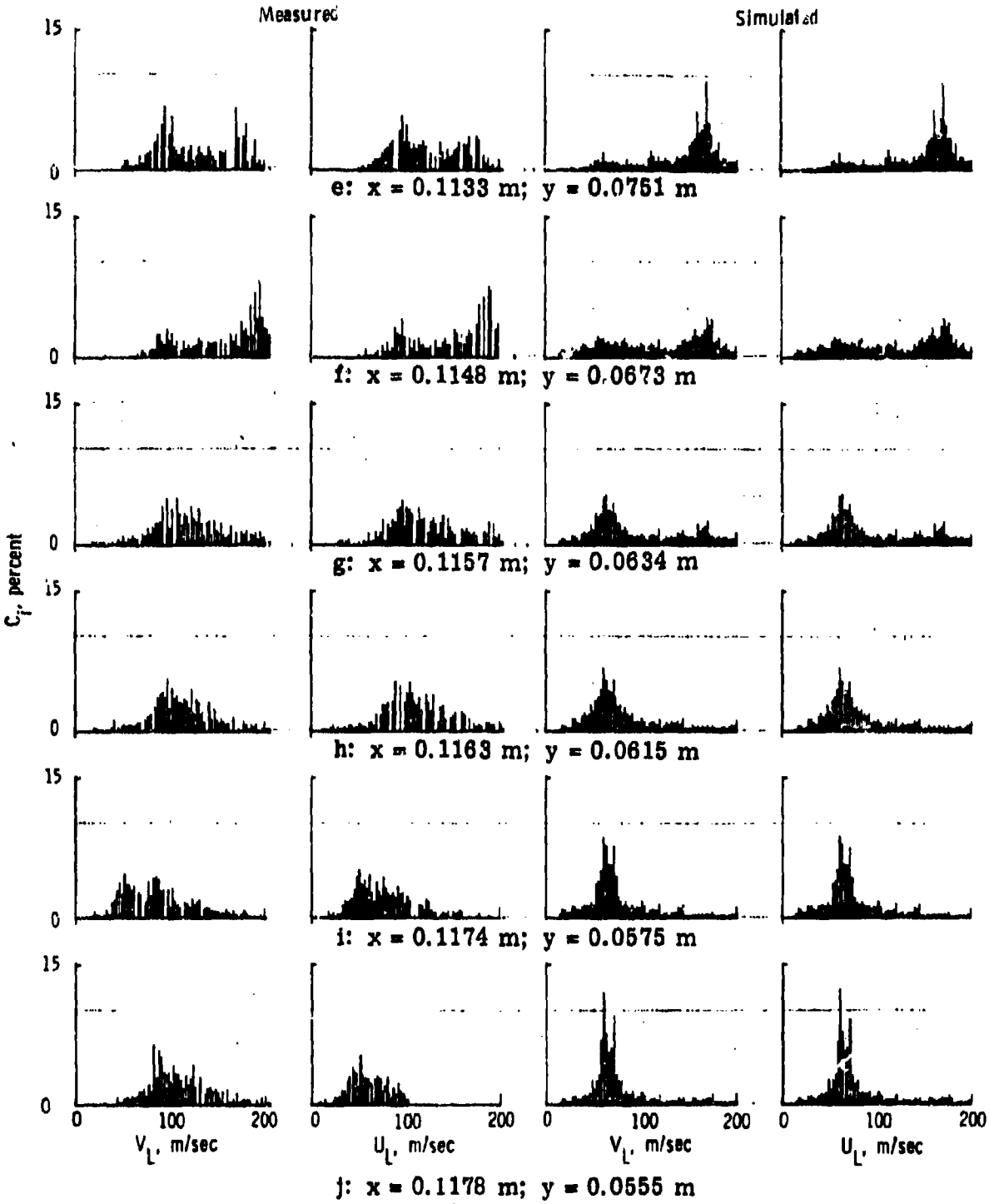


Figure 10.- Concluded.

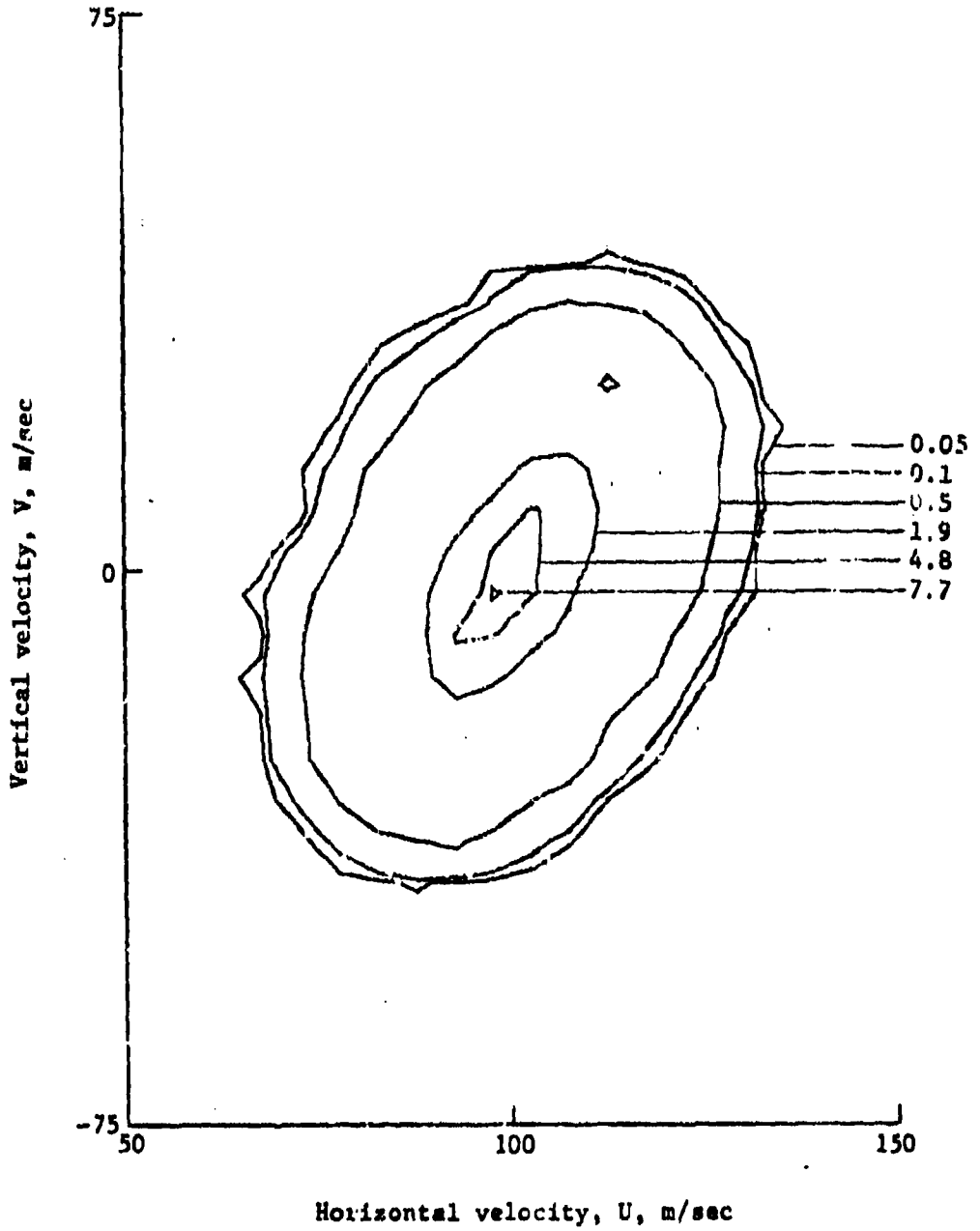


Figure 11.- Percentages of simulated pairs of measured velocities that fall within 2.5 m/sec of the U,V graph value.

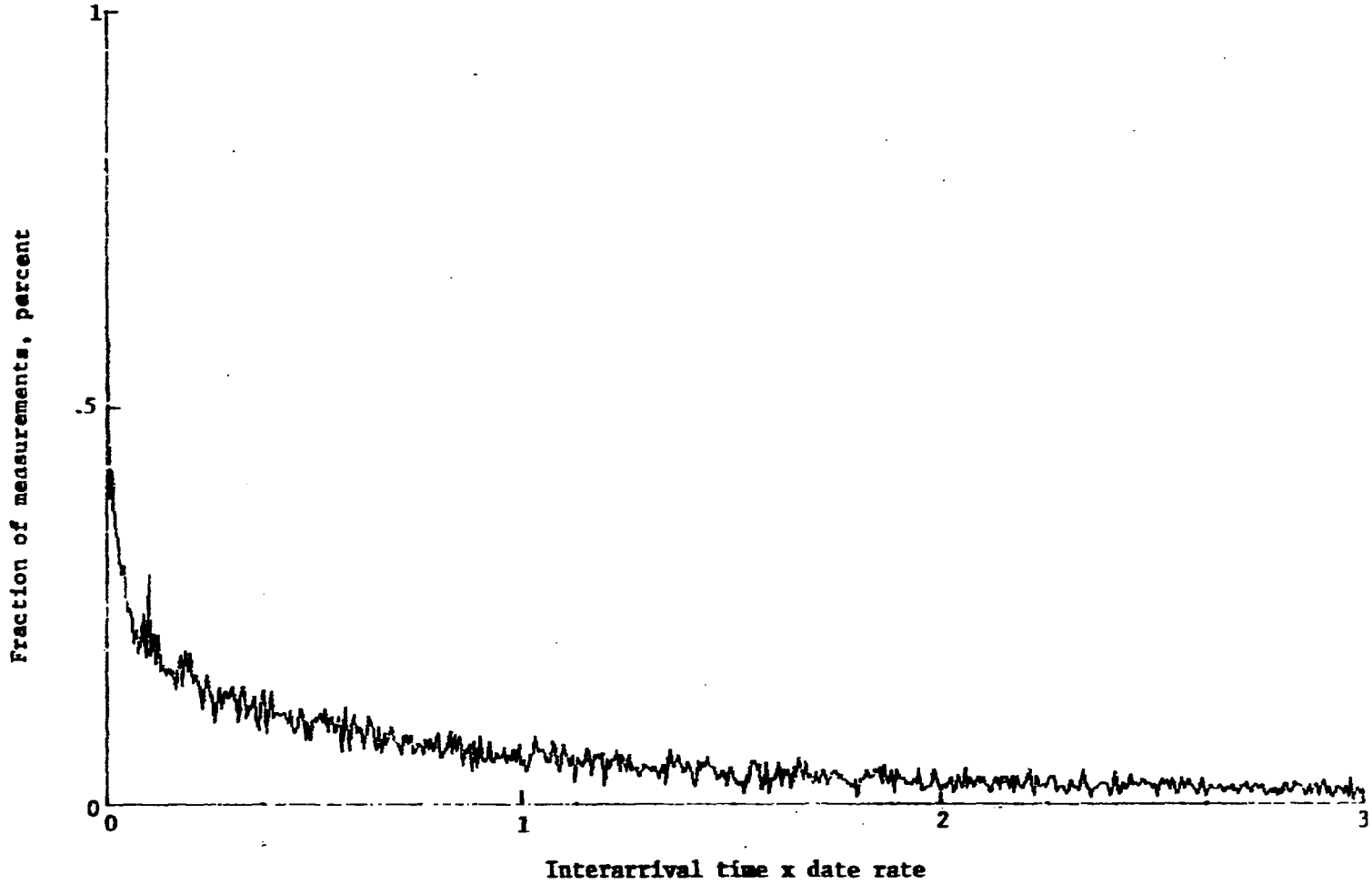
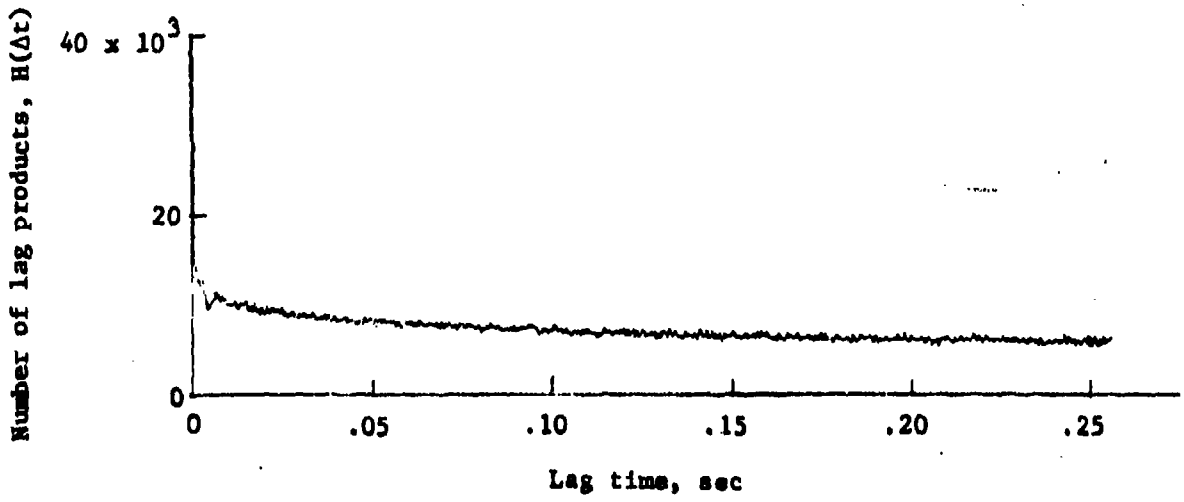
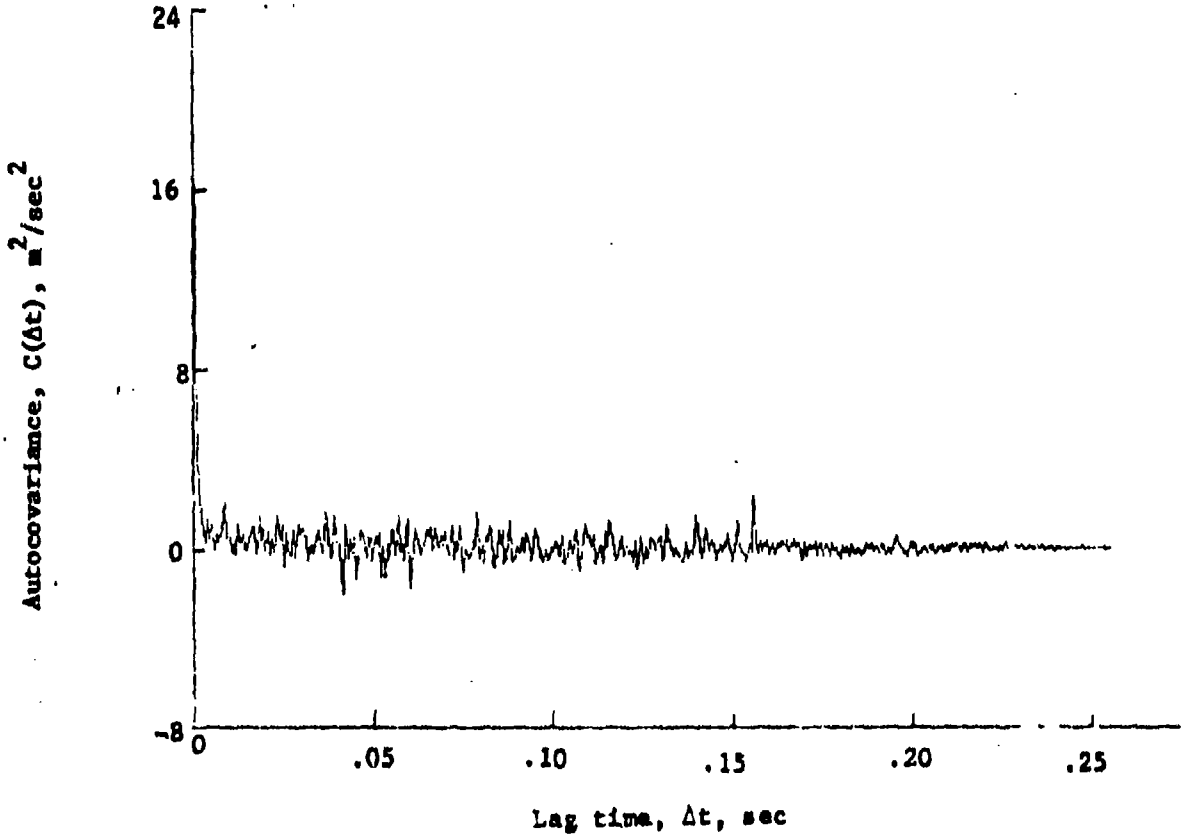


Figure 12. - Poisson distribution of interarrival times.

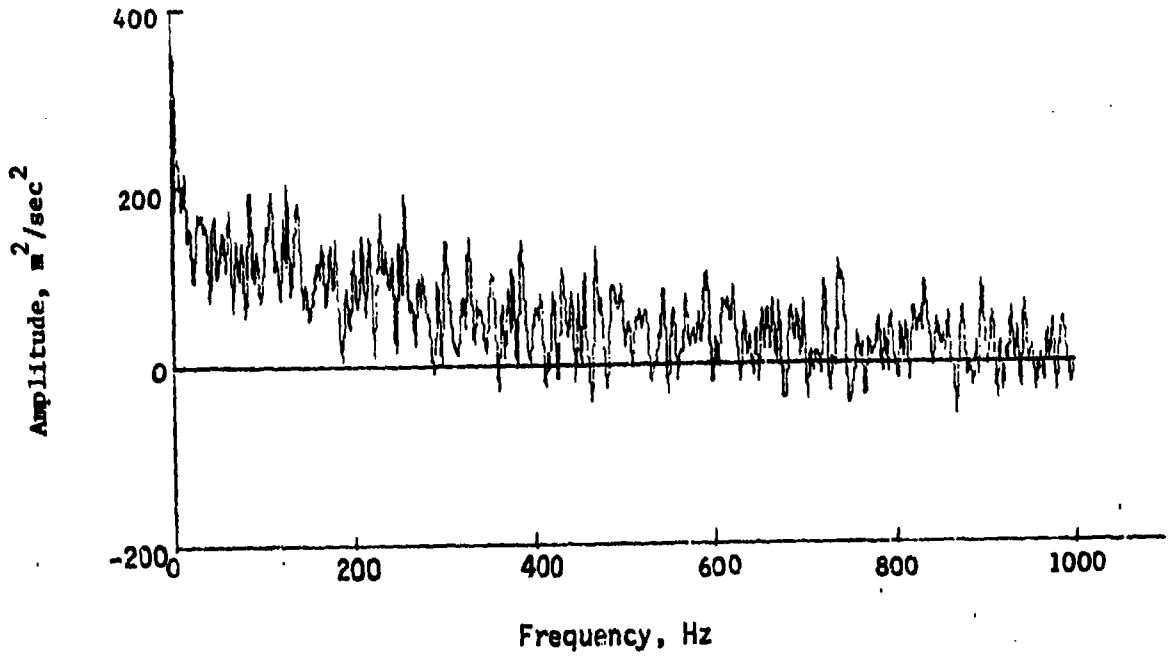


(a) Histogram of number of lag products

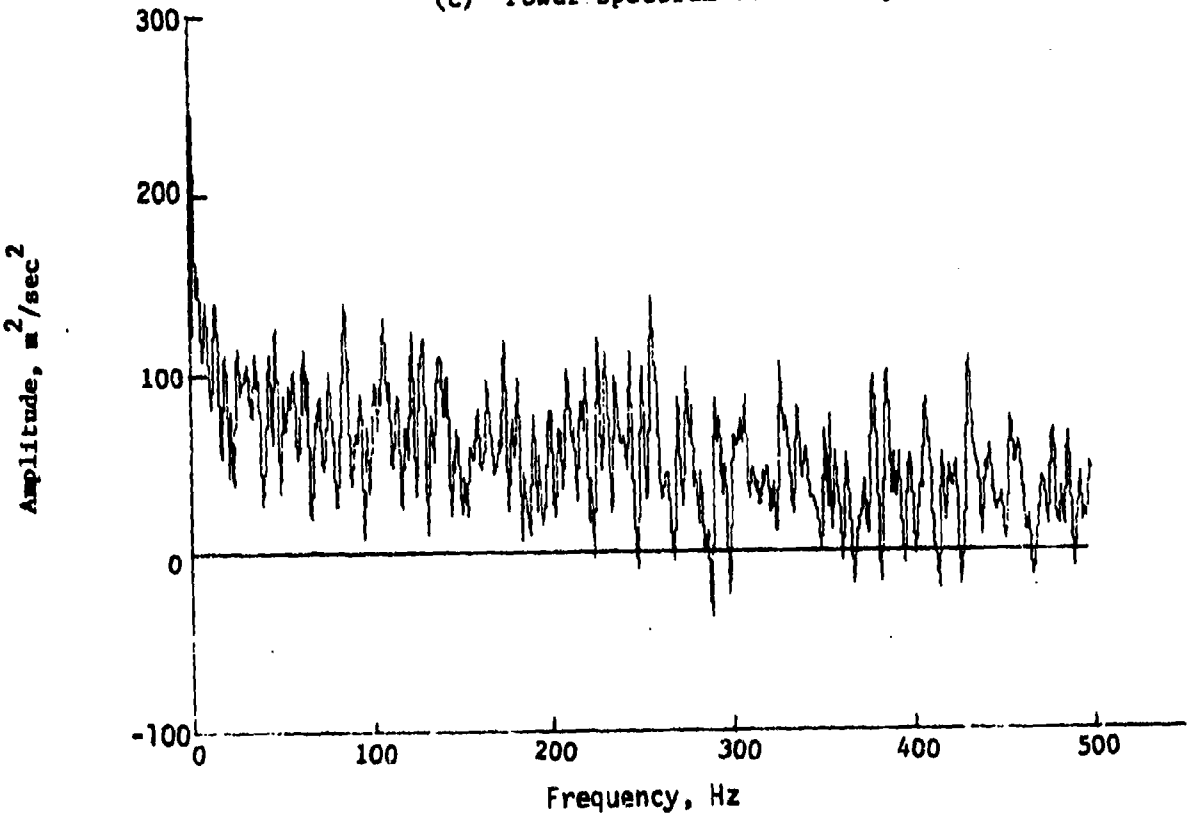


(b) Autocovariance.

Figure 13. - Calculation of power spectrum for stalled airfoil at Mach number 0.13.

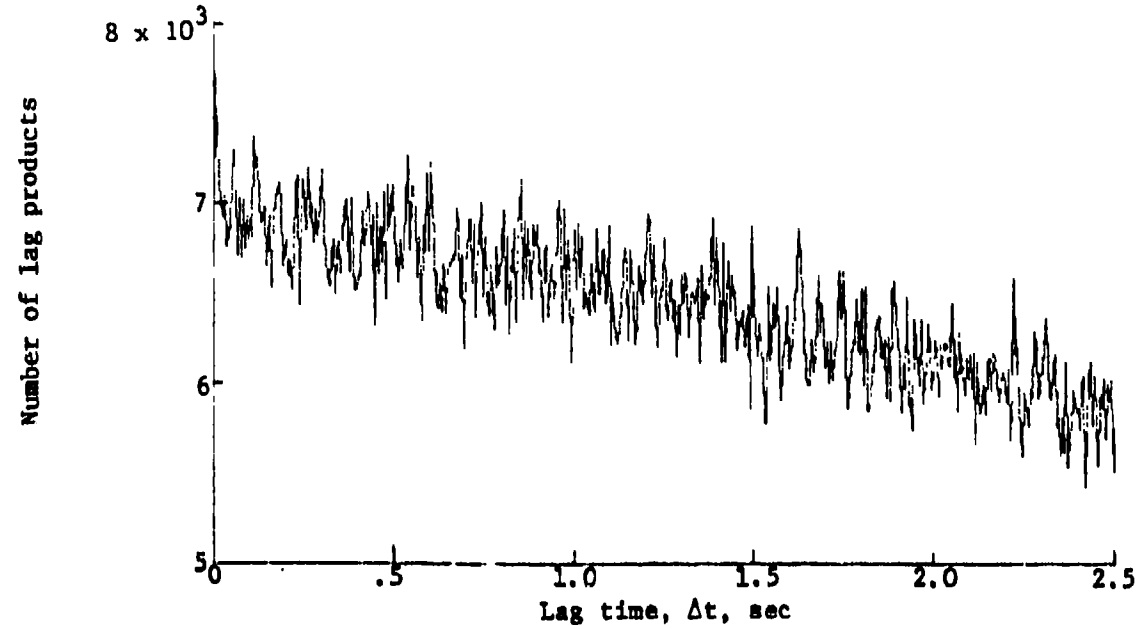


(c) Power spectrum of velocity.

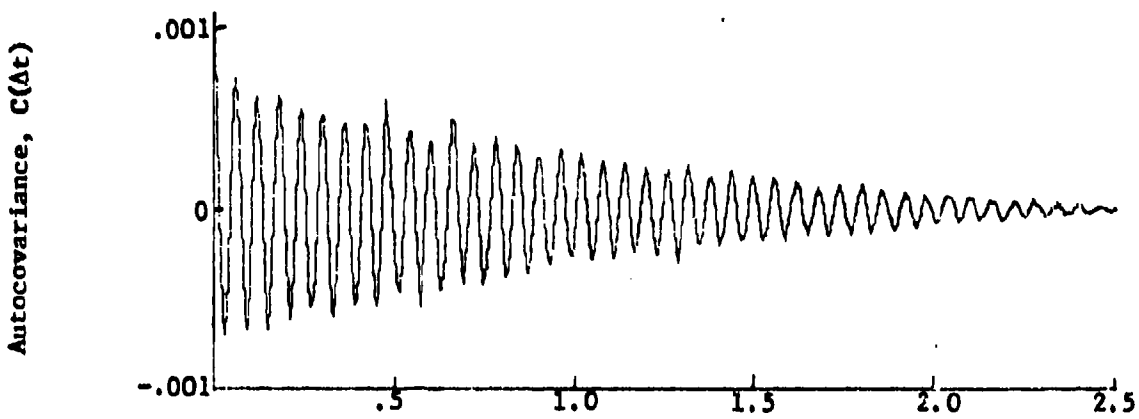


(d) Power spectrum for doubled minimum lag time

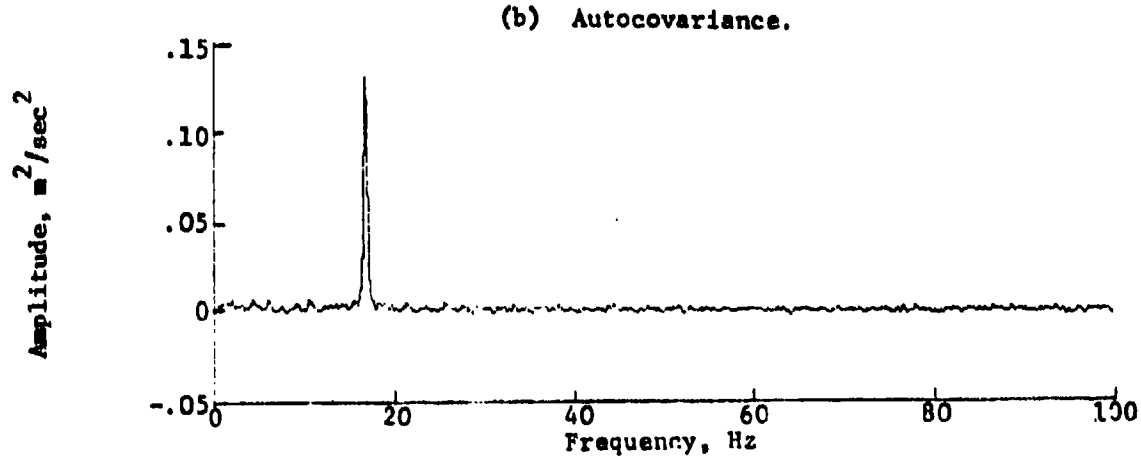
Figure 13. - Concluded.



(a) Histogram of number of lag products.



(b) Autocovariance.



(c) Power spectrum of velocity.

Figure 14. - Calculation of power spectrum from water tunnel with oscillating vane.

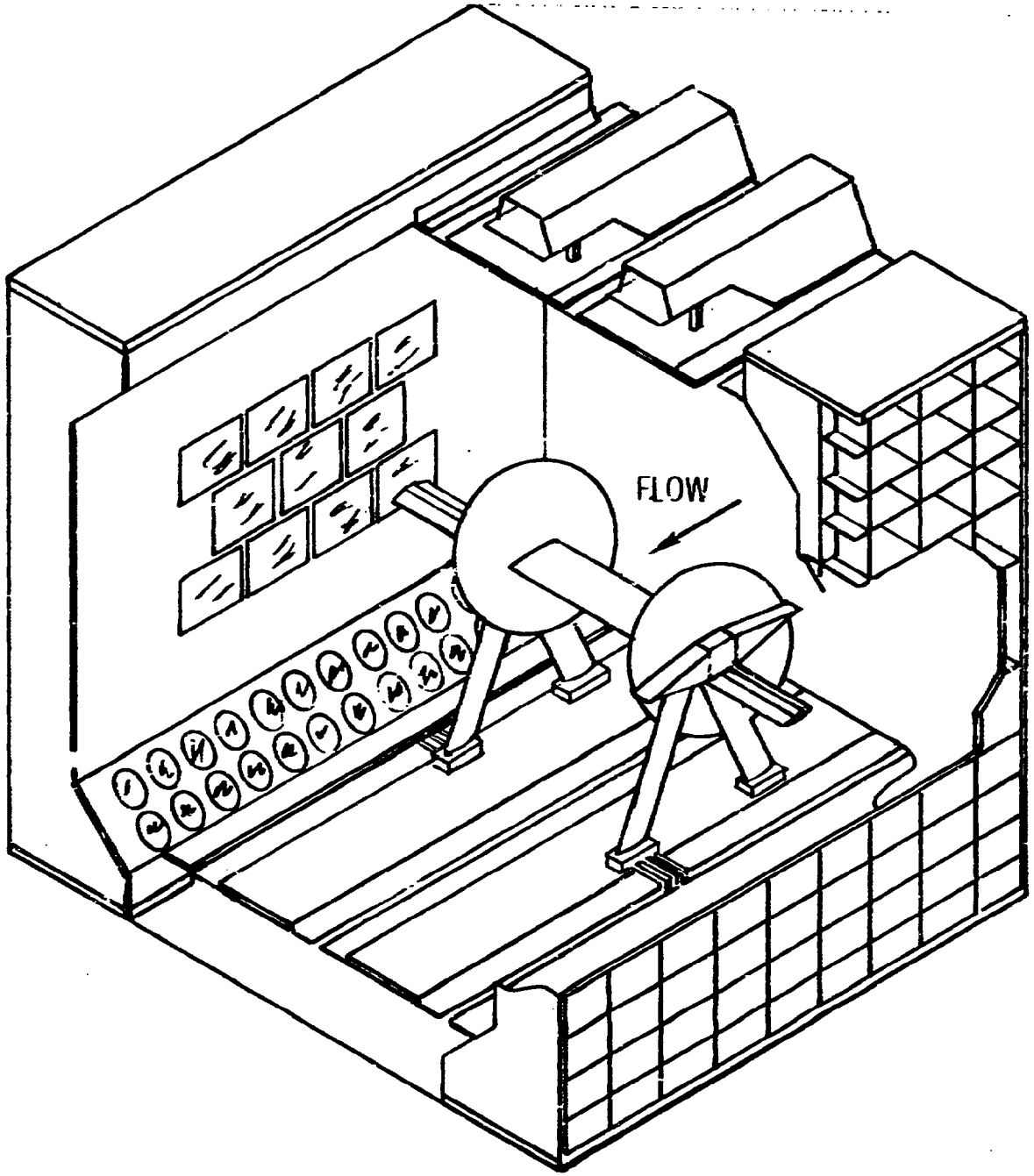


Figure 15.- Pitch Rig installed in the Langley Transonic Dynamics Tunnel.

Figure 16. - Pitch Rig installation in the Langley TDT.

RESOLVING UNDER-IDENTIFICATION THROUGH REPLICATION IN STRUCTURAL EXPERIMENTAL DESIGN

William S. Mallios

BDM Services Co.

ABSTRACT: Application of structural regression to experimental design often results in under-identification. A remedy, albeit unrealistic, is to assume the structural system is diagonally recursive. Reexamination of this assumption leads to a measure of the degree to which structure has been resolved in a non-recursive system, assuming identification. To assure identification, the experiment should be replicated with one replication used as an instrumental variable for the other.

1. INTRODUCTION.

In the structural regression system

$$A\,\underline{y} = \Gamma\,\underline{x} + \underline{\delta}\,, \qquad (1.1)$$

$\underline{y}$ is a $p \times 1$ vector of endogenous variables, $\underline{x}$ is a $q \times 1$ vector of exogenous variables, $A(p \times p) = (\alpha_{hh^*})$ is the direct effect[1] of y_{h^*} on y_h, $\alpha_{hh} = 1$, $\Gamma\ (p \times q) = (\gamma_{hi})$, γ_{hi} is the direct effect of x_i on y_h, and $\underline{\delta}(p \times 1)$ is the model error vector. Assume that $E(\underline{\delta}) = 0$ and

[1] See [5] for definitions of direct, indirect, and overall effects.

$E(\underline{\delta}\,\underline{\delta}') = \Sigma_\delta$; i.e., $\underline{\delta} : (\underline{0}, \Sigma_\delta)$, where Σ_δ is non-singular and contains finite diagonal elements. Assuming $|A| \neq 0$ and premultiplying (1.1) by A^{-1} yields the reduced regression system

$$\underline{y} = A^{-1}\Gamma\underline{x} + A^{-1}\underline{\delta} = B\underline{x} + \underline{\varepsilon}\,, \tag{1.2}$$

where $B\ (p \times q) = (\beta_{hi})$, β_{hi} is the overall x_i effect on y_h, and

$$\underline{\varepsilon} : (\underline{0}, \Sigma), \quad \Sigma = A^{-1}\Sigma_\delta A'^{-1}.$$

Let $\underline{y}_h$ and $\underline{x}_i$ denote $n \times 1$ vectors of observations on y_h and x_i. Then the h-th model of (1.1) is written as

$$\underline{y}_h = Y_h\,\underline{\alpha}_h + X_h\underline{\gamma}_h + \underline{\delta}_h = Z_h\underline{\theta}_h + \underline{\delta}_h, \tag{1.3}$$

where $Y_h(n \times p_h) = (\underline{y}_{h*})$, $X_h(n \times q_h) = (\underline{x}_i)$, $Z_h = (Y_h | X_h)$, and $\underline{\theta}'_h = (\underline{\alpha}'_h | \underline{\gamma}'_h)$. I_n denotes the $n \times n$ identity matrix in the assumption $\underline{\delta}_h : (\underline{0}, I_n\,\sigma^2_{\delta_h})$.

Letting X denote the $n \times q$ matrix of all controlled variables, the ordinary least squares (OLS) estimate of B is

$$\hat{B} = (X'X)^{-1}X'Y. \tag{1.4}$$

Σ is estimated by

$$S = (Y - X\hat{B})'(Y - X\hat{B})/(n - q). \tag{1.5}$$

Assuming identification [2, 3, 5], $\underline{\theta}_h$ and Σ_δ can be estimated indirectly through the reduced system by equating $\hat{B}$ to $B = A^{-1}$ and S to $\Sigma = A^{-1}\Sigma_\delta A'^{-1}$. Alternatively[2], $\underline{\theta}_h$ can be estimated directly through two stage least squares (2SLS) estimation as follows:

[2] See [2, 3] for other estimation techniques.

$$\hat{\underline{\theta}}_h = [(Z_h' X)(X'X)^{-1}(X'Z_h)]^{-1}(Z_h' X)(X'X)^{-1}X'\underline{y}_h; \qquad (1.6)$$

$$\text{var } \hat{\underline{\theta}}_h = [(Z_h'X)(X'X)^{-1}(X'Z_h)]^{-1}\,\sigma^2_{\delta_h}$$

where $\sigma^2_{\delta_h}$ is estimated by $s^2_{\delta_h} = (\underline{y}_h - Z_h\hat{\underline{\theta}}_h)'(\underline{y}_h - Z_h\hat{\underline{\theta}}_h)/(n - p_h - q_h)$. This estimate derives its name from a conceptual, two-fold application of OLS; i.e., in the reduced regression for Y_h, giver by $E(Y_h) = X\,B_h$, B_h is estimated by (1.4); after replacing Y_h by $\hat{Y}_h = X\,\hat{B}_h$ in (1.3), OLS is applied a second time which leads to the result in (1.6); in this process, the uncontrolled Y_h has been replaced by a consistent estimate, $\hat{Y}_h$, which is treated statistically as if it were controlled; see [3, p. 153].

2. COMMENTS ON THE ASSUMPTION OF A DIAGONAL Σ_δ.

In (1.3) a natural question is regarding the appropriateness of estimating $\underline{\theta}_h$ by $(Z_h'Z_h)^{-1}Z_h'\,\underline{y}_h$, the OLS estimate. When A is triangular and Σ_δ is diagonal, the structural system is termed diagonally recursive [2, 4, 5] and the OLS estimate is consistent. Regarding the plausibility of the diagonally recursive assumption, a triangular A might be realistic so long as the experiment is designed with this restriction in mind; i.e., a triangular A implies that, during the course of the experiment, no causal feedbacks occur between any two y_h and y_{h*} and that no variable has an indirect effect on itself [5]. However, an assumed diagonal Σ_δ has far reaching implications. Under a diagonal Σ_δ, extraneous variables (EVs) comprising δ_h are independent from one model error to the next; i.e., if u_h and u_{h*} denote any two EVs making up a component part of δ_h and δ_{h*}, respectively, then a diagonal Σ_δ implies that $E(u_h u_{h*}) = 0$; otherwise, $E(u_h u_{h*}) \neq 0$, $E(u_h \delta_h) \neq 0$, and $E(u_{h*}\delta_h) \neq 0$ together imply

that $E(\delta_h \delta_{h^*}) = 0$ which violates the diagonal Σ_δ assumption. Thus, aside from the structural models comprising the system, no other relevant models are associated with the experimental unit under a diagonal Σ_δ. For if there were, they would be defined by relations among EVs comprising model errors. But since $E(u_h u_{h^*}) = 0$ for $h \neq h^*$, there can be no relations among these[3]. Σ_δ thus provides a quantitative measure of structure resolution.

POSTULATE: Total ignorance regarding structure occurs when structural parameters are under-identified and a reduced regression analysis is the only recourse. Total resolution of structure (relative to a well defined experimental unit) is characterized by a diagonal Σ_δ which is validated through experimentation.[4] The degree of structure resolution is quantified on a [0, 1] scale in terms of an estimate of $|R|$, where R is the p x p matrix of model error correlations.

Note that the "invariance"[5] of the reduced regression under whatever the hypothesized causal scheme provides complete objectivity but total ignorance regarding structure. However, assuming identification, lack of this type of objectivity is no reason to reject structural regression. When two experimentors propose different causal schemes for the same set of data and the matter is not resolved in the ensuing analysis, continued experimentation will validate one or the other or reject both.

---

[3] Relevant structural relations among EVs contained wholly within one particular δ_h would likely indicate that $E(\delta_h) \neq 0$ and/or that the experimental unit needs redefinition.

[4] See [1, p. 260] for a test of the hypothesis that Σ_δ is diagonal.

[5] Invariance is used in the sense that A and Γ uniquely determine B in (1.2) though not conversely. Thus, and infinity of (A, Γ) structures could lead to the same B.

Consider, for example, a sugar beet experiment [5, p. 816], where the stand (y_2) of the crop was found to have a positive, direct effect on yield (y_1). The estimated model error correlation was - .45 (hence, estimated $|R| = .80$) which led to a conjecture of food competition between plants; i.e., apart from the average positive effect of stand on yield, a stand response above its expectation would tend to accompany a yield response below its expectation due to the greater competition by plants for food. The implication of this correlation is that additional structural relations remain to be hypothesized in future experiments and that these relations might involve measures of moisture content and plant food. Unfortunately, under-identification is generally the case in experimental design so that attention is redirected to methods of achieving identification.

3. DESIGNING THE EXPERIMENT TO REMEDY UNDER-IDENTIFICATION

Consider the following two model system describing an experiment in a completely randomized design:

$$\begin{aligned} y_1 &= \mu_1 + \tau_1 + \alpha_{12}\, y_2 + \delta_1 \\ y_2 &= \mu_2 + \tau_2 + \alpha_{21}\, y_1 + \delta_2 \,, \end{aligned} \tag{3.1}$$

where μ_h and τ_h denote mean and direct treatment effects, respectively. Since $q < p_h + q_h$, additional information is necessary to estimate structural parameters. For example, the reduced model errors corresponding to (3.1) are $\varepsilon_1 = (\delta_1 + \alpha_{12}\,\delta_2) \,/\, (1 - \alpha_{12}\alpha_{21})$ and $\varepsilon_2 = (\delta_2 + \alpha_{21}\delta_1)/(1 - \alpha_{12}\alpha_{21})$, whereupon, from $\Sigma = A^{-1}\Sigma_\delta A'^{-1}$,

$$\Sigma = \begin{bmatrix} \sigma^2_{\delta_1} + 2\alpha_{12}\sigma_{\delta_1\delta_2} + \alpha^2_{12}\,\sigma^2_{\delta_2}, & \sigma_{\delta_1\delta_2}(1 + \alpha_{12}\alpha_{21}) + \alpha_{21}\,\sigma^2_{\delta_1} + \sigma_{12}\,\sigma^2_{\delta_2} \\ & \sigma^2_{\delta_2} + 2\alpha_{21}\sigma_{\delta_1\delta_2} + \alpha^2_{21}\,\sigma^2_{\delta_1} \end{bmatrix} (1-\alpha_{12}\alpha_{21})^{-2}$$

Equating Σ to S in (1.5) yields three equations in five unknowns. If, however, there is no causal feedback with only a direct effect of y_2 on y_1 (i.e., $\alpha_{21} = 0$) and the ratio $\lambda^2 = \sigma^2_{\delta_1}/\sigma^2_{\delta_2}$ is known, then we have three equations in three unknowns, and α_{12} and $\rho = \sigma_{\delta_1\delta_2}/\sigma_{\delta_1}\,\sigma_{\delta_2}$ are estimated by

$$a_{12} = 1 - \lambda^2 s_{22} s_{12}^{-1} \text{ and } r = \lambda s_{22} s_{12}^{-1} + (\lambda s_{22} s_{12}^{-1}) - \lambda^{-1}.$$

Since knowledge of α_{21} and λ is not often available, another recource is to assume that (3.1) applies to the first replication of the experiment and that

$$y'_h = \mu'_h + \tau_h + \alpha_{hh*} y'_{h*} + \delta'_h \qquad (3.2)$$

applies to the second replication. Note that while α_{hh*} and treatment effects remain the same between replications, the block effects may differ. Moreover, it will generally be the case that $E(\delta_h \delta'_{h*}) = 0$ for $h, h* = 1, 2$. Subtracting y'_h from y_h in (3.1) and (3.2) yields

$$y_h = y'_h + (\mu_h - \mu'_h) + \alpha_{hh*}(y_{h*} - y'_{h*}) + (\delta_h - \delta'_h) \qquad (3.3)$$

Consider, first, estimation of parameters of the first model in (3.1). Using the result in (3.3), (3.1) can be replaced by

$$y_1 = \mu_1 + \tau_1 + \alpha_{12}\, y_2 + \delta_1$$

$$y_2 = y_2' + (\mu_2 - \mu_2') + \alpha_{21}\,(y_1 - y_1') + (\delta_2 - \delta_2') \qquad (3.4)$$

Taking the y_h' as controlled variables amounts to replicating the experiment and using one replication as an instrumental variable [2, 3, 5] for the other[6]. All parameters in (3.2) are identified as is made obvious by referring to the corresponding reduced system, given as follows:

$$y_1 = \Big\{\big[\mu_1 + \alpha_{12}\,(\mu_2 - \mu_2')\big] + \alpha_{12}\, y_2' - \alpha_{12}\alpha_{12}\, y_1' + \tau_1 + \big[\delta_1 + \alpha_{12}\,(\delta_2 - \delta_2')\big]\Big\}/\phi$$

$$y_2 = \Big\{\big[\alpha_{21}\,\mu_1 + (\mu_2 - \mu_2')\big] + y_2' - \alpha_{21} y_1' + \alpha_{21}\,\tau_1 + \big[\alpha_{21}\,\delta_1 + (\delta_2 - \delta_2')\big]\Big\}/\phi$$

where $\phi = 1 - \alpha_{12}\alpha_{21}$. It is clear that structural coefficients are over-identified.

As for estimating the τ_2, (3.1) is replaced by

$$y_1 = y_1' + (\mu_1 - \mu_1') + \alpha_{12}\,(y_2 - y_2') + (\delta_1 - \delta_1')$$

$$y_2 = \mu_2 + \tau_2 + \alpha_{21}\; y_1 + \delta_2$$

and 2SLS can be applied directly as in (3.4).

There is a price to be paid in using the replication method to produce identification. Firstly, the sample size is halved which

[6] In the same manner that lags are used in econometric models, one replication can be considered as a lag for the other replication.

reduces power. Secondly, if y_h and $\hat{y}_h$ (the consistent estimate of y_h obtained through the reduced system) are not highly correlated, the resulting structural estimates may be highly inefficient. However, the alternatives are (1) complete reliance on a reduced analysis (which should always accompany and complement a structural analysis) and (2) OLS estimation which generally leads to inconsistent estimates, but which may provide certain estimates with low mean square error.

REFERENCES

[1] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, New York: John Wiley & Sons, Inc. 1958.

[2] Goldberger, A. S., Econometric Theory, New York: John Wiley & Sons, Inc., 1964.

[3] Johnston, J., Econometric Methods, New York: McGraw-Hill, 1963.

[4] Klein, L. R., "On the Interpretation of Theil's Method of Estimating Economic Relationships", Metroeconomica, 7 (1955), 147 - 153.

[5] Mallios, W. S., "The Analysis of Structural Effects in Experimental Design", Journal of the American Statistical Association, 65 (1970), 808 - 827.

THE SAMUEL S. WILKS MEMORIAL MEDAL
BANQUET REMARKS
Frank E. Grubbs, Program Chairman of the Conference

The twenty-third year or occasion for the Design of Experiments Conference in Army Research, Development and Testing marks another very significant milestone for Statistical Methods in the Army. Each year I like to reflect back over previous conferences, and it is easy to see how much we owe a great debt to the memory of Sam Wilks for his vision in getting Army statisticians together on a yearly basis for the common good of all. Indeed, we continue to benefit considerably from our previous 22 conferences, which have promoted much good statistical work in the US Army. Don't you agree? The associations with our statistical friends from the universities have kept us up to date and provided much stimulus toward many timely accomplishments. These conferences have done a lot of good by simply getting us all together on problems of common interest and we cover so many fields of interest! Again, I am reminded we have not stuck to the title, "Design of Experiments", in all detail, but that is good as the field of statistical topics changes fast and we must always move on to new things or areas. I could go on and on concerning the good these conferences have accomplished, but I must mention that the success of these conferences would not have been so great were it not for our most dedicated friend, Francis Dressel, who as we all know again deserves a vote of thanks at this time, for his effective, continuing contributions (so sorry he couldn't make it this year.) Also, this is the first time we have been privileged to have our conference here at Monterey and we appreciate such nice facilities, and also Doug Tang, Wally Foster and Bob Launer are to be thanked for the very significant part they played again this year.

We now turn to the Samuel S. Wilks Memorial Medal.

The Samuel S. Wilks Memorial Medal Award, initiated jointly in 1964 by the US Army and the American Statistical Association, is administered for the Army by the American Statistical Association, a non-profit, educational and scientific society founded 138 years ago in 1839. The Wilks Award is given each year to a statistician - often a good 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 matter 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 name of the recipient on the reverse, and a citation and honorarium related to the magnitude of the Award funds, which were donated by Philip W. Rust of the Winnstead Plantation, Thomasville, Georgia. The Annual Army Design of Experiments Conferences, at which the Wilks Medal is given each year, are sponsored by the Army

Mathematics Steering Committee on behalf of the Office of the Chief of Research and Development and Acquisition, Department of the Army.

Previous recipients of the Samuel S. Wilks Memorial Medal include John W. Tukey of Princeton University (1965), Major General Leslie E. Simon (1966), William G. Cochran of Harvard University (1967), Jerzy Neyman of the University of California, Berkeley (1968), Jack Youden (1969) formerly of the National Bureau of Standards, George W. Snedecor (1970) formerly of Iowa State University, Harold Dodge (1971) formerly of the Bell Telephone Laboratories, George E. P. Box of the University of Wisconsin (1972) - and with us today, H. O. (HO) Hartley of Texas A&M University (1973) - and our keynote speaker, Cuthbert Daniel (1974) - private statistical consultant, Herbert Solomon of Stanford University (1975) - who just trekked to the United Kingdom for two years with ONR, and Solomon Kullback of George Washington University (1976).

This brings us up to this year, and I call on Jeff Kurkjian, University of Alabama, Chairman of the S. S. Wilks Memorial Medal Committee to discuss this year's committee work and give the presentation.

SAMUEL S. WILKS' MEMORIAL MEDAL COMMITTEE: MEMBERSHIP, CHARTER, SELECTION PROCEDURE

Badrig Kurkjian, University of Alabama

The 1977 Committee was made up of Badrig Kurkjian, Chairman, Francis Anscombe, Jerome Cornfield, Cuthbert Daniel, Fred Frishman, Frank Grubbs, Joan R. Rosenblatt, and Herbert Solomon. Three of these members were former employees of the US Army with virtually career-long experience with the Army Design Conference. Three others have considerable experience consulting with the Army on technical problems and policy matters associated with the business of the Army Mathematics Steering Committee. Moreover, the Committee contained three former Medalists--Cuthbert Daniel, Frank Grubbs, and Herbert Solomon.

One could summarize the charge to the Committee by stating simply that the recipient of the Wilks' Medal should be a person who has emulated Sam Wilks to a significant extent--that is, a scholar, a contributor to statistical methodology and one who unstintingly devoted significant effort to the public interest, in particular the U. S. Army Design Conference in Sam's case.

Each year the Committee considers nominees from prior years as well as those forwarded to the Committee from various sources within the statistical community in the Army and elsewhere. This year, the ballot contained twelve nominees, each of whom is a nationally, or internationally, renowned statistician. As might be expected each year, the voting is usually very close and two ballots are required to select the recipient. However, this year the Wilks' Medalist, Dr. Churchill Eisenhart, Senior Research Fellow, National Bureau of Standards, was the clear winner on the first ballot. The Committee had no difficulty in recognizing Dr. Eisenhart's professional career match with that of Sam Wilks.

The Army Design Conference was privileged to have Professor G. E. P. Box, University of Wisconsin and in-coming President of the American Statistical Association, present Dr. Eisenhart with the Medal, the official Citation, and a modest monetary honorarium.

REMARKS OF CHURCHILL EISENHART ON ACCEPTING THE 1977 SAMUEL S. WILKS MEMORIAL MEDAL

Chairman Grubbs, President-Elect Box, Fellow Statisticians, Ladies and Gentlemen:

This is for me a very happy occasion as I express my very great pleasure in accepting the 1977 Samuel S. Wilks Memorial Medal that honors my teacher, long-time friend, and the initiator of these Experiment Design Conferences. I especially appreciate the high honor of being presented this award, having served as a member of the Wilks Memorial Medal Committee of the American Statistical Association from 1965 through 1970.

I have spoken in great detail about Sam Wilks at two preceding Conferences of this series--the 10th and the 20th: about his extensive contributions toward the advancement of statistical methods in Army research, development and testing, and about his many other important contributions in the national interest. I shall limit myself on this occasion to sketching how very, very helpful Sam was to me in the early stages of my career. Generosity in helping others in spite of his own heavy schedule was one of Sam's outstanding characteristics.

I was Sam's first student in statistics. He arrived in Princeton in September 1933 in time to supervise my Senior Thesis on "The Accuracy of Computations Involving Quantities Known Only to a Given Degree of Approximation". The first part was an attempt to present a fairly complete survey of the accuracy of the general processes of arithmetic without recourse to probability theory and the methods of statistics, which were introduced and applied in the second part.

Sam also supervised the preparation of my first two publications in statistics. The first was a short note in the December 1935 issue of the *American Journal of Science* criticizing the statistical approach employed in a paper appearing in the May 1935 issue--too harshly, my geologist friend, W. C. Krumbein, says. The object of the paper on which I commented was to suggest a numerical measure of the degree of "likeness" of two or more "heavy mineral suites" with respect to their mineral contents. The measure of agreement or "likeness" advocated was such that the value obtained in a particular instance depended upon the order in which the respective minerals were listed: if listed alphabetically by their names in English, one value would result; if listed alphabetically in some other language, a different value would be found; and, if in order of their respective densities, still another value.

I suggested an approach via the χ^2 test of the homogeneity of frequency data arranged in an $r \times c$ table, and referenced R. A. Fisher's *Statistical Methods for Research Workers*. I would never have had the courage to submit this critical note for publication had Sam not been standing behind me all the way.

My second statistical paper, "A Test for the Significance of Lithological Variations", published in the December 1935 issue of the *Journal of Sedimentary Petrology*, was an exposition, for geologists, of the χ^2 test for homogeneity, with three worked examples utilizing data from the paper discussed in the note. This seems to have been the first exposition of χ^2 methods in the literature of geology.

Several months before those two papers appeared in print, I had left Princeton, at Sam's recommendation, for University College, London. I went there to work toward a Ph.D. in Statistics under Jerzy Neyman and Egon S. Pearson in the Department of Statistics. I also attended the lectures that R. A. Fisher (of the Galton Laboratory for National Eugenics) was giving on Experiment Design and on the History of Biometry; and at his request, prepared a little brochure on the use of ranked normal deviates in the analysis of data expressed as ranks, for the guidance of some of Professor Cyril Burt's graduate students in psychology. At the Annual Karl Pearson Memorial Dinner at University College in the spring of 1959, Egon Pearson introduced me as "one of the few persons who worked with Fisher, Neyman and a Pearson and managed to survive".

While I was at University College, a circumstance occurred that enabled me to help Sam for a change: Professor George G. Chambers of the University of Pennsylvania, had died on 24 October 1935, shortly after his graduate course "Modern Theory of Statistical Analysis" got underway. Sam was commissioned to complete the teaching of this course. He wrote me a hurried note saying that he was in dire need of up-to-date problems in statistical theory and methods for the students in his new class. Would I please send him some quickly. From time to time throughout the remainder of that academic year, I sent off to Sam a bundle of homework and test problems that we had been given in the courses that I was taking under Neyman, Pearson, B. L. Welch and Fisher.

Sam's next turn to help me came in the fall of 1937, when I took up my post as Station Statistician at the Wisconsin Agricultural Experiment Station. To find one's self *the* expert on statistics in a major research organization immediately after finishing one's doctoral program, without a period of "internship" training in applied work, with no senior expert at hand to consult, is a trying experience--to be avoided, if possible. At Wisconsin, however, I had the advantage that I did not have to "sell" statistical methods to the staff of the Experiment Station. There were already on the campus several agricultural research workers who had taken courses in statistics under George Snedecor at Iowa State or studied biometry under Forrest Immer at the Minnesota Agricultural

Experiment Station. These fellows were for the most part quite self-sufficient in statistics. Nonetheless, they were a source of difficulty for me: They would bring me hard problems to which the straight forward procedures that they had learned from Snedecor or Immer did not apply. I tackled these as best I could, and sent a draft to Sam in Princeton for his approval, correction, or other counsel. Only then did I turn over my "solution" to the "client".

More of a problem to me were the members of the Experiment Station staff who had acquired a smattering of statistical techniques of experiment design from lectures given there a previous summer by Cyril H. Goulden of the University of Manitoba. As an admirer of Goulden and his writings I have not the slightest doubt that what he presented in his lectures was entirely correct; but some of his listeners seem to have missed some of the essential details.

Thus, soon after my arrival, I was confronted with the results of a field trial of 24 varieties each replicated 4 times in a 4 x 4 rectangular arrangement of 16 cells with 6 varieties in each cell. (I do not recall the exact number of varieties involved, nor the exact size of the rectangular design, but the choices here will serve to bring out the problem I faced.) The disposition of the 4 replicates of each variety was such that each variety occurred once and only once in each cell-row and each cell-column.

I got a lot of argument from my consultees when I tried to convince them that, in spite of the last-mentioned restrictions, this arrangement was NOT a Latin Square; could not be analyzed as such; that the best that could be done would be to do a Randomized Blocks analysis with the cell-rows as "blocks", and again with the cell-columns as "blocks", and then use whichever analysis led to the smaller residual mean square for "error".

In view of the considerable unhappiness of the consultees at this verdict, and being not entirely sure that something better could not be done, I sent the whole package off to Sam in Princeton. He replied by return mail saying that in this particular instance I was entirely correct, inasmuch as the experimenters had failed to group the 24 varieties into 4 "bundles" of 6 varieties each. Had they done this and arranged the 4 replicates of these bundles in accordance with a 4 x 4 Latin Square, they would have had a Split-Plot Latin Square--a design that I didn't recall Fisher having discussed in his lectures. They then would have been able to do a regular Latin Square analysis with respect to the 6 different (but fixed) "bundles", leading to two "error" mean squares, one appropriate to comparing varieties in the same bundle, and one for comparing varieties in different bundles.

The point of all this is that he always took the trouble and the time to respond promptly and very helpfully by return mail--in this instance at a time when he was already enormously busy with his teaching, his work for the College Entrance Examination Board and his new duties as Editor of the *Annals of Mathematical Statistics*.

I could go on, but I believe that I have said enough to reveal that Samuel Stanley Wilks was the distinguished mathematical statistician who was my closest teacher, who launched me into my career, and who was also a wise and greatly loved friend and counselor from the moment of my first meeting him.

I shall cherish this medal bearing his name and his likeness.

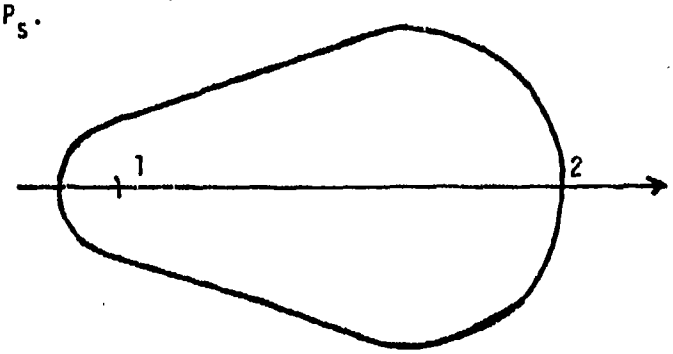
THREE DIMENSIONAL CURVE FITTING TECHNIQUES TO EXPRESS SUPPRESSION AS A FUNCTION OF RANGE AND ASPECT ANGLE

Chaunchy F. McKearn and David E. Brown
Combat Developments Experimentation Command
Fort Ord, California 93941

ABSTRACT.

During the 2nd Quarter FY 1978, the Combat Developments Experimentation Command will conduct the next in a series of suppression experiments, Supex III. The primary objective of this experiment is to determine the probability of suppression, P_s, as a function of range, r, and aspect angle, θ. Artillery projectiles will be set off in all directions and at varying ranges from the players, who will be observing through a periscope in an uncovered foxhole. What is needed is a surface fitting technique that will permit the surface, $P_s = g(r,\theta)$ to be determined from the data produced. The level curve for any fixed value of P_s must be a smooth curve which is perpendicular to the line of observation at the two points at which the curve intersects this line.

The results of previous experiments indicate that P_s considered as a function of only offset distance, x, $P_s = f(x)$, has an exponential or logarithmic form. These results also indicate that the probability of suppressing, P_s, is not symmetric to the front and rear of the observer. The curve below shows the general desired form of a level curve for a fixed value of P_s.



1. Location of observer. 2. Direction of observations.

The difficulty is in arriving at the form of an equation such that any curve for a fixed value of θ, i. e., $P_s = g(r,\theta_0)$, would be exponential or logarithmic and the level curve for a fixed value of P_s, i. e., $P_{s_0} = g(r,\theta)$, is a closed curve with continuous derivatives with $dx/d\theta = 0$ for $\theta = 0$ and π (to insure smoothness and vertical tangents).

I. INTRODUCTION. Combat Developments Experimentation Command (CDEC) has conducted a series of suppression experiments to measure the probability of suppression, P_s, as a function of miss distance. Generally, the players being suppressed represented antitank guided missile gunners and the suppressive weapons included both direct and indirect fire weapons from the M16 rifle up to the 8 in. Howitzer. This report concerns only the indirect fire point detonating high explosive rounds.

In order to collect empirical data on the phenomenon of suppression, the subjects were placed in protective foxholes as shown in Figure 1 and observed down range through periscopes. They were task loaded by requiring them to report the position of a target tank in reference to a row of numbered panels along its path at a range of 1500 meters. The gunners were required to track the target tank for fifteen consecutive seconds to receive credit for hitting the target tank.

The periscopes were instrumented in such a manner that when they were raised or lowered it was automatically recorded on the central computer. In addition, each periscope was electrically connected to a pop-up silhouette immediately in front of the foxhole so that when the periscope was raised the pop-up silhouette came up and when the periscope was lowered the silhouette went down. This pop-up silhouette was within the gunner's field of view and represented the gunner in an unprotected position. This assisted the subjects in perceiving the danger they would be in if they were located at the pop-up silhouette's position. If a piece of shrapnel hit the silhouette, a loud buzzer was set off in the subject's foxhole indicationg that had he been at the pop-up's location he would have been killed or wounded.

The artillery rounds were placed on the ground within the player's view at various ranges and statically detonated in a random manner. The data collected from these tests indicated that the probability of suppression as a function of miss distance could be reasonably well represented by an exponential curve of the form

$$P_s = Ae^{bx},$$

where

P_s = probability of suppression
x^s = distance between the foxhole and the detonation point, and
A and b are curve fitting parameters.

Figure 2 lists the curve parameters for the various munitions tested in CDEC's last suppression experiment, SUPEX II.

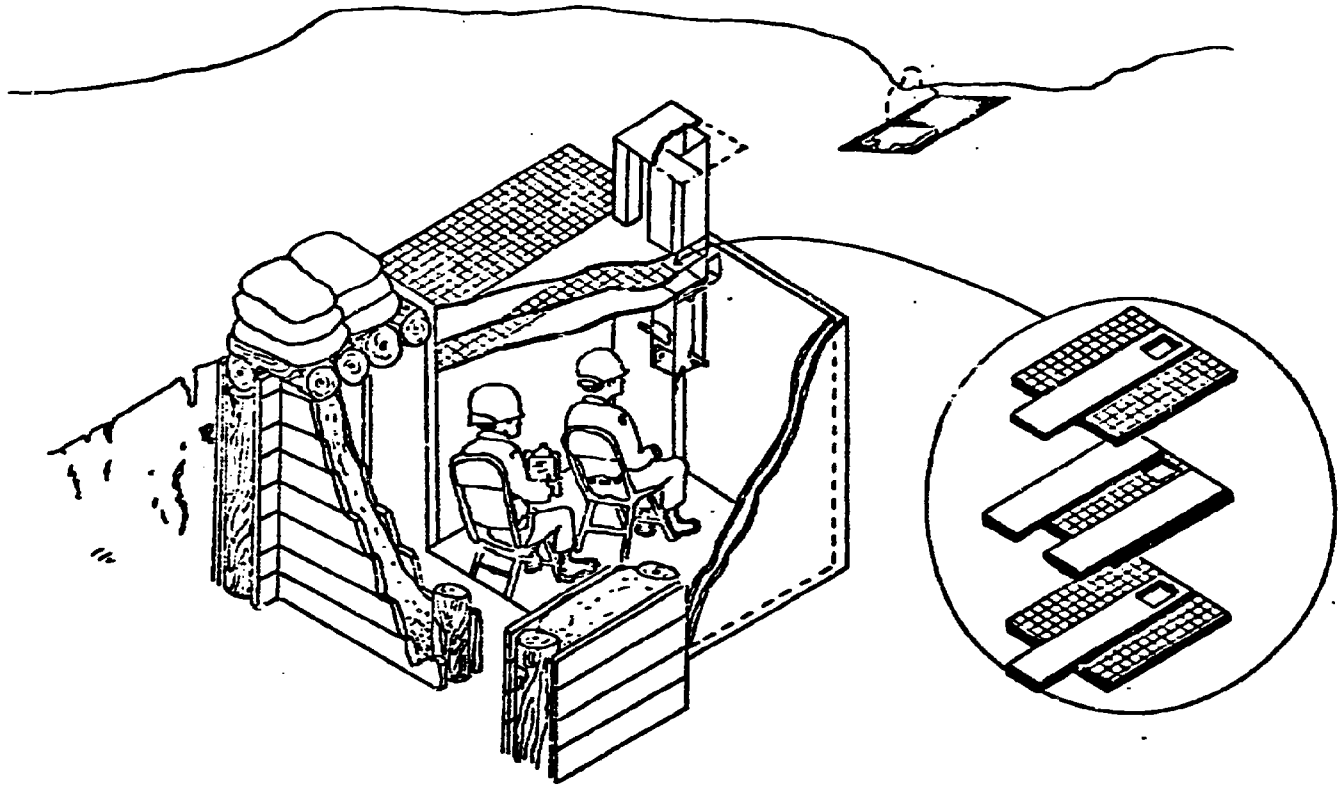


Figure 1: SUPEX II Range Layout and Foxhole Construction

EXPONENTIAL CURVE PARAMETERS

WEAPON	A	B
60mm Mortar	1.61911	-.02453
81mm Mortar	1.50512	-.01262
105mm Howitzer	1.64851	-.01306
105mm HEP - T	1.70799	-.01317
2.75 in Rocket	1.77098	-.01530
155mm Howitzer	3.26843	-.01773
8 in Howitzer	1.58806	-.00450

FIGURE 2: EXPONENTIAL CURVE PARAMETERS FOR EXPRESSING THE PROBABILITY OF SUPPRESSION AS A FUNCTION OF MISS DISTANCE.

II. EXPERIMENTAL DESIGN.

Last July CDEC hosted a Suppression Working Meeting to determine what the next step in CDEC's Suppression Program should be. Many of the attendees, mostly modelers, expressed the concern that CDEC's suppression data only addressed suppression caused by detonations directly to the observer's front. What was needed was a function of range and aspect angle, $P_s = g(r,\theta)$. To accomplish this the SUPEX III experiment is currently being planned and is scheduled to begin in April 1978.

The range for this experiment will be laid out as shown in Figure 3. Four foxholes will be located at the center of the wagon wheel with one foxhole oriented along each of the four principle axes. Five rounds will be placed along each of the twelve wagon wheel spokes and set off in a random manner. When all of the trials are completed there will be sixteen observations at each range at all twelve aspect angles.

III. STATEMENT OF THE PROBLEM.

The data obtained from SUPEX III should permit the development of a three dimensional surface expressing P_s as a function of range and aspect angle similar to that shown in Figure 4. Past experience indicates that for each aspect angle one could expect the data to fit a truncated exponential and for a fixed value of P_s one should obtain a level curve that is somewhat elliptical or egg-shaped with continuous derivatives.

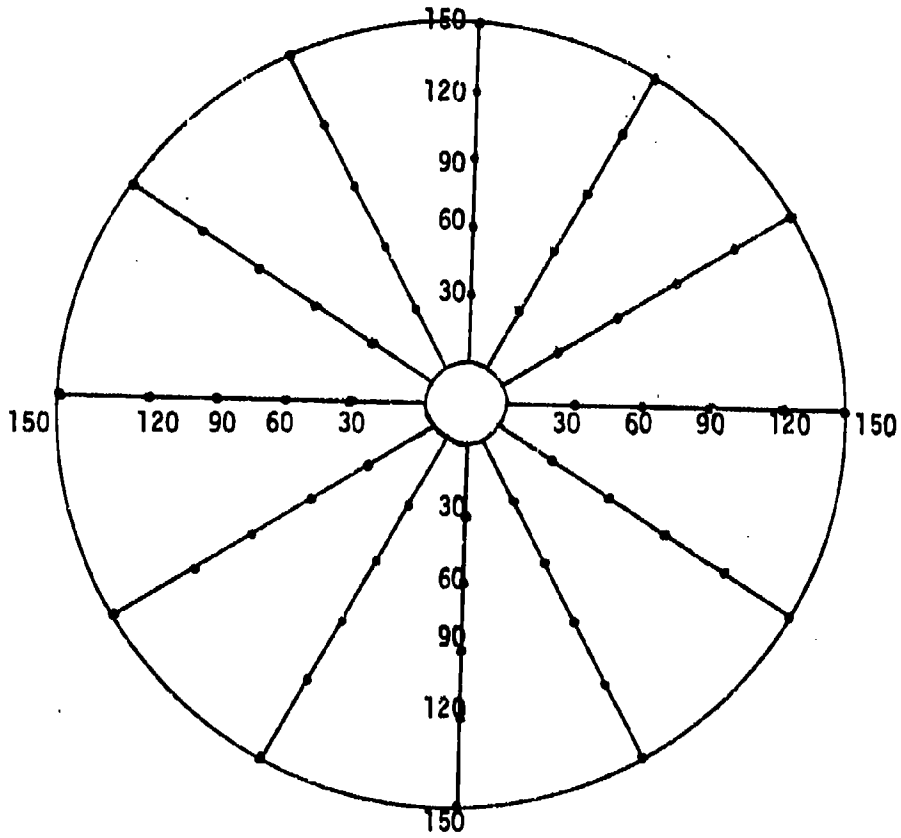


FIGURE 3: Range Layout for the 105mm Howitzer Trials

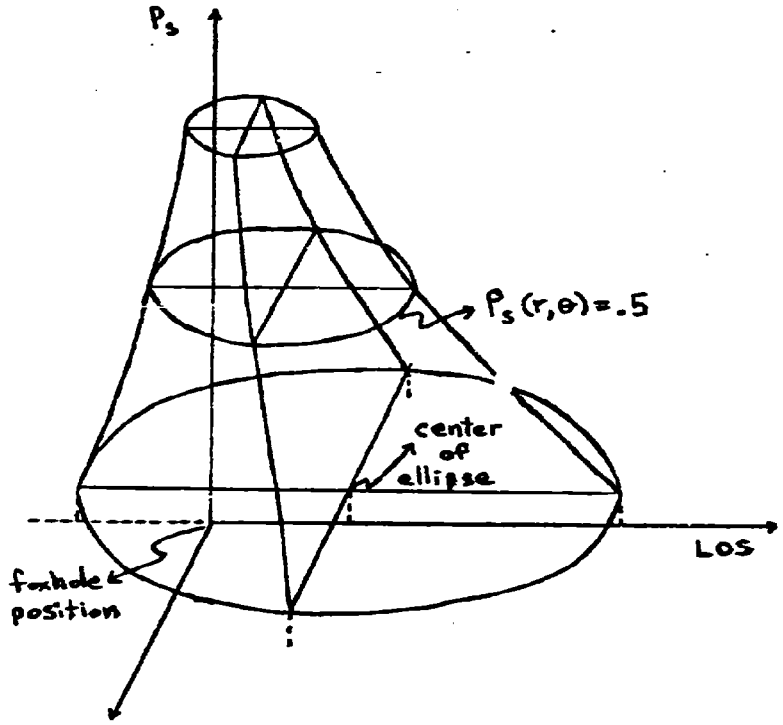


FIGURE 4: Probability of Suppression, P_S, as a Function of Range, r, and Aspect Angle, θ .

One candidate function is as follows:

$$P_s = Ae^{rb(1-\gamma\cos\theta)}$$

P_S = probability of suppression

r = miss distance

A,b = shaping constants

γ = excentricity

θ = aspect angle

The difficulty with this function is that in order for the level curves for a given value of P_S to assume the desired egg shape, γ will have to be a function of θ. This makes it difficult to determine all of the parameters by such conventional methods as least squares because the function is no longer linear in its parameters. What is needed is a method for non-linear regression that can handle a function like this or a different function which has the desired characteristics and is linear in its parameters.

ON VALIDATING CRITERION REFERENCED TESTS

Milton H. Maier and Stephen F. Hirshfeld

US Army Research Institute for the Behavioral and Social Sciences

INTRODUCTION

Skill Qualification Tests (SQT) have been developed to replace Military Occupational Specialty (MOS) proficiency tests as measures of ability to perform Army enlisted jobs. SQTs are performance-based, criterion-referenced measures of job proficiency, consisting of precisely defined tests of tasks, all of which are critical and necessary to performance of the job. The criterion-referenced approach provides an explicit relationship between job requirements and test content in that job requirements dictate content of SQTs. The SQT development process requires that tests be reviewed by subject matter experts and validated on representative job incumbents to assure that test content is job relevant. Test standards of acceptable levels of performance are also based on job requirements and test content. Performance standards are based on behaviorally derived absolute scoring standards, and are not based on performance relative to other soldiers who take the test. For these reasons SQTs are justifiably viewed as criterion-referenced tests of job proficiency.

This paper provides a description of the SQT program, its evolution, underlying assumptions, requirements, construction and validation processes, and methods of statistical analysis. It concludes with a set of questions characterizing some of the major issues still under review.

Army training during background in the late 1960's and early 1970's experienced a major revolution. Performance-based training and testing, based on critical job tasks and criterion-referenced standards of performance, were being implemented in entry-level training courses. Training objectives were operationally defined by the performance tests given during the course, and the tests were made public to students as well as instructors. Because of the direct relevance of these tests to the job, they were capable of focusing Army training activities.

By maintaining accountability, tests become effective instruments for institutional change. Test content helps implement doctrine about the way jobs are to be performed, and is helpful in defining training requirements and standards. The public nature of the tests helps focus attention on the critical elements of the job and enables effective use of soldiers' time in preparing for tests, thus improving individual readiness.

So impressive was the success of performance-based training and testing that the Army made the policy decision to change from the existing mode of "norm-referenced, paper-and-pencil testing," to the criterion-referenced mode of proficiency testing. These new criterion-referenced

tests, called Skill Qualification Tests (SQT), are having a profound impact on the entire Army community. The new testing procedures are forcing training managers, personnel managers, and research support personnel to rethink and often redefine their functions.

REQUIREMENTS OF SKILL QUALIFICATION TESTS

The basic requirement of SQTs is that the tests are job relevant. The test content must be based on job requirements, and the test scores must be accurate measures of ability to perform critical job tasks.

Training and Personnel Management. SQTs are used by both training and personnel management to help make important decisions affecting the career development of soldiers. Both training and personnel management need timely and accurate information about how well individuals are performing - training management to determine training requirements of individuals, and personnel management to help determine who to promote, reclassify, or reassign. Although both training and personnel management have a need for the same kind of information, their immediate requirements are not identical.

Training managers base their immediate training requirements on the specific tasks performed in their units. Therefore, from this point of view relevance of the tests for specific job assignments is the primary consideration, and it is defined in terms of the tasks that soldiers perform in their assignments. The set of tasks performed in an assignment is generally a subset of tasks required in a specialty. The task is a convenient unit for determining training requirements because tasks are observable, have initiating and terminating cues, and have standards of performance that can be reasonably well specified. Decisions about proficiency can be made at the task level, and training managers can identify the specific tasks on which soldiers need training. If the test measures performance on the specific tasks for which the training managers have responsibility, then the tests are serving their basic purpose.

Personnel managers are also concerned with the job performance of individual soldiers; but rather than focusing on soldiers' specific assignments, personnel managers need to know how well soldiers can perform all the tasks in a specialty. For example, performance in a specialty, such as Infantryman or Wheeled Vehicle Mechanic, cannot necessarily be inferred from the set of tasks found in any one assignment. Personnel managers, therefore, have a need for information based on a standard set of tasks for each specialty. All soldiers in a specialty need to be evaluated on the same set of tasks to enable fair decisions about which

soldiers to promote, retain, or reclassify. The need for a standard set of tasks in each specialty imposes additional testing requirements for feasibility and acceptability. The test scores should not be affected by when or where the test is taken, nor by whom it is administered and scored. The testing conditions, as well as performance standards, should be standardized.

The requirement for Army-wide standardization at the present state of the art in testing means that initially most of the test content is in the paper-and-pencil mode rather than hands-on performance tests. Paper and pencil tests generally lack the apparent job relevance of hands-on performance tests, and therefore an additional requirement is imposed to assure that the tests are acceptable to examinees, supervisors and commanders as valid measures of job proficiency.

Job relevance of the tests is the basic requirement for both training and personnel management, even though the definition of job relevance may have somewhat different meanings for the two purposes. For training purposes the focus is on the subset of tasks performed in the specific job assignment, whereas for personnel purposes the interest is on the entire set of tasks in the specialty.

Because of the strategic importance of Skill Qualification Tests to both training and personnel management, high level policy decisions were made about test content, validation, and scoring. The general requirements of the program are that tests must be fair and feasible.

Fairness and Feasibility of the Tests. Fairness means that all soldiers have an equal opportunity to demonstrate their true level of job competence. Test content must be based on actual job requirements, and testing conditions must be sufficiently constant throughout the Army so that scores obtained from administrations under varied conditions are not noticeably different. Tests given in Alaska, Panama, and Korea must all be administered under similar conditions, and, in addition, all persons administering and scoring the tests must be able to do so accurately and objectively. An additional requirement is that the tests must be acceptable to soldiers and knowledgeable experts as fair measures of ability to perform critical job tasks. Therefore, fairness attends to requirements of both training and personnel management.

Feasibility requires that the tests be suitable for administration in all types of units; equipment, terrain, personnel and all testing material must be readily available. Another aspect of feasibility is that testing time must be reasonable, with up to one day allowed for testing each soldier.

Form of Testing. The requirements that Skill Qualification Tests be fair and feasible put severe limitations on the use of hands-on performance tests. The history of performance testing is that scoring accuracy and standardization are difficult to obtain. The resolution of the fairness and feasibility requirements is to have several kinds of testing. Under present policy decisions, all Skill Qualification Tests contain a written component, and some Skill Qualification Tests contain a hands-on component. Four hours of testing is allowed for the written component, and up to four hours is allowed for the hands-on portion.

Hands-on performance tests are most desirable. They are a form of structured observation where a scorer evaluates an individual on a set of performance measures (observable behaviors). Advantages of hands-on testing are obvious: It tests actual performance, has high fidelity to the job, allows for immediate feedback, and has high face validity to examinees. However, considerable developmental effort is required to insure scoring reliability and standardization of conditions. It also is expensive in terms of equipment, personnel, and time, i.e., feasibility is often a problem. In order to ensure feasibility there is a natural tendency to truncate tests of tasks by shrinking the boundaries. Unfortunately, this may be at the expense of the validity of the test. For these reasons it is extremely difficult, if not impractical, to initiate a large scale hands-on testing system for an organization as large as the Army. Therefore, a hands-on component constitutes a subset of an SQT.

The decision to include a written component imposes careful consideration and analysis of what criterion-referenced measurement means in this context. Since the focus of Skill Qualification Tests is on ability to perform critical job tasks, that aspect must be retained. Each written test of a task is to consist of a set of items, where each item is designed to measure an essential behavior or step in performing the task. For tasks that require primarily mental skills, such as the supply and administration fields, written tests of tasks are often similar to the behaviors required on the job, and the standards for ability to perform the test of the tasks can be reasonably close to those on the job. For other tasks that require psychomotor skills, written test items only simulate actual job behaviors, and the setting of realistic standards indicating ability to perform the tasks is a more arbitrary process. To help approximate realistic job conditions, written items may have multiple correct responses and variable number of alternatives. This added flexibility increases the difficulty in developing appropriate methods for setting standards. The determination of reasonable standards for written tests of tasks is one of the most difficult issues in the SQT program.

Criterion-Referenced Measurement of Task Performance. Because Army jobs and training programs are structured in terms of critical tasks, the appropriate level of analysis for the SQT should also be based on tasks. The concept of "scorable unit" was invented to help assure criterion-referenced measurement of task performance. A scorable unit is designed to measure ability to perform a specific task, or in the case of complex tasks, a well defined subtask.

Each written scorable unit consists of a set of items, where each item is designed to measure an essential behavior or step in performing the task. Each item is scored pass-fail, and a prescribed number of items must be passed to be GO on the written scorable unit. A GO is counted as ability to perform the task. The current resolution to setting standards for written scorable units is to require that an *a priori* number of items be passed. For example, if a scorable unit contains five items, then four must be passed to obtain a GO.

Hands-on scorable units consist of a set of performance measures, where each performance measure is scored pass-fail, and a prescribed number of performance measures must be passed to be GO on the scorable unit. A GO on the scorable unit is interpreted as ability to perform the task. The standards of GO generally are comparable to what is required on the job.

The requirement that all scorable units be acceptable as fair measures of ability to perform tasks is applied to both the hands-on and written tests. Juries of experts must agree that the written items and hands-on performance measures reflect ability to perform the tasks. Perhaps a safer statement would be that failure to pass the items indicates that the person is not able to perform the task.

Establishing a Correspondence Between Test Content and Job Tasks. The most critical requirement of SQTs is their job relevance. The procedures for establishing job relevance are described in this section. Test content of all SQTs is a sample of critical tasks from the domain of job tasks in the specialty. In this way the tests have a specifiable and explicit link to the job. For each Army job there exists a Soldier's Manual that lists the tasks for which a soldier in that specialty is responsible. Therefore, this set of tasks becomes the operational definition of the job. Tests to measure performance on specific job tasks listed in the Soldier's Manual are developed from appropriate task analyses, and the tests for each task are operational definitions of performance on the tasks. Performance on the individual tasks is summed to obtain a total score, which in turn serves as the operational definition of job competence. Modern instructional technology, with its

emphasis on specification of objectives and verification that those objectives are attained, supports the above process for establishing the content and focus of SQTs, and thereby lends added credibility to these procedures.

Though the task is the basic level of analysis, the validity of task proficiency measurement depends on the adequacy of the test of the task. By means of detailed task analyses, the set of performance measures or behaviors required for successful performance of the task are identified. These lists of performance measures are all available in the Soldier's Manual. Each item developed to test for task proficiency must occupy a clearly specified relationship to a performance measure required in task performance. Assuming that the set of items developed for a test of a task has been selected in accordance with the procedures described above, one may assume with reasonably high confidence that successful performance of each tested behavior is a necessary condition for successful performance of the task. How to score the set of items in a written scorable unit to obtain estimates of ability to perform tasks is a complex question. Measurement error is always a problem that must be allowed for. Whether being scored GO on a test of a task requires passing all items included in the test of the task, or some number less than perfection, depends on the nature of the task, the fidelity with which the task can be tested in a written mode, the complexity of the format (e.g. multiple correct responses), and the number of items within the cluster. Use of subject matter experts in reaching such a determination is mandatory.

In the case of a hands-on test of a task, measurement error arising from the use of words is minimized. However, other measurement problems arise. One is that a full performance test of a task generally is not feasible. It may be too costly in terms of time, equipment, and personnel. Therefore, a truncated test of the task is often developed by eliminating some of the performance measures or steps required for the full performance test. By truncating the test, though, it is possible that the tested portion is necessary to successful task performance, but is not sufficient.

Validate Tests Prior to Administration. A first question to be resolved is how to define validity. The starting point is the usual definition of validity, i.e., that the tests measure what they are intended to measure. In the case of Skill Qualification Tests, the intent is to measure ability to perform critical job tasks. The content of the tests, therefore, becomes the crucial factor in establishing validity. The content must be thoroughly reviewed by experts to ensure that the right behaviors and decisions are assembled in each scorable unit. The first requirement, then, is consistent agreement among experts that the content of the test is based on ability to perform critical job tasks. A

second requirement is that the scorable units discriminate between performers (masters) and nonperformers (nonmasters). A third requirement applies only to written scorable units. All items in a written scorable unit must be consistent estimators of mastery on the task covered by the entire scorable unit. Thus, the conceptualizing of validity focuses on consistency: Consistency between the content of the test and the job tasks, consistency among expert reviews, and consistency in identifying mastery.

DEVELOPMENT PROCESS

Skill Qualification Tests are constructed and validated by Army agencies that have resident expertise in the job specialties. Generally these are the Army schools, but they also include other agencies, such as the Health Services Command. Since the test content must reflect job tasks, the test developers must have detailed task analyses available that identify the behaviors essential to successful performance of the tasks.

The development process for Skill Qualification Tests may be conceptualized in four steps:

1. Identify job tasks for testing; these tasks require special training or are frequently failed.

2. Identify behaviors or steps essential for performing each task; the intent is to identify the steps that are necessary and sufficient for successful task performance.

3. Develop scorable units (tests of tasks) to measure essential behaviors for the tasks; items in scorable units must have explicit relationship to task steps, and the scorable unit as a whole must correspond to performance of the task; items are scored pass-fail (1 or 0), and scorable units are scored GO/NO-GO (also 1 or 0) to reflect mastery or nonmastery of the task according to the prescribed standards; the number of scorable units scored GO is a measure of job proficiency.

Content of the Skill Qualification Tests is fixed after these three steps are completed. Experts review (a) the tasks selected for testing to make sure they are critical to the job; (b) the behaviors required to perform the task to make sure they are necessary and sufficient; and (c) the scorable unit to make sure that the items correspond to the behaviors. After the experts agree on the appropriateness of the test to job requirements, the test content cannot be changed.

4. Try out scorable units on soldiers.

This step serves only to establish the measurement properties of the tests. Items found to be unsatisfactory through the tryout can be revised, as long as the test content is not changed.

STATISTICAL ANALYSIS OF TRYOUT DATA

The tryout step was originally conceived of as the validation of Skill Qualification Tests, and the earlier steps as test construction. Experience gained during the past two years, however, has shown that for criterion-referenced tests, validation encompasses the entire development process.

The guiding principle of the developmental process is consistency of measurement. Experts must agree on the relevance of the test content to job requirements and the appropriateness of tests items to task behaviors. In the tryout on soldiers, the scorable units must be consistent indicators of ability to perform the task. For written scorable units, each item in a scorable unit is first correlated with an independent estimate of ability to perform the task, and then with the other items in the scorable unit. The external estimates of ability to perform the task are self-ratings obtained through standard questions. Up to 30 soldiers are included in the sample to determine consistency of measurement for each scorable unit. The analysis consists of computing an Agreement Index for each item and scorable unit:

Self-rating

		Performer	Nonperformer
Item or Scorable Unit	Pass or GO	a	b
	Fail or NO-GO	c	d

a,b,c, and d are cell frequencies

Agreement Index = ad - bc; if Agreement Index ≥ 0, then the item or scorable unit is satisfactory; if Agreement Index < 0, then the item or scorable unit is unsatisfactory, and must be examined for revision.

A second analysis involves examining patterns of Agreement Indices for items in a scorable unit. Items that have positive Agreement Indices are satisfactory, and items with negative Agreement Indices must be examined for revision.

SQT ISSUES STILL UNDER REVIEW

1. Is the Agreement Index an appropriate statistic to evaluate the quality of written items and scorable units?

2. For written scorable units, standards of performance are set arbitrarily, e.g., 3 of 4 items must be passed to be GO on a scorable unit. Are there statistical techniques to indicate level of mastery that can be readily employed by test developers who are not trained in statistics?

3. Are there alternative procedures for collecting and analyzing data on the satisfactoriness of written items and scorable units, which are also sensitive to the requirement of fixed test content?

4. Are there more appropriate ways of combining scores from items and scorable units into a total test score that indicates level of job proficiency?

ANALYSIS OF MAN-MACHINE INTERFACE INFORMATION IN CURRENT COMMUNICATIONS SYSTEMS

R. J. D'Accardi and H. S. Bennett, US Army Communications Research and Development Command, Fort Monmouth, New Jersey

C. F. Tsokos, Department of Mathematics, University of South Florida, Tampa, Florida

ABSTRACT. Experiments dealing with man-machine interface problems occurring in tactical communications systems have been conducted at Ft. Monmouth, NJ. The thrust of the study was to characterize the human element of a sophisticated system by varying the environmental factors of ambient light and acoustic noise and observing quantitative changes in operator performance. Specifically, the number of errors committed by a communications systems operator were observed as a function of the environmental factors. The equipments used were the standard teletypewriter terminal and an optical display terminal.

The object of this presentation is threefold: First, we discuss the importance of human-factors in system development and briefly review the experimental design. Secondly, we present a non-linear regression model and error matrices which can be used to predict operator performance as a function of the environmental factors of ambient light and acoustic noise, and thirdly, time series models are presented for the optical display terminal to illustrate the usefulness of characterizing, within reason, the error performance of a terminal operator working in a wide variety of environments.

I. THE IMPORTANCE OF HUMAN FACTORS IN SYSTEM DEVELOPMENT.

It is interesting to note that Human Factors studies in the Army can be traced back to World War I. It was found at that time that in the fledgling British Air Service 90% of all fatal accidents were the result of individual pilot deficiencies and only 2% were killed in combat (the remaining 8% were due to materiel deficiencies). This fact led the US Army to establish a laboratory designed to study problems (including the human factors aspects) connected with flying. It was called the Research Board of the Army Signal Corps Air Service and was established in October 1917. It was quickly followed by the School of Aviation Medicine in 1918 (now the School of Aerospace Medicine at Brooks Air Force Base, Texas) and the Physiological Research Laboratory (now the Aerospace Medical Research Laboratories at the Wright Patterson Air Force Base). By the time World War II began the human factors field has been taken over by industrial engineers and industrial psychologists (e.g. Taylor, Gantt, and the Gilbreths). It was World War II, however, with its quantum jump in the technological complexity of man/machine systems, which set the mold and pattern for modern present day human factors engineering.

The mission of modern day studies of man/machine environmental factors has five aspects:

a) It is connected with the contributions of the man/machine interface to the entire or over-all performance of the system. As the systems become more complex, they also become more vulnerable to catastrophic failures (shades of the power blackouts!) and the man/machine interface is a critically vulnerable point in such systems.

b) It must be concerned with the translation of broad system operational requirements into specific man/machine interface functional requirements. For example, how do you "get down to cases" in an air defense vigilance task with operator requirements when all you know is that a "bogie" must not get through even though it may occur only once in a 24 hr. day.

c) The human factors engineer must be involved in the promulgation of training and personnel selection criteria. If, in a complex system, the status of the man/machine interface is critical to the over-all system performance then the qualifications, job description, and needed skills for the human elements of the system (including maintenance as well as operation) must be a major duty of human factors engineering.

d) As most modern complex systems are relatively costly, it behooves the human factors and systems engineer to model, whenever possible, the system under consideration. Such models must be flexible enough to incorporate a realistic (and usually non-ergodic) representation of the man/machine interfaces. Analyses of data secured from these models also is the concern of the human factors engineer.

e) Finally, although modeling may be the norm for analysis of complex systems, the human factors engineer must never lose touch with the real world. Therefore, whenever feasible, he should be involved in actual system performance tests and in the analyses of the resulting data.

The work being reported on in this paper is in line with several of the above listed missions, and, in particular, the one under subparagraph "d" above. Before considering the details of the research, one should consider the environment in which the interface under study is immersed. This environment is described as a hierarchal command/control system. A generalized C^2 system model must make provision for sensing, filtering, analysis, decision making, and feedback at each level in the hierarchy of command. However, since each level in the hierarchy must feed information upwards in the chain of command and effector-action commands downward, the resultant loops are imbedded in a hierarchal fashion.

Let us trace one such loop. Imagine a line or field of sensors (REMBASS) at the FEBA intended to warn of enemy overload approach. In addition, let us visualize airborne reconnaissance (drone and manned), behind the lines intelligence operations, prisoner interrogation, signal intercept operations and the like. All of this "sensor" information must be filtered, classified, and appropriately analyzed and correlated for presentation to a commander for decision as to appropriate effector action (retreat, advance, hold, encircle, etc.). Once the effector action is ordered, the resulting movements and actions must be reported through the original information gathering network, as well as through command channel status reports, so that further or modified effector action may result. Thus, we have a reentrant feedback loop continuously in action. If we visualize this situation (sensing, decision, effector action, feedback) as occuring at least at each level of command (company, battalion, division, corps), then the significance of the imbedded or hierarchal nature of the multiple feedback loops becomes evident.

How does the particular man/machine interface being reported upon in the paper fit into the above? At almost every stage of information flow there is a point where multiple channels of information must be consolidated and summarized so as to form a new message. A common denominator at these points is the message center, and in particular, field or forward area message centers. The operators in such message centers operate under a combination of stressful environmental factors — acoustic noise, poor light, fear of bodily harm, etc. The subject study attempts to simulate under controlled conditions the first two factors and to substitute for fear of bodily harm a fear-of-failure situation by giving the operators who are taking part in the simulation a series of tasks which are greater in amount than the time allotted for their accomplishment.

This is the general scenario and motivation for the study. Now let us proceed with a discussion of the results which were to be realized from the data gathered. Since in a simulation one cannot hope to achieve all the detailed conditions possible, it was the purpose of this study to come up with a predictor model which would allow for insertion of other permutations and combinations of the considered conditions and then to predict operator performance under these new conditions.

II. DESIGN OF THE EXPERIMENT.

The details of the experimental design were reported in the Proceedings of the Twenty-first Conference on the Design of Experiments in Army Research Development and Testing, ARO Report 76-2, pp 13-29, May 1976. What follows in this section is a general summary of the experiment.

The significance of acoustic noise and ambient light on operator performance was investigated using both an optical display transmission device, and a standard teletypewriter. Primarily, the visual display terminal is a developmental equipment intended to visually present messages on a CRT display where an operator can see and correct his message prior to transmission.

The experiment consisted of testing the transcription accuracy of six experienced communications-center operators under 16 different combinations of environment. Ambient light was varied at four levels, ranging from 24 ft-candles to 3 ft-candles, and acoustic noise was concurrently varied at four sound-pressure levels ranging from 55 dBa to 95 dBa. The 55 dBa level was considered the quiet condition and the 95 dBa level represented an extremely annoying and distracting "pink" noise. The chosen ambient light levels of 24, 12, 6, and 3 ft-candles, respectively, represented successively deteriorating lighting conditions.

The messages for the experiment consisted of forty random-letter word groups of five characters each. They were derived through a random number generator and an alpha-numeric conversion. No message was a duplicate nor were they duplicated by any of the operators on either terminal equipment. The aim of the experiment was to vary the environmental variables and to observe the transcription accuracy of each operator utilizing the visual display terminal as a function of time. The response variable, accuracy (number of committed errors), was the measure of transcription errors that each operator committed per four second interval. The results were compared to an acceptable operator norm, i.e., typing a message format on a standard teletype terminal under the same conditions. Each operator was tested in four sessions, each session programmed for eight random environmental combinations, four for each terminal equipment. See Table 1. The tests were alternated between the optical display unit and the standard teletypewriter to reduce the effects of learning. A thirty minute familiarization period was given each operator prior to the tests.

TABLE 1

TREATMENT SCHEDULE PER OPERATOR

Session	Run	Environmental Treatment* Combinations: Optical Display Terminal	Environmental Treatment* Combinations: Teletype Terminal
I	1	1,4	3,1
	2	4,3	4,4
	3	3,2	2,2
	4	2,1	1,3
II	5	3,1	4,1
	6	4,4	1,2
	7	2,2	3,4
	8	1,3	2,3
III	9	4,1	2,4
	10	1,2	3,3
	11	3,4	1,1
	12	2,3	4,2
IV	13	2,4	1,4
	14	3,3	4,3
	15	1,1	3,2
	16	4,2	2,1

*Treatment = (Ambient Light Level, Acoustic Noise Level)

Ambient Light

Level	Value
1	24 ft-candles
2	12 ft-candles
3	6 ft-candles
4	3 ft-candles

Acoustic Noise

Level	Value
1	55 dBa
2	70 dBa
3	80 dBa
4	95 dBa

III. A NON-LINEAR REGRESSION MODEL FOR MAN-MACHINE INTERFACE.

In this section an acceptable model to predict operator performance is presented so that one can determine the environmental combination of ambient light and acoustic noise which generally causes a minimum number of committed errors. Various linear, multiple linear, and non-linear models were tested for both terminals. The criterion used for choosing the best model was the minimum SSE (sum of squares for error) where

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

and Y_i = observed errors,

$\hat{Y}_i$ = predicted errors.

The general model that best describes the observed data is of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2$$

$$+ \beta_5 X_2^2 + \beta_6 X_1 X_2^2 + \beta_7 X_1^2 X_2 + \beta_8 X_1^2 X_2^2 + \beta_9 X_1^3 + \beta_{10} X_2^3 + \varepsilon_j$$

where Y = average number of errors (operator performance) per cell,

X_1 = ambient light level,

X_2 = acoustic noise level,

β_i = model coefficients, $i = 0,1 \cdots 10$,

ε_j = experimental error, $j = 1, \cdots n$, (the extent to which the observed data and the model disagree, where ε_j's are independent and $\varepsilon \sim N(0, \sigma^2 I)$), and

$n = 16$.

The estimated values of the coefficients, error variance, correlation, and appropriate F statistic for both terminals are summarized in the following table:

Parameter	Optical Display Terminal	Teletypewriter Terminal
β_0	34.7500	-7.793
β_1	.5092	-6.365
β_2	-1.0840	1.018
β_3	- .0399	.1588
β_4	.0359	.1663
β_5	.0137	- .02055
β_6	.0002373	- .0007769
β_7	.001990	- .004906
β_8	-.000011	.00002257
β_9	.003293	.001425
β_{10}	.000053	.0001133
SSE	5.136	3.389
S_ϵ^2	1.027	.6779
$F_{(MODEL)}$	2.735	6.536
$R^2_{y\hat{y}}$	.8455	.9289
$R_{y\hat{y}}$	.9195	.9638

In the case of the optical display terminal, the F statistic indicates a possible overabundance of variables. In the case of the teletypewriter terminal, the small SSE, large $R^2_{y\hat{y}}$, and relatively small F statistic indicate an acceptable model.

Now, we begin to investigate the possibility of eliminating those variables that do not significantly contribute to the dependent variable. The procedure used to form the reduced models was the "forward selection procedure" which begins with the variable X_i, that has the highest correlation $\rho_{x_i y}$ with y. Next, the partial correlation coefficients of the remaining x_j and y, $\rho(x_j y|x_i)$, $j \neq i$, are calculated. The x_j with the greatest $\rho(x_j y|x_i)$ is selected to enter the regression equation. This process is continued, and as each variable is entered into the equation, the multiple correlation coefficient $R^2_{y\hat{y}}$ and the partial F test value for the most recent entry are examined. In the first case, one checks to assure a relatively insignificant change in $R^2_{y\hat{y}}$, and, secondly, whether or not the inserted variable has taken up a significant amount of variation over the previous variables in the regression model. When the partial F test becomes insignificant (the SSE is sufficiently reduced) and $R^2_{y\hat{y}}$ is not very different from the "full model", the process is terminated. The reduced model, therefore, contains all significant variables plus the first two insignificant variables to accomodate any error due to the estimates.

Based on the general model previously stated the appropriate reduced models which characterize operator performance for both terminals are as follows:

i) for the optical display terminal:

$$Y = \beta_0^1 + \beta_1^1 x_2 + \beta_2^1 x_1 x_2^2 + \beta_3^1 x_1^3 + \beta_4^1 x_2^3 + \varepsilon$$

where: $\beta_0^1 = 10.63$,

$\beta_1^1 = -0.1239$,

$\beta_2^1 = 0.000028$,

$\beta_3^1 = -0.0002202$,

$\beta_4^1 = 0.000008367$,

with SSE $= 8.678$,

$S_e^2 = 0.7889$,

$F_{(MODEL)} = 7.783$,

$R^2_{y\hat{y}} = 0.7389$,

$R_{y\hat{y}} = 0.8596$

ii) for the teletypewriter terminal:

$$Y = \beta_0^1 + \beta_1^1 X_1 + \beta_2^1 X_1 X_2 + \beta_3^1 X_1^2 X_2 + \beta_4^1 X_1^3 + \beta_5^1 X_2^3 + \varepsilon$$

where: $\beta_0^1 = 3.211$,

$\beta_1^1 = -1.365$,

$\beta_2^1 = 0.03532$,

$\beta_3^1 = -0.001288$,

$\beta_4^1 = 0.002123$,

$\beta_5^1 = -0.000004273$,

with: $SSE = 7.63$,

$S_\varepsilon^2 = 0.763$,

$F_{(MODEL)} = 10.5$,

$R^2_{y\hat{y}} = 0.84$,

$R_{y\hat{y}} = 0.9165$.

The reduced models now provide the capability to predict the number of transcribed errors given the desired combination of ambient light and acoustic noise.

IV. OPTIMAL LIGHT AND SOUND LEVELS.

One can now attempt to find the light-sound combination that causes the least number of errors to be committed. The method used was simply to evaluate the predicted value of Y for ordered pairs, (X_1, X_2), where X_1 assumes all integer values from 1 to 26, and X_2 assumes even integers values from 50 to 100. These ranges of X_1 and X_2 where chosen based upon the levels of X_1 and X_2 used in the experiment. Thus, the reduced models are used to provide a reasonable extrapolation outside the tested environmental limits.

The predicted Y values, i.e., the predicted number of errors, were calculated for the environmental combinations described in section II for the optical display data (using the reduced model) to obtain the matrix of table 2. Visual examination of this matrix shows that the minimum number of errors, i.e., 4.4, will occur at a light level of 24 ft-candles and a concurrent acoustic noise level of 54 dBa, or, if we are willing to extrapolate slightly outside the region from which data has been obtained, the absolute minimum, 3.8, occurs at 26 ft-candles and 50 dBa. Thus, one can conclude that the minimum number of errors committed on the optical terminal (in the region for which data was taken) occurs at the minimum sound and maximum light combinations, that is, 26 ft-candles/55 dBa.

Table 2. PREDICTED ERROR PERFORMANCE FOR VARIOUS LEVELS OF AMBIENT LIGHT AND ACOUSTIC NOISE FOR THE OPTICAL DISPLAY TERMINAL.

AMBIENT LIGHT LEVEL - Foot-candles of illumination	ACOUSTIC NOISE LEVEL -dBa																										
	50	52	54	56	58	60	62	64	66	68	70	72	74	76	78	80	82	84	86	88	90	92	94	96	98	100
26	3.8	3.8	3.8	3.9	4.0	4.1	4.2	4.3	4.5	4.7	4.9	5.1	5.3	5.6	5.8	6.1	6.4	6.8	7.2	7.5	7.9	8.4	8.8	9.3	9.8	10.4
25	4.1	4.1	4.1	4.2	4.3	4.4	4.5	4.6	4.8	4.9	5.1	5.3	5.5	5.8	6.1	6.3	6.7	7.0	7.3	7.7	8.1	8.5	9.0	9.5	10.0	10.5
24	4.4	4.4	4.4	4.5	4.6	4.6	4.7	4.9	5.0	5.2	5.3	5.5	5.8	6.0	6.2	6.5	6.8	7.2	7.5	7.9	8.3	8.7	9.1	9.6	10.1	10.6
23	4.6	4.7	4.7	4.7	4.8	4.9	5.0	5.1	5.2	5.4	5.5	5.7	5.9	6.2	6.4	6.7	7.0	7.3	7.6	8.0	8.4	8.8	9.2	9.6	10.1	10.6
22	4.9	4.9	4.9	4.9	5.0	5.1	5.2	5.3	5.4	5.5	5.7	5.9	6.1	6.3	6.5	6.8	7.1	7.4	7.7	8.1	8.4	8.8	9.3	9.7	10.2	10.6
21	5.1	5.1	5.1	5.1	5.2	5.3	5.3	5.4	5.6	5.7	5.8	6.0	6.2	6.4	6.7	6.9	7.2	7.5	7.8	8.1	8.5	8.9	9.3	9.7	10.2	10.6
20	5.3	5.3	5.3	5.3	5.4	5.4	5.5	5.6	5.7	5.8	6.0	6.1	6.3	6.5	6.7	7.0	7.3	7.5	7.8	8.2	8.5	8.9	9.3	9.7	10.2	10.6
19	5.4	5.4	5.4	5.4	5.5	5.5	5.6	5.7	5.8	5.9	6.1	6.2	6.4	6.6	6.8	7.0	7.3	7.6	7.9	8.2	8.5	8.9	9.3	9.7	10.1	10.6
18	5.6	5.6	5.6	5.6	5.6	5.6	5.7	5.8	5.9	6.0	6.1	6.3	6.4	6.6	6.8	7.1	7.3	7.6	7.9	8.2	8.5	8.9	9.2	9.6	10.1	10.5
17	5.7	5.7	5.7	5.7	5.7	5.7	5.8	5.9	5.9	6.1	6.2	6.3	6.5	6.7	6.9	7.1	7.3	7.6	7.8	8.1	8.5	8.8	9.2	9.6	10.0	10.4
16	5.8	5.8	5.7	5.7	5.8	5.8	5.8	5.9	6.0	6.1	6.2	6.3	6.5	6.7	6.9	7.1	7.3	7.5	7.8	8.1	8.4	8.7	9.1	9.5	9.9	10.3
15	5.9	5.8	5.8	5.8	5.8	5.8	5.9	5.9	6.0	6.1	6.2	6.3	6.5	6.6	6.8	7.0	7.2	7.5	7.7	8.0	8.3	8.6	9.0	9.4	9.7	10.2
14	5.9	5.9	5.9	5.8	5.8	5.9	5.9	6.0	6.0	6.1	6.2	6.3	6.5	6.6	6.8	7.0	7.2	7.4	7.7	7.9	8.2	8.5	8.9	9.2	9.6	10.0
13	6.0	5.9	5.9	5.9	5.9	5.9	5.9	5.9	6.0	6.1	6.2	6.3	6.4	6.6	6.7	6.9	7.1	7.3	7.6	7.8	8.1	8.4	8.7	9.1	9.4	9.8
12	6.0	5.9	5.9	5.9	5.9	5.9	5.9	5.9	6.0	6.1	6.1	6.2	6.4	6.5	6.6	6.8	7.0	7.2	7.5	7.7	8.0	8.3	8.6	8.9	9.3	9.7
11	6.0	5.9	5.9	5.9	5.8	5.9	5.9	5.9	5.9	6.0	6.1	6.2	6.3	6.4	6.6	6.7	6.9	7.1	7.3	7.6	7.8	8.1	8.4	8.7	9.1	9.5
10	6.0	5.9	5.9	5.8	5.8	5.8	5.8	5.8	5.9	5.9	6.0	6.1	6.2	6.3	6.5	6.6	6.8	7.0	7.2	7.4	7.7	7.9	8.2	8.5	8.9	9.2
9	6.0	5.9	5.8	5.8	5.8	5.8	5.8	5.8	5.8	5.9	5.9	6.0	6.1	6.2	6.3	6.5	6.6	6.8	7.0	7.3	7.5	7.8	8.0	8.3	8.7	9.0
8	5.9	5.9	5.8	5.8	5.7	5.7	5.7	5.7	5.7	5.6	5.8	5.9	6.0	6.1	6.2	6.3	6.5	6.7	6.9	7.1	7.3	7.6	7.8	8.1	8.4	8.8
7	5.9	5.8	5.8	5.7	5.7	5.6	5.6	5.6	5.7	5.7	5.7	5.8	5.9	6.0	6.1	6.2	6.4	6.5	6.7	6.9	7.1	7.4	7.6	7.9	8.2	8.5
6	5.9	5.8	5.7	5.6	5.6	5.6	5.6	5.5	5.6	5.6	5.6	5.7	5.7	5.8	5.9	6.1	6.2	6.3	6.5	6.7	6.9	7.2	7.4	7.7	8.0	8.3
5	5.8	5.7	5.6	5.6	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.5	5.6	5.7	5.8	5.9	6.0	6.2	6.3	6.5	6.7	6.9	7.2	7.4	7.7	8.0
4	5.8	5.7	5.6	5.5	5.4	5.4	5.4	5.3	5.3	5.4	5.4	5.4	5.5	5.5	5.6	5.7	5.8	6.0	6.1	6.3	6.5	6.7	6.9	7.2	7.5	7.7
3	5.7	5.6	5.5	5.4	5.4	5.3	5.3	5.2	5.2	5.2	5.2	5.3	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.1	6.3	6.5	6.7	6.9	7.2	7.5
2	5.6	5.5	5.4	5.3	5.3	5.2	5.2	5.1	5.1	5.1	5.1	5.1	5.2	5.2	5.3	5.4	5.5	5.6	5.7	5.9	6.1	6.2	6.5	6.7	6.9	7.2
1	5.6	5.4	5.3	5.3	5.2	5.1	5.1	5.0	5.0	5.0	5.0	5.0	5.0	5.1	5.1	5.2	5.3	5.4	5.5	5.7	5.8	6.0	6.2	6.4	6.7	6.9

A similar matrix of predicted errors was computed for the reduced teletypewriter model, and is shown in table 3. In this case, visual examination shows that the minimum number of predicted errors occur at a light level of about 16-17 ft-candles and at a concurrent sound level of about 55 dBa. In both cases (optical display and teletypewriter) the results of the minima were expected. It is to be noted, however, that in a tactical situation the environmental factors of ambient light and acoustic noise are far from optimal. Thus, one can conclude from the matrices that for a wide variety of the environmental factors X_1 and X_2, one can predict how well experienced communicators will perform.

Table 3. PREDICTED ERROR PERFORMANCE FOR VARIOUS LEVELS OF AMBIENT LIGHT AND ACOUSTIC NOISE FOR THE TELETYPEWRITER TERMINAL.

AMBIENT LIGHT LEVEL - Foot-candles of Illumination	ACOUSTIC NOISE LEVEL -dBa																										
	50	52	54	56	58	60	62	64	66	68	70	72	74	76	78	80	82	84	86	88	90	92	94	96	98	100
26	6.9	6.9	6.9	7.0	7.0	7.0	7.0	7.0	7.0	6.9	6.9	6.9	6.8	6.8	6.7	6.7	6.6	6.5	6.4	6.3	6.2	6.1	6.0	5.8	5.7	5.5
25	5.6	5.7	5.8	5.9	5.9	6.0	6.1	6.1	6.2	6.2	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.2	6.2	6.2	6.1	6.0	6.0	5.9	5.8
24	4.6	4.7	4.8	5.0	5.1	5.2	5.3	5.5	5.6	5.7	5.7	5.8	5.9	6.0	6.0	6.1	6.1	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.1	6.1
23	3.7	3.9	4.0	4.2	4.4	4.6	4.8	4.9	5.1	5.2	5.4	5.5	5.6	5.7	5.8	5.9	6.0	6.1	6.2	6.3	6.3	6.4	6.4	6.4	6.5	6.5
22	2.9	3.2	3.4	3.6	3.9	4.1	4.3	4.5	4.7	4.9	5.1	5.3	5.4	5.6	5.7	5.9	6.0	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.8	6.9
21	2.4	2.6	2.9	3.2	3.4	3.7	4.0	4.2	4.4	4.7	4.9	5.1	5.3	5.5	5.7	5.9	6.1	6.3	6.4	6.6	6.7	6.9	7.0	7.1	7.2	7.3
20	1.9	2.2	2.5	2.9	3.2	3.4	3.7	4.0	4.3	4.6	4.8	5.1	5.3	5.6	5.8	6.0	6.2	6.4	6.6	6.8	7.0	7.2	7.3	7.5	7.6	7.7
19	1.6	2.0	2.3	2.6	3.0	3.3	3.6	3.9	4.2	4.5	4.8	5.1	5.4	5.6	5.9	6.1	6.4	6.6	6.8	7.1	7.3	7.5	7.7	7.8	8.0	8.2
18	1.4	1.8	2.1	2.5	2.9	3.2	3.5	3.9	4.2	4.5	4.8	5.2	5.5	5.7	6.0	6.3	6.6	6.8	7.1	7.3	7.6	7.8	8.0	8.2	8.4	8.6
17	1.3	1.7	2.1	2.5	2.8	3.2	3.6	3.9	4.3	4.6	4.9	5.3	5.6	5.9	6.2	6.5	6.8	7.1	7.3	7.6	7.9	8.1	8.3	8.6	8.8	9.0
16	1.3	1.7	2.1	2.5	2.9	3.3	3.6	4.0	4.4	4.7	5.1	5.4	5.8	6.1	6.4	6.7	7.0	7.3	7.6	7.9	8.1	8.4	8.6	8.9	9.1	9.3
15	1.4	1.8	2.2	2.6	3.0	3.4	3.8	4.1	4.5	4.9	5.2	5.6	5.9	6.3	6.6	6.9	7.2	7.5	7.8	8.1	8.4	8.7	8.9	9.2	9.4	9.6
14	1.5	1.9	2.3	2.7	3.1	3.5	3.9	4.3	4.7	5.0	5.4	5.8	6.1	6.4	6.8	7.1	7.4	7.7	8.0	8.3	8.6	8.9	9.1	9.4	9.6	9.9
13	1.7	2.1	2.5	2.9	3.3	3.7	4.1	4.5	4.8	5.2	5.6	5.9	6.3	6.6	6.9	7.3	7.6	7.9	8.2	8.5	8.8	9.0	9.3	9.5	9.8	10.0
12	1.9	2.3	2.7	3.1	3.5	3.9	4.3	4.6	5.0	5.4	5.7	6.1	6.4	6.7	7.1	7.4	7.7	8.0	8.3	8.5	8.8	9.1	9.4	9.6	9.8	10.1
11	2.1	2.5	2.9	3.3	3.7	4.1	4.4	4.8	5.2	5.5	5.8	6.2	6.5	6.8	7.1	7.4	7.7	8.0	8.3	8.6	8.8	9.1	9.3	9.6	9.8	10.0
10	2.4	2.8	3.1	3.5	3.9	4.2	4.6	4.9	5.3	5.6	5.9	6.2	6.6	6.9	7.2	7.4	7.7	8.0	8.3	8.5	8.8	9.0	9.2	9.4	9.7	9.9
9	2.6	3.0	3.3	3.7	4.0	4.4	4.7	5.0	5.3	5.7	6.0	6.3	6.5	6.8	7.1	7.4	7.6	7.9	8.1	8.4	8.6	8.8	9.0	9.2	9.4	9.6
8	2.9	3.2	3.5	3.8	4.2	4.5	4.8	5.1	5.4	5.6	5.9	6.2	6.5	6.7	7.0	7.2	7.4	7.7	7.9	8.1	8.3	8.5	8.6	8.8	9.0	9.1
7	3.1	3.4	3.7	3.9	4.2	4.5	4.8	5.0	5.3	5.6	5.8	6.0	6.3	6.5	6.7	6.9	7.1	7.3	7.5	7.7	7.8	8.0	8.1	8.3	8.4	8.5
6	3.2	3.5	3.7	4.0	4.2	4.5	4.7	5.0	5.2	5.4	5.6	5.8	6.0	6.2	6.4	6.5	6.7	6.9	7.0	7.1	7.3	7.4	7.5	7.6	7.7	7.8
5	3.3	3.6	3.8	4.0	4.2	4.4	4.6	4.8	5.0	5.1	5.3	5.5	5.6	5.8	5.9	6.0	6.1	6.3	6.4	6.4	6.5	6.6	6.7	6.7	6.8	6.8
4	3.4	3.6	3.7	3.9	4.1	4.2	4.4	4.5	4.6	4.7	4.9	5.0	5.1	5.2	5.3	5.4	5.4	5.5	5.5	5.6	5.6	5.7	5.7	5.7	5.7	5.7
3	3.4	3.5	3.6	3.7	3.8	3.9	4.0	4.1	4.2	4.2	4.3	4.4	4.4	4.5	4.5	4.5	4.6	4.6	4.6	4.6	4.6	4.5	4.5	4.5	4.4	4.3
2	3.2	3.3	3.4	3.4	3.5	3.5	3.5	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.5	3.5	3.4	3.4	3.3	3.2	3.1	3.0	2.9	2.8
1	3.0	3.0	3.0	3.0	3.0	3.0	2.9	2.9	2.9	2.8	2.8	2.7	2.6	2.6	2.5	2.4	2.3	2.2	2.1	1.9	1.3	1.7	1.5	1.3	1.2	1.0

V. TIME-SERIES MODELING OF MAN/MACHINE INTERFACES.

The best non-linear regression model presented in the previous section dealt with the prediction of the number of committed errors as a function of two environmental variables, namely, light and sound. More often, the communications engineer is interested in such factors as performance, and efficiency as a function of time. Thus, utilizing time-series models, it may be possible to characterize a group of operators either singly or as a whole for predicting the number of committed errors at times t_1, t_2, t_3, ..., t_n, in the future. The time-series approach for this type of information is somewhat unique in that not many attempts have been made to implement this methodology in analyzing time-dependent man-machine interface data. In view of this uniqueness, there are a number of shortcomings that were experienced. One of the most serious limitations was the sample size. However, enough information is available so that one can initiate the time-series methodology into this particular subject area. This approach is extremely useful because it characterizes, within reason, the error performance of any communications terminal equipment operator working in wide variety of environments.

Incorporated into the design of the experiment was a four-second time interval counter. This provided a running count of the number of transcribed errors in each four-second time period for the duration of the test. Thus, thirty-two non-deterministic time-series were created (sixteen per terminal, one corresponding to each combination of environmental factors). Of the time series so obtained the two most critical environmental combinations are presented, namely, (1,4) and (4,4) (refer to Section II). Criticality was determined by the degree of non-stationarity of the series, or in other words, the amount of filtering required to bring the process into statistical equilibrium.

Clearly, the time-series characterization of the data is very promising from the point of view of affording to the communications system designer and planner a means to predict the human element of the total communications system architecture. The following stochastic formulations obtained were very adequate in characterizing the underlying process of error performance:

a. for the (1,4) environment, <u>teletypewriter terminal</u>, we obtained the mixed autoregressive-moving averages (ARMA) model:

$$\hat{X}_t = -0.046 + 0.660X_{t-1} + 0.367X_{t-2} + Z_t + 0.449Z_{t-1} + 0.223Z_{t-2} + 0.422Z_{t-3},$$

b. for the (1,4) environment, optical display terminal, the third order autoregressive (AR) model obtained was:

$$\hat{X}_t = + 0.254X_{t-1} + 0.133X_{t-2} + 0.355X_{t-3} + 0.258X_{t-4} + Z_t,$$

c. for the (4,4) environment, teletypewriter terminal, we obtained another mixed model, (ARMA):

$$\hat{X}_t = 0.006 + 1.785X_{t-1} - 0.570X_{t-2} - 0.215X_{t-3} + Z_t + 0.950Z_{t-1} + 0.191Z_{t-2} + 0.016Z_{t-3},$$

d. and, finally, for the (4,4) environment, optical display terminal, the third order moving-averages (MA) process obtained was:

$$\hat{X}_t = 2.158 + Z_t - 0.453Z_{t-1} + 0.023Z_{t-2} + 0.051Z_{t-3}.$$

To illustrate the adequacy of the models figures 1 and 2 graphically display the observed and simulated information for the optical display terminal (ODT). These particular presentations were chosen because of the projected role of the ODT in future communications systems. The details of the teletypewriter terminal analysis and a comparison to the ODT will be presented at a later date.

One of the implied features of this research is that for each environmental combination, no common realization, either ARMA, MA, or AR, was obtained to characterize operator performance. One can conclude, therefore, that even with an adequately developed procedure for analysis, more than one characterization may be required to evaluate the human subsystem in sophisticated communications systems. The procedures developed clearly provide a realistic view of the complex man-machine interface that occurs in current communications systems.

BIBLIOGRAPHY

1. "Design of Experiments Dealing with Man-Machine Interfaces in Current Communication Systems", R. J. D'Accardi, H. S. Bennett, and R. S. Hennessy. Proceedings of the Twenty-first Conference on the Design of Experiments in Army Research Development and Testing. ARO Report 76-2, May 1976.

2. "Probability and Statistics for Engineers and Scientists", Walpole, R.E., and R. H. Myers, Macmillan, New York, 1972.

3. "Time Series Analysis, Forecasting and Control, Box, G.E.P., and G.M. Jenkins, Holden-Day, San Francisco, 1970.

4. "Applied Regression Analysis", Draper, N. R. and H. Smith, John Wiley, New York, 1966.

FILTERED MODEL:

$$y_t = -0.746\, y_{t-1} - 0.613\, y_{t-2} - 0.258\, y_{t-3} + Zt$$

AUTOREGRESSIVE FORECASTING MODEL:

$$\hat{x}_t = 0.254\, x_{t-1} + 0.133\, x_{t-2} + 0.355\, x_{t-3} + 0.258\, x_{t-4} + Z_t$$

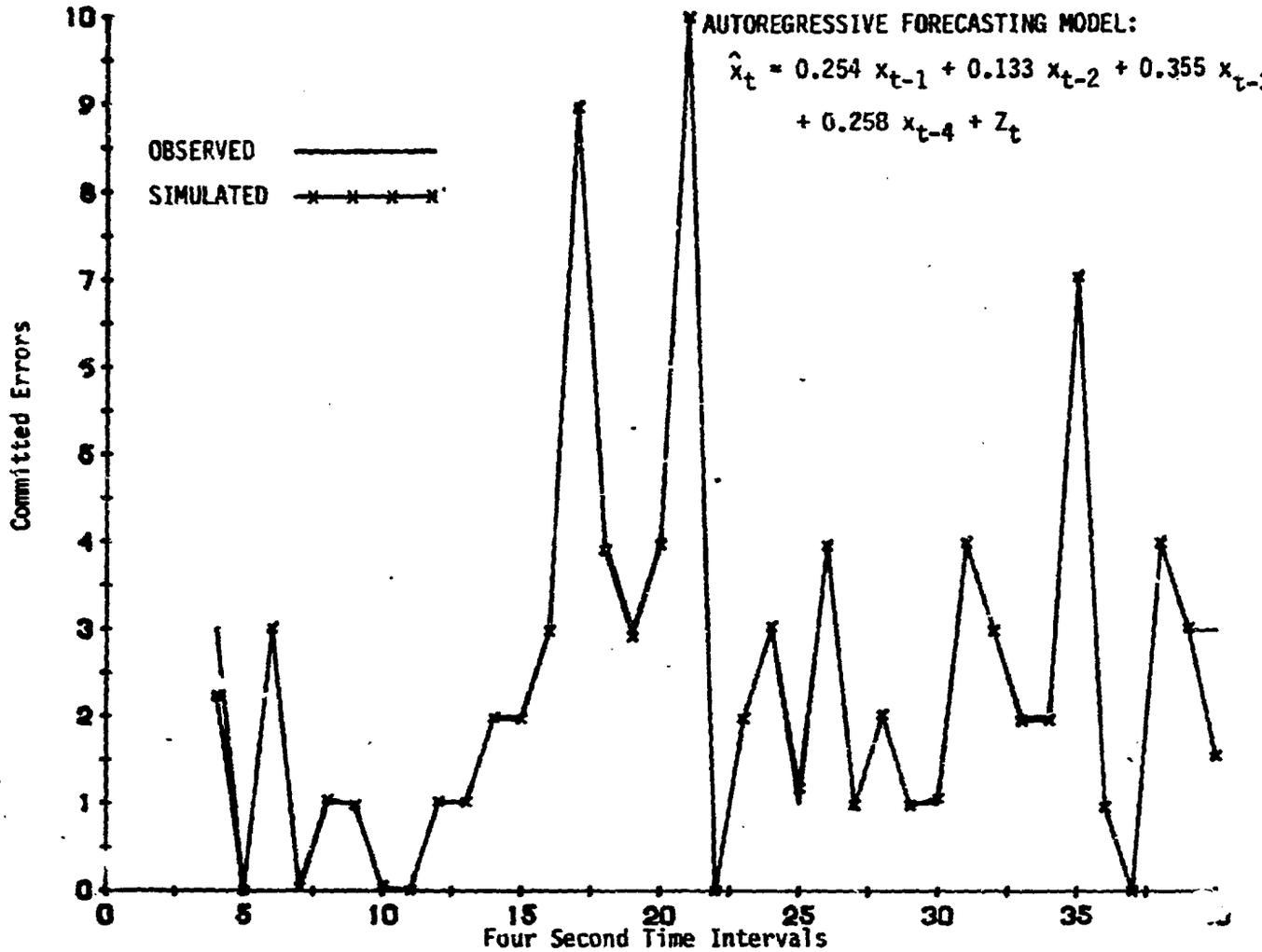


FIGURE 1 SIMULATED MAN/MACHINE INTERFACE SERIES USING THE AUTOREGRESSIVE MODEL vs. THE OBSERVED INFORMATION FOR THE OPTICAL DISPLAY TERMINAL, 1,4 ENVIRONMENT

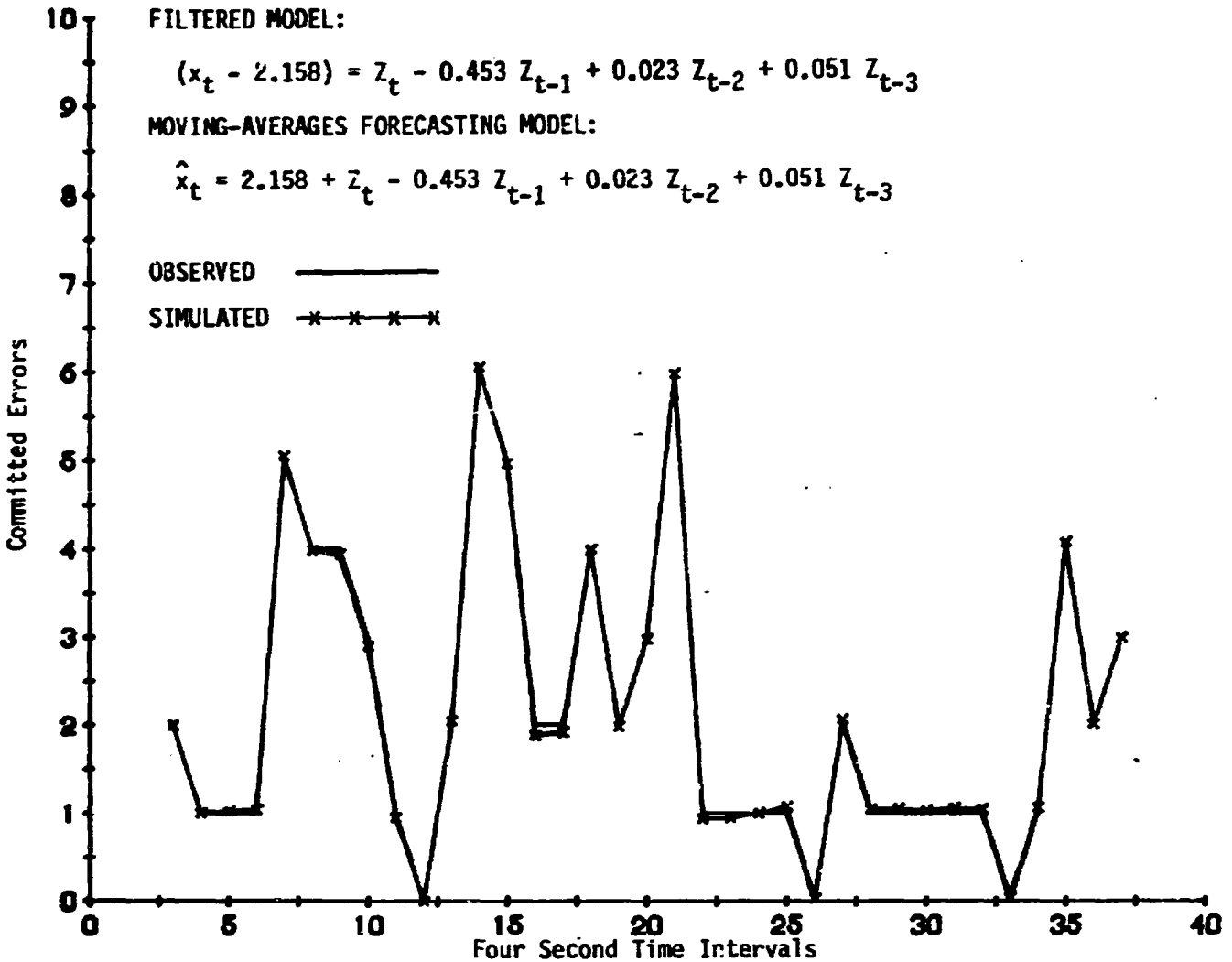


FIGURE 2 SIMULATED MAN/MACHINE INTERFACE SERIES USING THE MOVING-AVERAGES MODEL vs. THE OBSERVED INFORMATION FOR THE OPTICAL DISPLAY TERMINAL, 4,4 ENVIRONMENT

IMPROVED QUANTIFICATION OF PLAYER EFFECTS IN EXPERIMENTAL DESIGN

William Mallios, Robert Batesole,
Donald Leal, and Thieu Tran
BDM Scientific Support Laboratory
U.S. Army Combat Developments Experimentation Command
Fort Ord, California

ABSTRACT. This paper discusses and illustrates a methodological alternative to standard experimental designs for use in certain applications. Focus is on the quantification of random subject or player effects to be used in place of dummy (1, 0) variables in the usual linear model assumed for analysis of variance. The advantages of this approach are: (1) increasing the efficiency of the analysis; (2) providing explanation of player differences; (3) forming a base for the evaluation of adjusted treatment effects; and (4) logical formulization for extrapolations to other populations of individuals for increased utility of the results.

The reader is assumed to have some familiarity with the statistical analysis of experimental data.

I. A MIXED EFFECTS MODEL.

Consider the Mixed Effects Model

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad (1.1)$$

where τ_i is the fixed differential effect of the i-th treatment, i=1,..., p; the β_j, j=1,..., q, are random block effects which are assumed normally[1] and independently distributed with $E(\beta_j) = 0$ and variance $(\beta_j) = \sigma^2_\beta$; i.e., β_j: NID $(0, \sigma^2_\beta)$; for the model error, ε_{ij}, it is assumed that ε_{ij}: NID $(0, \sigma^2_\varepsilon)$. The dependent variable is y_{ij}, while μ is the base from which differential effects are measured.

[1] The normality assumption is required for tests of significance, not for estimation.

II. THE INDIVIDUAL AS A BLOCK. In many applications, the block is defined by an individual who is subjected to some or all of the ρ treatments in succession. For example, in clinical trials, cross-over designs[2] are used to compare drugs (treatments) by subjecting each individual to each drug in succession.[3] A second example is in military field testing where each participant is subjected to fire by each of different weapon types (treatments), the dependent variable is some measure of suppression per treatment.

III. TREATMENT BY BLOCK INTERACTIONS. In designing experiments characterized by these examples, it is hoped that treatments and blocks do not interact. Assuming non-significance, this interaction becomes the inherent model error, ϵ_{ij}. If, on the other hand, this interaction is expected to be significant, then replication should be incorporated in the experiment so that this interaction can be estimated. Replication, however, poses problems in the examples just discussed. In the drug experiment, replicating the individual may induce complex carry-over effects of drugs. In the military experiment, replication may induce learning and/or boredom effects so that repeated use of an individual within a treatment level does not constitute a replication in the statistical sense of the word. Consequently, many experiments are designed under the assumption of no block by treatment interaction when prior logic is to the contrary. In fact, it is not unreasonable to expect individuals to react differently to treatments in many situations. For example, in military field experimentation, subjecting different individuals to the same experimental situation will produce different responses depending upon an individual's military experience, degree of enthusiasm, mental aptitude, physical sensitivity (hearing, eyesight, etc.), and physical endurance. Treatments which are sensitive to any of these attributes will lead to an interaction of block (player) by treatment since these attributes will differ from player to player.

2Cross-over designs (see (1)) are sometimes favored over parallel designs wherein individuals are maintained on the same treatment over the entire period of experimentation; i.e., individuals are nested within treatments. With cross over designs, differences between individuals are neutralized in comparing treatments, given certain assumptions are met.

3The ordering of drugs will generally differ between groups of individuals so as to allow for estimation of carry over effects.

IV. QUANTIFYING THE BLOCK EFFECT IN TERMS OF A DUMMY VARIABLE. In model (1.1), the block effect quantifies the individual in terms of a dummy (0,1) variable. The intent of these variables is to isolate the between individuals source of variation so as to increase the efficiency of the analysis. Note, however, that under model (1.1), no attempt is made to distinguish between differences in physiological or psychological states within individuals; i.e., the state of the individual may vary during times when different treatments are administered to him. The result of the (0,1) dummy (block) variable analysis is the estimation of an "average" effect for each individual. If these states vary substantially during experimentation, the efficiency of the analysis corresponding to model (1.1) decreases relative to the case where "adjustments" thru the use of covariables that quantify these states are made for. Moreover, variations in these states during experimentation, which cannot be realistically controlled only measured, can lead to serious biases in comparing treatments when a predominance of a particular state exists within a treatment over another treatment.

V. OTHER METHODS OF QUANTIFYING THE INDIVIDUAL. How does one quantify the individual other than through dummy variables? A general answer is through covariables while a specific answer lies in the particular application. In drug experiments, measures drawn from the blood and/or urine serve to quantify the individual. In the military field test, the individual partially quantifies himself in terms of his responses to psychological questions.

If the psychological or physiological states are not expected to vary significantly within individuals over the course of the experiment, quantification of the individual may be required only once, say prior to the application of the first treatment. Replacing the block dummy variables with the covariables quantifying the individual serves several purposes. Firstly, the covariables explain differences between individuals whereas dummy variables do not. Secondly, assuming that only a few covariables are required to adequately quantify the person, the replacement of the dummy variables by the covariables adds to the error degrees of freedom and hence to the power of the test. Thirdly, more freedom is allowed to estimate treatment by player covariable(s) interaction in

lieu of the previously mentioned treatment by block interaction. Finally, the possibility exists to use these covariables as a logical formulization for extrapolation to other populations of players thereby enhancing the utility of the results.

Then in place of model (1.1),

$$y_{ik} = \alpha_o + \tau_i + \sum_{k=1}^{r} \alpha_k x_{ik} + \delta_{ik} \qquad (5.1)$$

may be applied, where x_{ik} denote covariables which quantify the individual, k=1,..., r; the α_k are regression coefficients; and the δ_{ik} are model errors with the usual assumptions for δ_{ik} accompany the model for tests of significance.

Model (5.1) holds if the states fluctuate widely within individuals, though in this case, quantification of the individual should take place just prior to each treatment application, not after. If the individual is quantified following treatment, there is the possibility of treatments affecting the covariables. In this event, direct and indirect treatment effects may have to be considered through a system of structural regression equations; e.g., model (5.1) and

$$X_{ik} = \alpha'_o + \tau_{ki} + \delta_{ik} \qquad (5.2)$$

could form a system where τ_i of (5.1) is the direct i-th treatment effect on y, τ_{ki} of (5.2) is the direct i-th treatment effect on x_k, and $\tau_i + \alpha_k \tau_{ki}$ is the overall i-th treatment effect on y; see Mallios (2)

For the case of significant block by treatment interactions with player quantification taking place prior to each treatment application, the model would take the form

$$y_{ik} = \gamma_o + \tau_i + \sum_{k=1}^{r} \gamma_k X_{ik} + \sum_{i,k}^{P,r} (\tau\gamma)_{ik} X_{ik} + \varepsilon'_{ik} \qquad (5.3)$$

where the γ are regression coefficients. The $(\tau\gamma)_{ik}$ allow for the γ_{ik} to differ between treatments. Here, note that with this formulization, replication within a treatment is not necessary, since the repitition aspect is through the communality, provided overlapping exists between treatments, of the x_{ik} responses.

VI. QUANTIFYING A PLAYER'S PROPENSITY TO PARTICIPATE IN A SUPPRESSION EXPERIMENT. An experiment was conducted to evaluate individuals' assessment of danger when fired upon under different conditions while situated in foxholes. The seven treatments included overhead fire by different small arms with varying bursts at varying ranges. Initially, 31 participants were rehearsed and prebriefed on experimental objectives and techniques. Thereupon, all 31 were situated in separate foxholes and were simultaneously subjected to each treatment over seven distinct trials. Following each trial, each player gave an assessment as to whether the particular treatment was "very dangerous", "quite dangerous", "fairly dangerous", or "not very dangerous".

Prior to each trial, each player answered the series of questions in Table 6.1. Their answers were intended to give measure to the player's propensity to participate in the experiment. In very loose terms, the answers give measure to player motivation.

Note that most of the questions are directed at short term attitude changes rather than long term changes; e.g., the player could be bored, tired, or hungry on one trial but not another. Table 6.1 presents the percentage of yes responses over all players and all trials. Due to the high percentage of yes responses, questions 4 and 8 were deleted from further consideration.

Since the questions were answered on a per trial basis, it must be established that the questions had the same meaning between trials or that relations between questions remained the same over trials. Accordingly, based on quantifying "yes", no answer, and "no" responses according to -1, 0, and 1, a 10 by 10 covariance matrix, say S_i, based on questionnaire responses was calculated for each trial. Let S_i estimate Σ_i. Then relations between questions differ between trials if the hypothesis

$$H_o : \Sigma_1 = \ldots = \Sigma_7$$

is rejected. Using the likelihood ratio criterion (see (3)), H_o was not rejected. Consequently, all the questionnaire data were pooled into one covariance matrix (based on differing mean vectors per trial), say S.

The matrix S was converted to C, the matrix of simple correlations, and a principal component analysis (3) was performed. The three eigenvectors associated with the three largest eigenroots are given in Table 6.1. These eigenvectors are part of a principal component analysis and provide a redimensioning of the original questions to isolate the inherent pattern in the responses to the questions. Thus, the eigenvector associated with the largest eigenroot represents the linear combination of the original responses which had the most variability. These eigenvectors are then used to generate the values of the covariables. On a subjective basis, these eigenvectors are designated as indices relating to experiment validity, to player discomfort, and to trial structure.

For the first index, scores for the 31 players are given for a particular trial in Table 6.2. These scores reflect the degree to which participants felt the experiment was valid prior to the particular trial.

VII. REPLACING BLOCK EFFECTS WITH COVARIABLES IN A DISCRIMINANT ANALYSIS.

In this experiment, the dependent variable - level of danger - is categorical so that discriminant analysis[4] is a natural recourse with treatments and blocks as predictors.

In the first analysis treatments were forced in as predictors while player (block) effects were allowed to enter, if significant, through stepwise discriminant analysis (4). In the second analysis, block effects were replaced by covariables produced from eigenvectors[5] corresponding to C and by interactions of the covariables produced from the first three eigenvectors with treatments. Again, treatments were forced in as predictors while the other variables were scanned for significance as before.

One result was that following scanning of variables for significance, the U statistic (a measure of the goodness of the discriminate) dropped from .59 in the first analysis to .46 in the second. Thus, quantifying the player per treatment allow for a great explanation of the variability.

---

[4] Although the following example employs a discriminant analysis, this quantification technique has and can be used in the general linear model.

[5] An eigenvector when multiplied with the vector Xn of (1,0) responses will produce the scores which become the measure(s) of the covariable(s) to be used as the predictors in the model.

The model exercise of the discriminant function through Bayes Theokin[6] is given in Table 7.1. Presented therein are the probabilities of the four danger categories given the particular treatment and given a particular score for the first index. For example, for treatment 7, a high score for index 1 was contrasted with a low score. Of these who thought the experiment was valid (large negative scores for index 1), 57% though treatment 7 was very dangerous, 36% considered it quite dangerous, 6% considered it faily dangerous, while 1% said it was not very dangerous. These probabilities are contrasted with those associated with individuals who thought the experiment was not valid.

The obvious implication here is that the players "propensity to participate" going into a particular trial has an overwhelming effect on the outcome. Without adjustments for these states, experimental results would have fallen somewhere between the two sets of results in Table 7.1. Thus, it can be seen that the quantification of the player in this way not only provided a more efficient analysis, but also some insight into the dynamics of the experiment which would ultimately lead to better experimental technique.

REFERENCES

(1) W. G. Cochran and G. M. Cox (1957), Experimental Designs, Wiley, N.Y.

(2) W.S. Mallios (1970), The analysis of structural effects in experimental design. JASA, Vol. 65, pp 808-827.

(3) T.W. Anderson (1958), An Introduction To Multivariate Statistical Analysis, Wiley, N.Y.

(4) U.C.L.A. Biomedical Computer Programs, P Series, (1976) W.J.Dixon Editor, University of California Press, Los Angeles.

TABLE 6.1 PLAYER QUESTIONNAIRE AND FIRST THREE PRINCIPLE COMPONENTS

			INDICIES RELATING TO		
		% RESPONDING YES	EXPERIMENT VALIDITY	DISCOMFORT	TRIAL STRUCTURE
1.	On the whole, are these trials realistic?	1. 50	-0.48	-0.06	-0.10
2.	Are these trials fatiguing?	2. 24	0.32	-0.47	-0.03
3.	Are these trials boring?	3. 47	0.33	0.09	0.50
4.	Was your task adequately described to you?	4. 99	deleted from further consideration		
5.	Does this experiment benefit the taxpayer?	5. 50	-0.36	-0.34	0.07
6.	Are there too many distractions during the trials?	6. 66	0.01	0.25	0.23
7.	Is the number of trials per day about right?	7. 66	-0.28	0.30	-0.49
8.	Are the controller personnel considerate of you?	8. 100	deleted from further consideration		
9.	Do you want to be replaced as a player?	9. 33	0.28	0.25	-0.17
10.	Is the weather seriously bothering you?	10. 14	0.22	-0.56	-0.32
11.	Is the experiment safe?	11. 80	-0.32	0.06	0.55
12.	Could this experiment be improved a lot?	12. 58	-0.34	0.34	-0.03

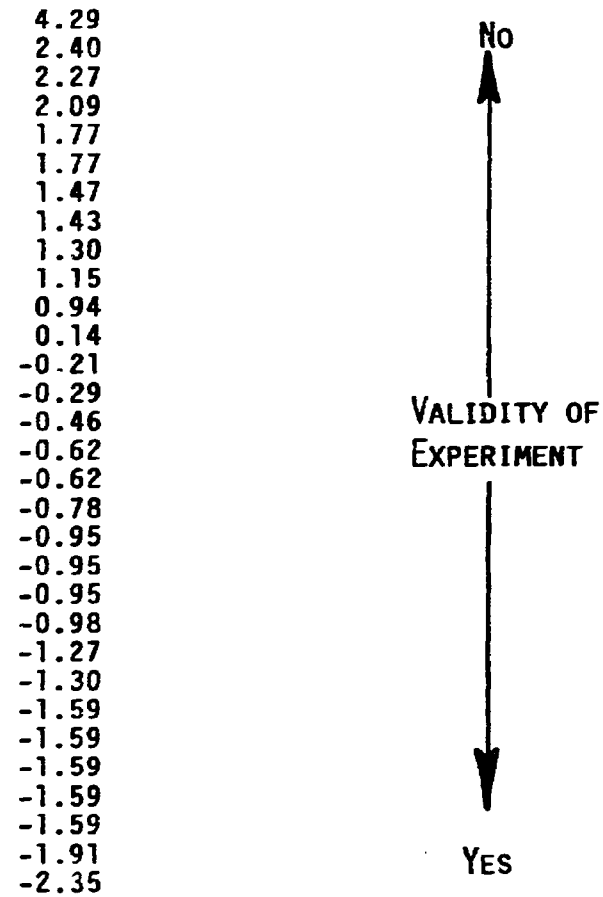


FIGURE 6.2 PLAYER SCORES FOR VALIDITY INDEX

TABLE 7.1 PROBABILITIES OF DANGER LEVELS FOR GIVEN TREATMENTS

	TREATMENTS				POSITIVE RESPONSES REGARDING VALIDITY OF EXPERIMENT				NEGATIVE RESPONSES REGARDING VALIDITY OF EXPERIMENT			
					DANGER LEVEL							
	WEAPON	BURST	TIME BETWEEN BURST	RANGE	VERY	QUITE	FAIRLY	NOT VERY	VERY	QUITE	FAIRLY	NOT VERY
1.	M16	3	1	100M	.03	.75	.18	.04	01	06	86	07
2.	M16	1	1	100M	.01	07	72	20	00	02	07	91
3.	M60	3	12	200M	.01	06	26	67	00	01	01	98
4.	M60	6	12	200M	.01	17	71	11	00	13	12	75
5.	M60	EYELIC		400M	.01	15	76	08	00	00	77	22
6.	M2	EYELIC		400M	.99	01	00	00	01	99	00	00
7.	M2	6	4	800M	.57	36	06	01	01	06	77	15

ERRORS IN LINEAR FITS DUE TO FUNCTION MISMATCH AND NOISE WITH SPLINE APPLICATIONS

G. W. Lank, W. B. Kendall, P. A. Gartenberg
MARK Resources, Inc., Marina del Rey, California

INTRODUCTION

In producing trajectory estimates from noisy radar data it is generally necessary to smooth the radar data by fitting a deterministic function to it. The choice of function depends on how much is known about the trajectory. However, usually all that is known is that range as a function of time will be a "smooth" function with "small" values for its higher derivatives. Then a reasonable and practical choice for the deterministic function is a polynomial of low order. This is the function which has zero for all derivatives beyond a certain order, and thus will be a good approximation to any true range function which has sufficiently small higher derivatives over the smoothing interval.

A smoothing function related to polynomials, but which has wider applicability, is the polynomial spline. This function consists of a series of polynomials which are used over contiguous time intervals to represent the true range function. The individual time intervals are chosen to be sufficiently short for all higher-order derivatives to be negligible (i.e., over each short interval the range data very nearly follow a polynomial) and smoothness of the overall fit is achieved by constraining the individual polynomials to match their neighbor's value, slope, and perhaps higher derivatives, at the boundaries (knots) between polynomials. This function has the advantages that it can be used to smooth data over intervals which are far too long to use a low-order polynomial, but at the same time it is much more constrained (and, therefore, much smoother) than a higher-order polynomial.

PROBLEM TO BE SOLVED

This brings us to the problem addressed here. When fitting splines to noisy measurements there are two distinct sources of error which prevent the fitted smooth function from being equal to the true noise-free underlying function: (1) Even in the absence of noise the underlying (trajectory) function may not be of the form of a spline, so that a perfect fit is impossible. (2) Noise in the (range) measurements of the underlying function prevent a perfect fit. Quantitative results for the effects of these two error sources can be gotten as follows.

FORMULATION

Assume we observe a function, such as range versus time, at M discrete times which are not necessarily uniformly spaced. The observations of the function have additive noise present in each sample. The noise is Gaussian, zero mean, independent from sample to sample, and has the same variance σ^2 at each sample. The observed noisy function is to be fitted in time by the weighted sum of F basis functions. In general if the function were to be observed noise free, its form would not necessarily be exactly equal to a weighted sum of the F basis functions. A set of basis functions which is used in practice is those functions which yield a polynomial spline.

THE ERROR AVERAGED OVER TIME

The statistics of the sum E_T of *all* the squared errors at the sampled times (i.e., the sum of the squared differences between the resultant weighted sum of the basis functions and the actual noise-free function) is found. It is found that E_T has a biased χ^2 distribution with F degrees of freedom

with the variance corresponding to each degree of freedom given by σ^2. The bias is the sum of the squared errors which would exist at the sampled times if no noise were present. It is due to the fact that the noise-free function is not necessarily exactly equal to a weighted sum of the F basis functions.

The probability density of E_T is specifically given by

$$p(E_T) = \begin{cases} \dfrac{x^{F/2-1}e^{-x/2}}{2^{F/2}\sigma^2\Gamma(F/2)} & x > 0 \\ 0 & 0 \leq x \end{cases}$$

where

$x = (E_T - E_b)/\sigma^2$,

E_b = bias,

F = number of basis functions,

$\Gamma(\cdot)$ = the gamma function.

The significant characteristic of E_T as far as the noise is concerned is that for a given bias E_b the probability density of E_T depends only upon σ^2 and F, and not on the specific functional form of the basis functions used. It is also independent of the number of sampled points. Furthermore, the ensemble average of E_T is

$$\bar{E}_F = \sigma^2 F + E_b.$$

Thus, the larger the number F of degrees of freedom, the larger will be the expected error.

If the structural forms of the basis functions are changed in order to make the bias E_b smaller, then the probability density $p(E_T)$ will be unaffected except for a shift of the function to lower values of E_T. This shift equals the difference between the original and the new value of E_b. This is true as long as the number F remains constant. Thus, for constant F it may be possible to reduce the errors in the noise-free function estimate by making functional changes in the basis functions. Doing this will not affect the statistics of the error due to the presence of noise. The effects of noise and E_b on the resultant error are thus statistically independent.

THE ERROR AT SPECIFIC TIMES

The squared error between the fit and the actual function *at any given time* (not necessarily at a sampled time) has a non-central χ^2 distribution with one degree of freedom. The noncentrality parameter is the squared error between the weighted sum of the F basis functions and the function to be fitted when no noise is present. The variance for the one degree of freedom is the mean squared error due to the effect of the noise.

It has been found that the variance at a specific time cannot be obtained without knowledge of the basis functions, and even then it cannot be obtained in closed form. However, it can be evaluated readily by numerical computer techniques. This has been done for the case of polynomial splines. The polynomials' first P-1 derivatives(s)*

*If we have P=0 then P-1 is -1. In this case neither the function nor its derivatives are continuous at the knots (i.e., independent polynomials are fit between adjacent knots).

are assumed continuous at N knots. The values of the polynomials at the knots at the beginning and end of the spline are not constrained. Each of the polynomials making up the spline is of degree D. The knots are not assumed to be uniformly spaced. Also, the times at which sampling takes place are not uniformly spaced, nor do they have to occur at the times at which the knots are placed. The number of degrees of freedom in this case is given by

$$F = P + (N-1)(D+1-P) .$$

Examples of mean squared error versus time have been obtained for this case using a computer program. Examples are shown in Figures 1 through 8. The examples are all for third-degree polynomial splines (D=3). Cases have been obtained using three knots and also six knots. Values of P used were from zero to three, which covers the range of continuities which can exist at the knots of a third-degree polynomial spline.

In all cases the M discrete times at which the function is sampled are uniformly spaced. The value of M used was large, as this is the situation of general interest. The total time of observation used for all plots was one unit of time. Plots of mean-squared error multiplied by (M/σ^2) versus time were made. For any total time of observation and any *large* M these plots can be used to obtain the mean-squared error versus time. This is done by multiplying the ordinate by the actual σ^2/M and the abscissa by the actual total time of observation.

CONCLUSION

The errors in spline fits to noisy data have been analyzed, and their probability distribution has been determined. Closed-form results

were obtained for the statistics of the squared error averaged over time. Numerical results for the statistics of the squared error as a function of time have been presented.

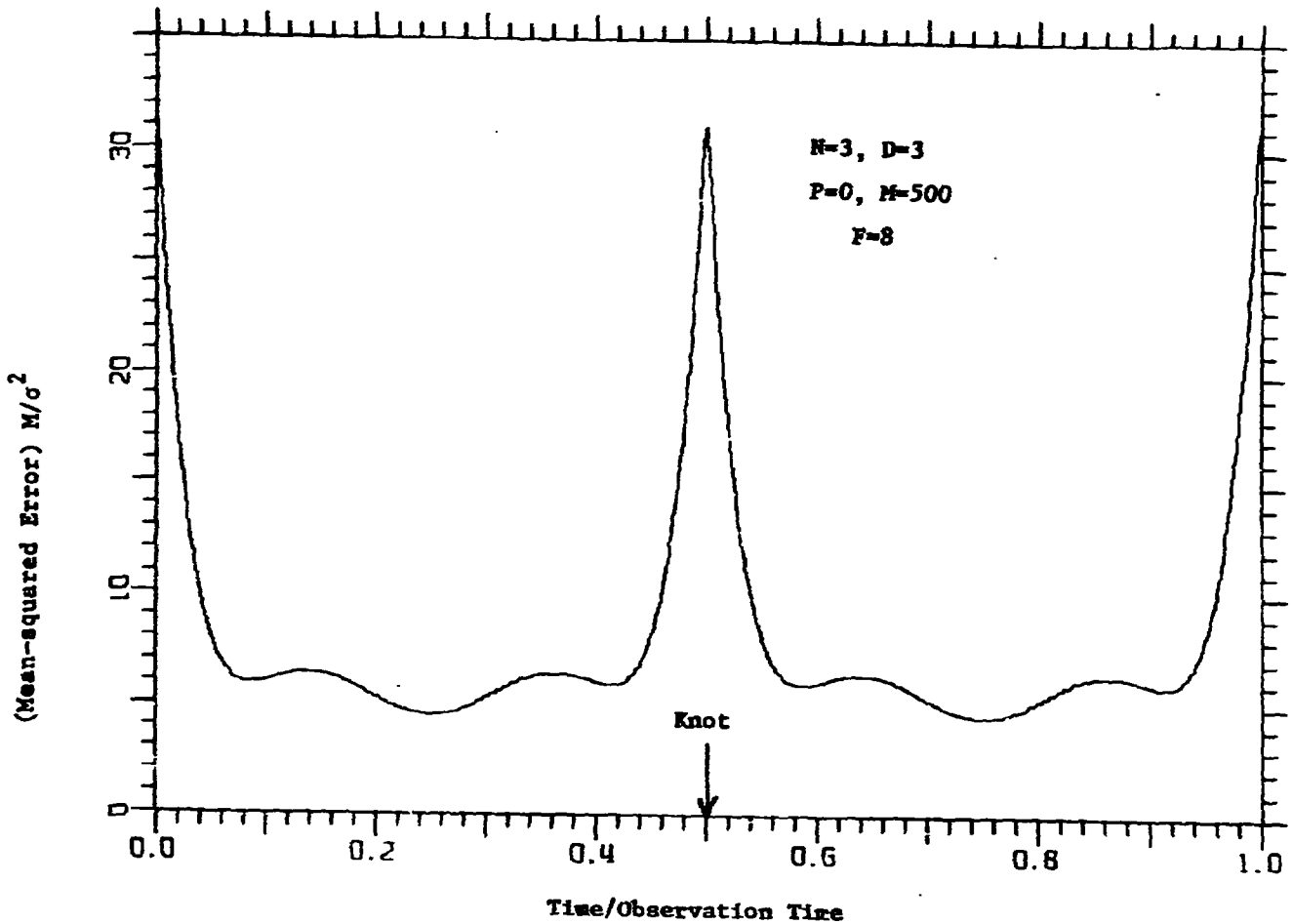


Figure 1. Mean-squared Error for Three Knots (Third-Degree Polynomials with no Constraints).

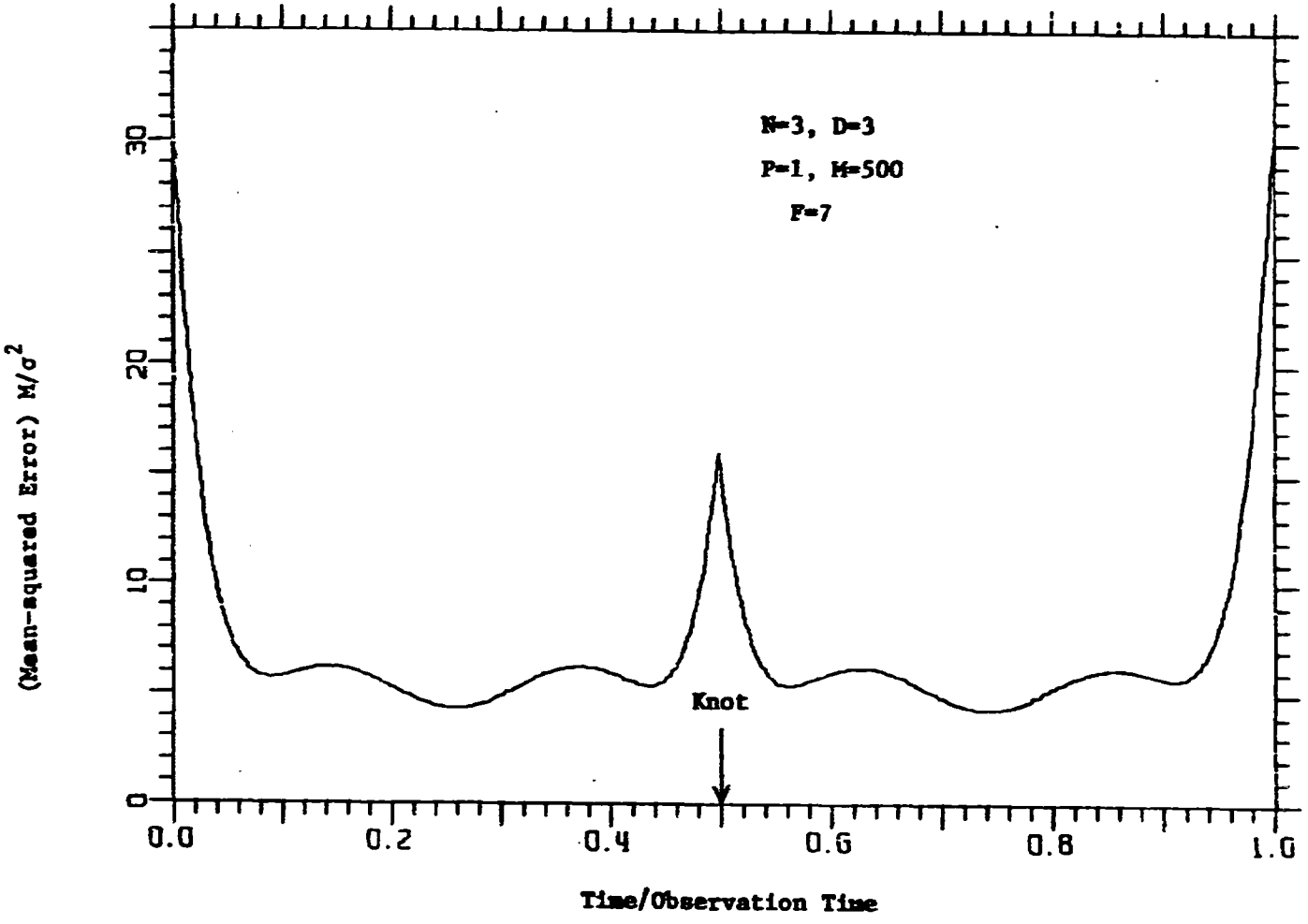


Figure 2. Mean-squared Error for Three Knots (Third-Degree Continuous Polynomials).

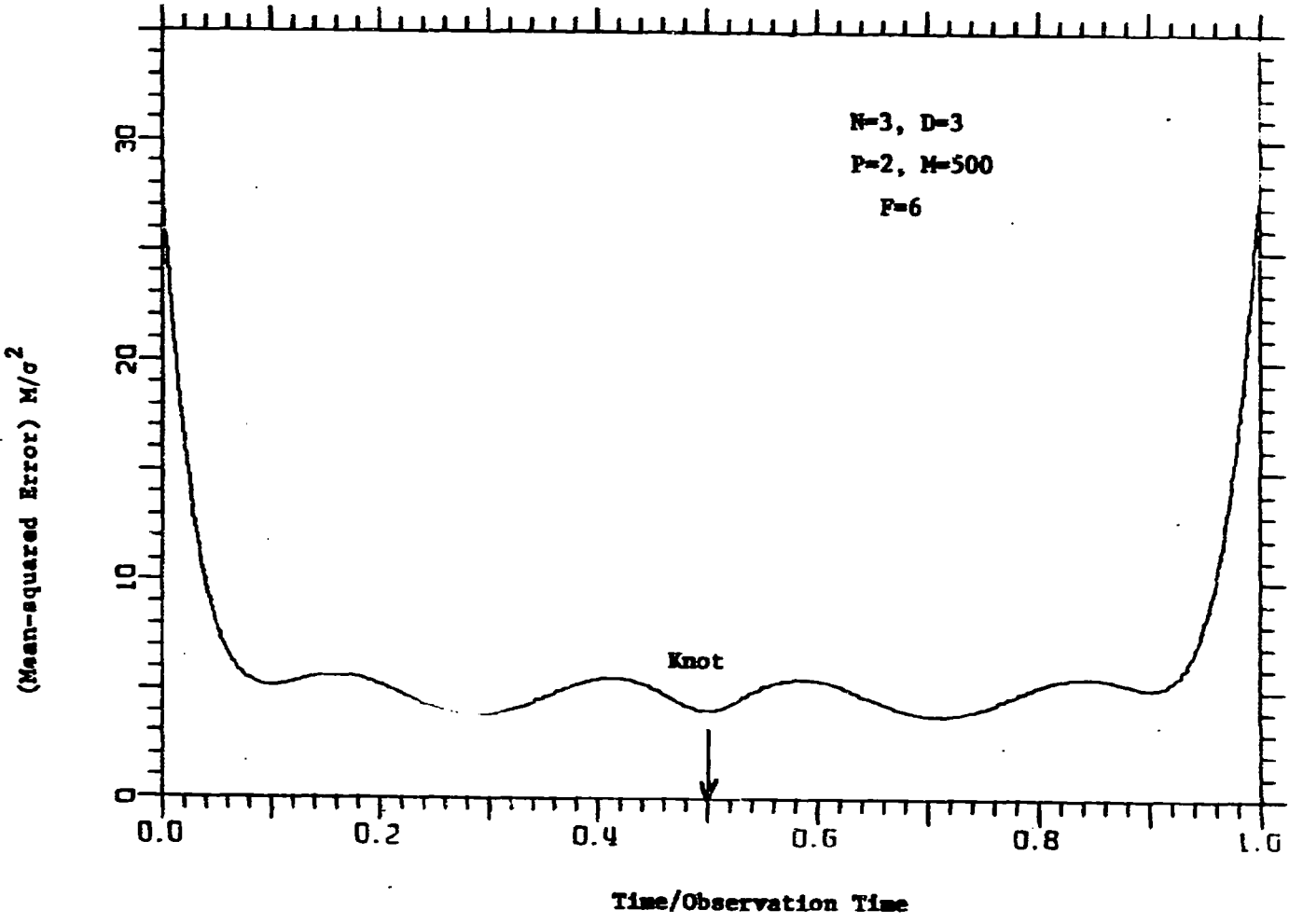


Figure 3. Mean-Squared Error for Three Knots (Third-Degree Polynomials Continuous to First Derivative).

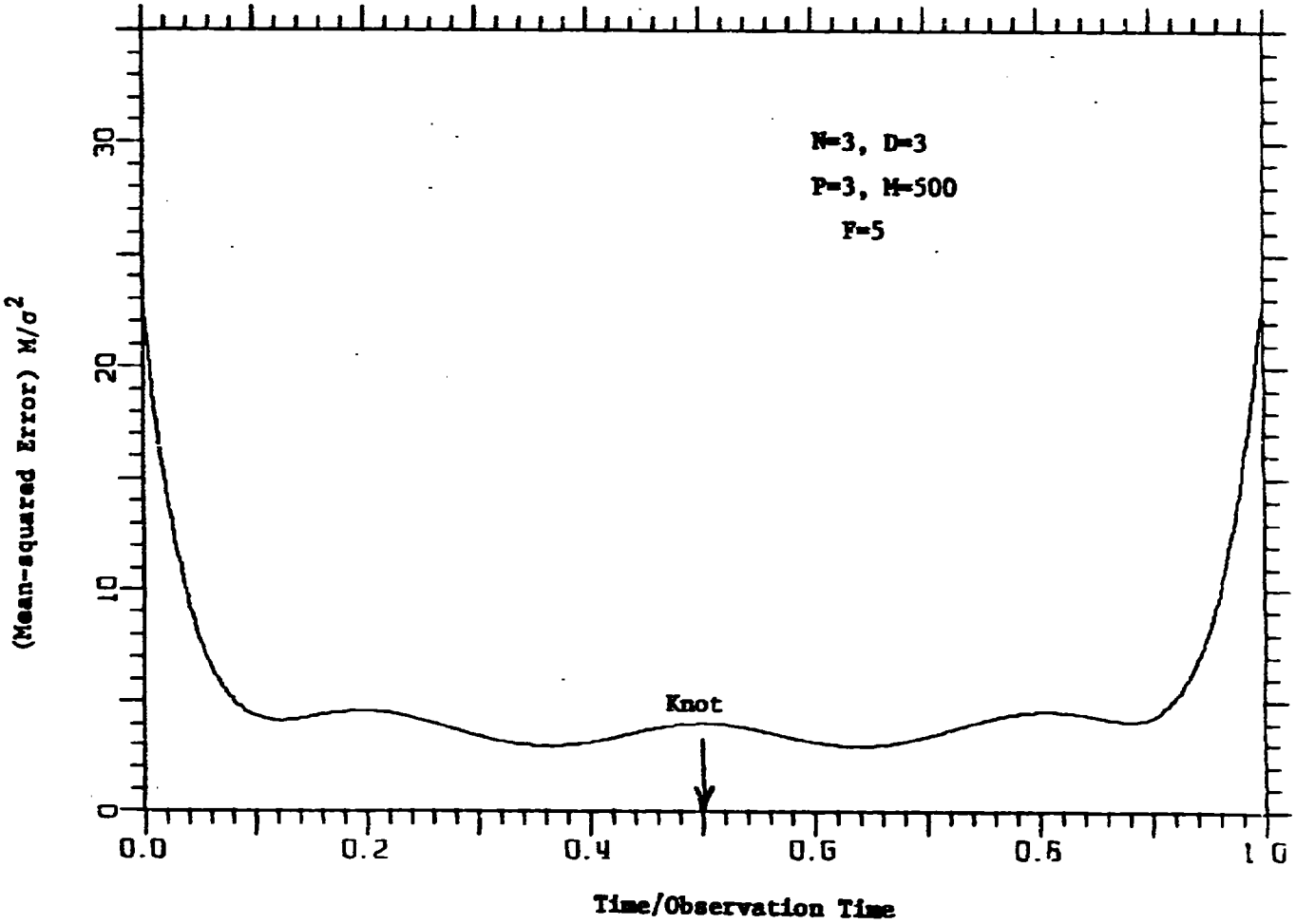


Figure 4. Mean-squared Error for Three Knots (Third-Degree Polynomials Continuous to Second Derivative).

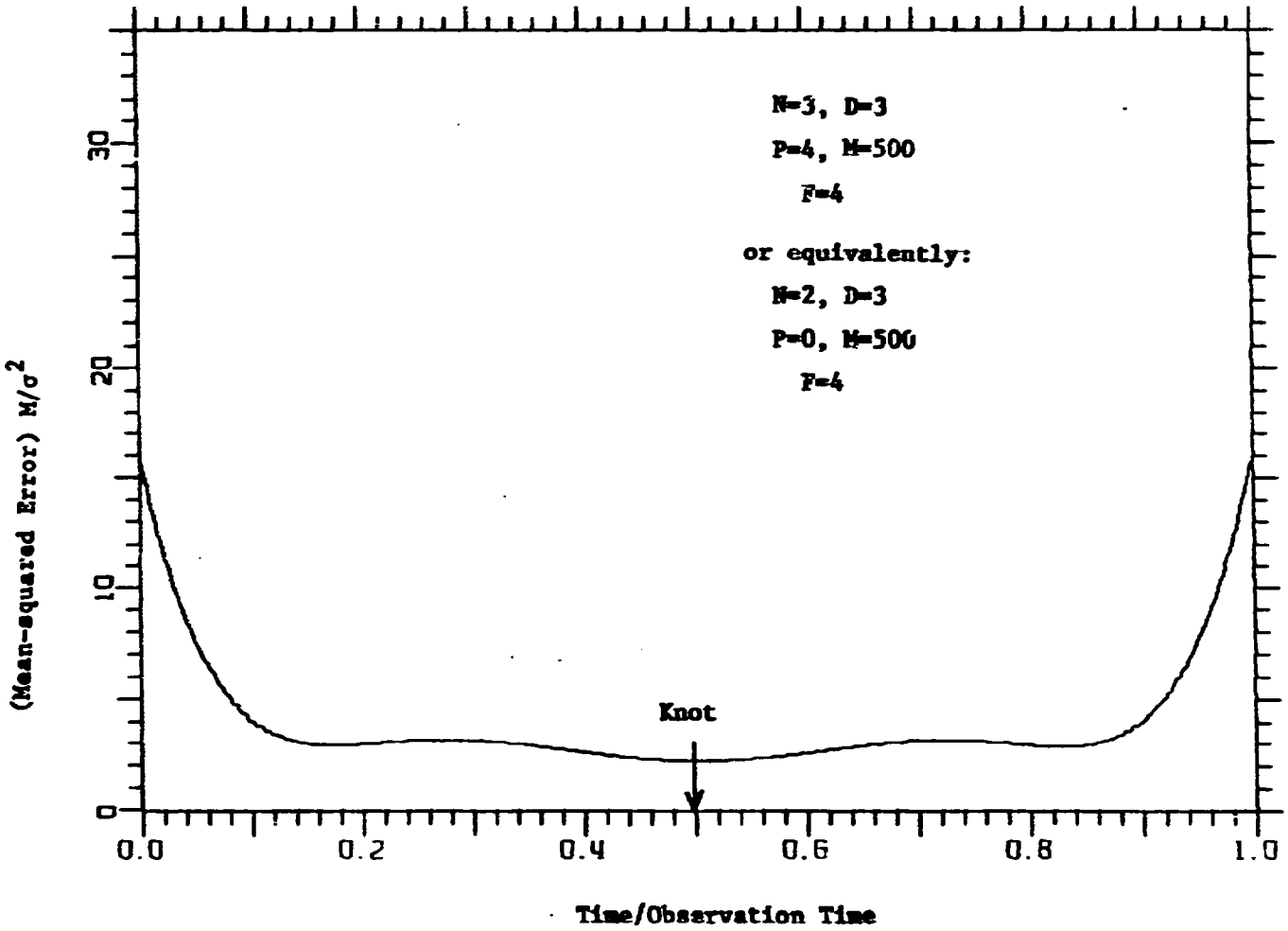


Figure 5. Mean-squared Error for Three Knots (Third-Degree Polynomials Continuous to Third Derivative).

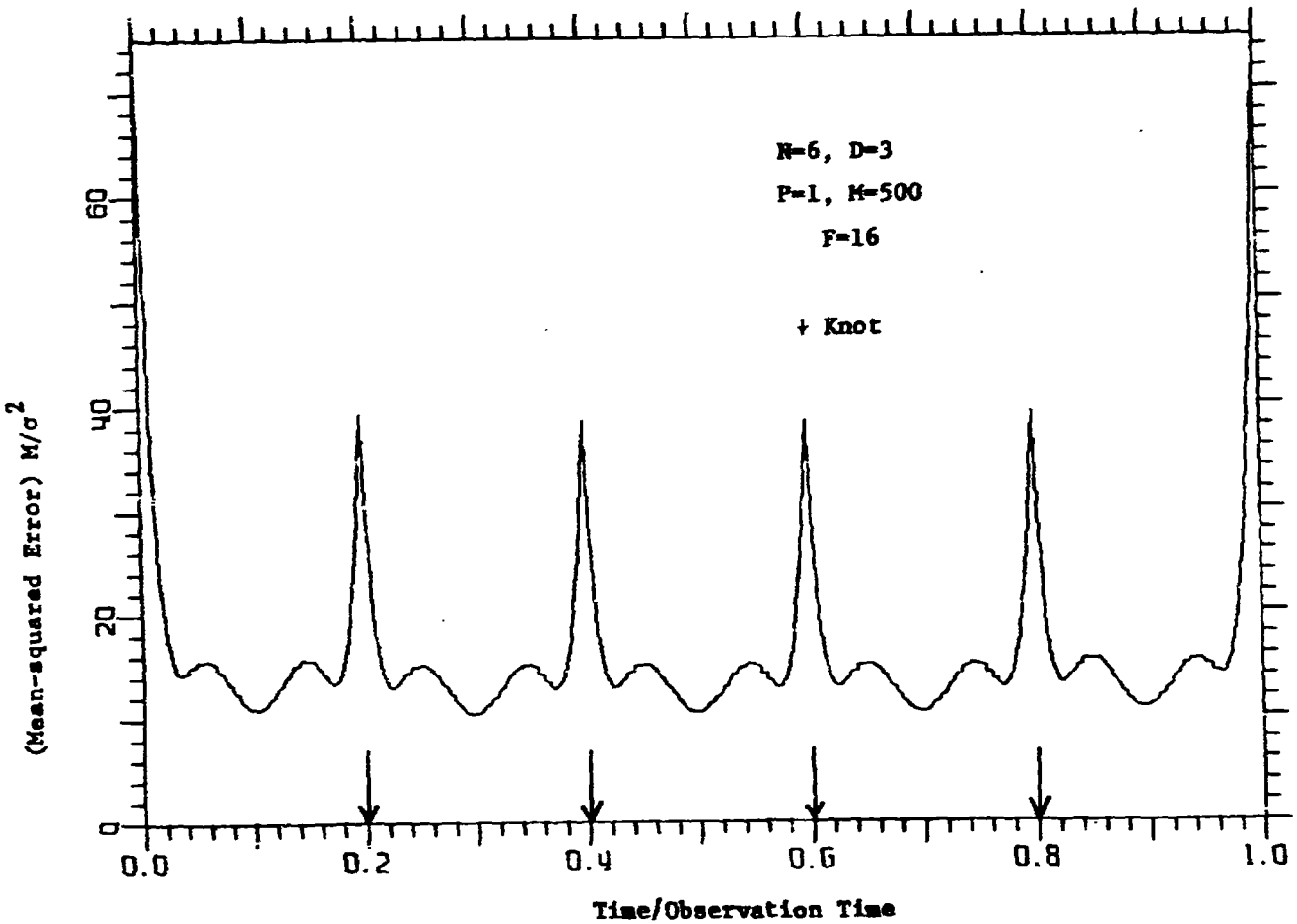


Figure 6. Mean-squared Error for Six Knots (Third-Degree Continuous Polynomials).

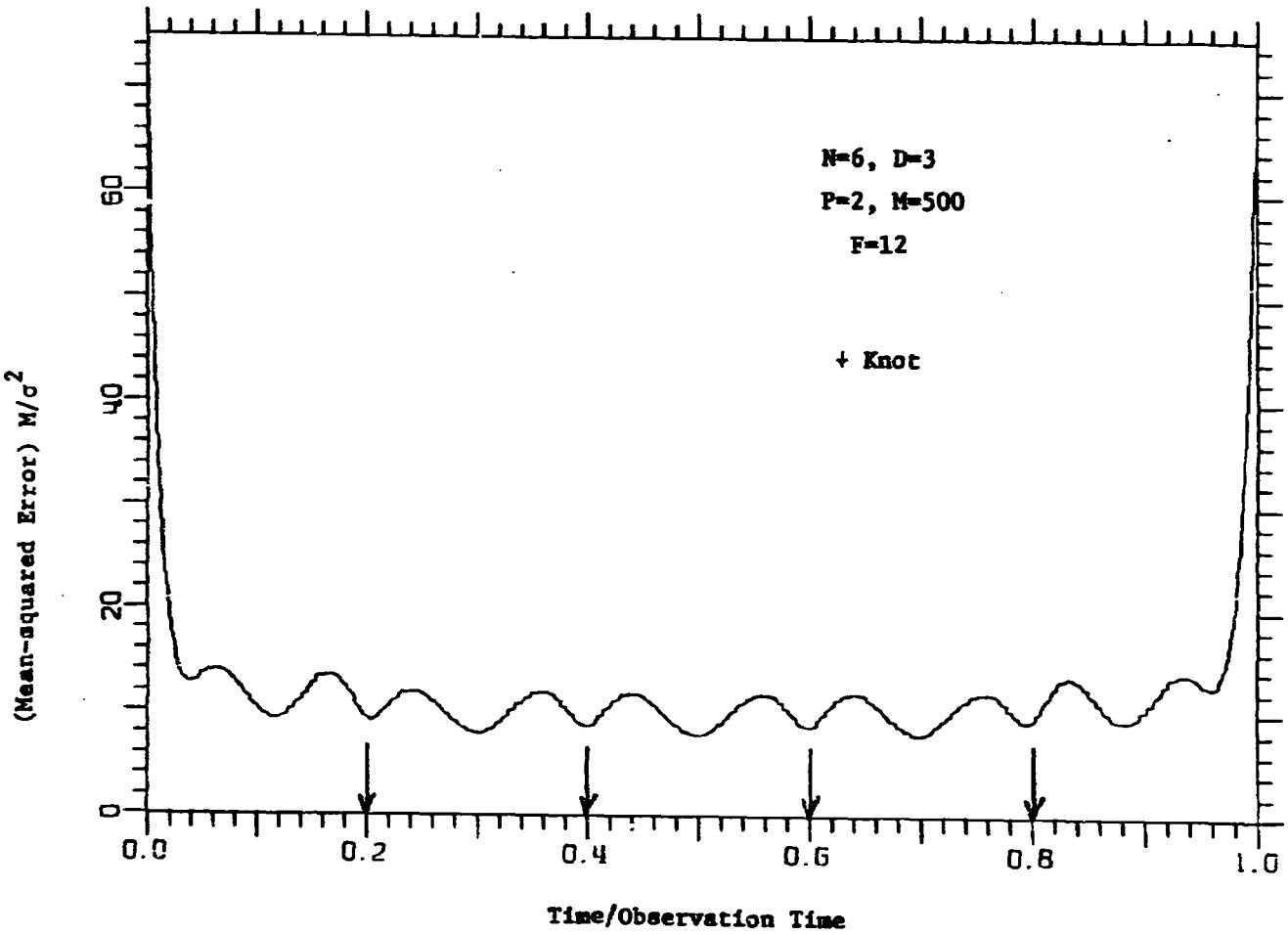


Figure 7. Mean-squared Error for Six Knots (Third-Degree Polynomials Continuous to First Derivative).

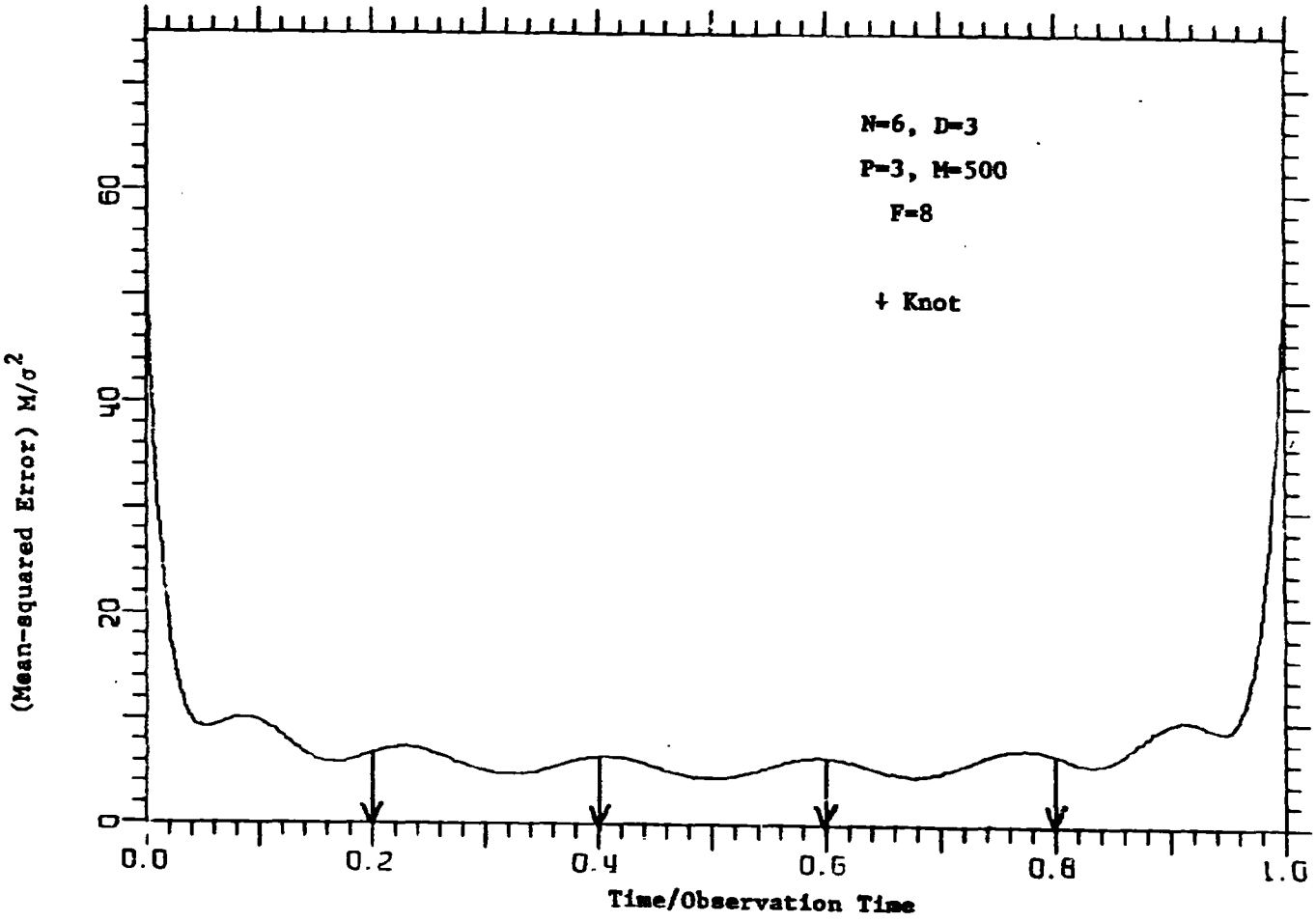


Figure 8. Mean-squared Error for Six Knots (Third-Degree Polynomials Continuous to Second Derivative).

AUTOREGRESSIVE MODELS OF AIRCRAFT MOTION AND AIR DEFENSE PREDICTION

Walter J. Dziwak*
ARRADCOM, Dover, New Jersey

ABSTRACT

One of the primary functions of a gun air defense system is to accurately predict the target position a time-of-flight into the future. If the future position is known accurately, then one can be reasonably assured of a hit on target provided that the remaining fire control errors as well as errors arising from uncertainties in ballistic and meteriological conditions are small.

The availability of aircraft flight data made it possible for the first time to analyze aircraft motion statistically and to build models of aircraft motion. These models are, of necessity, statistical because those components of aircraft motion induced by wind gusts, terrain features, and evasive maneuvers are generally unknown and must therefore be treated as random variables.

It was found that models of rate of change of target acceleration as autoregressive moving average processes lead to prediction schemes which enhanced the predictability of target future position, especially at extended ranges (long time-of-flight). Furthermore, these models were found to exhibit a remarkable degree of robustness; a lack of sensitivity due to changes in the coefficients of the autoregressive models as well as to changes in aircraft maneuvers seems to be an inherent feature of these models.

Other variables, chosen to be more explicitly tied to the dynamics of aircraft motion and less dependent on the choice of coordinates, were also modeled as autoregressive processes. Again, the results were encouraging, indicating that significant improvements in predictive capability inhere in the autoregressive models.

*Much of the work done on this project was contributed by Max Mintz, Steve Heuling, Stan Goodman.

I. INTRODUCTION

The most common prediction schemes in use for many years in the air defense community were the so-called linear and quadratic prediction equations

$$\vec{x}(t + t_F) = \vec{x}(t) + \dot{\vec{x}}(t)t_F$$

$$\vec{x}(t + t_F) = \vec{x}(t) + \dot{\vec{x}}(t)t_F + \tfrac{1}{2}\ddot{\vec{x}}(t)t_F^2$$

referenced to some inertial coordinate system.* The realization that these equations fare poorly against highly maneuvering targets led to the development of numerous predictors, including polynomial types after Blackman,(1) constant energy and defense of a known point after H. Weiss(2) as well as the variaty derivable from the Weiner as well as the more general Kalman-Bucey filter equations. Unfortunately, the performance of these predictors against real targets remained largely unknown. With the availability of attack aircraft data in 1974, however, their relative predictive capabilities could be determined.(3) The results led to the conclusion that no one predictor is best for all classes of attack maneuvers for a particular aircraft. Furthermore, it was found that the single largest contributor to the prediction error lay not in the availability of accurate knowledge of target state, but rather in the unpredictable pilot induced maneuvers.

Rather than try to formulate a new set of deterministic equations as in (1) and (2), one is thus led to consider statistical models of target motion. Although there is no a-priori reason for believing that autoregressive models will lead to better predictors, their consideration appears reasonable in view of the exhaustive efforts already directed to alternative schemes.

*For the short times of flight involved (1-4 sec), the rotation of the earth can be neglected. Thus, a coordinate frame fixed to a stationary weapon system can be viewed as an inertial reference frame.

II. AR MODELS

Let x_i be some generalized coordinate. The variable x takes on the value x_i at time $(i-1)\Delta$ where Δ is some time increment. Autoregressive models, then, are governed by the following assumptions:

(1) $x_n = \beta_1 x_{n-1} + \beta_2 x_{n-2} + \ldots + \beta_p x_{n-p} + u_n$

$-\infty < n < \infty$ (i.e, $\forall n$)

(2) $E[u_n] = 0, \quad \forall n$

$E[u_n u_m] = \alpha^2 \delta_{mn}, \quad \forall m,n$

where $E[\cdot]$ indicates an ensemble average over $\cdot$.

If $\beta_p \neq 0$, then the model is an autoregressive (AR) model of order p.

Since $\{u_n\}$ satisfies (2) for all n, and x_n is given by (1), u_n is uncorrelated with x_{n-1}, x_{n-1}, x_{n-2}, ... and $E[x_m] = 0$ for all m.

If

(3) $r(k) \equiv E[x_{n+k} x_n]$

then one can determine the β_i's in terms of the covariance functions r(k) as follows:

Multiply (1), successively by x_{n-1}, x_{n-2}, ... x_{n-p} to obtain p equations of the form.

(4) $x_n x_{n-j} = \beta_1 x_{n-1} x_{n-j} + \beta_2 x_{n-2} x_{n-j} + \cdots + \beta_p x_{n-p} x_{n-j} + u_n x_{n-j}$

Taking the expectation value of both sides of (4), one obtains p equations in p unknowns. The r(k)'s are assumed known. Defining r_p and $\underline{R}$ by

$$r_p \equiv \begin{pmatrix} r(1) \\ \vdots \\ r(p) \end{pmatrix} \quad ; \quad R \equiv \begin{pmatrix} r(0) & r(1) & \ldots & r(p-1) \\ r(1) & r(0) & \ldots & r(p-2) \\ \vdots & & & \vdots \\ r(p-1) & \ldots & & r(0) \end{pmatrix}$$

the p equations can be expressed more concisely by

(5) $\underline{R}\,\beta = r_p$

or

(6) $\beta = \underline{R}^{-1} r_p$

where β is the column matrix $(\beta_1 ; \beta_2 , \cdots , \beta_p)^T$

III. ESTIMATION OF β FROM ACTUAL TIME SERIES DATA

The determination of β is dependent upon one's ability to compute r(k) from an ensemble of time series data. In practice, however, such data is often unavailable for obvious reasons; time, money, and resources often do not permit the accumulation of such data. This is especially true in the present discussion where replication of aircraft flight paths becomes both time consuming and costly. One is thus lead to consider replacement of ensemble averages with averages over time. Thus, in place of (3), one estimates r(k) with

$$\hat{r} = \frac{1}{N} \sum_{i=1}^{N-k} x_{i+k} x_i \quad .$$

The matrix β is then formed by replacing r(k) with $\hat{r}(k)$ in the relation

$$\beta = \underline{R}^{-1} r_p \quad .$$

IV. MODEL IDENTIFICATION

The determination of the "proper" choice of p is an important practical question. The following result[4] is often useful:

If for a given choice of p the estimated value of β_p derived from the sample covariance function satisfies $|\beta_p| \ll 1/\sqrt{N}$, then one can assume that $\beta_p = 0$ and hence check the model with order p-1. However, in order for this result to hold, one must strengthen the assumptions with respect to $[u_n]$. Specifically, one assumes that the $[u_n]$ are independent and identically distributed random variables.

Fortunately, the autoregressive models considered here turned out to be of order no higher than six.

V. OTHER MODELS

Models other than autoregressive models were also considered. These included the moving average and autoregressive moving average models. These are characterized, respectively by

$$(7) \quad x_n = u_n + \alpha_1 u_{n-1} + \cdots + \alpha_q u_{n-q}$$

with

$$E[u_n] = 0$$

$$E[u_n u_m] = \sigma^2 \delta_{mn}$$

and by

$$(8) \quad x_n = \beta_1 x_{n-1} + \cdots + \beta_p x_{n-p} + u_n + \alpha_1 u_{n-1} + \cdots + \alpha_q u_{n-q}$$

with

$$E[u_n] = 0$$

$$E[u_n u_m] = \sigma^2 \delta_{mn} .$$

Equation (7) is a qth order moving average process and (8) is a (p,q) autoregressive moving average process.

Analysis of the aircraft data indicates that aircraft motion is adequately modeled as an autoregressive process rather than either a moving average or autoregressive moving average processes.

VI. PREDICTION

For a process described by the autoregressive model

$$x_n = \beta_1 x_{n-1} + \beta_2 x_{n-2} + \ldots + \beta_p x_{n-p} + u_n$$

one wishes to estimate x_{n+k} from the known values x_{n-1}, x_{n-2}, $\ldots$, x_{n-p}. This is accomplished by first noting the following:

$$(9) \quad E[x_n / x_{n-1}, x_{n-2}, \ldots, x_{n-p}] = \beta_1 x_{n-1} + \beta_2 x_{n-2} + \ldots + \beta_p x_{n-p}$$

$$(10) \quad E[x_{n+k} / x_{n-1}, x_{n-2}, \ldots, x_{n-p}] = \eta_1(k)\, x_{n-1} + \eta_2(k) x_{n-2} + \ldots + \eta_p(k) x_{n-p} \equiv \hat{x}_{n+k}$$

where $k \geq 0$.

The prediction procedure to be derived will be recursive. The resulting equations can be easily implemented and concisely expressed in matrix form. For this purpose, we relabel the generalized coordinates x_n as follows:

$$y_1(n) = x_n$$
$$y_2(n) = x_{n-1}$$
$$\vdots$$
$$y_p(n) = x_{n-p+1}$$

Thus,

$$(11) \quad \begin{bmatrix} y_1(n) \\ y_2(n) \\ \vdots \\ y_p(n) \end{bmatrix} = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} y_1(n-1) \\ y_2(n-1) \\ \vdots \\ y_p(n-1) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u_n$$

Define

(12) $$\underline{y}(n) = \begin{bmatrix} y_1(n) \\ y_2(n) \\ \vdots \\ y_p(n) \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad ; \quad \Gamma = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Then

(13) $$\underline{y}(n) = \Phi \underline{y}(n-1) + \Gamma u_n$$

Now

$$E[\underline{y}(n)/x_{n-1}, \ldots, x_{n-p}] = \begin{bmatrix} \hat{x}_n \\ x_{n-1} \\ \vdots \\ x_{n-p+1} \end{bmatrix}$$

and

$$E[\underline{y}(n-1)/x_{n-1}, \ldots, x_{n-p}] = \begin{bmatrix} x_{n-1} \\ \vdots \\ x_{n-p} \end{bmatrix}$$

so if

$$\hat{\underline{y}}(n-1) = E[\underline{y}(n-1)/x_{n-1}, \ldots, x_{n-p}]$$

and

$$\hat{\underline{y}}(n) = E[\underline{y}(n)/x_{n-1}, \ldots, x_{n-p}]$$

then

(14) $$\hat{\underline{y}}(n) = \Phi \hat{\underline{y}}(n-1)$$

In general, if

$$\hat{\underline{y}}(n+k) = E[\underline{y}(n+k)/x_{n-1}, \ldots, x_{n-p}]$$

then

(15) $$\hat{\underline{y}}(n+k) = \Phi^{k+1} \hat{\underline{y}}(n-1)$$

which is the scheme by which prediction is accomplished for a pth order autoregressive process.

VII. PREDICTION USING AIRCRAFT FLIGHT DATA

As seen in Section II, u_n is characterized as a white noise sequence. That is,

$$E[u_n] = 0$$

$$E[u_n u_m] = \sigma^2 \delta_{nm}$$

Analysis of the aircraft data indicated that if $\dddot{x}$ is the rate of change of target acceleration in some inertial coordinate frame, then it satisfies approximately the statistical assumptions of AR models. One can then predict future position with the aid of the following assumptions:

$$(16) \quad x_n = x_{n-1} + \Delta \dot{x}_{n-1} + \Delta^2/2 \ddot{x}_{n-1} + \Delta^3/6 \dddot{x}_{n-1}$$

$$(17) \quad \dot{x}_n = \dot{x}_{n-1} + \Delta \ddot{x}_{n-1} + \Delta^2/2 \dddot{x}_{n-1}$$

$$(18) \quad \ddot{x}_n = \ddot{x}_{n-1} + \Delta \dddot{x}_{n-1}$$

with $\ddot{x}$ modeled as an autoregressive process:

$$(19) \quad \dddot{x}_n = \beta_1 \ddot{x}_{n-1} + \beta_2 \ddot{x}_{n-2} + \ldots + \beta_p \ddot{x}_{n-p} + u_n$$

Proceeding as in Section VI,

$$\begin{bmatrix} x_n \\ \dot{x}_n \\ \ddot{x}_n \\ \dddot{x}_n \\ x_{n-1} \\ \vdots \\ x_{n-p+1} \end{bmatrix} = \left[\begin{array}{ccc|ccccc} 1 & \Delta & \Delta^2/2 & \Delta^3/6 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \Delta & \Delta^2/2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \Delta & 0 & \cdots & 0 & 0 \\ \hline 0 & 0 & 0 & \beta_1 & \beta_2 & \cdots & \beta_{p-1} & \beta_p \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & & 1 & 0 \end{array}\right] \begin{bmatrix} x_{n-1} \\ \dot{x}_{n-1} \\ \ddot{x}_{n-1} \\ \dddot{x}_{n-1} \\ \vdots \\ \dddot{x}_{n-p} \end{bmatrix}$$

or

$$\hat{\bar{x}}_n = \left[\begin{array}{c|c} \underline{A} & \underline{B} \\ \hline \underline{0} & \phi \end{array}\right] \bar{x}_{n-1}$$

Thus,

$$(20) \quad \hat{\bar{x}}_{n+k} = \left[\begin{array}{c|c} \underline{A} & \underline{B} \\ \hline \underline{0} & \phi \end{array}\right]^{k+1} \bar{x}_{n-1}$$

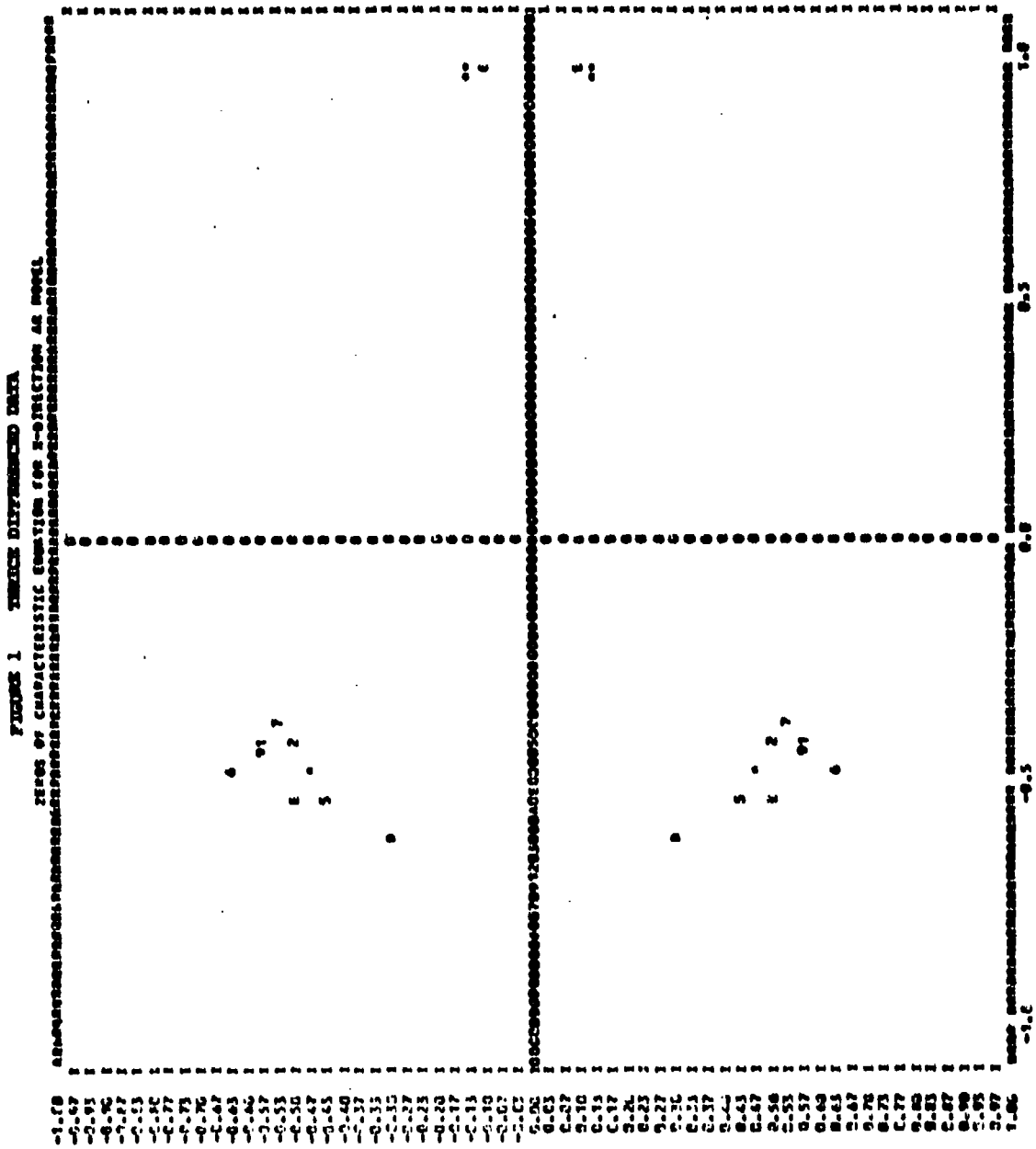
which is the analogue of equation (15).

The reason for partitioning the transition matrix as above is that this results in substantial savings in computational labor.

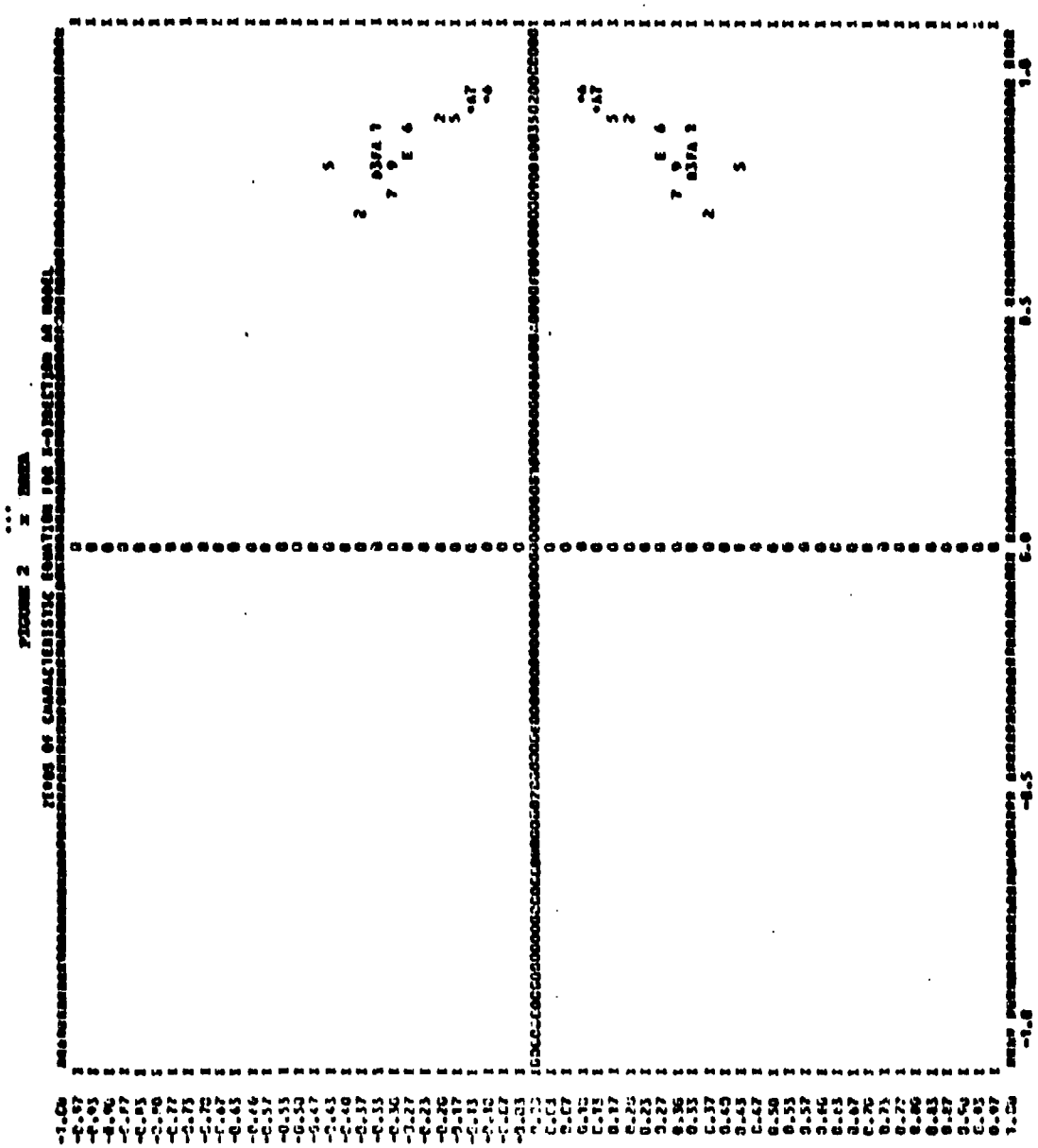
From the raw aircraft data that was available, two independent sets of $\ddot{x}$ data were produced. One set was derived from smoothed accelerometer data from on board the aircraft, and the other by thrice differencing smoothed position data. Prediction for the second set, i.e., the thrice differenced data, is accomplished by computing $(x_{n+3}-3x_{n+2}+3x_{n+1}-x_n)$ then proceeding as in (20).

The autoregressive coefficients for the two classes of predictors will be different. This is exhibited in Figures 1 and 2 which show the distribution of the roots of the characteristic equations for the x-axis. Here, each symbol is associated with a separate and distinct flight path. It is interesting to note that the groupings of the roots are quite different for the two models. (This is also true for the y and z axes which are not shown.) However, the roots show a marked similarity within a particular class. This suggests a commonality in the statistical description of the data, although the full import of this feature was not investigated.

Comparing the performance of the two classes of models, it was found that the predictors developed from the thrice differenced data do not perform as well as the $\ddot{x}$ predictors, but the differences are not substantive for short T_f.



294



VIII. VELOCITY-GAMMA-SIGMA AR MODEL

Here we define a new set of dynamic variables $(\bar{v},\gamma,\sigma)$ where $\bar{v}$ is the aircraft velocity, γ the angle between $\bar{v}$ and the horizontal plane, and σ the angle between the projection of $\bar{v}$ onto the horizontal plane and y-axis.* The mean values of the generalized coordinates are not zero as in the previous cases, so they are removed adaptively from a one second time window. Thus, the AR models are

$$(21)\quad \hat{v}_n = \bar{v}_n + \beta_1(v_{n-1} - \bar{v}_{n-1}) + \cdots + \beta_5(v_{n-5} - \bar{v}_{n-5})$$

where $\bar{v}_n$ is the adaptive mean

$$(22)\quad \bar{v}_n = \frac{1}{N}\sum_{i=1}^{N} v_{n-i}\ , \quad N = 10$$

the data rate being 10 data points/sec. The variable $\bar{v}$, it turns out, is adequately modeled as a fifth order AR process.

Similarly,

$$(23)\quad \hat{\gamma}_n = \bar{\gamma}_n + \beta_1(\gamma_{n-1} - \bar{\gamma}_{n-1}) + \cdots + \beta_6(\gamma_{n-6} - \bar{\gamma}_{n-6})\ .$$

Sigma, however, is modeled as a first-differenced AR process thereby reducing the dependence on the orientation of the x-y coordinate axes.

$$(24)\quad \hat{\sigma}_n = \sigma_{n-1} + \Delta\sigma_n + \beta_1(\Delta\sigma_{n-1} - \Delta\bar{\sigma}_{n-1}) + \cdots + \beta_5(\Delta\sigma_{n-5} - \Delta\bar{\sigma}_{n-5})$$

where

$$(25)\quad \Delta\sigma_n = \sigma_{n-1} - \sigma_{n-2}$$

and

$$(26)\quad \Delta\bar{\sigma}_n = \frac{1}{N}\sum_{i=1}^{N} \Delta\sigma_{n-i}\ , \quad N = 10\ .$$

The relation between the target position and the dynamic variables $(\bar{v},\gamma,\sigma)$ is non-linear so prediction must proceed recursively via the following equations:

$$(27)\quad \hat{x}_n = x_{n-1} + \Delta(v_{n-1}\cos\gamma_{n-1}\sin\sigma_{n-1})$$

$$(28)\quad \hat{y}_n = y_{n-1} + \Delta(v_{n-1}\cos\gamma_{n-1}\cos\sigma_{n-1})$$

$$(29)\quad \hat{z}_n = z_{n-1} + \Delta(v_{n-1}\sin\gamma_{n-1})\ .$$

* An inertial x,y,z set is used as before.

IX. RESULTS

The predictors developed in Sections VII and VIII were compared with some of the common predictors that have been in use over the years. The comparisons were made by generating a histogram of the number of shells fired from a hypothetical gun air defense system which fall within specified bins or regions of the target after some time-of-flight T_f. No attempt was made to model error sources in the fire control system or to generate realistic ballistic trajectories. The shells were assumed to be free of the earth's gravitational field and meteriological effects. The projectile velocity was taken at 1000 m/sec. For the purpose of comparing predictors, the added complexity of introducing fire control errors and accurately modeling ballistic trajectories seems unwarranted and does not shed light on the relative efficacy of the predictors under comparison.

Comparison of the A-R predictors was made with the following standard models:

Linear

$$\hat{x}(t + T_f) = x(t) + \dot{x}(t)T_f$$

Quadratic

$$\hat{x}(t + T_f) = x(t) + \dot{x}(t)T_f + \ddot{x}(t)T_f^2/2$$

First Order Markovian in Acceleration*

$$\hat{x}(t + T_f) = x(t) + \dot{x}(t)T_f + \ddot{x}(t)\cdot\frac{e^{-\omega T_f} + \omega T_f - 1}{\omega^2}$$

with $\omega = .1$

Previous studies[3] comparing a larger class of predictors of which the three above are a subset were made with the conclusion that no single predictor is best over the range of flight paths considered here. Thus, inclusion of this larger class is unnecessary since nothing new will be learned that is not already known.

---

* The model for this process is $\dddot{x} = -\omega\ddot{x} + u$ from which the above equation is derived.

Tables 1 and 2 typify the type of data that was obtained in comparing A-R predictors with the three above. The following explanations of these tables are in order. Minimum miss distance is the distance of closest approach between the target and projectile. Regular one point misses refer to the miss after time T_f. The bottom row of numbers designates a distance between target and projectile. Column 5 differs from all other columns in that it lists the total number of projectiles falling within 5 m of the target. The remaining columns designate the number of projectiles falling within a bin of certain width. For example, column 3 gives the number of projectiles falling with 2 to 3 meters of the target, column 7 the number of projectiles falling within 10 to 15 meters of the target, etc. The last column gives the total number of rounds fired in a given time interval (T_f).

As is evident from the figures, the A-R predictor performs better than the quadratic predictor. Of particular interest, however, is the region where $T_f > 3$ sec., where predictors have traditionally fared poorly. Here, we see that with the A-R predictors, some rounds fall in close proximity of the target (ie, within 15m); in contrast, no rounds fall in the region with the quadratic predictor.

These observations hold in general. That is, they can be made for the entire class of flight paths investigated (12 in number), as well as for the linear and Markovian predictors. Furthermore, the A-R thrice-differenced predictors, as well as the $\dot{V}$-γ-σ models, also perform markedly better than either of the standard predictors.

Table 3 gives the performance of the A-R thrice-differenced predictor for flight pass 13 (same as for Table 2) and Table 4 the performance of the $\dot{V}$-γ-σ predictors, also for the same flight pass. Observe that the thrice-differenced predictor does not perform quite as well as the '$\dot{x}$' predictor, an observation alluded to in Section VII. In addition, the $\dot{V}$-γ-σ predictors do not fare as well as the '$\dot{x}$' predictors.

TABLE 1

MINIMUM MISS DISTANCE INFORMATION FOR QUADRATIC PREDICTION MODEL

FLIGHT 5 PASS 13

FIRST DATA POINT IS 13.6 LAST DATA POINT IS 40.2

	RMSE	RE
MIN DISTANCE	127.30033	93.17196
REG 1-POINT	153.10290	112.47088

HISTOGRAM OF MIN DISTANCE MISSES

TOF BETWEEN	0.	1.	2.	3.	4.	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	TOTAL
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	1	2	2	*	9	23	36	4	0	0	0	0	0	0	0	0	0	0	0	72
2.0 AND 3.0 SEC	0	0	0	0	*	1	3	7	2	2	4	27	6	0	0	0	0	0	0	0	52
3.0 AND 4.0 SEC	0	0	0	0	*	0	0	0	0	0	0	5	10	11	20	5	0	0	0	0	51
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	1	30	27	15	11	5	2	91

HISTOGRAM OF REGULAR 1-POINT MISSES

TOF BETWEEN	0.	1.	2.	3.	4.	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	TOTAL
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	1	2	2	*	8	22	35	7	0	0	0	0	0	0	0	0	0	0	0	72
2.0 AND 3.0 SEC	0	0	0	0	*	1	3	7	2	2	4	27	6	0	0	0	0	0	0	0	52
3.0 AND 4.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	1	20	24	6	0	0	0	0	51
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	6	20	32	17	13	1	91

TABLE 2

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE XYZ-TRIPLE-BOT PREDICTION MODEL
FLIGHT 5 PASS 13 MODELS ASSUMED TO HAVE 0.0 MEANS.

FIRST DATA POINT = 13.6 LAST DATA POINT = 42.2

	PENALTY	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
CLOSEST APPROACH:	0.38547	50.0558	32.2069	28.0951	-0.0746	10.8067	-1.4532	39.9933	-11.2875
PREDICTED MISS:	0.37702	63.4652	40.4081	28.3547	-0.0749	34.1259	9.4177	43.3291	-12.1809

HISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	TOTAL
0.0 AND 1.0 SEC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	64	14	3	5	86	6	0	0	0	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	0	5	5	12	13	10	6	7	4	1	0	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	1	2	6	4	5	5	16	10	9	0	0	0	0	0	0	58
4.0 AND 5.0 SEC	0	0	0	0	0	1	0	0	0	0	8	18	20	16	3	0	0	0	0	66
	1.	2.	3.	4.	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	269 TOTAL

HISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	TOTAL
0.0 AND 1.0 SEC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	62	16	3	5	86	6	0	0	0	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	0	3	6	12	13	10	6	7	3	2	0	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	0	3	4	3	0	2	13	23	10	0	0	0	0	0	0	58
4.0 AND 5.0 SEC	0	0	0	0	0	0	0	0	1	0	0	10	23	22	9	1	0	0	0	66
	1.	2.	3.	4.	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	269 TOTAL

CLOSEST APPROACH ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0 AND 2.0 SEC	1.9344	1.1599	0.6602	0.1124	0.9250	-0.0270	1.5653	0.0598
2.0 AND 3.0 SEC	15.4135	12.8256	14.2106	-1.2098	2.4140	-0.5248	5.4595	-1.0627
3.0 AND 4.0 SEC	46.5434	40.4213	21.4314	5.3143	16.0539	6.4062	38.0691	15.9193
4.0 AND 5.0 SEC	90.0691	83.8295	51.4853	-4.1594	15.6086	-11.0937	72.2363	-59.2246

PREDICTED MISS ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0 AND 2.0 SEC	1.9768	1.1936	0.6394	0.1018	0.7028	-0.0635	1.7336	0.0944
2.0 AND 3.0 SEC	15.5355	12.9850	14.3231	-1.2075	2.4080	-0.5028	5.5140	-1.0684
3.0 AND 4.0 SEC	54.8483	50.1546	20.2520	3.7578	30.5885	-10.6077	40.7742	22.9605
4.0 AND 5.0 SEC	116.5054	109.3422	52.4511	-2.7798	62.6016	48.2261	83.0567	-69.0978

TABLE 3

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE XYZ-THRICED-DIFFERENCED PREDICTION MODEL
FLIGHT 5 PASS 13 MODELS ASSUMED TO HAVE 0.0 MEANS.

FIRST DATA POINT = 13.6 LAST DATA POINT = 42.2

	PENALTY	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
CLOSEST APPROACH:	0.19159	102.4251	70.4256	42.0393	5.5274	22.1368	-6.1926	90.7389	-35.1693
PREDICTED MISS:	0.17973	124.2098	84.6701	42.3372	5.6579	56.9777	12.1446	101.9273	-39.0502

HISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	10	15	10	*	50	23	12	6	1	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	0	1	1	*	2	3	3	9	6	7	8	15	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	*	0	0	0	0	2	3	13	9	13	13	0	0	0	0	0	53
4.0 AND 5.0 SEC	0	0	0	0	*	0	1	0	0	0	0	3	3	4	27	21	18	6	1	0	84
TOTAL	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	282

HISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	6	12	9	*	46	26	13	6	1	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	0	0	1	*	2	3	1	9	8	7	8	15	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	*	0	0	0	0	1	2	8	8	16	18	0	0	0	0	0	53
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	0	0	1	3	5	10	17	26	17	4	1	84
TOTAL	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	282

CLOSEST APPROACH ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0 AND 2.0 SEC	7.8278	6.2827	4.4791	-0.0642	3.3173	-0.5555	5.4962	-0.7989
2.0 AND 3.0 SEC	36.2803	33.3908	30.8373	-4.5901	9.8978	4.2049	20.4079	9.2006
3.0 AND 4.0 SEC	79.2568	72.2961	33.5244	14.6565	28.4377	6.4852	65.9473	-19.9962
4.0 AND 5.0 SEC	173.9665	162.8643	67.8380	12.2751	32.5733	-26.9261	156.8481	-135.6152

PREDICTED MISS ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0 AND 2.0 SEC	8.1468	6.6834	4.4020	-0.1275	2.5829	-0.6235	6.3674	-1.0273
2.0 AND 3.0 SEC	38.7835	33.9535	31.0993	-4.8623	6.4568	1.6792	22.2554	10.3691
3.0 AND 4.0 SEC	82.3828	83.0291	31.4515	11.5825	46.0143	-22.2144	68.5931	32.4054
4.0 AND 5.0 SEC	214.1117	203.1194	69.0063	14.8901	97.6180	54.4109	177.6308	-156.9607

TABLE 4

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE VELOCITY-GAMMA-SIGMA PREDICTION MODEL
FLIGHT 5 PASS 13 V-G-S MODELS HAVE ADAPTIVE MEANS FROM 1.0 SEC WINDOW

FIRST DATA POINT = 14.2 LAST DATA POINT = 42.2

	PENALTY	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
CLOSEST APPROACH:	0.29879	53.2170	35.8898	25.1588	3.0531	15.4385	-5.3467	44.2802	-16.1812
PREDICTED MISS:	0.29004	66.8547	43.2319	26.0133	3.9300	29.0505	10.6593	54.3041	-21.3427

HISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	8	11	10	22	*	76	13	3	0	0	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	1	1	4	*	8	14	7	5	6	2	11	0	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	*	0	1	1	5	6	6	16	8	0	0	0	0	0	0	0	52
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	2	0	9	22	23	12	6	0	0	0	0	74
																					271
																					TOTAL

HISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	3	14	18	22	*	73	16	3	0	0	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	0	2	0	*	8	13	7	3	7	4	11	0	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	*	0	1	1	4	5	5	19	8	2	7	0	0	0	0	0	52
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	1	1	6	9	24	18	13	2	0	0	0	74
																					271
																					TOTAL

CLOSEST APPROACH ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0 AND 2.0 SEC	4.3962	3.7077	3.5403	0.4804	1.5148	0.4288	2.1207	1.1859
2.0 AND 3.0 SEC	20.8261	17.0185	18.6178	-6.6524	4.9751	1.8734	7.8958	2.8934
3.0 AND 4.0 SEC	53.6345	45.5019	29.6574	17.4046	19.4508	-8.2946	40.2339	-10.1129
4.0 AND 5.0 SEC	89.5282	82.6616	37.8961	3.1181	24.2162	-17.0323	77.4129	-55.5990

PREDICTED MISS ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1.0 AND 2.0 SEC	6.5073	3.8594	3.5969	0.6091	1.0821	0.7994	2.4914	1.2283
2.0 AND 3.0 SEC	21.1411	17.3869	18.9135	-6.5745	4.0586	2.1094	8.5297	2.7425
3.0 AND 4.0 SEC	59.0046	49.4975	31.6644	19.1232	17.8085	9.9974	36.4946	-17.2386
4.0 AND 5.0 SEC	116.5178	106.2879	38.7467	5.1545	53.4276	36.2522	96.0239	-69.5382

X. ROBUSTNESS

The sensitivity of the predictive performance of a particular A-R model to changes in the autoregressive coefficients is important because, in practice, one does not have a-priori knowledge of the statistics of target motion from which one can compute these coefficients. One is thus lead to pose the following questions: How do the predictors perform when the coefficients associated with a particular axis are used for all three coordinate axes, and how well can one predict with a single model for all available flight passes. In answering these questions, it was found that the A-R models exhibit a remarkable degree of robustness. Table 5 gives the performance of a standard thrice-differenced A-R predictor. Table 6, generated for the same flight path, was produced by using the x-coefficients for all three coordinates. Notice that little degration in performance was incurred by using a single set of coefficients. Table 7 was generated by using a model developed for a different flight pass. Again, the predictors perform quite well. Using a single model for all flight passes, one is led to the conclusion that a single set of A-R coefficients can be used for prediction against a given aerial target.

TABLE 5

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE XYZ-THRICED-DIFFERENCED PREDICTION MODEL
FLIGHT 5 PASS 5 MODELS ASSUMED TO HAVE 0.0 MEANS.

FIRST DATA POINT = 4.1 LAST DATA POINT = 32.2

	PENALTY	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
CLOSEST APPROACH:	0.32223	33.9571	25.2983	22.4264	-0.5291	20.2780	-1.4444	15.4577	-1.3645
PREDICTED MISS:	0.29997	36.5575	27.7615	22.1898	0.9294	24.6253	-5.0759	15.4162	2.6860

HISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN				*																
0.0 AND 1.0 SEC	12	12	2	5	*	41	5	0	0	0	0	0	0	0	0	0	0	0	0	46
1.0 AND 2.0 SEC	8	13	7	9	*	40	7	0	0	0	0	0	0	0	0	0	0	0	0	47
2.0 AND 3.0 SEC	0	0	1	2	*	4	11	5	8	8	2	11	2	0	0	0	0	0	0	51
3.0 AND 4.0 SEC	0	0	0	1	*	1	4	1	1	4	5	31	8	5	0	0	0	0	0	60
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	2	2	3	5	24	28	3	0	0	0	0	0	67
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	271 TOTAL

HISTOGRAM OF PREDICTED MISSES

TOF BETWEEN				*																
0.0 AND 1.0 SEC	0	5	22	10	*	41	5	0	0	0	0	0	0	0	0	0	0	0	0	46
1.0 AND 2.0 SEC	4	14	7	9	*	39	8	0	0	0	0	0	0	0	0	0	0	0	0	47
2.0 AND 3.0 SEC	0	0	1	0	*	4	8	7	7	9	2	12	2	0	0	0	0	0	0	51
3.0 AND 4.0 SEC	0	0	0	0	*	0	0	1	3	4	5	31	11	5	0	0	0	0	0	60
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	1	2	1	22	31	10	0	0	0	0	0	67
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	271 TOTAL

CLOSEST APPROACH ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	2.8417	2.3650	2.0146	1.4113	1.6235	0.1599	0.9839	-0.4466
1.0 AND 2.0 SEC	3.3051	2.7924	1.4479	-0.0270	2.1889	0.1553	2.0090	0.1552
2.0 AND 3.0 SEC	25.6591	21.3177	22.5258	-4.3138	9.8027	-0.7644	7.4084	1.3749
3.0 AND 4.0 SEC	43.1946	39.5950	31.3195	16.6914	23.7078	21.3908	17.9668	13.5756
4.0 AND 5.0 SEC	49.7871	47.0585	27.6603	-14.7539	32.8869	-24.6352	25.1419	-18.5251

PREDICTED MISS ERRORS:

TOF BETWEEN	T-RMSE	T-ME	X-RMSE	X-ME	Y-RMSE	Y-ME	Z-RMSE	Z-ME
0.0 AND 1.0 SEC	3.2410	3.0381	2.1284	1.5828	1.5775	0.4060	1.8669	-1.5530
1.0 AND 2.0 SEC	3.5467	3.0936	1.4586	0.2581	2.1139	-0.4712	2.4460	0.8427
2.0 AND 3.0 SEC	26.3545	22.2728	22.5596	-3.8767	12.3200	-1.7190	5.8175	2.4195
3.0 AND 4.0 SEC	45.2430	42.4852	30.3029	18.2866	22.4300	17.9737	25.0112	17.5222
4.0 AND 5.0 SEC	55.9259	53.0328	27.8796	-10.9335	43.3803	-35.2664	19.2022	-6.1939

TABLE 6

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE XYZ-THRICE-DIFFERENCED PREDICTION MODEL

X COEFFICIENTS USED FOR ALL X-Y-Z MODELS.

FLIGHT 5 PASS 5

FIRST DATA POINT IS 4.8 LAST DATA POINT IS 32.2

	PENALTY	RMSE	ME
CLOSEST APPROACH:	0.3073764	32.72018	24.13045
PREDICTED MISS:	0.2852770	35.33394	26.65469

HISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	10	14	8	4	*	41	5	0	0	0	0	0	0	0	0	0	0	0	0	0	46
1.0 AND 2.0 SEC	4	14	8	8	*	37	11	0	0	0	0	0	0	0	0	0	0	0	0	0	48
2.0 AND 3.0 SEC	0	1	0	1	*	3	8	7	7	5	6	13	2	0	0	0	0	0	0	0	51
3.0 AND 4.0 SEC	0	0	0	0	*	0	3	3	6	5	11	19	8	5	0	0	0	0	0	0	60
4.0 AND 5.0 SEC	0	0	0	0	*	0	1	4	5	4	3	24	20	7	0	0	0	0	0	0	68
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	273 TOTAL

HISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	4	23	9	*	40	6	0	0	0	0	0	0	0	0	0	0	0	0	0	46
1.0 AND 2.0 SEC	1	12	8	12	*	35	13	0	0	0	0	0	0	0	0	0	0	0	0	0	48
2.0 AND 3.0 SEC	0	1	0	1	*	2	7	5	10	6	6	13	2	0	0	0	0	0	0	0	51
3.0 AND 4.0 SEC	0	0	0	0	*	0	1	2	4	9	7	22	9	6	0	0	0	0	0	0	60
4.0 AND 5.0 SEC	0	0	0	0	*	0	1	0	0	3	1	28	26	9	0	0	0	0	0	0	68
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	273 TOTAL

THE XYZ COEFFICIENTS ARE:

1.	0.40035	1.00255	0.18831	-0.48269	-0.18828

TABLE 7

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE XYZ-THRICE-DIFFERENCED PREDICTION MODEL

FILTERS ARE FROM FLIGHT 5 PASS 1.

FLIGHT 5 PASS 5

FIRST DATA POINT IS 4.8 LAST DATA POINT IS 32.2

	PENALTY	RMSE	ME
CLOSEST APPROACH:	0.3067623	32.24388	23.86668
PREDICTED MISS:	0.2821292	34.47070	26.18990

HISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	9	14	8	4	*	39	6	0	0	0	0	0	0	0	0	0	0	0	0	0	45
1.0 AND 2.0 SEC	8	11	9	8	*	37	11	0	0	0	0	0	0	0	0	0	0	0	0	0	48
2.0 AND 3.0 SEC	0	0	2	2	*	4	7	6	7	5	7	13	2	0	0	0	0	0	0	0	51
3.0 AND 4.0 SEC	0	0	0	0	*	0	3	4	4	7	14	17	6	5	0	0	0	0	0	0	60
4.0 AND 5.0 SEC	0	0	0	0	*	0	2	5	2	2	6	21	28	2	0	0	0	0	0	0	68
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	272 TOTAL

HISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	2	25	8	*	39	6	0	0	0	0	0	0	0	0	0	0	0	0	0	45
1.0 AND 2.0 SEC	5	8	10	11	*	37	11	0	0	0	0	0	0	0	0	0	0	0	0	0	48
2.0 AND 3.0 SEC	0	0	1	1	*	2	8	7	7	5	7	13	2	0	0	0	0	0	0	0	51
3.0 AND 4.0 SEC	0	0	0	0	*	0	0	3	3	6	12	24	6	6	0	0	0	0	0	0	60
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	2	2	7	25	25	7	0	0	0	0	0	0	68
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	272 TOTAL

THE XYZ COEFFICIENTS ARE:

1.	0.37651	0.22399	0.29397	-0.35217	-0.32915
2.	0.18825	1.02238	0.47344	-0.34466	-0.42317
3.	0.23706	0.94619	0.40063	-0.29396	-0.39185

XI. DISPERSION SYNTHESIS

In a realistic combat environment, more than one gun air defense system will be employed for defense of a given area. If a communications link is established between the systems, one can enhance hit probability by firing the guns at points in space dictated by some optimization criteria. Optimization for the location of the bursts was done for the case where four guns are employed. This was done as follows:

Orient the y-axis along the target flight path. The burst pattern is then defined in the x-z plane as in Fig. 3 below.

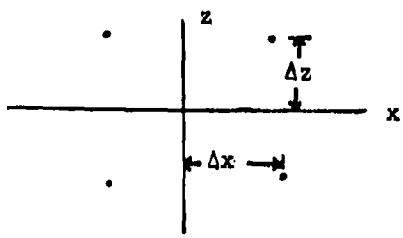


Fig. 3

The pattern is defined by:

$$\Delta x = a_1 + a_2 T_f + a_3 T_f^2$$

$$\Delta z = b\Delta x \, .$$

The four parameters were obtained via optimization using the performance criterion

$$F = \frac{1}{N} \sum_{i=1}^{N} e^{-d_i^2/2\sigma^2}$$

where d_i is the minimum miss of the four shells and $\sigma = 5.0$, the radius of the "hit" circle.* This particular form for the performance criterion was chosen in order that the number of rounds falling within 5m of the center of the target be maximized.

* For attack aircraft, 5m is roughly the effective radius of the target.

Table 8 typifies the results obtained for all flight passes. One finds that more rounds fall within the 5m distance of the target, although the percentage of rounds falling within this distance is not necessarily larger. However, there is a decrease in the RMS of the distance of closest approach as expected. (Compare with Table 2.)

TABLE 8

MINIMUM MISS DISTANCE INFORMATION FOR AUTOREGRESSIVE XYZ-TRIPLE-DOT PREDICTION MODEL
FLIGHT 5 PASS 13 MODELS ASSUMED TO HAVE 0.0 MEANS.

DEF. OF DISPERSION PATTERN: DELX = -5.8764 + (1.6691)*TOF + (1.4261)*TOF**2, DELZ = (0.4500)*DELX

FIRST DATA POINT = 14.5 LAST DATA POINT = 42.2

	ALL POINTS PENALTY	ALL POINTS RMSE	ALL POINTS ME	MIN MISS PENALTY	MIN MISS RMSE	MIN MISS ME
CLOSEST APPROACH:	0.2593482	54.063639	36.033346	0.3593552	39.177767	24.434735
PREDICTED MISS:	0.2382280	66.771682	44.343584	0.3200334	54.643323	34.803723

TOTAL NUMBER OF POINTS (CLUSTERS OF FOUR ROUNDS)

NISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	0	2	42	*	250	116	2	0	0	0	0	0	0	0	0	0	0	0	0	368
2.0 AND 3.0 SEC	0	4	1	6	*	26	66	43	21	21	13	22	0	0	0	0	0	0	0	0	212
3.0 AND 4.0 SEC	0	1	0	0	*	3	16	10	12	14	20	63	58	43	0	0	0	0	0	0	232
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	3	2	4	7	35	33	80	82	18	0	0	0	0	264
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	1076 TOTAL

NISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	0	0	12	*	169	197	2	0	0	0	0	0	0	0	0	0	0	0	0	368
2.0 AND 3.0 SEC	0	1	4	6	*	24	63	46	23	20	13	23	0	0	0	0	0	0	0	0	212
3.0 AND 4.0 SEC	0	0	0	0	*	0	8	8	9	10	14	39	78	65	1	0	0	0	0	0	232
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	0	3	15	24	62	104	47	9	0	0	0	264
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	1076 TOTAL

MINIMUM MISS OF CLUSTER OF FOUR ROUNDS

NISTOGRAM OF CLOSEST APPROACH MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	0	2	35	*	89	3	0	0	0	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	4	2	5	*	23	18	7	3	2	0	0	0	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	1	0	0	*	3	11	5	5	6	6	10	9	5	0	0	0	0	0	0	58
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	3	2	4	5	20	7	14	10	1	0	0	0	0	66
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	269 TOTAL

NISTOGRAM OF PREDICTED MISSES

TOF BETWEEN	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	
0.0 AND 1.0 SEC	0	0	0	0	*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.0 AND 2.0 SEC	0	0	0	12	*	89	3	0	0	0	0	0	0	0	0	0	0	0	0	0	92
2.0 AND 3.0 SEC	0	1	1	5	*	21	19	7	4	2	0	0	0	0	0	0	0	0	0	0	53
3.0 AND 4.0 SEC	0	0	0	0	*	0	7	3	3	2	3	16	16	8	0	0	0	0	0	0	58
4.0 AND 5.0 SEC	0	0	0	0	*	0	0	0	0	0	2	10	16	12	19	6	1	0	0	0	66
	1.	2.	3.	4.	*	5.	10.	15.	20.	25.	30.	50.	70.	100.	150.	200.	250.	300.	350.	400.	269 TOTAL

XII. CONCLUSIONS

The results presented heretofore are by no means conclusive. Prediction using A-R models for dynamic variables linked more closely to the dynamics of aircraft motion is presently under investigation. The brief discussion on dispersion synthesis is by no means the last word on the subject and a game theoretic approach to the problem seems to be in order. Degradation in predictive capability under conditions of sensor and ballistic errors remains to be determined. Nevertheless, the results appear encouraging. The predictors discussed here, which were designed and tested against actual aircraft data, outperform any class of predictors developed heretofore. As more data becomes available, additional tests of model robustness can be made, using an already developed predictor against a new set of flight data.

One is compelled to conclude that with some engineering intuition and judgment, increased system performance can be had for a cheap price by properly analyzing and modeling threat data.

REFERENCES

1. Blackman, Filtering and Prediction, Addison Wesley, 1965.

2. Final Report, A Parametric Study of the Advanced Forward Area Air Defense Weapon System (AFAADS), Vol. 1, Systems Analysis, Data Systems Division, Litton Systems, Inc., MS 00678A, Oct. 1970.

3. S. Goodman, A. McGoldrick, Air Defense Prediction Algorithm Study, U. S. Army Armament Command, Frankford Arsenal, Philadelphia, Pa., GA-TR-76027 May 1976.

4. Anderson, T. W., Time Series Analysis, Wiley, 1971.

A SENSITIVITY EVALUATION OF A LARGE SCALE TACTICAL SYSTEM AVAILABILITY UNDER VARYING SUPPORT RESOURCE LEVELS

ROBERT A. HALL AND HOWARD M. BRATT
AVIATION RESEARCH AND DEVELOPMENT COMMAND
Fort Eustis, Virginia

Introduction

A major problem for Army decision makers and consequently Army operations research analysts is the estimation of needed resources to support large scale tactical systems. This problem is further compounded by the question, what if I reduce the particular resource by XX%?

This paper presents one method of handling this problem, that is, through the use of simulation. The US Army, through the Applied Technology Laboratory, has developed computer mathematical programs that simulate the experiences of a system in the field. These computer programs are known as the Army Reliability and Maintainability Simulation (ARMS) model.* ARMS is a highly complex set of computer programs that simulates the operational and maintenance policies of a quantity of aircraft in the field. ARMS flies the aircraft; breaks parts; fixes the parts, either on-aircraft or off-aircraft, if off-aircraft, at one of four different maintenance levels; inspects the aircraft and queues and limits the aircraft resources. Use of ARMS allows the analyst to define his system to the detail he requires or to the level to which he has data. This definition includes malfunction rates, probability of remove and replace, times to repair, number of men needed to perform the repair, time-between-overhaul, if applicable, and off-equipment repair (higher level maintenance). Also defined are mission scenarios by minute segments, scheduled calls for aircraft, continuous missions, random missions, effects of flight essential failures, maintenance concepts, manpower limits, and shift hours.

The fielded system chosen for study is the CH-47C. This is a highly complex aircraft system that will provide a highly active system for study.

When estimating aircraft resources, there are three broad areas that may be examined: GSE, manpower, and parts availability. This paper examines all three areas showing the independent effects of a reduction in each parameter.

A question arises, how do you measure the effects of a percentage reduction in a resource? There are as many answers to this question as there are interested groups wanting such an answer. We have chosen one main variable for examination based on the assumption that the object of maintenance is to get aircraft ready for launch. If aircraft are ready when called, then the resources supporting that aircraft are sufficient.

*Reference 1

The CH-47C Model

An ARMS model version of the tandem rotor, medium lift CH-47C helicopter had been developed and validated several years ago under contract with The Boeing Company.* This model with only minor changes became the basic vehicle used in the current study. The CH-47C consisted of 164 elements and 11 subsystems. For each element the following data was provided:

- Maintenance Actions per Operating Hour
- Flight Criticality
- Mission Equipment Essentiality
- Probability that a Maintenance Action would be Discovered at Time of Failure
- Probabilities that an Undiscovered Maintenance Action would be Discovered at Subsequent Scheduled Inspections and Mission Events
- For Flight Critical Elements, the Consequences of a Failure During Flight, Probability Distribution Including Forced Landing, Attrition, Abort Mission and Continue Mission
- The Probability that a Maintenance Action would Cause a Remove and Replace Event Rather than a Repair in Place
- The Mean Time, Using the Exponential Distribution, to Repair in Place
- Administrative Time Delay (RIP)
- Ground Support Equipment Required (RIP)
- Military Occupation Speciality (MOS) Code of Each Type of Mechanic Required and Number of Each Required (RIP)
- The Probability that this Maintenance Action would Require a Functional Test Flight (RIP)
- For Remove and Replace (R/R) Maintenance Actions, the Supply Delay Time to Obtain and Prepare the Replacement Part
- Administrative Delay Time for such Things as Processing Paper Work, Scheduling the Maintenance Action, etc.
- Ground Support Equipment and Maintenance Facilities Required for the R/R Action
- The MOS Codes Required to Perform the R/R Action
- The Probability a Spare Component would be in Stock when Requested (this parameter was used in the study)
- The Restock Time, Delay to Obtain a Part on Order (3 days was used in this study)
- Probability that this R/R Action would Result in the Requirement for a Maintenance Check Flight
- There were 16 Elements with Scheduled Time Between Removal (TBO) which varied from 2400 hours to 300 hours

*Reference 2

There were 8 scheduled maintenance events modeled in the study:

Daily Inspection, every 24 Hours if the Aircraft had Flown and in 72 Hours if the Aircraft had not Flown on the Previous 3 Days (not accomplished if the aircraft was down for maintenance or lack of spare parts)
12.5 Hour Spectrographic Oil Analysis Sample
25 Hour Preventative Maintenance Intermediate Inspection
25 Hour Spectrographic Oil Analysis Sample
100 Hour Preventative Maintenance Periodic (PMP) Inspection
90 Day Fire Extinguishing System Inspection
6 Month Pitot/Static System Inspection
12 Month Engine Fire Extinguisher Inspection

There were 2 maintenance shifts at organizational level:

Shift I	Start at 0600	Stop at 1400
Shift II	Start at 1400	Stop at 2200

Manpower quantity was a parameter in the study. The "Super Crew Chief" concept, i.e., a mechanic trained in all maintenance disciplines. This concept was necessary to parameterize the manpower function in the study. It could be thought of as the same as supporting a very large number of vehicles which, because of the size of the fleet, require a large number of each type of mechanic. The number of mechanics used in the base case was 40 and this number was gradually reduced in subsequent runs as discussed in the portion of the paper that describes the experiment. Ground Support Equipment, another parameter used in the study, was also generalized for the same reasons as applied to the manpower. In the base case, 15 pieces of GSE were provided and this number, also, was subsequently reduced during the experiment.

The third parameter used in the study was the probability of spare parts being available when required for remove-and-replace actions. 100% availability was used in the base case and the percentage was reduced in subsequent runs of the model. Another variable that impacts the sensitivity of parts availability is the period of time chosen for the delivery of unavailable parts once they have been ordered. A time period of 72 hours was chosen as a constant (no distribution function) for the resupply time when parts probabilities were less than 100% in the experimental model runs. Any user of the data reported in this paper must recognize that the assumption of a 72 hour supply time has a significant impact on the sensitivity of the results relative to the spares parameter. For example, a resupply time approaching zero hours with a 50% probability of spares availability would have the effect of providing almost 100% spares within a few minutes of the time they were requested. The mission chosen for the CH-47C helicopter is called the resupply mission in which the helicopter is carrying external

loads of munitions to a forward gun site. The following segments and elapsed times were used:

Ground preflight and engine start and taxi	30 Min
Flight	90 Min
Post flight, taxi and park	30 Min
Refuel	30 Min

In simulation modeling in general and especially when an attempt is being made to quantify the optimum quantity of logistic support resources necessary to obtain a desired effectiveness, it is essential that the number of aircraft requested significantly exceed the maximum capability of the system; in simulation parlance this is called "Loading the System." The following mission schedule was requested, 7 days a week, for 4 weeks:

Take-Off Time	Max Number of Aircraft	Min
0700	4	1
0830	4	1
1000	4	1
1130	4	1
1300	4	1
1430	4	1
1600	4	1
1730	4	1
1900	4	1

The Max/Min numbers are to be interpreted as Max = the desired number of aircraft per mission and Min = the minimum number of aircraft that will be permitted to fly the mission. From this data, 1008 launches are scheduled per a 28 day month to fly 252 missions. In the base case, 753 aircraft launches were accomplished and all of the 252 scheduled missions were flown with at least one aircraft on the mission. 74.7% of the scheduled launches were met.

*In simulation modeling it is necessary to provide a simulated period of stabilization running prior to the start of the data collection period. The stabilization period is sized to assure that those functions and interactions which occur during the simulation become stabilized before final statistics are collected. That is, people are working and being demanded, parts are being used and ordered, delays are occurring for lack of resources, etc. To speed the stabilization, an initial quantity of flight hours is distributed across the aircraft fleet and time scheduled removal components. All runs consist of a 2 week stabilization period.

*Reference 3

Remembering that the ARMS model is stochastic and that probability distributions are widely used in the internal decision process, it is to be understood that any one replication represents only one realization in a distribution of possible outcomes. Therefore, for statistical validity as well as for parameter smoothing, replications of the runs at each data point using different random number streams are required. The number of replications required to achieve statistical confidence will vary with the scheduled activity within the simulated scenario and also with the length of the simulation period. For the data used in this report, 10 replications were made at each data point and a 28 day simulation period was used. A mathematical average was made of the replicated values. From this data, trend lines were computed for each test parameter using a second degree polynomial regression program. Another parameter that could have been used in this study would be the number of aircraft in the fleet. For this study the number of CH-47C aircraft remained constant at 16.

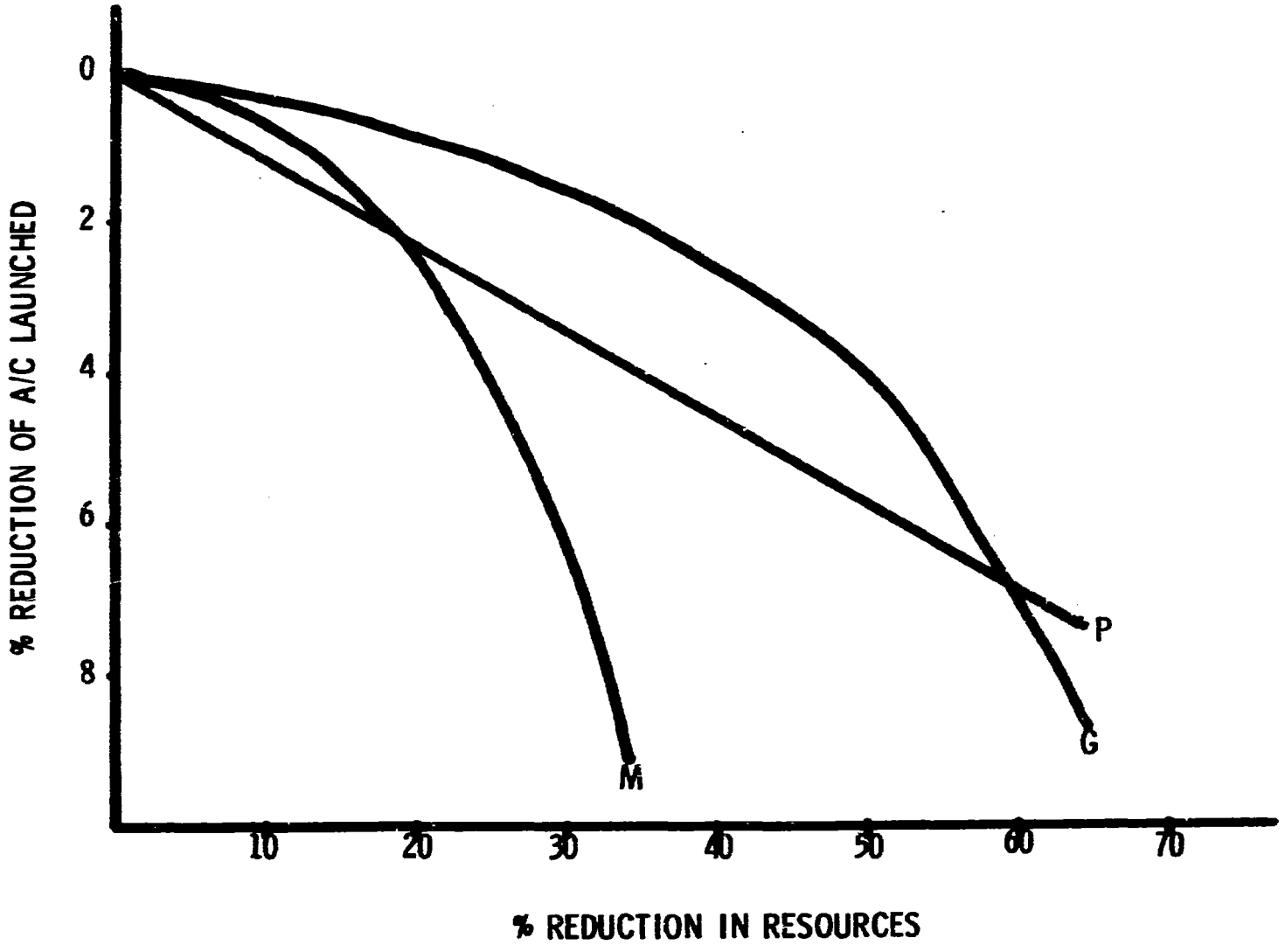
Having achieved our baseline point, we began running the cases to show the effects of varying the support resources. Holding GSE and Parts Availability constant, we made simulation runs decreasing Manpower by 10%, 20%, 30%, and 40% from the baseline point. The aircraft launched showed an immediate impact with a 10% decrease in manpower causing a .5% decrease in our parameter. As the Manpower decreased, its effects rapidly increased until the 30%-40% decrease caused a jump from 8% less aircraft launched to 18% less aircraft launched. This indicates that any further decrease in available Manpower would severely limit the capability of the CH-47C to perform its missions.

The Ground Support Equipment simulation cases were handled the same way. While holding Manpower and Parts Availability constant, the GSE was reduced 10%, 20%, ... 70%. The graph of the results was generally the same, however, GSE had a more gradual initial impact than Manpower showed. GSE did not have an accelerating effect until it had been decreased approximately 50% of its initial strength.

Parts Availability was decreased 90%, 80%, ... 50% while Manpower and GSE were held constant. However, the effects of Parts Availability did not follow the same general slope as Manpower and GSE. Its shape is more a straight line than curved. The effects of Parts Availability apparently are linear, at least through a large reduction in the Parts Availability. This may be due to the large number of variables involved in this area, such as inventory restock delay time (3 days for this paper) or probability of remove and replace.

After viewing the results curves from our experiment, the analyst can give answers to the question of how a reduction in support resources can affect his particular system. However, the analyst must realize that in any given situation, he must do more than was done in this paper. We did

RESOURCES IMPACT



our experiment with certain simplifying assumptions and certain variables as constants. The analyst should review these and see if they are adequate for his particular situation. If they are not, he should change them so they are applicable. He must also take into consideration what the decision maker wants. For instance if it is how to reduce the cost of resources while causing the least impact on the system, the analyst must calculate the total cost involved in reducing resources. For instance, he may be able to reduce a high cost resource that has a large impact on his system and compensate for the decrease with an increase of another resource and still save the necessary dollars as opposed to just reducing the resource with the least impact on his system, hoping he will get the necessary cost saving, which may not happen. Also, the analyst may find that for reasons beyond his control, he may not be able to reduce the resources that his analysis tells him should be reduced.

All the above situations are just reasons why the question of "resources impacting availability" is not an easy one. However, in answering these questions, the analyst does have a tool that will help him do his job, that is the ARMS model. If used correctly, it can be a great help.

REFERENCES

1. BRATT, Howard M., "Aircraft R&M Simulation (ARMS) Model," 1975 Army Numerical Analyses Conference.

2. DOUGHERTY, J. J., "CH-47C/HLH R&M Simulation Analysis," Final Report Applied Technology Laboratory, US Army Research & Technology Laboratories (AVRADCOM), Fort Eustis, VA.

3. BRATT, Howard M., "Homeostatic Criteria for Assessment of the Morphological State of Large Scale Hybrid Analog-Simulation Models," 21st Conference of Army Mathematicians.

CH-47C SIMULATION MODEL SCENARIO-BASELINE

Scenario Simulated

One platoon of 16 Army helicopters.
Flying program consisted of 7 flying days per week.
Mission duration is 1.5 hours.
Launch schedule during each flying day

0700	4 aircraft	0830	4 aircraft
1000	4 aircraft	1130	4 aircraft
1300	4 aircraft	1430	4 aircraft
1600	4 aircraft	1730	4 aircraft
1900	4 aircraft		

Other Flight Considerations

Standby aircraft ready at all times during the scheduled flying intervals.
Mission flight is possible up to 30 minutes after scheduled flight time. After this interval, flight is scrubbed.

Maintenance Concept Simulated

PMP occurs at intervals of 100 hours. PMI occurs at intervals of 25 hours.
Preventive maintenance daily (PMD) inspections occur daily if the aircraft has flown or every 72 hours if not flying.
Maintenance personnel are available between 0600 and 2200 during the 7-day flying period per week.
The aircraft consists of 164 elements. There are 16 time change components within this total.
Organizational maintenance simulated only.

BASELINE CH-47 MODEL PARAMETERS

16 aircraft per platoon

100-hour PMP, requiring 6 men for 7.5 hours each

25-hour PMI, requiring 3.5 men for 3.4 hours each

Daily inspection performed @ 1830 each day requiring 2 men for 2.4 hours each

TBO values ranging from 300 hours to 2400 hours

Utilization of 71.9 hours per aircraft per month

Maintenance action rate of .0.5 per flight hour

Flying and maintenance take place 7 days a week

2 maintenance shifts of 8 hours each used daily

CH-47C

01	AIRFRAME
02	COMMUNICATION/NAVIGATION
03	DRIVE
04	ELECTRICAL
05	EQUIPMENT
06	LANDING GEAR
07	FLIGHT CONTROL
08	HYDRAULIC
09	ROTOR
10	INDICATING
11	POWERPLANT

CH-47C

SUBSYSTEM 06 LANDING GEAR

ELEMENT

01	FITTING AND BRACES
02	POWER STEERING
03	SHOCK STRUTS
04	SWIVEL LOCK SYSTEM
05	WHEEL ASSEMBLIES
06	WHEELS AND PARKING BRAKE

CH-47C
BASELINE PARAMETERS

	Called	Cancelled	Late
Missions	252	0	12

	Called	Launched	Cancelled
Aircraft	1008	753	255

	Shift		Max Que	
	#1	#2	#1	#2
MOS UTIL	56.7	56.2	14	8

GSE	Total Calls	MA's Waiting
	1044	3

RESOURCES (BASELINE)

MANPOWER - 40

GROUND SUPPORT EQUIPMENT - 15

PARTS AVAILABILITY - 100%

USE OF LOGNORMAL CONFIDENCE BOUNDS ON RELIABLE LIFE WHEN THE TRUE LIFE DISTRIBUTION IS NOT LOGNORMAL

Eugene E. Coppola
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, NY 12189

1. Reliable Life and Its Lower Confidence Bound

Reliable life is that time S during which a specified proportion R of a population of devices will operate continuously without failure. The proportion R is called the reliability. Reliable life is especially important for devices which can fail catastrophicly; that is, failure of the device can result in the destruction of the device and possibly surrounding equipment and also possible injury or death to operating personnel. Cannon components such as tubes and breeches fall into this category. For such catastrophicly-failing devices, it is important that the device be operated only during the time for which the probability of successful operation (R) is high. For cannon components, R is generally specified to be 0.999. The reliable life for cannon components is also known as safe life, and we will use the two terms interchangeably.

For a new device, reliable life is not known and must be estimated from test data. For cannon components a confidence requirement is added. That is, it must be shown with a specified confidence level C that the actual reliable life exceeds a given value. For cannon components, C is generally specified as 0.9. In practice, because of the confidence requirement, point estimates of safe life are not used; instead a lower confidence bound on safe life at level C is used. The lower confidence bound will be called lower confidenced safe life (LCSL).

For cannon components, catastrophic failures are caused by fatigue cracks. Consequently, safe life is important only for fatigue failures. There are other ways that cannon components can fail (e.g., excessive wear in tubes) but these are fail-safe types of failure and hence are ignored in safe life determination. Fatigue testing, even with the laboratory simulation techniques employed today, is very expensive and time consuming. This greatly limits the amount of data that can be collected for any one type of device. The generally accepted method today is to test six specimens to failure and to base safe life calculations on these.

Because data is limited and the specified reliability is so high, non-parametric and distribution-free methods do not give good results. Consequently, it is necessary to assume that the failure times follow a distribution of known mathematical form. The lognormal and Weibull distributions are commonly used for this purpose, although there has never been

enough data from any one particular device to make a determination of the true failure distribution. There are some theoretical justifications behind both the lognormal and Weibull distributions, but we shall not consider them here.

2. Fracture Mechanics Model of Gun Tube Fatigue

Using Paris' equation for rate of crack growth and experimental results, Throop and others at Watervliet Arsenal have developed a deterministic model of fatigue crack growth in gun tubes [3,7]. After some manipulation of Throop's equations [5], the following equation relating crack depth b to number of cycles N results:

$$N = \frac{1}{Gk}\left(b_o^{-k} - b^{-k}\right) \qquad (1)$$

where: b_o = the initial crack depth, assumed to be present after a few rounds of firing

$G = \dfrac{C(\alpha S\sqrt{\pi})^{2k+1}}{E\sigma_y K_{IC}}$

k = a parameter dependent on material properties and stress intensity

S = maximum hoop stress at the bore

α = a parameter depending on crack shape and on the residual stresses introduced by the autofrettage process

E = Young's modulus

σ_y = yield strength

K_{IC} = fracture toughness for a crack in a tangential stress field

C = a parameter varying with k to maintain dimensional homogeneity and possibly depending on material properties

From equation 1, one can calculate the number of cycles N required for the crack to reach a critical depth at which fatigue failure occurs, provided one knows the relevant material properties. The material properties, however, vary from tube to tube, that is, they are random.

3. Computer Simulation of Fatigue Failure

Using Throop's model, Racicot [5] performed Monte-Carlo simulations to generate a large number of pseudo-fatigue lives that could then be analyzed statistically. However, there is not sufficient data at this time to determine the distributions of the material properties (b_o, k, S, α, E, σ_y, K_{IC}, and C) that appear in Throop's model. Racicot therefore assumed that each of the material properties had the same type of distribution and that this type of distribution was either normal, lognormal or Weibull. The parameters

of the assumed material-properties distributions were estimated from experimental data. The present author [2] has extended Racicot's results by considering more general cases. For each run of the author's simulation program, the program was instructed to pick at random a distribution type for each of the material properties. The parameters of the material-property distributions were then estimated from experimental data. This method would hopefully allow some independence from unwarranted assumptions. The present author has also considered cases where the material properties are correlated; Racicot assumed that most of the material properties were statistically independent. In this manner, we obtained several sets of simulated fatigue data, each of which could be examined statistically.

One question of great interest was whether the simulated data could be described by various parametric distribution families. This problem was approached in the standard way. For each parametric family, the parameters were estimated from the simulated data to obtain an approximating distribution. The approximating distribution could then be compared to the simulated data by goodness-of-fit statistics. We used three goodness-of-fit statistics: Kolmogov-Smirnov (KS), Cramer-von Mises (CVM) and Anderson-Darling (AD). (See reference 6 for definitions and uses of these.)

As an example of the sort of results obtained, Figure 1 shows the frequency histogram for the data of Run #1, consisting of 10,000 simulated fatigue lives. Figure 2 shows the empirical cumulative distribution function (cdf) of the simulated data, along with the approximating distributions from several parametric distribution families. None of them really gives a good fit. In Table 1 we show the goodness-of-fit statistics calculated for several distribution families. The lognormal distribution gives the best fit (the smaller the goodness-of-fit statistic, the better the fit). The Birnbaum-Saunders runs a close second. Weibull and exponential distributions do not fit nearly as well.

These results were typical for the simulated data; the lognormal or the Birnbaum-Saunders gave the best fit. They were quite close together and generally did much better than the other distributions. In his studies, Racicot concluded that the lognormal gave the best fit (he did not consider the Birnbaum-Saunders) and recommended that the lognormal distribution be used in the future for fatigue life studies. The only problem is that the present author has shown that although the lognormal distribution usually does give better fits, the fit is not totally acceptable. In fact, the goodness-of-fit statistics in most cases were significantly too large, thus leading to a rejection of lognormality. It then becomes important to know how well procedures derived from the assumption of lognormality work even though the fatigue life distribution is probably not lognormal.

4. Birnbaum-Saunders vs. Lognormal

Before we consider the adequacy of the lognormal, we should explain why we are not performing a similar analysis for the Birnbaum-Saunders distribution, even though the Birnbaum-Saunders and the lognormal fit about equally well. Actually, the closeness of the Birnbaum-Saunders and the

lognormal was anticipated on theoretical grounds. The cdf of the Birnbaum-Saunders distribution is given by:

$$F_1(x;\alpha,\beta) = \begin{cases} \Phi[\frac{1}{\alpha}(\frac{x}{\beta})^{\frac{1}{2}} - \frac{1}{\alpha}(\frac{x}{\beta})^{-\frac{1}{2}}] & x > 0 \\ 0 & x \leq 0 \end{cases}$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$

is the standard normal cdf. α and β are positive unknown parameters.

The lognormal cdf is given by:

$$F_2(x;\sigma,\theta) = \begin{cases} \Phi[\frac{1}{\sigma}\ln(\frac{x}{\theta})] & x > 0 \\ 0 & x \leq 0 \end{cases}$$

where σ and θ are positive unknown parameters.

Now suppose that X is a random variable with cdf $F_1(\cdot;\alpha,\beta)$. Let $Y = (X/\beta)^{1/\alpha}$. The cdf of Y is easily seen to be:

$$G(y;\alpha,\beta) = \Phi(\frac{y^{\alpha/2} - y^{-\alpha/2}}{\alpha}) \quad \text{for } y > 0$$

and 0 for $y \leq 0$. Now let α approach 0. For any $y > 0$,

$$\lim_{\alpha \to 0} \frac{y^{\alpha/2} - y^{-\alpha/2}}{\alpha} = \ln y .$$

Consequently, for all y,

$$\lim_{\alpha \to 0} G(y;\alpha,\beta) = \Phi(\ln y) ,$$

which is the standard lognormal distribution.

The above suggests that for small α, the Birnbaum-Saunders distribution $F_1(x;\alpha,\beta)$ can be approximated by the lognormal distribution $F_2(x;\alpha,\beta)$. The opposite is also true: For small σ, the lognormal $F_2(x;\sigma,\theta)$ can be approximated by the Birnbaum-Saunders $F_1(x;\sigma,\theta)$. The difference $F_1(x;\alpha,1) - F_2(x;\alpha,1)$ is shown in Figures 3 and 4. The approximation is quite good for small α. In fact, data from gun tube fatigue tests suggests that α

will usually be small (less than 0.3). We also observed small α for the simulation data. In practice, therefore, the Birnbaum-Saunders and the lognormal will be so close as to be nearly interchangeable. However, the lognormal is much easier to deal with in practice. So we have chosen to ignore the Birnbaum-Saunders even though it fits about as well as the lognormal.

5. Methods of Confidencing Safe Life

In the past, there have been 3 main schemes for calculating LCSL used at Watervliet Arsenal. The first assumes the underlying failure distribution is lognormal; the other two assume that the underlying distribution is Weibull. In the following we assume that $x_1, \ldots, x_N$ are identically distributed, independent fatigue lives obtained from testing.

Method I: Lognormal MLE Method

This method is based on the maximum likelihood estimates (MLE's) of the lognormal distribution. (See Ref. 4, p. 264-268 for a fuller exposition of this method.) Let:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} \ln x_i$$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\ln x_i - \bar{y})^2$$

The LCSL $\tilde{S}_I$ is given by:

$$\tilde{S}_I = \exp(\bar{y} - K_N(R,C)\hat{\sigma})$$

where $K_N(R,C)$ is a tolerance factor dependent on R,C and N.

Method II: Weibull BLIE Method

This method is based on the Best Linear Invariant Estimators (BLIE's) of the extreme-value distribution. For this method, the underlying fatigue life distribution is assumed to be 2-parameter Weibull. The extreme-value distribution enters the picture because the logarithm of a random variable with a 2-parameter Weibull distribution has an extreme-value distribution. (See Mann, Schafer, and Singpurwalla [4] for a fuller exposition of this method.) One calculates two numbers $\tilde{\eta}$ and $\tilde{\xi}$ which are the BLIE's of extreme value location and scale and are basically just weighted sums of the logarithms of the failure times, the weights depending on sample size N. The LCSL is given by:

$$\tilde{S}_{II} = \exp(\tilde{\eta} - L_N(R,C)\tilde{\xi})$$

where $L_N(R,C)$ is a tolerance factor (not the same as the tolerance factor in Method I) depending on R,C and N.

While Method I and Method II are superficially similar, it has been found in practice that Method II will generally give a much smaller LCSL than Method I. The author conjectures that some general law is at work that requires $S_I \geq S_{II}$ with high probability but he has not been able to show this.

The third method involves a Bayesian scheme devised by Clarke [1]. This method involves much laborious computation and we will not consider it here. Most often, this Bayesian method gives an LCSL intermediate in value between S_{II} and S_I.

6. Adequacy of Methods of Confidencing Safe Life

Because of the random method of selection of material-property distribution, each run of the simulation program effectively establishes a possible fatigue life population from which we can select random samples. Each population has its own true safe life, which can be estimated fairly well. We can then perform simulation studies for each population to determine how well the methods given above for constructing LCSL actually work.

The most important property of a lower confidence bound is that it underestimates the true quantity with a given probability C, the confidence level. Both Methods I and II are derived from assumptions on the underlying failure distribution. Let us call the "nominal confidence level" the confidence level C one would have if the appropriate assumptions were true, and the "actual confidence level" the probability C_A that the method in question produces on LCSL less than the true safe life. If the assumption from which the method is derived is true, then $C_A = C$. As mentioned above, the assumption of an underlying lognormal or Weibull distribution is probably not true. So we will probably have $C_A \neq C$. If $C_A > C$, the method gives conservative bounds, that is, we are actually underestimating more often than we think we are. Because we are dealing with devices that can fail catastrophicly, a conservative method is more to be desired than a non-conservative one.

The lognormal MLE and Weibull BLIE methods (I and II) give conservative bounds for all runs. The actual confidence levels C_A were estimated from 1000 simulated replicates of samples for various R and N and for nominal confidence $C = 0.9$. Some results are shown in Figures 5 through 8. The estimated true confidence levels are of course random variables, which accounts for the jaggedness of the curves in these figures. However, the main point here is not so much to determine the true confidence level but to determine if $C_A > C$. For all of our C_A's, except for a few in Run #2 with $R = 0.9$, we do indeed have $C_A > C$ with a 90% confidence. We can therefore conclude that the lognormal MLE and Weibull BLIE methods do give conservative confidence bounds.

An additional interesting fact emerges from these graphs. It appears that while both methods are conservative, the Weibull BLIE method is more conservative (that is, it gives a larger C_A) than the lognormal MLE method. This would suggest that the lognormal MLE method is to be preferred to the Weibull BLIE method.

7. Conclusions

The lognormal distribution, while generally yielding better fits to the simulated fatigue data than the other distributions considered, is probably not the exact fatigue life distribution. Methods derived from the lognormal are generally conservative and can be used. However, the lognormal may be overly conservative for large reliabilities and better methods probably exist. Methods derived from the Weibull distribution are extremely conservative for large reliabilities and should be avoided.

References and Bibliography

1. Clarke, R. W., "A General Computational Algorithm for Bayesian Confidence Bounds," Watervliet Arsenal Technical Report No. WVT-6911, Watervliet, NY, May 1969.

2. Coppola, E. E., "Probabalistic Models of Gun-Tube Fatigue Based on a Fracture-Mechanics Model," Benet Weapons Laboratory Technical Report, June 1977.

3. Davidson, T. E., and J. F. Throop, "Practical Fracture Mechanics Applications to Design of High Pressure Vessels," Watervliet Arsenal Technical Report No. WVT-TR-76047, Watervliet, NY, Dec. 76.

4. Mann, N. R., R. E. Schafer and N. D. Singpurwalla, Methods for Statistical Analysis of Reliability and Life Data, New York: John Wiley & Sons, 1974.

5. Racicot, R. L., "A Probabilistic Model of Gun Tube Fatigue," Benet Weapons Laboratory Technical Report No. ARLCB-TR-77029, May 1977.

6. Stephens, M. A., "EDF Statistics for Goodness of Fit and Some Comparisons," Journal of the American Statistical Association, 69, (Sep 74), pp. 730-737.

7. Throop, J. F., and G. A. Miller, "Optimum Fatigue Crack Resistance," ASTM Special Technical Publication 467, Philadelphia, PA, 1970, pp. 154-168.

Table 1

Goodness-of-Fit Results

Run 91

Distribution Type	Parameters Location	Shape	Scale	Goodness-of-Fit Statistics KS	CVM(x10^4)	AD
Normal	11737	-	2570	0.072	16.6	10.6
Lognormal	-	4.594	11463	0.051	9.5	6.2
Extreme-Value	12894	-	2004	0.13	61.9	48.9
Weibull	-	5.893	12642	0.10	38.7	29.5
Exponential 1-parameter	-	-	11738	0.47	594.7	282.5
Exponential 2-parameter	5630	-	6108	0.30	287.3	149.9
Double-Exponential	10581	-	2004	0.063	16.5	10.8
Inverse-Weibull	-	5.893	10393	0.087	28.0	20.2
Birnbaum-Saunders	-	0.2178	11466	0.052	9.6	6.2

FIGURE 2
CUMULATIVE DISTRIBUTION
Run #1

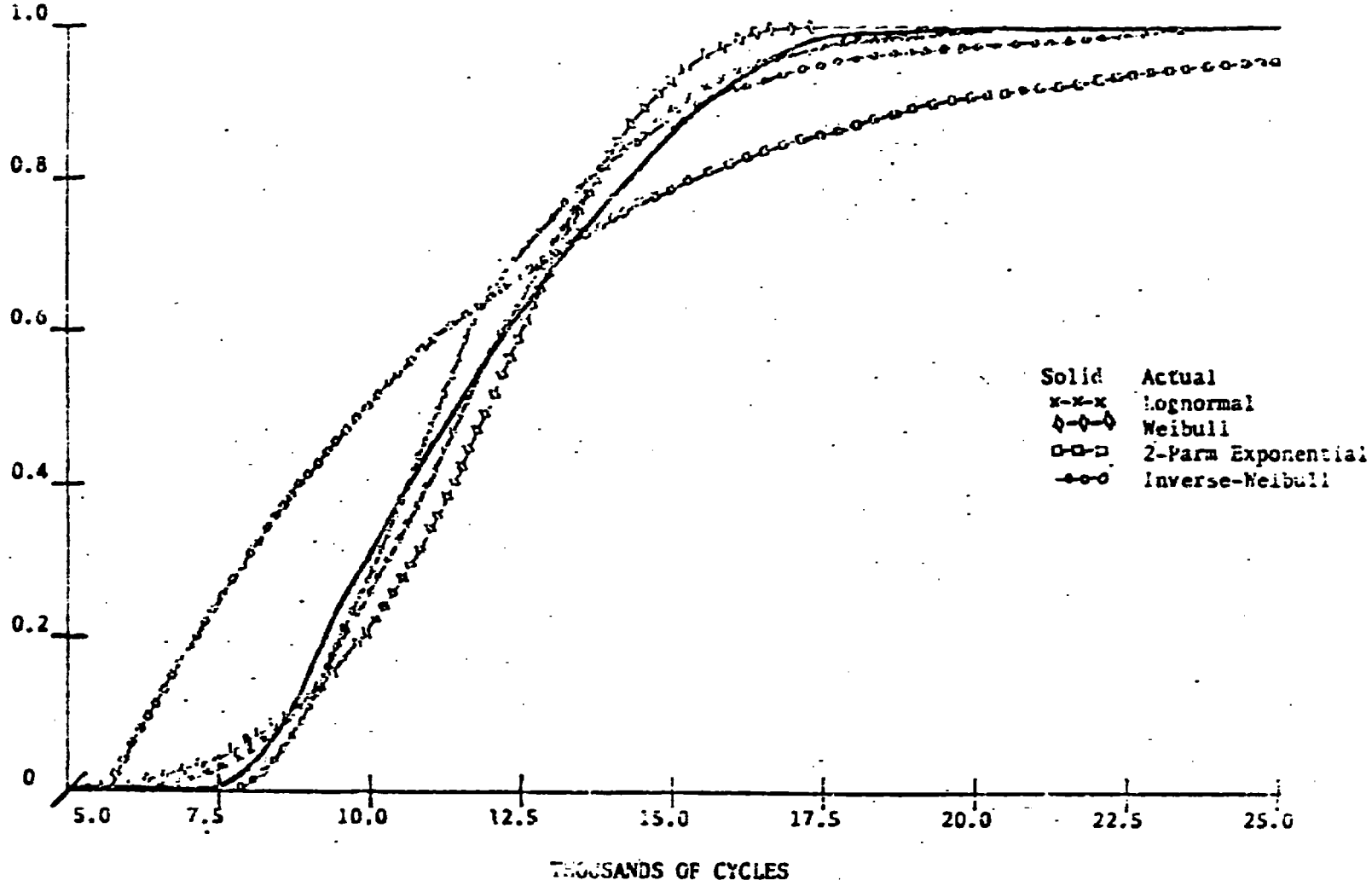


FIGURE 3
The Birnbaum-Saunders Distribution and Its
Lognormal Approximation

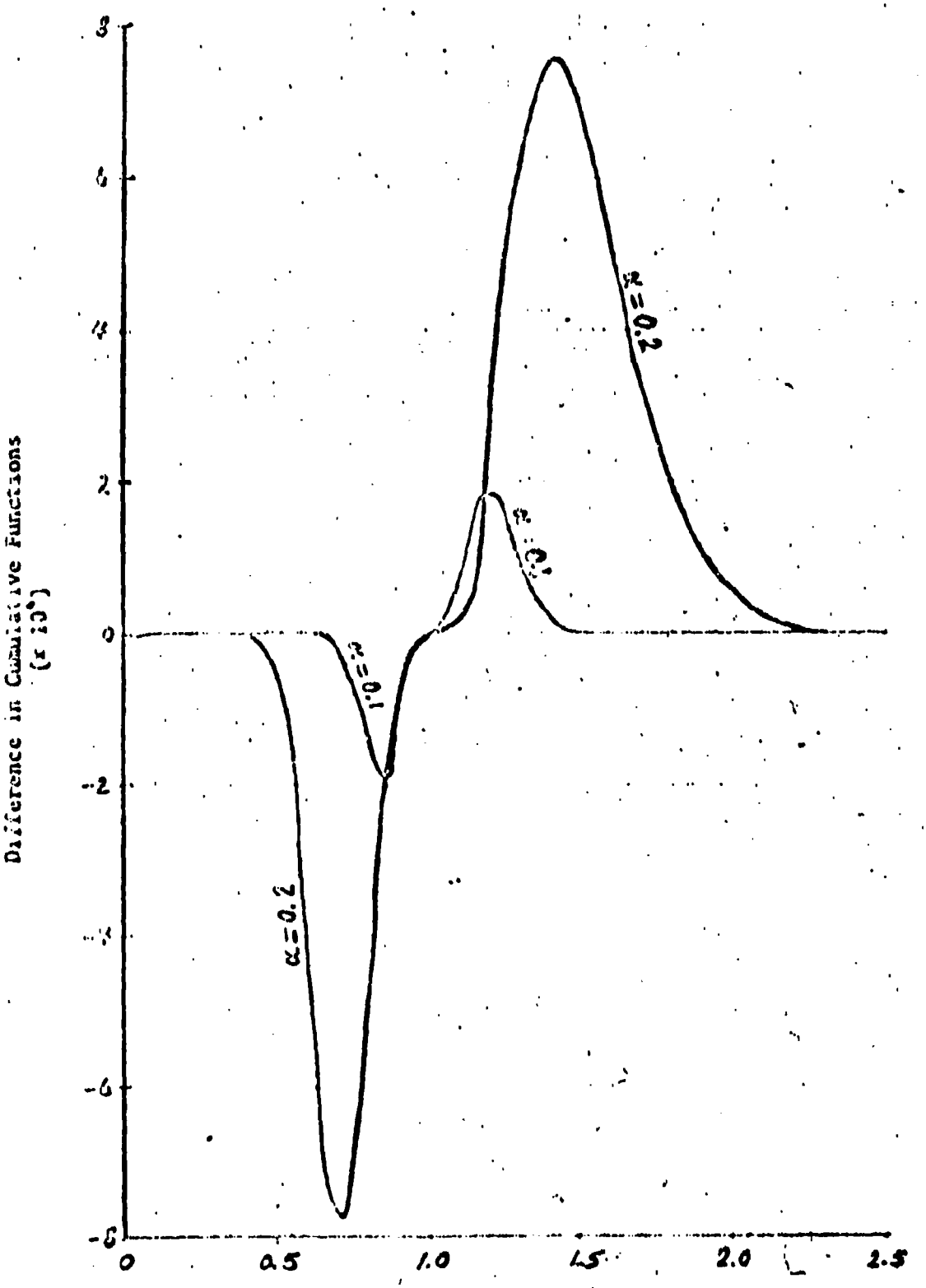


FIGURE 4
The Birnbaum-Saunders Distribution and its
Lognormal Approximation

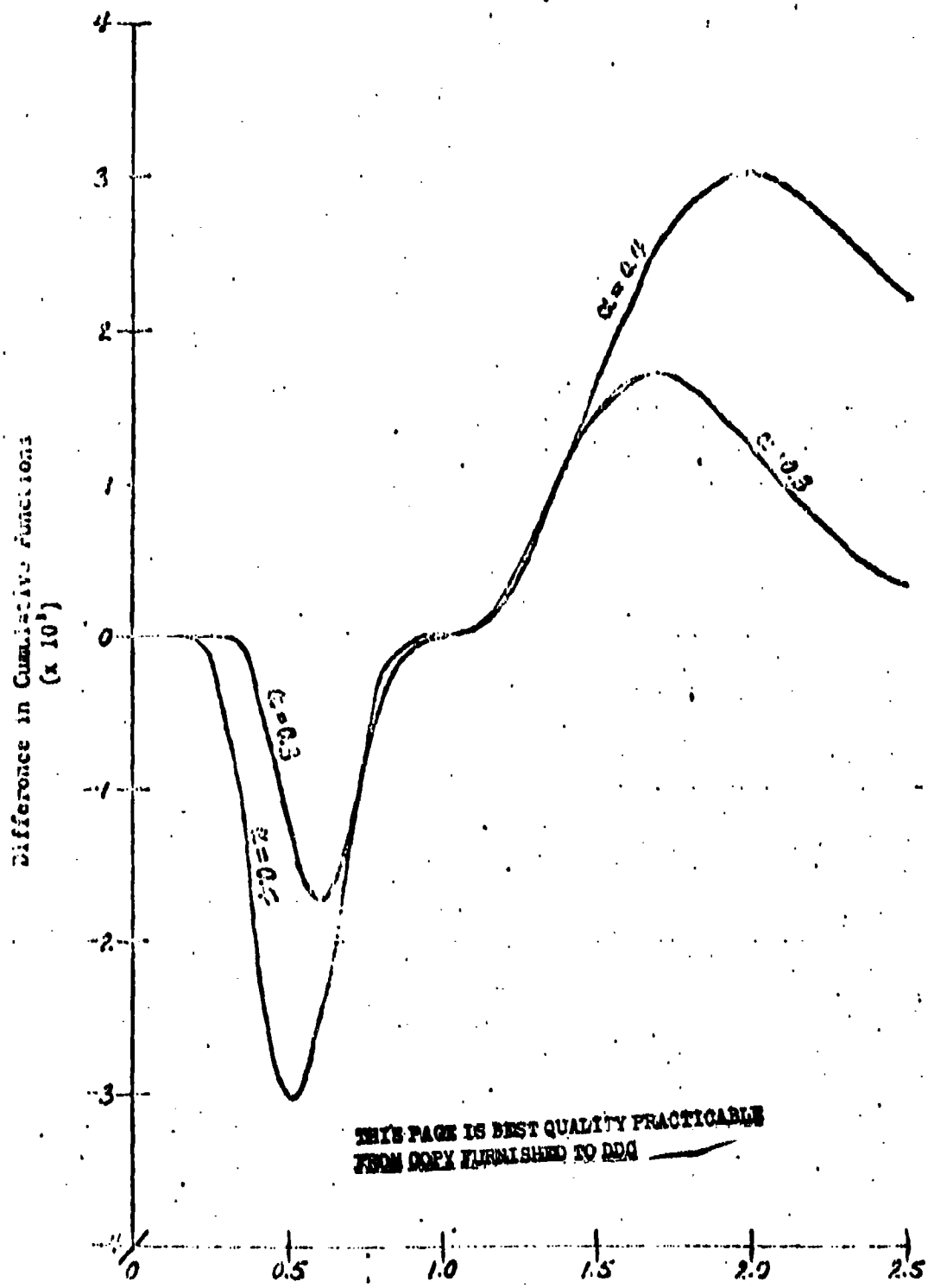


FIGURE 5

Run #1

Lower Confidenced Safe Life, Lognormal MLE Method .

Nominal Confidence 90%

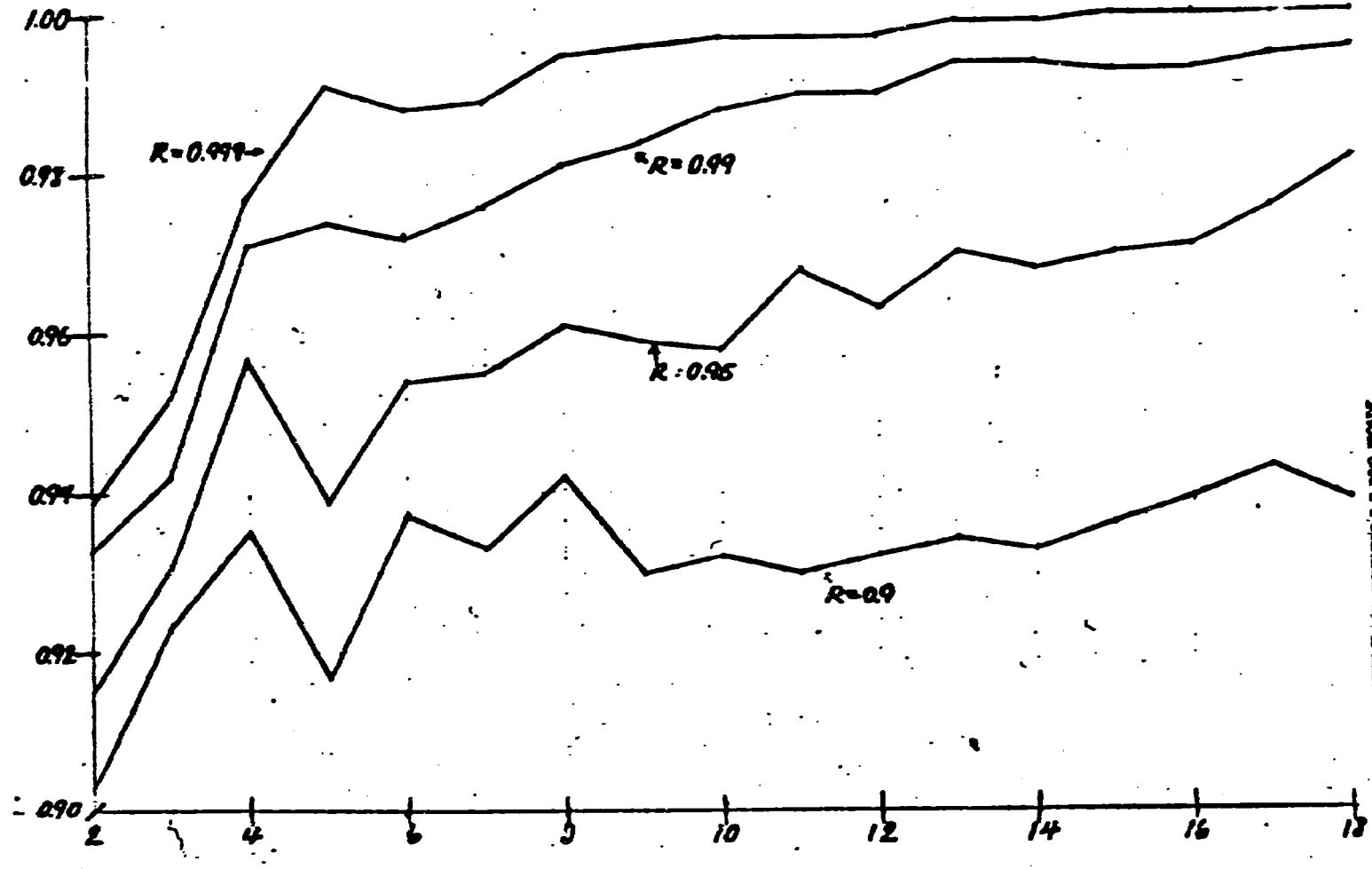


FIGURE 6
Run #2
Lower Confidenced Safe Life, Lognormal MLF Method
Nominal Confidence 90%

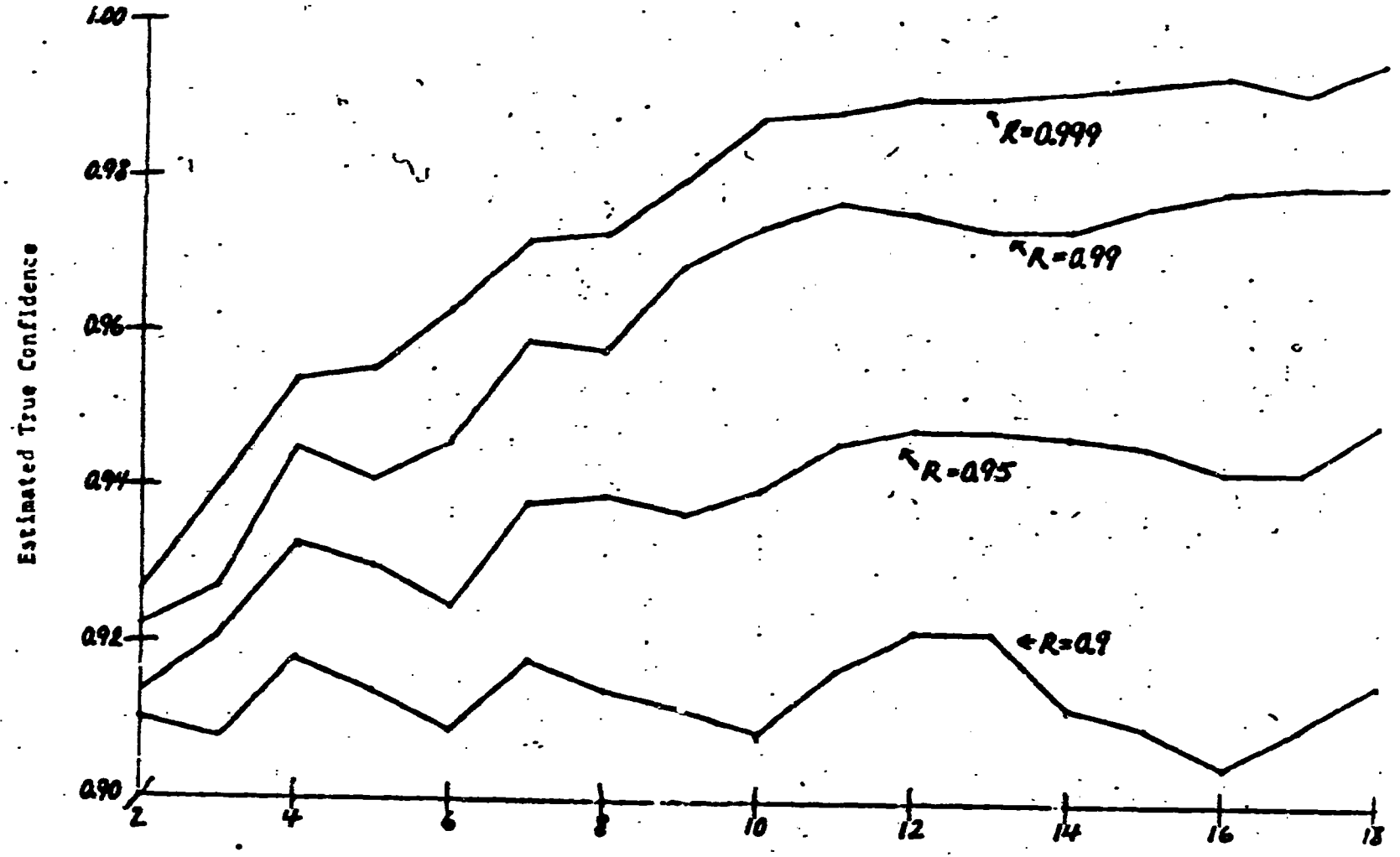


FIGURE 7
Run #1
Lower Confidenced Safe Life, Weibull BLIE Method
Nominal Confidence 90%

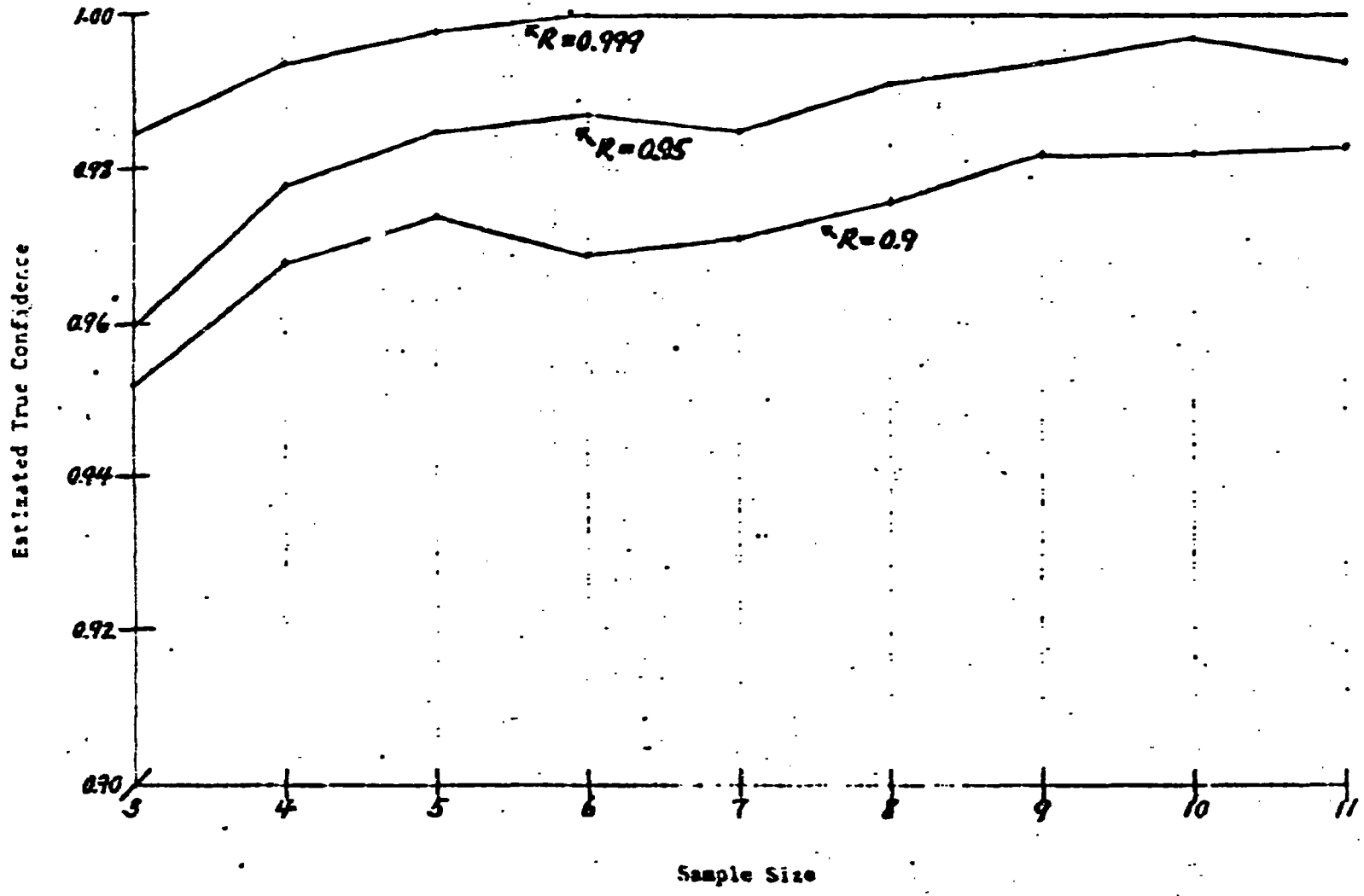
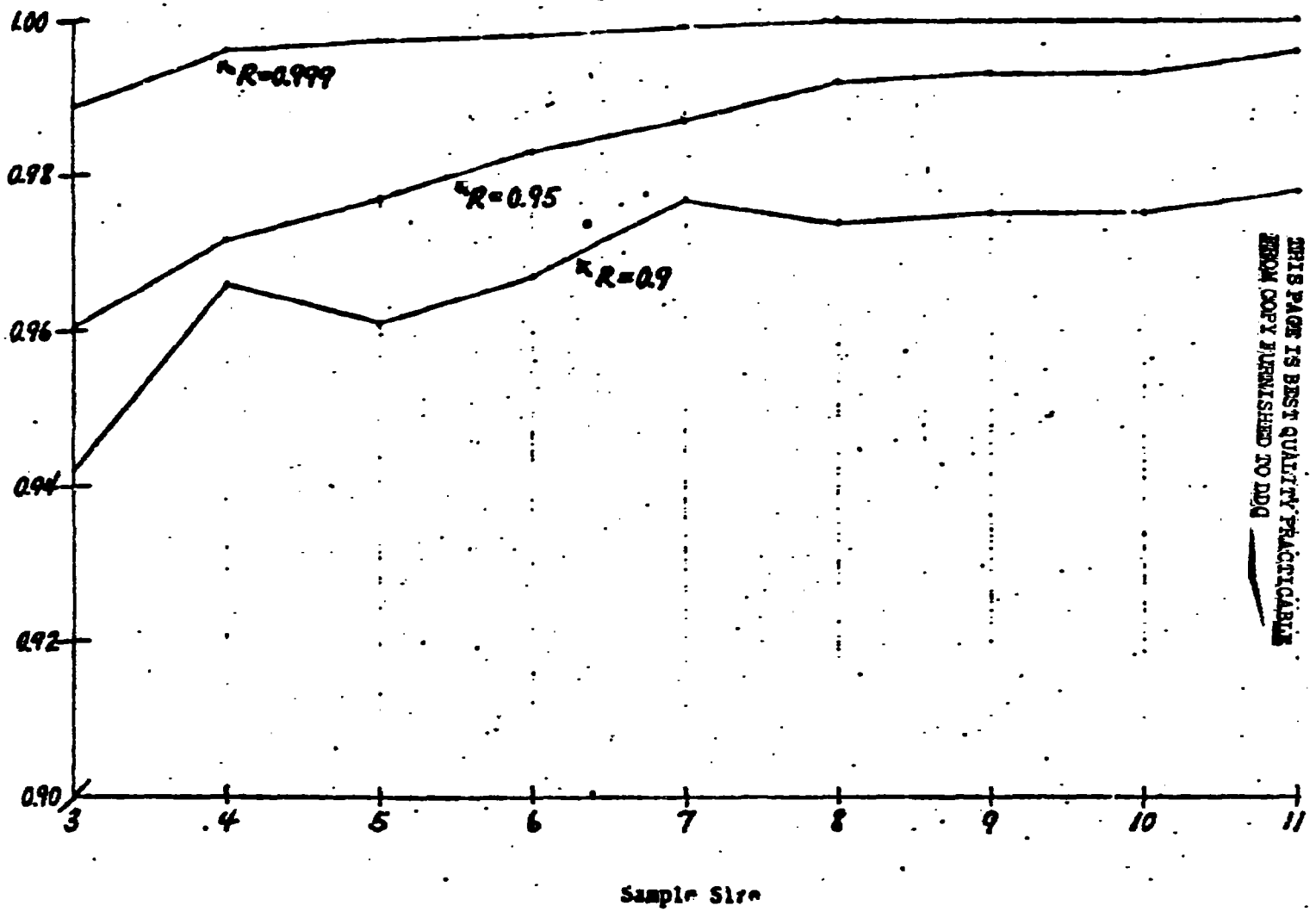


FIGURE 8

Run #3

Lower Confidenced Safe Life, Weibull BLIE Method

Nominal Confidence 90%



DOUBLE TESTING IN BINOMIAL DATA

G. R. Andersen
Battlefield Systems Integration, HQ DARCOM
Alexandria, Virginia 22333

ABSTRACT:

Suppose that a sample $w_1, w_2, \ldots, w_N$ of size N is drawn at random from some infinite population. Each element of this sample is to be classified as defective or non-defective according to one or more tests. To be specific, denote by T_0 a (preliminary) test which, although it classifies each element of the sample as defective or non-defective it may do so incorrectly. Denote by T_1 a (primary) test which also classified members of the sample, but does so without error. T_0 is often called a fallible test, while T_1 is called an infallible test. This paper discusses some aspects of the problem of estimating the probability p, that an element of the population is non-defective, on the basis of the sample $w_1, \ldots, w_N$, when all the members of this sample are subjected to the T_0 -test, but only a subsample of size n ($n < N$) is tested according to T_1. This problem has been referred to in the literature (e.g., Tenenbein,(1)) as "estimating from Binomial data with misclassifications."

For convenience, we will identify the symbols T_0 and T_1, representing the tests, with numerical valued functions which assign the value o to a defective and the value 1 to a non-defective sample item.

This paper will only be concerned with those tests T_0 which are necessary for T_1, in the sense that $T_0(w_i) = 0$ implies with probability one that

---

(1) Tenenbein, A., "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications", Journal of the Amer. Statist. Assoc., Sept 1970, Vol. 65

$T_1(w_1) = 0$. That is, passing the T_0 test is a necessary condition for passing the T_1 test. Examples of such tests are numerous; they are sometimes thought of as screening tests. In the field of reliability, think of attempting to judge the reliability of items in a stockpile by applying a cheap (nondestructive) test to a large sample followed by an expensive test applied to some of the items which pass the first test.

The difference then between this problem and the one studied in the Tenenbein paper is that here the size of the second sample, the subsample, is random. This is because here the subsample is drawn from those members of the original sample which pass the T_0 -test; whereas, in Tenenbein's paper the size of the subsample does not depend on the number of members in the initial sample which pass the fallible test.

Of course, if every sample member that passed a (necessary) T_0 -test was subjected to the T_1 -test, then the appropriate $\hat{p}$ would be the classical estimate. In the application that prompted this study both the fallible T_0 -test and the infallible T_1 -test were costly. Therefore, long before the test was run, the initial sample size N for the T_0 -test and a nonrandom subsample size, ν, for the T_1 -test had to be specified. Hence, the classical estimate of p would result only if, by chance, S_N, the number of T_0 -successes, did not exceed ν. However, the size of the subsample, in general, could only be stated to be $n = \text{minimum } (S_N, \nu)$. Therefore, the need arose to find a way of judging which values of N and ν to choose. As usual certain "precision-in-estimation" statements were required, so the question was, first of all, what is the best estimator of p in this setup and, secondly, what should N and ν be in order to guarantee that a certain level of precision will be achieved in estimation, subject to constraints on the costs of testing.

1. SUMMARY OF RESULTS: A precise statement of the problem considered here is given in Section 2* and the maximum likelihood estimation, $\hat{p}_{N\nu}$, of p together with some exact distribution results are given in Section 3. The relationship between the results of this note and those in A. Tenenbein's paper (1) is explained in Section 3, Remark 3.4, where the exact and asymptotic variance of $\hat{p}_{N\nu}$ is presented. (The exact variance is not obtained in the problem considered in (1)). Basically, in the context of Tenenbein's work, this amounts to showing how much the asymptotic variance of $\hat{p}_{N\nu}$ is reduced when the preliminary T_0 is necessary for T_1 (and so can misclassify in only one direction as opposed to both directions as in (1)); this reduction in the variance of $\hat{p}_{N\nu}$ cannot be obtained from Tenenbein. The asymptotic properties of the estimator and an associated statistic are derived in Section 4. Both random and nonrandomly standardized forms of the central limit theorem are given for $\hat{p}_{N\nu}$ and the statistic giving the exact number of successes in the second sample of size min (S_N, ν).

Approximate confidence intervals for p are derived in Section 5. Realizations of these confidence intervals have different functional forms depending on whether the observed number, S_N, of T_0 successes is greater than, or less than, or equal to, ν.

In Section 6, the required modification to A. Tenenbein's (1) results on sample size determination based on cost and precision are given for necessary tests.

*This article and the others noted below will appear in a paper which is being prepared for printing in a national journal.

ANALYSIS OF CENSORED SURVIVAL DATA[1]

Norman Breslow

University of Washington
Seattle, WA. 98195

ABSTRACT

Recent developments in the methodology of censored survival data are briefly reviewed. These include estimation of the survival curve, non-parametric tests for the comparison of r survival curves, tests for trend, and the regression analysis of survival data. A final section provides some additional references to the recent literature.

I. INTRODUCTION

Censored survival data arise in a wide variety of statistical investigations. In clinical trials one measures duration of response from start of treatment until relapse or death due to disease. Observations on response time are censored for those subjects still in remission at the study's end, as they are for patients lost-to-follow up during the course of the study. Animal carcinogenesis studies, such as used by the Food and Drug Administration to determine the safety of food additives, provide another example of censored survival data. Here the endpoint is the age at diagnosis

1. Paper prepared for the 23rd Conference on Design of Experiments in army research, development and testing held at the Naval Postgraduate School in Monterrey October, 1977.

of a particular kind of cancer, while censorship occurs because of death due to other causes including sacrifice. In tests of the reliability of missile components, failure times are measured from the start of testing until failure of the component, with censorship imposed by the failure of other components or the necessity of analyzing the data before all items have failed. While all of these types of data occur widely in practice, the presentation below emphasizes the clinical trial since that is the area of application with which the author is most familiar.

Figure 1 illustrates the results for the control group in a clinical trial reported by Heyn et al (1974). This trial was designed to investigate the effects of combined chemotherapy as an adjunct to surgery and radiation in the treatment of childhood rhabdomyosarcoma. The endpoint for analysis was the re-appearance of tumor, whether at the site of original treatment or through distant metastasis, so that children who remained disease-free at the time the data were analyzed had censored observations. In addition to the control arm IA, there were two groups of children who received the drugs actinomycin-D (ACT-D) and vincristine (VCR): group IB were concurrently randomized with the controls, both these groups having apparently had their tumors completely resected; while IIA consisted of patients with microscopic residual disease at the margin of surgical resection.

Interim data from all three arms are presented in Table 1. Note that the censored observations for arm IA, those in the column

labelled "disease-free", are smaller in the table than they are in the figure. This is because the figure was drawn from data computed at a later point in time, when additional follow up was available for most patients who had not already died.

Analysis of censored survival data such as presented in Table 1 has several goals. First one wants an estimate of the survival curve, the probability of surviving t units of time, for each of the comparison groups. Statistical tests are required next to determine whether the observed differences between the curves are real or are simply chance effects. If real, a method of quantifying the nature of the differences is desirable. Finally there may be available concomitant observations, including continuous measurements such as age at diagnosis, whose joint effects on survival are important to determine.

2. ESTIMATION OF SURVIVAL CURVES

The first step in the analysis of censored survival data is to form a series of 2 x r contingency tables as shown in Table 2. One table is formed for each of the K distinct times $0 = t_0 < t_1 < t_2 \cdots < t_K$ at which deaths (or failure, relapses, etc.) occur. The column totals n_{ik} refer to the total number of subjects in the i^{th} group who remain "at risk", i.e. alive and under observation, just prior to time t_k. The tabular entries d_{ik} and s_{ik} denote the numbers of these who die at t_k, and survive t_k, respectively. Table 3 illustrates the calculation of the first three such tables for the data in Table 1. Here $r = 3$ and $t_1 = 2$, $t_2 = 3$ and $t_3 = 9$ months. Note

that the tables for increasing t_k refer to a constantly diminishing population "at risk" as additional subjects die or are withdrawn (censored) from further observation.

Kaplan and Meier (1958) derived the maximum likelihood non-parametric estimate of the survival curve based on censored data. This may be calculated recursively from the entries in the 2 x r contingency tables shown in Table 2. Starting from $\hat{P}(t_0) = 1$, and suppressing the group index i, the recursion formula is

$$\hat{P}(t_k) = \hat{P}(t_{k-1})(\frac{s_k}{n_k}) \qquad (1)$$

for $k = 1,2,\ldots,K$. In other words, the probability of surviving past t_k is estimated as the probability of surviving past t_{k-1} times the conditional probability of surviving past t_k given survival to t_k. The curves remain flat between failure times. Because of the multiplicative structure (1), Kaplan and Meier refer to their estimate as the product limit (PL) estimate. In case there is no censorship in the data, this reduces to the familiar empirical distribution function.

Table 4 shows the calculation of the relapse-free survival curve from the interim data in Table 1 for treatment group IA. The corresponding curves calculated from final study data for all three treatment curves are shown in Figure 2. Numbers above each curve at annual intervals in this figure refer to numbers of patients still at risk in each group. These are an important means of judging the stability of the estimates. Such estimates can in fact be quite

unstable in the tail of the survival distribution where few subjects remain at risk.

A more formal method of judging the stability of the PL estimate is to calculate its variance. Kaplan and Meier provided a variance formula for their estimate, which may also be expressed recursively. Starting from $\hat{V}\{P(t_0)\} = 0$ this is defined by

$$\hat{V}\{P(t_k)\} = \hat{V}\{P(t_{k-1})\}\left(\frac{s_k^2}{n_k^2}\right) + \{\hat{P}(t_k)\}^2\left(\frac{d_k}{n_k s_k}\right). \qquad (2)$$

Breslow and Crowley (1974) show that in large samples $\hat{P}(t)$ is approximately normally distributed with mean equal to the true survival function $P(t)$ and a variance which may be estimated from (2). Note that neither $\hat{P}(t)$ nor $\hat{V}\{P(t)\}$ will change after the last uncensored response time in each group, even though additional subjects continue to be withdrawn from observation. In this region the estimated variance often does not accurately reflect the variability in the estimated survival, which will be substantial unless large numbers remain on study.

3. COMPARISON OF SURVIVAL CURVES: THE LOG RANK TEST

A very simple but powerful non-parametric test for the comparison of r survival curves with censored data may also be calculated from the series of 2 x r contingency tables shown in Table 2. This test exploits the fact that, under the null hypothesis of no difference in the underlying survival distributions and conditional upon fixed values for the marginal totals in the 2 x r table, the

vector $\underset{\sim}{d}_k = (d_{1k},\ldots,d_{rk})'$ of observed deaths at the t_k has an r-dimensional hypergeometric distribution. Consequently the null expectation of the number of deaths in group i at t_k is

$$e_{ik} = E(d_{ik}) = n_{ik}\left(\frac{D_k}{N_k}\right),$$

i.e. the number at risk in the i^{th} group times the death rate for all r groups combined. An illustration of this calculation is given in Table 3 for the interim study data. The covariance matrix $\underset{\sim}{V}_k$ of $\underset{\sim}{d}_k$ has, under the null hypothesis, an (i,j) component equal to

$$||\underset{\sim}{V}_k||_{ij} = \begin{cases} \dfrac{n_{ik}(N_k-n_{ik})D_kS_k}{N_k^2(N_k-1)}, & i = j \\ -\dfrac{n_{ik}n_{jk}D_kS_k}{N_k^2(N_k-1)}, & i \neq j \end{cases}$$

The main idea behind the test is to sum up the statistics calculated from each of the K 2 x r tables into a vector

$$\underset{\sim}{O} = \Sigma_k \underset{\sim}{d}_k$$

of observed numbers of deaths in each group, a vector

$$\underset{\sim}{E} = \Sigma_k \underset{\sim}{e}_k$$

of expected numbers of deaths, and a summary covariance matrix

$$\underset{\sim}{V} = \Sigma_k \underset{\sim}{V}_k .$$

Since the K 2 x r tables refer to overlapping sets of subjects they are not, strictly speaking, statistically independent.

Nevertheless Cox (1975) has shown that the conditional distributions for the observation vectors $\underset{\sim}{d}_k$ may be formally regarded as independent, so that $\underset{\sim}{V}$ is an appropriate covariance matrix for $\underset{\sim}{O}-\underset{\sim}{E}$. $\underset{\sim}{V}$ is a singular covariance matrix of dimension r-1. This corresponds to the fact that $\Sigma O_i = \Sigma E_i$ is the total number of deaths observed in all r groups. However by defining $\underset{\sim}{O}^*$ and $\underset{\sim}{E}^*$ to be the first r-1 components of $\underset{\sim}{O}$ and $\underset{\sim}{E}$, and by $\underset{\sim}{V}^*$ the (r-1) x (r-1) upper left hand corner of $\underset{\sim}{V}$, a test statistic for testing equality of the r survival curves is obtained as

$$T_1 = (\underset{\sim}{O}^*-\underset{\sim}{E}^*)'\underset{\sim}{V}^{*-1}(\underset{\sim}{O}^*-\underset{\sim}{E}^*).$$

This is distributed as chi-square on r-1 degrees of freedom under the null hypothesis.

The test T_1 was first proposed for survival data by Mantel (1966). Cox (1972) later derived it from likelihood theory under the proportional hazards (PH) model, in which the instantaneous death rates in the r groups are in constant ratio throughout the follow up period. (This model is discussed further below). Peto and Peto (1972), considering only the case r = 2, argued that it was an asymptotically efficient test under Cox's model and dubbed it the "log rank" test.

A conservative approximation to T_1 which requires no matrix inversion is given by the familiar chi-square formula

$$T_2 = \sum_{i=1}^{r} (O_i-E_i)^2/E_i.$$

While $T_2 \leq T_1$, in fact the two will be quite close provided that

there are few ties among the uncensored survival times (i,e. most of the D_k in Table 2 are unity) and that the patterns of censorship operating in the r groups are not grossly different. See Peto and Pike (1973) and Crowley and Breslow (1975) for discussion of this approximation.

Table 5 illustrates the manner of presentation of the summary and test statistics for the interim study data. Note the calculation of the ratio O/E of observed to expected numbers of deaths in each treatment group. These are very useful as measures of treatment effect since their ratios, e.g. $O_1/E_1 \div O_2/E_2$, estimate the relative death rates in the respective treatment groups (Breslow, 1975).

4. ALTERNATE WEIGHTING SCHEMES: THE GEHAN/BRESLOW TEST

The summary statistics O-E weight the observed differences $d_k - e_k$ in each table in a manner which is appropriate to the PH model already mentioned. However this is not the only possible weighting scheme. Multiplying the observed differences before summing by N_k, the total number of subjects in the k^{th} table, gives more weight to the earlier times t_k when larger numbers are at risk. This leads to the scores

$$W_i = \sum_{k=1}^{K} \{N_k d_{ik} - n_{ik} D_k\},$$

covariance matrix

$$\underset{\sim}{V}_w = \sum_{k=1}^{K} N_k^2 \underset{\sim}{V}_k$$

and test statistic

$$T_3 = \underset{\sim}{W}^{*\prime} \underset{\sim}{V}_w^{*-1} \underset{\sim}{W}^*,$$

where the asterisks (*) denote the corresponding r-1 dimensional quantities. A conservative approximation to this statistic not requiring matrix inversion is

$$T_4 = \sum_{i=1}^{r} W_i^2 / G_i,$$

where

$$G_i = \sum_{k=1}^{K} \{N_k D_k S_k n_{ik} / (N_{k-1})\}.$$

The scores W_i may also be obtained from a pairwise comparison of the observations in the i^{th} treatment group with those in the remaining r-1 groups. Each such pair is assigned the value +1 (or -1) according as the <u>true</u> survival time for the first pair member is known to be smaller than (or larger than) that for the second member. Ties or indeterminate comparisons due to censorship are assigned 0 values. Gehan (1965) suggested the use of such scores for the comparison of two samples (r=2), noting that the resulting test T_4 essentially reduced to the familiar Wilcoxon rank sum test when there was no censorship. Breslow (1970) extended this work to the case of r>2 samples, proposing also covariance matrix $\underset{\sim}{V}_w$ and the statistic T_3. This latter statistic is valid for situations where the patterns of censorship operative in the r treatment groups are unequal, as in animal carcinogenesis studies where there is differential toxic mortality. The conservative approximation T_4 is

strictly valid only where there is equality of censorship.

In practice the tests T_1 and T_3 often yield rather similar numerical values. However this is not always true and some comments on the proper interpretation when only one statistic is significant are in order. Since T_3 weights early values more heavily, it may achieve significance when there is an early separation between the survival curves which later come together or even cross over. T_1 gives more weight to the latter part of the curves, and would detect differences in the curves which only appeared later on. Such behavior often indicates an interaction between treatment and time on the instantaneous death rates, which is worthy of investigation in its own right.

5. TESTING FOR TREND

In many situations the r treatment groups will correspond to r different levels or dosages of some quantitative variable x, say $x_1<x_2<\ldots<x_r$. In such cases the global chi-square tests T_1 and T_3 are notoriously lacking in power. One would prefer instead a single degree of freedom test for trend in survival with increasing dose.

Fortunately, such tests for trend are readily calculated from the summary statistics already at hand. In the case of the $\underline{O}$ and $\underline{E}$ analysis, one uses

$$T_5 = \frac{\{\underline{x}'(\underline{O}-\underline{E})\}^2}{\underline{x}'\underline{V}\underline{x}}$$

as a single degree of freedom chi-square for a linear trend of O-E with x, and

$$T_6 = T_1 - T_5$$

as a chi-square on r-2 degrees of freedom for deviations from linearity (Tarone, 1975).

Similarly, when using the W scores,

$$T_7 = \frac{\{\underset{\sim}{x}'\underset{\sim}{W}\}^2}{\underset{\sim}{x}'\underset{\sim}{V}_{\underset{\sim}{W}}\underset{\sim}{x}}$$

provides a test for linear trend of these scores with x and

$$T_8 = T_3 - T_7$$

a test for deviations from linearity.

6. ADJUSTMENT BY STRATIFICATION

When it is thought that the r comparison groups may differ with respect to factors which influence survival, an adjusted or stratified analysis which corrects for the confounding effects of such variables is in order. Such an analysis is carried out very simply, as follows.

First, divide the population into strata which are more or less homogeneous internally with respect to the confounding variable or variables. Of course there is a limitation on the number of confounders which may be simultaneously accommodated in this fashion since if strata become too large in number, and small in size, a large loss of comparative information may result.

Next, perform separate survival analyses within each stratum. This means calculation of the survival curves and especially the summary statistics $\underset{\sim}{O}$, $\underset{\sim}{E}$, $\underset{\sim}{V}$, $\underset{\sim}{W}$ and $\underset{\sim}{V}_w$ defined earlier. These summary statistics are then cumulated by simple addition over strata.

Finally, calculate the adjusted test statistics T_1, T_2, T_5, and T_6 just as before using the cumulated summary statistics $\underset{\sim}{O}$, $\underset{\sim}{E}$ and $\underset{\sim}{V}$ in place of the stratum specific ones. Likewise calculate T_3, T_4, T_6 and T_7 using the adjusted or cumulated $\underset{\sim}{W}$ and $\underset{\sim}{V}_w$.

7. REGRESSION ANALYSIS OF SURVIVAL DATA: THE PH MODEL

If the number of confounding concomitant variables is very large, the stratified analysis approach quickly breaks down due to large numbers of strata with just one or a few subjects in each. Furthermore, it may be of interest to quantify the relationship between survival times and concomitant variables, some of which may be continuous. This situation calls out for some kind of regression model.

A usual (normal theory) regression approach would specify that the survival times, or some transform such as their logarithm, were equal to a linear combination of the concomitant variables plus some random error term. While not impossible, the generalization of such models for use with censored data may be quite awkward and computationally involved. Thus considerable interest was aroused by Cox (1972) when he proposed an alternative type of regression model formulated in terms of the effect of the regression

variables on death rates rather than times of death. Statistical analysis under this model turned out to be much more tractable than for those others proposed earlier. Furthermore, it avoided any parametric assumptions about the shape of the underlying survival curve.

Cox's model is defined in terms of the time t specific death rate or hazard function $\lambda(t|\underset{\sim}{z})$ for an individual having a p-vector of covariates $\underset{\sim}{z}$. Specifically he assumes

$$\lambda(t|\underset{\sim}{z}) = \exp(\underset{\sim}{\beta}'\underset{\sim}{z})\lambda_0(t) ,$$

where $\underset{\sim}{\beta}$ is an unknown p- vector of parameters (regression coefficients), while $\lambda_0(t)$ is the unknown hazard or death rate function for an individual with a standard ($\underset{\sim}{z}=\underset{\sim}{0}$) set of covariates. A consequence of this model is that the ratio of hazard functions for two individuals with different sets of covariates,

$$\frac{\lambda(t|\underset{\sim}{z}_1)}{\lambda(t|\underset{\sim}{z}_2)} = \exp\{\underset{\sim}{\beta}'(\underset{\sim}{z}_1-\underset{\sim}{z}_2)\},$$

does not depend on time, whence the title proportional hazards (PH) model.

Several authors (Cox, 1972, 1975; Kalbfleisch and Prentice, 1973; Breslow, 1974, 1975) have developed the likelihood analysis of the PH model from rather distinct points of view. Providing that there are no ties in the uncensored data, all derive for the log-likelihood function of $\underset{\sim}{\beta}$ the expression

$$L(\underset{\sim}{\beta}) = \sum_{k=1}^{K} \{\underset{\sim}{\beta}' \underset{\sim}{z}_k - \ell n \sum_{j \in R(t_k)} \exp(\underset{\sim}{\beta}' \underset{\sim}{z}_j)\},$$

where $R(t_k)$ is the risk set of subjects still alive and under observation at $t_k - 0$; $\underset{\sim}{z}_k$ is the covariate vector for the individual who dies at t_k; and the outer summation is over all K true (uncensored) times of death. In case of ties, the three approaches yield somewhat different likelihoods; see also Efron (1977).

Taking the vector of first partial derivative of L, setting equal to $\underset{\sim}{0}$ and solving the resulting non-linear equations yields a maximum likelihood estimate $\hat{\underset{\sim}{\beta}}$ for the regression coefficients. A covariance matrix for this estimate is obtained in the usual fashion by inversion of the negative of the matrix of second partials of L. The integral

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du$$

defines the cumulative hazard function for the standard covariate set. Once $\hat{\underset{\sim}{\beta}}$ is obtained this may be estimated by

$$\hat{\Lambda}_0(t) = \sum_{t_k \leq t} \{ \sum_{j \in R(t_k)} \exp(\hat{\underset{\sim}{\beta}}' \underset{\sim}{z}_j)\}^{-1},$$

where the outer summation is again over true survival times t_k less than or equal to t. The corresponding estimate of the survival function

$$P_0(t) = \exp\{-\Lambda_0(t)\}$$

is

$$\hat{P}_0(t) = \prod_{t_k \leq t} \left\{ 1 - \frac{1}{\sum_{j \in R(t_k)} \exp(\hat{\beta}' z_j)} \right\} .$$

Notice that when $\hat{\beta} = 0$ this reduces to the PL estimate of Kaplan and Meier, calculated from the entire set of observations considered as one homogeneous sample.

8. FURTHER READING

Much of the above material is presented in greater detail in my review article (Breslow, 1975) on the PH model and its applications to survival data. Some additional applications of this model to epidemiologic data are outlined in a forthcoming paper (Breslow, 1978). Peto, Pike, et al (1976, 1977) present a thorough discussion of the use of the model in the design and analysis of clinical trials.

A computer program for calculating the PL estimate and all the test statistics presented in sections 2-5 above is available from Thomas, Breslow and Gart (1977).

Several authors have pointed out that the W scores defined in section 4 do not lead to the most efficient generalization of Wilcoxon's test to censored data. They all propose essentially the same statistic as an alternate generalization. See Efron (1965), Peto and Peto (1972), and Prentice (1978).

A comparison of the efficiencies of the test statistics using Monte Carlo techniques is made by Lee et al (1975). Efron (1977)

discusses the efficiency of the $\ln$-likelihood function L for the PH model from a more abstract viewpoint.

Extensions of the PH regression model for use with grouped or heavily tied data are discussed by Cox (1972), Kalbfleisch and Prentice (1973), Thompson (1977) and Prentice and Gloeckler (1978).

9. REFERENCES

Breslow, N. (1970): A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 57, 579-594.

Breslow, N. (1974): Covariance analysis of censored survival data. Biometrics 30, 89-100.

Breslow, N. (1975): Analysis of survival data under the proportional hazard model. Int. Statist. Rev. 43, 43-57.

Breslow, N. (1978): The proportional hazards model: applications in epidemiology. Comm. Statist (to appear).

Breslow, N. and Crowley, J. (1974): A large sample study of the life table and product limit estimates under random censorship. Ann. Statist 2, 437-453.

Cox, D. R. (1972): Regression models and life tables (with discussion). J. Roy Statist. Soc. B 34, 187-220.

Cox, D. R. (1975): Partial likelihood. Biometrika 62, 269-279.

Crowley, J. and Breslow, N. (1975): Remarks on the conservatism of $\Sigma(O-E)^2/E$ in survival data. Biometrics 31, 957-961.

Efron, B. (1965): The two sample problem with censored data. Proc. Fifth Berkeley Symp. 4, 831-854.

Efron, B. (1977): The efficiency of Cox's likelihood function for censored data. J. Amer. Statist. Assoc. 72, 557-565.

Gehan, E. (1965): A generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 52, 203-223.

Heyn, R., Holland, R., Newton, W. A., Tefft, M., Breslow, N., and Hartmann, J. (1974). The role of combined chemotherapy in the treatment of rhabdomyosarcoma in children. Cancer 34, 2128-2141.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267-278.

Kaplan, E. L. and Meier, P. (1958): Non-parametric estimation from incomplete observation. J. Amer. Statist. Assoc. 53, 457-481.

Lee, E. T., Desu, M. M., and Gehan, E. (1975): A Monte Carlo study of the power of some two sample tests. Biometrika 62, 425-432.

Mantel, N. (1966): Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chem. Reports 50, 163-170.

Peto, R. and Peto, J. (1972): Asymptotically efficient rank invariant test procedures. J. Roy. Statist. Soc. A 135, 185-206.

Peto, R. and Pike, M. C. (1973): Conservatism of the approximation $\Sigma(O-E)^2/E$ in the log rank test for survival data or tumor incidence data. Biometrics 29, 579-583.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976, 1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Brit. J. Cancer 34, 585-612. II. Analysis and examples. Brit. J. Cancer 35, 1-37.

Prentice, R. L. (1978): Linear rank tests with right censored data. Biometrika (to appear).

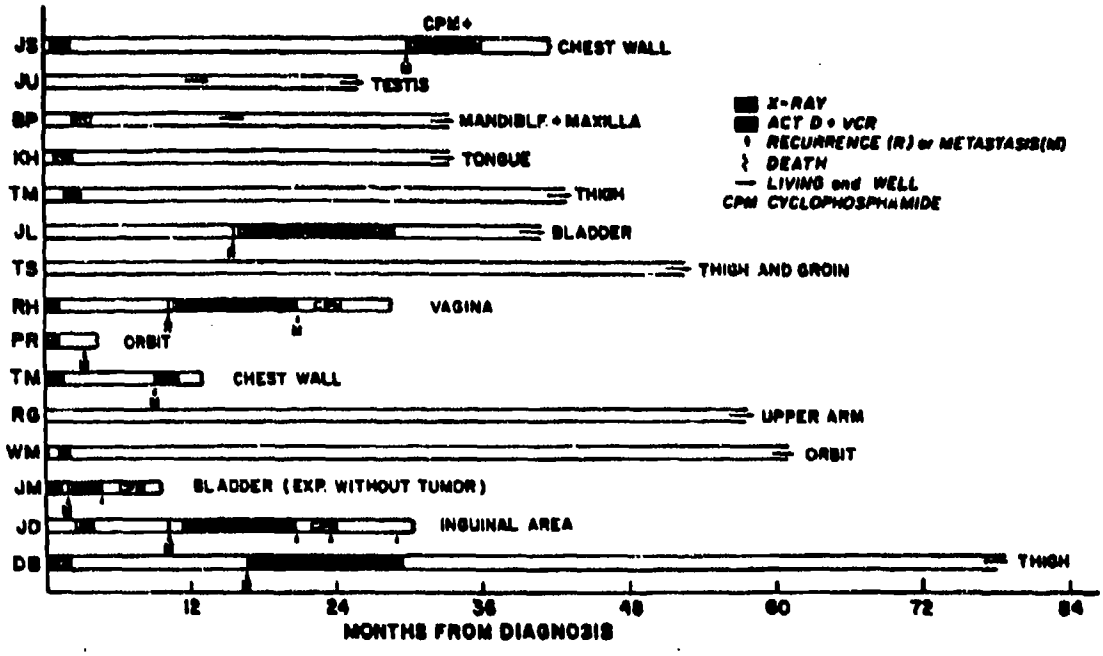
Prentice, R. L. and Gloeckler, L. (1978). Regression analysis of grouped survival data with application to the study of breast cancer survival times. Biometrics (to appear).

Tarone, R. E. (1975). Tests for trend in life table analysis. Biometrika 62, 679-682.

Thomas, D. G., Breslow, N. and Gart, J. (1977): Trend and homogeneity analyses of proportions and life table data. Computers Biomed. Research 10, 373-381.

Thompson, W. A. (1977): On the treatment of grouped observations in life studies. Biometrics 33, 463-470.

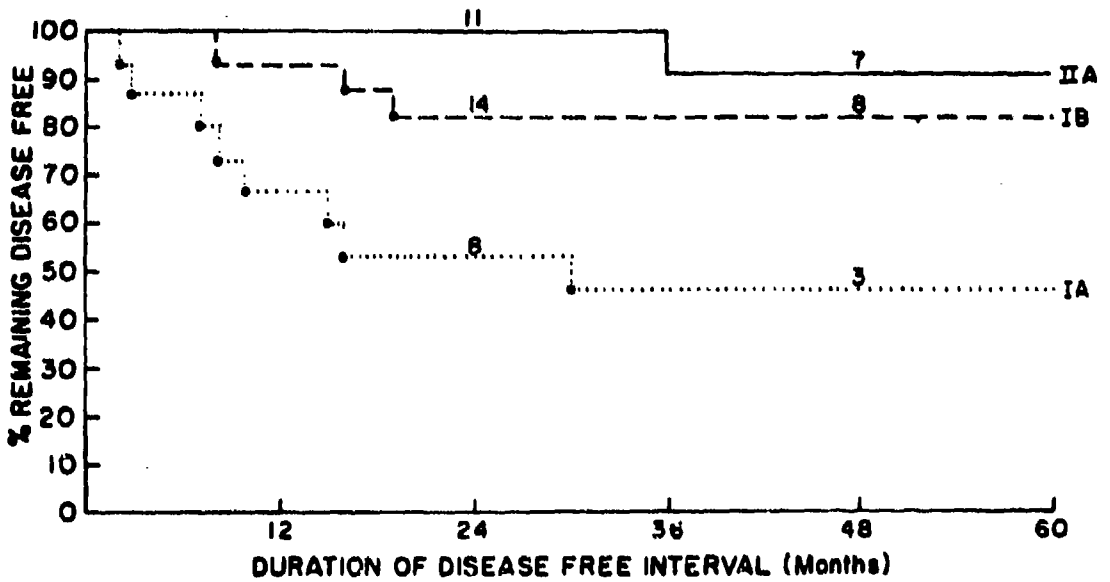
FIGURE 1



Course of patients in Part IA, control group

The above figure, as well as the next one, were first published in Volume 34:2128-2141, 1974 of the journal CANCER. They appeared in an article by Heyn, R., Holland, R., Newton, W. A., Tefft, M., Breslow, N., and Hartmann, J., entitled "The Role of Combined Chemotherapy in the Treatment of Rhabdomyosarcoma in Children". We appreciate the fact that the editor, Dr. J. E. Rhoads of CANCER and Dr. Ruth Heyn gave their permission to reproduce Figures 1 and 2 in these Proceedings.

FIGURE 2



The duration of the disease-free interval in patients from Part IA (control), IB (treated), and IIA (microscopic residual, treated). Shown above each curve at 24 and 48 months are the numbers of patients known to be disease-free after those time periods.

TABLE I

INTERIM DATA FROM CCG614: MONTHS FROM START OF TREATMENT TO RELAPSE OR LAST OBSERVATION FOR THREE TREATMENT GROUPS

TUMOR COMPLETELY RESECTED				MICROSCOPIC RESIDUAL	
SURG + X-RAY		SURG + X-RAY + ACT-D + VCR		SURG + X-RAY + ACT-D + VCR	
IA		IB		IIA	
Relapsed	Disease-Free	Relapsed	Disease-Free	Relapsed	Disease-Free
2	12	9	12	37	25
3	15	16	19		28
9	18	19	20		29
10	24		20		38
10	36		24		42
15	40		24		45
16	45		30		47
30			31		48
			34		50
			42		52
			44		
			53		
			59		
			62		

TABLE 2

FORMATION OF 2 x r CONTINGENCY TABLES COMPARING DEATH RATES AMONG r TREATMENT GROUPS AT EACH DISTINCT TIME OF DEATH

PATIENTS FOLLOWED TO TIME t_k

	Treatment Group				
	1	2		r	Totals
Deaths (at t_k)	d_{1k}	d_{2k}		d_{rk}	D_k
Survivors	s_{1k}	s_{2k}		s_{rk}	S_k
Total "at risk"	n_{1k}	n_{2k}		n_{rk}	N_k

TABLE 3

ILLUSTRATION OF 2 x r TABLES FOR

CCG614 INTERIM STUDY DATA

t = 2 months

	IA	IB	IIA	Total
Relapsed	1	0	0	1
Disease-Free	14	17	11	42
"At Risk"	15	17	11	43
Expectations:	0.349	0.395	0.256	1.000

t = 3 months

	IA	IB	IIA	Total
Relapsed	1	0	0	1
Disease-Free	13	17	11	41
"At Risk"	14	17	11	42
Expectations:	0.333	0.405	0.262	1.000

t = 9 months

	IA	IB	IIA	Total
Relapsed	1	1	0	2
Disease-Free	12	16	11	39
"At Risk"	13	17	11	41
Expectations:	0.634	0.829	0.537	2.000

TABLE 4

ESTIMATION OF SURVIVAL CURVE FOR GROUP IA

BY METHOD OF KAPLAN AND MEIER (1958)

Month t_k	Number At Risk n_{1k}	Number Surviving s_{1k}	Conditional Probability $\hat{p}(t_k)$	Survival Probability $\hat{P}(t_k)$
2	15	14	0.933	0.933
3	14	13	0.929	0.866
9	13	12	0.923	0.799
10	12	10	0.833	0.666
15	9	8	0.888	0.592
16	7	6	0.857	0.507
30	4	3	0.750	0.381

TABLE 5
SUMMARY STATISTICS FOR CCG614 INTERIM DATA

	Treatment Group		
	IA	IB	IIA
No. of pts. (N)	15	17	11
Relapses observed (O)	8	3	1
Relapses expected (E)	3.11	4.99	3.90
O/E	2.57	0.60	0.26

T_1 = 11.10, 2 d.f. , p = 0.004

T_2 = 10.77, 2 d.f. , p = 0.005

T_3 = 11.41, 2 d.f. , p = 0.003

THE JACKKNIFE: SURVEY AND APPLICATIONS

Rupert G. Miller, Jr.
Stanford University

1. Introduction

The jackknife technique is becoming so familiar to statisticians that it is almost not necessary to reintroduce it with each article. However, to establish notation and to aid any reader encountering the jackknife for the first time, a brief definition is given.

Let $Y_1, \ldots, Y_n$ be n independent random variables identically distributed according to the distribution function F, which depends on an unknown parameter θ. The aim of the statistical analysis is to estimate or test θ. The jackknife technique can be applied to any estimation procedure which for any sample size gives a point estimate $\hat{\theta}(Y_1, \ldots, Y_n) = \hat{\theta}$ of θ. The ith deleted estimate is the estimate obtained by applying the estimation procedure to the sample with the ith random variable removed, i.e.,

$$\hat{\theta}_{-i} = \hat{\theta}(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n) . \tag{1}$$

Corresponding to the ith deleted estimate is the ith pseudo-value

$$\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i} . \tag{2}$$

The jackknifed estimate of θ is the average of these pseudo-values

$i = 1,\ldots,n$ as each random variable is deleted in turn, i.e.,

$$\tilde{\theta} = \frac{1}{n}\sum_{1}^{n}\tilde{\theta}_i = n\hat{\theta} - \frac{(n-1)}{n}\sum_{1}^{n}\hat{\theta}_{-i} \quad . \tag{3}$$

If n is large, a variation of the jackknife can be invoked to save on the computation time. The modification is to divide the total sample into g groups of size k each $(n = g \times k)$, and to successively delete each one of the groups rather than single random variables. The ith deleted estimate is now

$$\hat{\theta}_{-i} = \hat{\theta}(Y_1, \ldots, Y_{(i-1)k}, Y_{ik+1}, \ldots, Y_n) \quad , \tag{4}$$

and the corresponding pseudo-value is

$$\tilde{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_{-i} \quad . \tag{5}$$

The jackknifed estimate is still the average of the pseudo-values, i.e.,

$$\tilde{\theta} = \frac{1}{g}\sum_{1}^{g}\tilde{\theta}_i = g\hat{\theta} - \frac{(g-1)}{g}\sum_{1}^{g}\hat{\theta}_{-i} \quad . \tag{6}$$

Quenouille (1949, 1956) introduced the jackknife as a method of bias reduction, and this aspect is surveyed in Section 2. On the other hand, Tukey (1958) saw the jackknife as a device for robust interval estimation, and developments along this line are summarized in Section 3. Robust point estimation has also been a rapidly developing field in recent years,

and the connection between it and the jackknife through the influence function is explored in Section 4. Application of the jackknife to various statistical problems is illustrated in Section 5.

In my 1974 review article, almost all methodological papers on the jackknife published before or during 1973 were listed. The reader is referred to this earlier article for an extensive bibliography of papers from that era. A few papers were missed (Collins (1970), Cronbach et al. (1972), Hollander and Wolfe (1973), Mosteller (1971), and Pennel (1972)), and these are included in the references to this paper. The final section of this paper is a bibliography of all methodological papers on the jackknife published between 1974 and 1977 which have come to my attention.

2. Bias Reduction

Quenouille (1956) pointed out that if the estimator $\hat{\theta}$ for a sample of size n has the expectation

$$E(\hat{\theta}) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots , \tag{7}$$

then the jackknifed estimator eliminates the leading bias term, i.e.,

$$E(\tilde{\theta}) = \theta + 0 + \frac{b_1}{n^2} + \ldots . \tag{8}$$

This idea was generalized in Schucany, Gray, and Owen (1971). Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ with expectations of the form

$$E(\hat{\theta}_1) = \theta + f_1(n)b(\theta) ,$$
$$E(\hat{\theta}_2) = \theta + f_2(n)b(\theta) . \tag{9}$$

Then, the estimator

$$\tilde{\theta}^* = \det\begin{pmatrix} \hat{\theta}_1 & \hat{\theta}_2 \\ f_1(n) & f_2(n) \end{pmatrix} \Big/ \det\begin{pmatrix} 1 & 1 \\ f_1(n) & f_2(n) \end{pmatrix} \tag{10}$$

is an exactly unbiased estimator for θ, i.e., $E(\tilde{\theta}^*) = \theta$.

The estimator $\tilde{\theta}^*$ is called the generalized jackknife. It includes the standard jackknife as a special case with the identifications

$$\hat{\theta}_1 = \hat{\theta} , \quad \hat{\theta}_2 = \frac{1}{n}\sum_1^n \hat{\theta}_{-i} ,$$
$$f_1(n) = \frac{1}{n} , \quad f_2(n) = \frac{1}{n-1} . \tag{11}$$

By extending the definition (10) to include three or more estimators, the second or higher order bias terms in the expansion (7) can be exactly eliminated by the generalized jackknife. For more detail on these generalizations the reader is referred either to Schucany, Gray, and Owen (1971) or Gray and Schucany (1972).

3. Robust Interval Estimation

In his 1958 abstract Tukey proposed that, as a method for robust confidence interval construction, $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ could be treated as n

independently, identically distributed random variables with mean θ. The pseudo-values are clearly identically distributed provided $Y_1,\ldots,Y_n$ are independently, identically distributed and the estimators $\hat{\theta}(Y_1,\ldots,Y_n)$ treat the random variables symmetrically, so the major question hinges on whether or not the pseudo-values behave as though they are approximately independent. In this direction a great deal of research has been devoted to learning when

$$\frac{\sqrt{n}(\tilde{\theta} - \theta)}{\sqrt{\frac{1}{n-1}\Sigma_1^n(\tilde{\theta}_i - \tilde{\theta})^2}} \xrightarrow{\mathcal{L}} N(0,1) \qquad (12)$$

as $n \to \infty$.

If $\tilde{\theta}$ is asymptotically normally distributed as indicated in (12), then the interval estimate $\tilde{\theta} \pm g^{\alpha/2}(\Sigma_1^n(\tilde{\theta}_i - \tilde{\theta})^2/n(n-1))^{1/2}$, where $g^{\alpha/2}$ is the $1 - (\alpha/2)$ percentile point of the standard normal distribution, gives a robust way of testing or bounding θ. Folklore says that in place of $g^{\alpha/2}$ it is better to use $t_{n-1}^{\alpha/2}$, where $t_{n-1}^{\alpha/2}$ is the percentile point from a t distribution with $n-1$ degrees of freedom. The rationale for this folklore ostensibly stems from a strong belief in the approximate independence and normality of the pseudo-values, but, with just one exception, all papers on distribution theory for the jackknife have focused on establishing asymptotic normality. In fact, it is often the case in practice that the jackknife t intervals are conservative (i.e., wider than necessary for the nominally listed coverage), so it may be a better policy to use the normal critical constant.

Under what circumstances is the Tukey proposal valid, and when is it invalid? Miller (1964, 1968) proved that (12) is valid if the estimator is a smooth function of a sample mean or means, i.e., $\hat{\theta} = f(\bar{Y})$. Estimators of this type include transformed averages and variances. This approach was extended by Arvesen (1969) and Arvesen and Layard (1975) to functions of U-statistics $\hat{\theta} = f(U)$ in order to handle variance component problems. The proposition (12) is also true for functions of regression estimators $f(\hat{\underset{\sim}{\beta}})$ in the general linear regression model $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$ as shown by Miller (1974b).

Brillinger (1964) took a different limiting approach by holding g fixed and letting $k \to \infty$ for the grouped jackknife (4) - (6). His proof shows that if $\hat{\theta} = f(\hat{\phi})$ where $\hat{\phi}$ is a root of the likelihood equation

$$\sum_{1}^{n} \frac{\partial \log p(y_i, \phi)}{\partial \theta} = 0 , \tag{13}$$

then as $k \to \infty$

$$\frac{\sqrt{g}(\tilde{\theta} - \theta)}{\sqrt{\frac{1}{g-1}\Sigma_1^g(\tilde{\theta}_i - \tilde{\theta})^2}} \xrightarrow{d} t_{g-1} . \tag{14}$$

This is the lone instance in which the t distribution, rather than the normal, has been established as the approximating distribution. Typically, however, one would prefer to have $g >> k$, so the asymptotics do not establish the t approximation in many problems.

Since the maximum likelihood estimator has a well-established asymptotic distribution theory involving Fisher's information, the need for

jackknifing in this context has been questioned. However, in recent work Reeds (1977) has answered these criticisms. Firstly, he has proved the asymptotic normality for $\tilde{\theta}$ when $g \to \infty$ with $k = 1$, and secondly he has shown that the jackknife gives the correct asymptotic variance for $\tilde{\theta}$ and $\hat{\theta}$ even if the model is incorrect. The Fisher information does not do this because it is computed theoretically on the basis of the assumed density $p(y,\phi)$. If the model is incorrect it may not be clear what $\hat{\phi}$ is estimating, but in problems like the location of a symmetric distribution it will be. Reeds' work applies as well to more general M-estimators and in this regard the reader should also see Brillinger (1976).

The basic ingredient needed in the estimator $\hat{\theta}$ for the Tukey proposal to work is for it to be a smooth function of each Y_i. The proofs depend upon power series expansions of the estimator in each of the random variables. The common motif of the estimators mentioned above is that asymptotically they are all functions of glorified means. By this I mean that they are asymptotically equivalent to a function of a (possibly weighted) sum of independent, identically distributed random variables. This is true for U-statistics, regression coefficients, maximum likelihood estimators, M-estimators, etc. In the cases where the jackknife is known not to work, such as for the median or other percentile estimators, this is not true.

Three remarks seem in order before closing the discussion on the use of the jackknife for robust interval estimation.

The first is that based on mean square error considerations in ratio and other problems and on uniqueness criteria, the choice $g = n$ seems

best. (See, for example, Rao (1965) and Rao and Webster (1966).) However, Hinkley (1977b) tentatively suggests that the accuracy of the t distribution approximation may be improved by taking k larger than one and that there is little loss in efficiency for small $k > 1$. The argument for this is that the skewness and kurtosis of $\Sigma_{(i-1)k+1}^{ik} Y_i/k$, which is the dominant linear term in $\tilde{\theta}_i$, may be considerably improved by selecting k slightly larger than one.

The second remark is that jackknifing does <u>not</u> correct for outliers. The reader should not confuse large sample robustness of the jackknife procedure for any underlying distribution with resistance to contaminating observations in small or moderate size samples. In fact, the jackknife appears to be rather sensitive to aberrant values. This sensitivity may make it a useful device for detecting outliers in complicated estimation problems. Trimming of the pseudo-values or application of other robust procedures to them may be a good way of correcting for the outliers. Hinkley (1976, 1977a) has started an investigation of this in the context of correlation coefficient estimation.

The third and final remark is that if you are going to use a grouped jackknife with $k > 1$, random selection of the groups is probably the most sensible approach when the $Y_1, \ldots, Y_n$ are identically distributed, but if the underlying random variables are not identically distributed, then one presumably has the opportunity to do better. In particular, the regression situation comes to mind. It may be possible to exploit the pattern in the independent variable vectors $\underset{\sim}{x}_i$ associated with the Y_i to form groups which give the jackknifed estimator a better mean squared

error or improved robustness in interval estimation. Hinkley and Miller have some inconclusive results along this line, but at this point it is difficult to see what the general principle of selection should be.

4. Connection with Influence Functions

To establish the connection between the jackknife and the influence function it is necessary to give a brief description of the latter. von Mises (1947) introduced the idea in his study of differentiable statistical functions, but it remained relatively unnoticed for two decades until investigators interested in robust estimation uncovered its usefulness (see Hampel (1974)).

In many estimation problems the unknown parameter θ can be considered to be a function $\theta = T(F)$ of the underlying distribution F, and its estimator $\hat{\theta}$ to be the same function of the sample distribution function. For example, in the case of the mean, $\theta = \int y dF(y)$ and $\hat{\theta} = \int y dF_n(y) = \Sigma_1^n Y_i/n$. The influence function $I(y,\theta)$ measures the amount of change in $T(F)$ for an infinitesimal change in the weight assigned by F to y. It is like a partial derivative of T with respect to a change in F at coordinate y. Specifically,

$$I(y,\theta) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_y) - T(F)}{\varepsilon} , \qquad (15)$$

where δ_y is the distribution function which places mass one at y.

Under regularity conditions the function T can be expanded in a series involving (15) and higher order derivatives. Specifically,

$$T(G) = T(F) + \int I(y,\theta)dG(y) + \ldots \quad . \tag{16}$$

In the case where G is the sample distribution function F_n, the expansion (16) and the identifications $\hat{\theta} = T(F_n)$, $\theta = T(F)$ give

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{1}^{n} I(Y_i,\theta) + \ldots \quad . \tag{17}$$

The random variables $I(Y_i,\theta)$ are independently and identically distributed with mean $\int I(y,\theta)dF(y) = 0$ and variance $\int I^2(y,\theta)dF(y)$. Since the higher order terms in (17) are $o_p(n^{-\frac{1}{2}})$ under the regularity conditions, the asymptotic distribution of $\hat{\theta}$ is given by

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \int I^2(y,\theta)dF(y)) \; . \tag{18}$$

The connection between the jackknife and the influence function is that the pseudo-values give finite difference sample estimates of the influence function. For

$$\varepsilon = -\frac{1}{n-1}, \quad 1-\varepsilon = \frac{n}{n-1}, \quad \text{and} \quad F = F_n \; , \tag{19}$$

the quantity $(1-\varepsilon)F + \varepsilon\delta_y$ at $y = Y_i$ becomes

$$(1-\varepsilon)F_n + \varepsilon\delta_{Y_i} = \frac{n}{n-1}F_n - \frac{1}{n-1}\delta_{Y_i} = F_{n-1,-i} \; , \tag{20}$$

where $F_{n-1,-i}$ is the sample distribution function based on $n - 1$

observations with the ith observation deleted. If the finite difference sample estimate of $I(y,\theta)$ at Y_i for $\theta = \hat{\theta} = T(F_n)$ is defined by

$$\hat{I}(Y_i,\hat{\theta}) = \left.\frac{T((1-\varepsilon)F + \varepsilon\delta_{Y_i}) - T(F)}{\varepsilon}\right|_{F=F_n,\varepsilon=-1/(n-1)} , \qquad (21)$$

then it follows that

$$\begin{aligned} \tilde{\theta}_i &= n\hat{\theta} - (n-1)\hat{\theta}_{-i} , \\ &= \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{-i}) , \\ &= \hat{\theta} + \hat{I}(Y_i,\hat{\theta}) , \end{aligned} \qquad (22)$$

because $\hat{\theta}_{-i} = T((1-\varepsilon)F_n + \varepsilon\delta_{Y_i})$.

If the influence function is sufficiently smooth so that $\hat{I}(y,\hat{\theta})$ converges to $I(y,\theta)$ for all y as $n \to \infty$, then each pseudo-value $\tilde{\theta}_i$ is approximately $\theta + I(Y_i,\theta)$. This means the jackknife will be behaving correctly asymptotically because $\tilde{\theta}$ will be asymptotically normally distributed with mean θ and variance $\int I^2(y,\theta)dF(y)/n$, which is the correct limiting distribution of $\hat{\theta}$ for any underlying distribution function F .

Huber (1972) had indicated that the jackknife should work properly asymptotically for robust estimators with smooth influence functions. An example is the trimmed mean, which has a continuous influence function. A little algebra shows that the sample variance of the pseudo-values for the trimmed mean approximately equals the Winsorized sample variance. The latter is the correct variance to use with the trimmed mean so the jackknife

is performing as it should (see Cox and Hinkley (1974), p. 350). On the other hand, the median and the Winsorized mean have discontinuous influence functions. It is well known that the jackknife doesn't work for the former, and it won't work for the latter either because it depends heavily on two order statistics.

Two recent developments are worth mentioning before closing this section.

Hinkley (1977a) has initiated an investigation into the second order derivatives to see if there is any information in them which might improve the performance of the jackknife. Specifically, he examines estimators which admit the expansion

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{1}^{n} I_1(Y_i,\theta) + \frac{1}{2}\frac{1}{n^2}\sum_{1}^{n}\sum_{1}^{n} I_2(Y_i,Y_j,\theta) + \ldots \qquad (23)$$

to see what effect the term involving the second derivative I_2 has on the jackknife.

The jackknife operates by deleting observations. Thus, as a finite difference approximation to the derivative $I(y,\theta)$, it subtracts mass at y. Mallows has proposed an alternative finite difference approximation which adds mass at y. In effect, this introduces a procedure which adds hypothetical observations to the sample. For a discussion of this the reader is referred to Devlin, Gnanadesikan, and Kettenring (1975). In a similar spirit Efron (1977) has proposed inferential procedures based on samples generated randomly according to the empirical distribution function of the sample. He has coined the term bootstrap for these procedures, and

he has demonstrated that the jackknife is just a linear approximation to the bootstrap.

5. Applications

i) Ratios. One of the earliest applications of the jackknife was to ratio problems. Let $X_1, \ldots, X_m$ be a sample with theoretical mean μ, and $Y_1, \ldots, Y_n$ be a sample with theoretical mean η. The problem is to estimate $\theta = \eta/\mu$, and the standard ad hoc estimator is $\hat{\theta} = \bar{Y}/\bar{X}$. Durbin (1959) showed that jackknifing $\hat{\theta}$ improves not only its bias but also its mean squared error in many cases. Later authors amplified on these results and compared the jackknifed estimator with other ratio estimators. For a full discussion of this application the reader is referred to Miller (1974a).

ii) Variances. The sensitivity of normal theory variance testing procedures to departures from normality is well established. Mosteller and Tukey (1968) and Miller (1968) proposed jackknifing $\hat{\theta} = \log s^2$ as a way of handling this problem in robust fashion. Shorack (1969) compared the jackknife estimator and some other robust procedures for the two sample problem. These ideas also extend to robustly handling the k sample problem and variance component problems. For a fuller discussion on this area the reader is referred to Miller (1974a).

iii) Correlation Coefficients. Another problem where the normal theory procedure is not robust is interval estimation for the correlation coefficient. The test that ρ equals zero is robust to non-normality, but for $\rho \neq 0$ the asymptotic variance of $\hat{\theta} = \tanh^{-1} r = (1/2)\ln\{(1+r)/(1-r)\}$ is not $1/(n-3)$ unless the underlying distribution is normal. Duncan and Layard (1973) studied jackknifing $\hat{\theta}$ and found that it works well for most distributions. Recent work on improving the jackknife in connection with the correlation problem is contained in Hinkley (1976, 1977a).

iv) Censored Data. Considerable progress has been made on the analysis of censored data within the last two decades. In four landmark articles the product-limit estimator of a distribution function was introduced by Kaplan and Meier (1958), the log-rank analysis for two sample tests on censored data appeared in Mantel and Haenszel (1959), the Wilcoxon rank test was adapted to censored data by Gehan (1965), and Cox (1972) presented his conditional likelihood analysis of a proportional hazards model. None of these procedures requires the services of the jackknife because the relevant standard errors can be estimated without difficulty. However, for more complicated censoring and truncation problems as in Turnbull (1974, 1976) estimation of the standard error becomes messier and the jackknife may be useful. Similarly, the standard error for the estimated probability of survival beyond a

specified time for the proportional hazards model with covariates is sufficiently complicated that the jackknife may be a good way of estimating it. Preliminary work on the performance of the jackknife in the presence of censoring appears in Miller (1975) and Route (1976).

v) Model Simulation. It is difficult to get analytic answers for probability models which are sufficiently intricate to accurately model realistic storage systems, queueing systems, etc. Usually it is necessary to simulate the system on a computer. The estimates of the important parameters of the system can sometimes be improved by jackknifing, and the variability of the parameter estimates can be assessed by jackknifing. Examples of this can be found in Gaver (1975, 1977) and Iglehart (1975).

6. References

[1] Arvesen, J. N. (1969). Jackknifing U-statistics. Annals of Mathematical Statistics 40, 2076-2100.

[2] Brillinger, D. R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. Review of the International Statistical Institute 32, 202-206.

[3] Collins, J. R. (1970). Jackknifing generalizability. Doctoral thesis, University of Colorado.

[4] Cox, D. R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187-220.

[5] Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). The Dependability of Behavioral Measurements. Wiley, New York. pp. 54-57.

[6] Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika 62, 531-545.

[7] Duncan, G. T. and Layard, M. W. J. (1973). A Monte-Carlo study of asymptotically robust tests for correlation coefficients. Biometrika 60, 551-558.

[8] Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. Biometrika 46, 477-480.

[9] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. Biometrika 52, 203-223.

[10] Gray, H. L. and Schucany, W. R. (1972). The Generalized Jackknife Statistic. Marcel Dekker, New York.

[11] Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association 69, 383-393.

[12] Hollander, M. and Wolfe, D. A. (1973). Nonparametric Statistical Methods. Wiley, New York. pp. 103-111.

[13] Huber, P. J. (1972). Robust statistics: A review. Annals of Mathematical Statistics 43, 1041-1067.

[14] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481.

[15] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22, 719-748.

[16] Miller, R. G., Jr. (1964). A trustworthy jackknife. Annals of Mathematical Statistics 35, 1594-1605.

[17] Miller, R. G., Jr. (1968). Jackknifing variances. Annals of Mathematical Statistics 39, 567-582.

[18] Mosteller, F. (1971). The jackknife. Review of the International Statistical Institute 39, 363-368.

[19] Mosteller, F. and Tukey, J. W. (1968). Data analysis, including statistics. Handbook of Social Psychology (eds. G. Lindzey and E. Aronson). Addison-Wesley, Reading, Mass.

[20] Pennel, R. (1972). Routinely computable confidence intervals for factor loadings using the "jackknife". British Journal of Mathematical and Statistical Psychology 25, 107-114.

[21] Quenouille, M. H. (1949). Approximate tests of correlation in time-series. Journal of the Royal Statistical Society, Series B 11, 68-84.

[22] Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika 43, 353-360.

[23] Rao, J. N. K. (1965). A note on the estimation of ratios by Quenouille's method. Biometrika 52, 647-649.

[24] Rao, J. N. K., and Webster, J. (1966). On two methods of bias reduction in the estimation of ratios. Biometrika 53, 571-577.

[25] Schucany, W. R., Gray, H. L., and Owen, D. B. (1971). On bias reduction in estimation. Journal of the American Statistical Association 66, 524-533.

[26] Shorack, G. R. (1969). Testing and estimating ratios of scale parameters. Journal of the American Statistical Association 64, 999-1013.

[27] Tukey, J. W. (1958). Bias and confidence in not-quite large samples (abstract). Annals of Mathematical Statistics 29, 614.

[28] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. Journal of the American Statistical Association 69, 169-173.

[29] Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society, Series B 38, 290-295.

[30] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. Annals of Mathematical Statistics 18, 309-348.

7. Jackknife References: 1974-1977

[1] Anderson, C. W. and Ray, W. D. (1975). Improved maximum likelihood estimators for the gamma distribution. Communications in Statistics 4, 437-448.

[2] Anderson, P. O. (1975). Large sample and jackknife procedures for small sample orthogonal least squares inference. Communications in Statistics 4, 193-202.

[3] Arvesen, J. N. and Layard, M. W. J. (1975). Asymptotically robust tests in unbalanced variance component models. Annals of Statistics 3, 1122-1134.

[4] Arvesen, J. N. and Salsburg, D. S. (1975). Approximate tests and confidence intervals using the jackknife. Perspectives in Biometrics (ed. R. M. Elashoff). Academic Press, New York. pp. 123-147.

[5] Bissell, A. F. and Ferguson, R. A. (1975). The jackknife - toy, tool, or two-edged weapon. The Statistician 24, 79-100.

[6] Braun, H. (1975). Robustness and the jackknife. Technical Report No. 94, Series 2, Department of Statistics, Princeton University.

[7] Brillinger, D. R. (1976). Approximate estimation of the standard errors of complex statistics based on sample surveys. New Zealand Statistician 11, 35-41.

[8] Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association 69, 364-367.

[9] Cox, D. R. and Hinkley, D. V. (1974). Theoretical Statistics. Chapman and Hall, London. pp. 261-265,350.

[10] Davis, W. W. (1977). Robust interval estimation of the innovation variance of an ARMA model. Annals of Statistics 5, 700-708.

[11] Duncan, G. T. (1976). An empirical study of jackknife - constructed confidence regions in nonlinear regression. Technical Report No. 119 (NSF Grant MPS 75-07539), Department of Statistics, Carnegie-Mellon University.

[12] Duran, B. S. (1976). A survey of nonparametric tests for scale. Communications in Statistics A5, 1287-1312.

[13] Efron, B. (1977). Bootstrap methods: Another look at the jackknife. Technical Report No. 32 (PHS Grant 5 R01 GM21215-02), Division of Biostatistics, Stanford University.

[14] Ferguson, R. A. and Fryer, J. G. (1974). The effect of jackknifing maximum likelihood estimates; some single parameter illustrations. Unpublished manuscript.

[15] Frawley, W. H. (1974). Using the jackknife in testing dose responses in proportions near zero or one - revisited. Biometrics 30, 539-545.

[16] Freedman, D. A. (1977). Estimating standard errors in a complex sample survey. Unpublished manuscript.

[17] Gaver, D. P. (1975). Methods for assessing variability, with emphasis on simulation data interpretation. International Symposium on Uncertainties in Hydrologic and Water Resource Systems.

[18] Gaver, D. P. (1977). Modeling and estimation of complex system availability. Proceedings of the Twenty-third Conference on the Design of Experiments in Army Research, Development and Testing.

[19] Gray, H. L., Schucany, W. R., and Watkins, T. A. (1975). On the generalized jackknife and its relation to statistical differentials. Biometrika 62, 637-642.

[20] Hartigan, J. A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. Annals of Statistics 3, 573-580.

[21] Hinkley, D. V. (1976). Robust jackknife correlations. Technical Report No. 19 (PHS Grant 1 R01 GM21215-01), Division of Biostatistics, Stanford University.

[22] Hinkley, D. V. (1977a). Improving the jackknife with special reference to correlation estimation. Technical Report No. 294 (NSF Grant MCS77-00959), School of Statistics, University of Minnesota.

[23] Hinkley, D. V. (1977b). Jackknife confidence limits using Student t approximations. Biometrika 64, 21-28.

[24] Hinkley, D. V. (1977c). On jackknifing in unbalanced situations. Technometrics 19, 285-292.

[25] Iglehart, D. (1975). Simulating stable stochastic systems, V: Comparison of ratio estimators. Naval Research Logistics Quarterly 22, 553-565.

[26] Jones, H. L. (1974). Jackknife estimation of functions of stratum means. Biometrika 61, 343-348.

[27] Layard, M. W. J. (1974). A Monte Carlo comparison of tests for equality of covariance matrices. Biometrika 16, 461-465.

[28] Lemeshow, S. and Epp, R. (1977). Properties of the balanced half-sample and jackknife variance estimation techniques in the linear case. Communications in Statistics A6, 1259-1274.

[29] Miller, R. G., Jr. (1974a). The jackknife - a review. Biometrika 61, 1-15.

[30] Miller, R. G., Jr. (1974b). An unbalanced jackknife. Annals of Statistics 2, 880-891.

[31] Miller, R. G., Jr. (1975). Jackknifing censored data. Technical Report No. 14 (PHS Grant 1 R01 GM21215-01), Division of Biostatistics, Stanford University.

[32] Pandey, T. N. and Hubert, L. (1975). An empirical comparison of several internal estimation procedures for coefficient alpha. Psychometrika 40, 169-181.

[33] Rao, P. S. R. S. (1974). Jackknifing the ratio estimator. Sankhyā C 36, 84-97.

[34] Reeds, J. A. (1977). Jackknifing maximum likelihood estimates. To appear in Annals of Statistics.

[35] Rey, W. (1975a). Mean life estimation from censored samples. Biometrie-Praximetrie 15, 145-159.

[36] Rey, W. (1975b). Applicability of the jackknife method. Technical Report R300, M. B. L. E. Research Laboratory, Brussels, Belgium.

[37] Rey, W. (1975c). Missing values in principal component analysis. Report R304, M. B. L. E. Research Laboratory, Brussels, Belgium.

[38] Rogers, W. T. (1976). Jackknifing disattenuated correlations. Psychometrika 41, 121-133.

[39] Route, R. A. (1976). An investigation of the applicability of the product-limit estimate to the statistical analysis of sonar detection distributions. Thesis, Naval Postgraduate School, Monterey, California.

[40] Schucany, W. R. and Woodward, W. A. (1977). Adjusting the degrees of freedom for the jackknife. Communications in Statistics A6, 439-442.

[41] Sen, P. K. (1977a). Some invariance principles relating to jackknifing and their role in sequential analysis. Annals of Statistics 5, 316-329.

[42] Sen, P. K. (1977b). On jackknifing in estimating the finite end-points of a distribution. Unpublished manuscript.

[43] Sharot, T. (1976a). Sharpening the jackknife. Biometrika 63, 315-321.

[44] Sharot, T. (1976b). The generalized jackknife: finite samples and subsample sizes. Journal of the American Statistical Association 71, 451-454.

[45] Thorburn, D. E. (1976a). Some asymptotic properties of jackknife statistics. Biometrika 63, 305-313.

[46] Thorburn, D. E. (1976b). Asymptotic properties of some different jackknife techniques. Unpublished manuscript.

[47] Thorburn, D. E. (1977). On the asymptotic normality of the jackknife. Scandinavian Journal of Statistics 4, 113-118.

[48] Wainer, H. and Thissen, D. (1975). When jackknifing fails (or does it?). Psychometrika 40, 113-114.

[49] Woodward, W. A. and Gray, H. L. (1975). Minimum variance unbiased estimation in the gamma distribution. Communications in Statistics 4, 907-922.

MODELING AND ESTIMATING THE AVAILABILITY OF COMPLEX SYSTEMS: THE JACKKNIFE, COMMON-CAUSE, AND INSPECTION MODELS

Donald P. Gaver
Operations Research Department
Naval Postgraduate School
Monterey, Ca. 93940

1. Introduction

An important property of any system of cooperative or interacting components or equipments is its availability. By this is meant, roughly speaking, the fraction of time during which the system is operative and thus able to perform its intended function, and is not down for maintenance or repair. This paper outlines various ways in which component, and then system, availability may be described, i.e. represented by mathematical models. In Section 4 it is shown how, in several cases, operational data may be used to estimate availability, and also to assess the uncertainty, or error, of the estimates. The technique used for this purpose here is called the jackknife; see Mosteller and Tukey (1977), and Gaver and Chu (1977), from which the present account is borrowed. In Section 6 models for redundant repairable systems susceptible to common cause (sometimes termed common mode) failures are described and analyzed. It is shown that redundancy loses effectiveness when common cause failures, perhaps caused by external events such as weather or human error, tend to occur. Finally in Section 7 a sample

model for a standby system subject to periodic inspection is introduced and examined. Although occasional inspection and testing of a standby unit, such as a military weapon or a reactor safety system, is important to detect inoperability, too-frequent inspection may well increase the likelihood of failure. The model suggests an optimum--or at least reasonable--inspection interval as a compromise.

2. Systems and Scenarios

Examples of the kinds of systems we have in mind are shipboard communications (for a study see Perrin (1975)), general aircraft, including the engines and avionics, nuclear reactor safety systems, electric power boilers and generators, telecommunications systems including those involving satellites, and computer systems.

Such systems are complex, being made up of various interacting components, usually including a human link in either an active or maintenance capacity. The effect of improper maintenance is addressed in the inspection model of Section 7 but is otherwise ignored. A range of operating scenarios must be considered. Some are

1) Equipments always active, except when failed and when maintenance is carried out; examples: a base-loaded electric power generator powered by a nuclear plant,

2) Equipments inactive (on "cold standby") unless needed; examples: most weapons such as missiles, nuclear reactor safety systems,

3) Equipments (modules) active unless they are in maintenance or spare stock; example: replaceable aircraft engines.

There are other scenarios also; many include a certain amount of redundancy, i.e. extra equipments to be relied upon in case one or more of those "on line" fail.

Some appropriate definitions of availability are as follows:

a) Availability is the (expected) fraction of time an equipment is workable or up. Such a definition obviously relates to productivity of a base-loaded power generation or propulsion system.

b) Availability is the probability that a system is up when needed. Such a definition is suitable for a "cold standby" system, such as a missile or other weapon, or a reactor safety system, or perhaps certain communication devices. To say that the system is "up when needed" may also imply that the system remains up for a significant time period thereafter.

3. A Single Equipment Model

Consider a single equipment, for instance or a component of a system or a system itself, such as a nuclear power plant. Describe the equipment times to failure or uptimes by random variables U_i, and the subsequent repair times by random variables D_i, $i = 1,2,\ldots$. Supposing that the system begins up, then the first cycle terminates at $U_1 + D_1$ with the system again up; the ith cycle duration is $U_i + D_i = C_i$. Then if $A_U(t)$ is the availability of the system at t, given that the system was initially beginning an up period, and if cycle times are independent, one arrives at the Volterra integral equation for $A_U(t)$

$$A_U(t) = 1 - F_U(t) + \int_0^t A_U(t-x)\, F_C(dx) \,, \tag{3.1}$$

F_U being the distribution function (d.f.) of U, and F_C the d.f. of a cycle length C. Renewal theory shows that if either U or D or C has an absolutely continuous component that then

$$\lim_{t \to \infty} A_U(t) = \frac{E[U]}{E[U] + E[D]} = A_U \tag{3.2}$$

provided the expected values $[E,U]$ and $E[D]$ are finite. This simple expression describes the long-run point availability of the system. Notice that nothing is said about the independence

of U and the subsequent D: examples exist to show that if U and D are positively related (correlated) the rate of approach to the value A_U is slower than if they are independent (see Gaver (1972)). If the equipment is an emergency unit (weapon or safety system) that is required at a random time T, and T has the exponential distribution $F_T(t) = 1 - e^{-st}$, where $s^{-1} = E[T]$, then the convolution properties of Laplace transforms show that availability at demand time is, for any $s > 0$,

$$A_U(s) = \int_0^\infty A_U(t)\, e^{-st} s\, dt = \frac{1 - \hat{F}_U(s)}{1 - \hat{F}_C(s)}\,, \qquad (3.3)$$

where $\hat{F}_U(s) = E[e^{-sU}]$, and $\hat{F}_C(s) = E[e^{-sC}]$. This expression can easily be evaluated for some familiar distributions (not the log normal), and $A_U(s)$ approaches A_U as $s \to 0$. Demand times occurring according to gamma distributions, or even more general laws, can be handled in similar fashion.

4. Single Equipment Availability Estimation by the Jackknife Method.

Suppose observations are available on the up and down times of a single equipment; denote these by $u_1, u_2, \ldots, u_n$, and $d_1, d_2, \ldots, d_m$ respectively. These may be used to make

inferences (statistical estimates) of system availability. One promising technique for dealing with this problem is the jackknife; see Mosteller and Tukey (1977), Chap. 8. Analysis shows that in large samples the jackknife method tends to remove estimator bias--its originally advertised purpose--and in addition supplies usefully accurate confidence limits. We report Monte Carlo simulation results that indicate the validity of such confidence limits for realistically smallish numbers of observations as well. In a later section we also show how the method extends to systems of independently failing, and independently maintained, equipments.

The approach proceeds by first examining the obvious point estimate of A_U:

$$\tilde{A} = \frac{\bar{u}}{\bar{u} + \bar{d}} , \tag{4.1}$$

where $\bar{u}$ and $\bar{d}$ are the means of the observed up and down times. We first rewrite it (transform) to consider

$$z = \ell n\left(\frac{\tilde{A}}{1 - \tilde{A}}\right) = \ell n\ \bar{u} - \ell n\ \bar{d} . \tag{4.2}$$

The purpose of this transformation is to allow consideration of a quantity more nearly symmetrical and even normal (Gaussian) than is $\tilde{A}$ itself. Note that although the log transformation is likely to be effective other possibilities exist as well;

the cube root or Wilson-Hilferty is a plausible alternative (see Kendall (1947)), apparently not yet much investigated.

Having computed z based on all observations one next recomputes z, but leaving out the jth pair of observations $(j = 1,2,\ldots,n)$:

$$z_{-j} = \ell n\left(\sum_{\substack{i=1 \\ i\neq j}}^{n} u_i\right) - \ell n\left(\sum_{\substack{i=1 \\ i\neq j}}^{n} d_i\right).$$

Here it is assumed that the number of up times and down times are equal. Next, compute the pseudovalues

$$z_j = nz - (n-1)z_{-j}, \qquad (j = 1,2,\ldots,n),$$

and their mean and variance:

$$\bar{z} = \frac{1}{n}\sum_{j=1}^{n} z_j, \qquad s_z^2 = \frac{1}{n-1}\sum_{j=1}^{n} (z_j - \bar{z})^2.$$

Now quote as point estimate of availability the quantity

$$\tilde{A}_{JK} = \frac{e^{\bar{z}}}{1 + e^{\bar{z}}}$$

obtained by inverting the log-logistic transformation. At the suggestion of J. W. Tukey (1958), treat the individual z_j's as approximately independent and Normal and so apply Student's t to establish approximate confidence limits first on

$\ell n(A_U/(1-A_U))$ and then on A_U itself: for two-sided $(1-\alpha)\cdot 100\%$ intervals find

$$H_\alpha = \bar{z} + t_{1-\alpha/2}(n-1)\cdot\sqrt{\frac{s_z^2}{n}}$$

$$L_\alpha = \bar{z} - t_{1-\alpha/2}(n-1)\cdot\sqrt{\frac{s_z^2}{n}}$$

so that, approximately,

$$L_\alpha \leq \ell n\left(\frac{A_U}{1-A_U}\right) \leq H_\alpha$$

and thus also

$$\frac{e^{L_\alpha}}{1+e^{L_\alpha}} \leq A_U \leq \frac{e^{H_\alpha}}{1+e^{H_\alpha}}$$

with $(1-\alpha)\cdot 100\%$ confidence.

Asymptotic techniques (n large) of R. Miller (1964) will show that this procedure tends to be valid. That it is also robust of validity--coverage of the true availability reasonably close to stated 95% for a variety of distributions of up and down times-- is borne out by simulations; see the following tables for $n = 15$ and $n = 25$. Distributions considered are these

A. U_i and D_i mutually independent and each exponential; $E[U_i] = \lambda^{-1}$, $E[D_i] = \mu^{-1}$.

B. U_i independently exponential and D_i independently gamma with shape parameter $k = 3$.

C. U_i independently exponential and D_i gamma and independent with the gamma proportional to the preceding exponential up time; shape parameter $k = 2$.

D. U_i independently "long-tailed h" (i.e. $U = Xe^{hX}$, with X exponential, adjusted for desired mean and variance), D_i independently exponential.

The long-tailed h distributions of D are introduced to represent data appearing nearly exponential for small-to-medium values but that has long tails. For more details see Gaver and Chu (1977), and Gaver (1978). Thus an attempt has been made by means of the above four distributional forms to deal with data of a reasonable and plausible variety. This is necessary, for there is little chance that the "correct distributions" can be identified from the data itself in an applied situation. Notice that in the case of data model A--ups and downs independently exponential--an exact solution is available in terms of the classical F-statistic, for $\bar{D}/\bar{U}$ is seen to be a ratio of independent chi-squares. Acting as if the "F" procedure is applicable in every case considered is clearly less valid than is the jackknife, as the tables show.

5. The Jackknife Applied to System Availability Estimation

The jackknife technique can also be applied to estimate the availability of systems of equipments; in fact, this may be its most important application. We indicate by some examples the effectiveness of the procedure.

K-Component Redundant--Identical Units

If K units are in parallel, and all must be down in order for the system to be down, then long-run unavailability is, under independence assumptions,

$$\bar{A} = \prod_{i=1}^{K} \frac{E[D_i]}{E[U_i] + E[D_i]} = \left(\frac{E[D]}{E[U] + E[D]} \right)^K \tag{5.1}$$

which would naturally be estimated by

$$\tilde{\bar{A}} = \left(\frac{\bar{d}}{\bar{u} + \bar{d}} \right)^K \tag{5.2}$$

But now

$$z = \ln \left[\frac{1 - (\tilde{\bar{A}})^{1/K}}{(\tilde{\bar{A}})^{1/K}} \right] = \ln \bar{u} - \ln \bar{d} \tag{5.3}$$

and so one merely jackknifes z as before and inverts to put $(1-\alpha)\cdot 100\%$ confidence limits on A:

$$1 - \left(\frac{1}{1 + e^{L_\alpha}} \right)^K \leq A \leq 1 - \left(\frac{1}{1 + e^{H_\alpha}} \right)^K \tag{5.4}$$

TABLE 1

Simulation Experiments Validating Jackknife Single-Unit Availability

95% Confidence Limits; Two-Sided (t = 2.145)

n = 15

Underlying Distributions	Coverage (%)	Average Length	Variance Length
A. (exponential)	JK: 95.0	1.94×10^{-2}	1×10^{-4}
$\lambda = 0.01$, $\mu = 1$	"F": 94.4	1.63×10^{-2}	4×10^{-5}
B. (exponential, gamma)	JK: 94.1	1.39×10^{-2}	3×10^{-5}
$\lambda = 0.01$, $\mu = 1$	"F": 97.7	1.64×10^{-2}	3×10^{-5}
k=3			
C. (exponential, corr. gamma)	JK: 92.2	8.4×10^{-3}	1.7×10^{-5}
$\lambda = 0.01$, $\mu = 1$	"F": 99.5	1.6×10^{-2}	0.98×10^{-5}
k=2			
D. (long-tailed h, exponential)	JK: 92.4	2.42×10^{-2}	1.8×10^{-4}
$\lambda = 0.01$, $\mu = 1$	"F": 88.0	1.75×10^{-2}	7×10^{-5}
h=0.2			

TABLE 2

Simulation Experiments Validating Jackknife Single-Unit Availability
95% Confidence Limits; Two-Sided (t = 2.064)

n = 25

	Underlying Distributions	Coverage (%)	Average Length	Variance Length
A.	(exponential) $\lambda = 0.01$, $\mu = 1$	JK: 96.2 "F": 95.9	1.27×10^{-2} 1.21×10^{-2}	2×10^{-5} 1×10^{-5}
B.	(exponential, gamma) $\lambda = 0.01$, $\mu = 1$ k=3	JK: 94.2 "F": 98.8	9.88×10^{-3} 1.22×10^{-2}	1×10^{-5} 1×10^{-5}
C.	(exponential, corr. gamma) $\lambda = 0.01$, $\mu = 1$ k=2	JK: 94.7 "F": 99.9	6.00×10^{-3} 1.20×10^{-2}	0.4×10^{-5} 0.3×10^{-5}
D.	(long-tailed h, exponential) $\lambda = 0.01$, $\mu = 1$ h=0.2	JK: 94.1 "F": 88.7	1.64×10^{-2} 1.27×10^{-2}	4×10^{-5} 2×10^{-5}

K-Component Redundant--Different Units

If the units are unlike it is plausible to jackknife

$$z = \ell n\, \tilde{\bar{A}} = \sum_{i=1}^{K} \ell n \left(\frac{\bar{d}_i}{\bar{u}_i + \bar{d}_i} \right) . \qquad (5.5)$$

The straightforward way of carrying this out is to compute as before the jth pseudovalue ℓ_{ij} of $\ell n\, \tilde{\bar{A}}_i$ for $i = 1,2,\ldots,K$, find its mean and variance, denoted by $\bar{\ell}_i$ and s_i^2. Last, combine to obtain

$$\bar{\ell} = \sum_{i=1}^{K} \bar{\ell}_i , \quad \text{and} \quad s^2 = \sum_{i=1}^{K} \frac{1}{n_i} s_i^2 . \qquad (5.6)$$

Upper and lower confidence limits on $\ell n\, \tilde{\bar{A}}$ are then of the form

$$H_\alpha = \bar{\ell} + t_{1-\alpha/2}\Big(\sum_{i=1}^{K} n_i - K\Big) \cdot s$$

$$L_\alpha = \bar{\ell} - t_{1-\alpha/2}\Big(\sum_{i=1}^{K} n_i - K\Big) \cdot s$$

These may be translated to limits on $\bar{A}$, and on A. An alternative procedure is one of linearization around the jackknifed point estimate of $\ell n\, \bar{A}$; for details see Gaver and Chu (1977).

Some Monte Carlo simulation results are exhibited in Table 3. Once again the results seem usefully valid and efficient.

TABLE 3

Simulation Experiments Validating Jackknife Two-Unit Redundant Unavailability

95% Confidence Limits

n = 25

(JK,1: Jackknife Procedure 1; JK,2: Jackknife Procedure 2)

Underlying Distributions	Coverage (%)		Average Length	Variance Length
A. (exponential up, exponential down)	JK,1:	93.9	3.3×10^{-4}	1.7×10^{-8}
$\lambda_1 = 0.01$, $\lambda_2 = 0.02$	JK,2:	94.7	3.7×10^{-4}	2.4×10^{-8}
$\mu_1 = 1 = \mu_2$				
B. (exponential up, gamma down)	JK,1:	94.5	2.8×10^{-4}	1.0×10^{-8}
$\lambda_1 = 0.01$, $\lambda_2 = 0.02$	JK,2:	94.6	2.8×10^{-4}	1.1×10^{-8}
$\mu_1 = \mu_2 = 1$; shape = 3				
D. (long-tailed, h up, exponential down)	JK,1;	93.0	4.6×10^{-4}	4.5×10^{-8}
$\lambda_1 = 0.01$, $\lambda_2 = 0.02$	JK,2:	93.0	5.1×10^{-4}	6.0×10^{-8}
$\mu_1 = \mu_2 = 1$, h = 0.2				

Two-Out-of-Three Voting

A final example is provided by a system of three units that is considered available if any two are simultaneously available. Thus

$$A = A_1A_2A_3 + \bar{A}_1A_2A_3 + A_1\bar{A}_2A_3 + A_1A_2\bar{A}_3$$

As usual, up and down time data are assumed to be available on all three units; we do not wish to assume them identical. One procedure is as follows. First compute the jackknifed point estimate of system availability. Next consider the log-logistic transformation $\ell = \ell n[\tilde{A}/(1-\tilde{A})]$, and expand to linear terms around the jackknifed point estimate $\tilde{A}'$, thus finding an expression for s_ℓ^2; for further details see Gaver and Chu (1977).

TABLE 4

Simulation Experiments Validating Jackknife Two-Out-of-Three Voting System

95% Confidence Limits

$n_1 = n_2 = n_3 = 15$

Underlying Distributions	Coverage %	Average Length	Variance Length
A. (exponential up, exponential down) $\lambda_1 = 0.01$, $\lambda_2 = 0.02$, $\mu_1 = \mu_2 = \mu_3 = 1$	94.5	3.0×10^{-3}	2.1×10^{-6}
B. (exponential up, gamma down) $\lambda_1 = 0.01$, $\lambda_2 = 0.02$, $\lambda_3 = 0.04$, $\mu_1 = \mu_2 = \mu_3 = 1$	94.2	2.2×10^{-3}	8.5×10^{-7}
D. (long-tailed, h up, exponential down) $\lambda_1 = 0.01$, $\lambda_2 = 0.02$, $\lambda_3 = 0.04$, $\mu_1 = \mu_2 = \mu_3 = 1$	92.9	4×10^{-3}	4.8×10^{-6}

6. Common-Cause Failure Models

The previously discussed models for availability of systems assumed that the component equipments failed and were repaired independently. Such an assumption is often inappropriate: common causes of failure, such as environmental shock or personnel error, may well be decisive. We present now a simple model for catastrophic common-cause failure.

A Repairable System Experiencing Common Cause Failure

Consider a system of m $(m \geq 1)$ identical equipments, each one of which fails independently with rate λ (exponential time to failure), and is repaired (after an exponential time) with rate μ. The system is also confronted by a common cause failure mechanism, such that when it is activated the system fails completely. The rate of occurrence of the latter is c. Rule that the system is operative or up so long as k out of m $(1 \leq k \leq m)$ units operate. The system fails as soon as at least $\ell = m-k+1$ units are down simultaneously. The problem addressed here is to calculate the expected time to system failure, where failure may occur either because of the individual machine chance failures, or because of the common-cause catastropic event.

Analysis of the model may be conducted in terms of the state variable $D(t)$; $D(t) = j$ means that j units are on repair at time t. Clearly $D(t)$ is a Markov process, and its state transition rates are specified as follows: given $D(t) = j$, then

Change	Probability
$D(t + dt) = D(t) + 1$	$\lambda_j dt$
$D(t + dt) = D(t) - 1$	$\mu_j dt$
$D(t + dt) = m$	$c dt$
$D(t + dt) = D(t)$	$1 - (\lambda_j + \mu_j + c)dt$

All other probabilities are negligible. Of course λ_j and μ_j may be specified so as to represent any kind of system; for instance, one in which there are limited numbers of repairmen and thus queueing occasionally occurs, or one in which not all units are simultaneously operative and susceptible to failure. Here we specify these parameters to be

$$\lambda_j = (m - j)\lambda, \qquad \mu_j = \min(j,r)\mu, \qquad j = 0,1,\ldots,m \tag{6.1}$$

where r $(1 \leq r \leq m)$ is the number of repairmen available to work simultaneously. Furthermore, $r = m$ in the numerical examples.

The process $\{D(t)\}$ is actually birth-and-death (see Feller (1957)) with an independent Markovian killing process. Denote by T_ℓ the elapsed time for the system to pass for the first time from $D(0) = 0$--no element down--to the state ℓ or greater, at which point system failure occurs. Note that

$$P\{T_\ell > t | D(0)=0\} = P\{T_\ell^* > t | D(0)=0\}e^{-ct} \tag{6.2}$$

where T^* is the first passage time to ℓ in the ordinary birth-and-death process that admits no catastrophes. Equation (6.2) simply expresses the fact that failure time exceeds t if and only if neither a chance failure nor a catastrophic failure occurs before t.

Now Laplace transform (6.2) to obtain

$$\int_0^\infty e^{-st} P\{T_\ell > t | D(0)=0\} = \int_0^\infty e^{-st} P\{T_\ell^* > t | D(0)=0\} e^{-ct} dt$$

$$= \frac{1}{s+c} \{1 - E[e^{-(s+c)T_\ell^*}]\} , \qquad (6.3)$$

the latter following by an integration by parts. The Laplace transform of T_ℓ^* is of the form

$$E[e^{-(s+c)T_\ell^*}] = \prod_{i=0}^{\ell-1} \phi_i(s+c) \qquad (6.4)$$

where

$$\phi_i(s+c) = \frac{\lambda_i}{\lambda_i + \mu_i + s + c - \mu_i \phi_{i-1}(s+c)} , \qquad i = 1,2,3,...$$

and

$$\phi_0(s+c) = \frac{\lambda_0}{\lambda_0 + s + c} ; \qquad (6.5)$$

see Karlin and Taylor (1975). By combining (6.3) and (6.4) one may calculate (6.2); then, allowing $s \to 0$ there results

$$E[T_\ell] = \frac{1 - E\left[e^{-cT^*_\ell}\right]}{c} .$$

Numerical Example

Let $m = 3$, $k = 1$, $\ell = 3$, meaning that the system of three equipments fails only when all are down simultaneously. Put $\lambda = 10^{-2}(\text{days}^{-1})$, $\mu = 1(\text{days}^{-1})$, and consider the effect of varying the catastrophe rate, c.

$E[T_\ell]$	c (Catastrophe Rate)
3.5×10^5	0
1×10^4	10^{-4}
1×10^3	10^{-3}

Obviously a catastrophe rate as great as 10^{-4} completely dominates the effect of the individual unit chance failures. Thus only if the catastrophe rate is of magnitude 10^{-5} or smaller will the present redundancy be at all effective.

7. An Inspected System

In this section we turn attention to an equipment that is not expected to operate constantly, but that is intended to be ready when needed. An example is a weapon such as a gun or missile; another example is an alarm or safety system, perhaps associated with a nuclear power system.

Attempts to insure the operability or readiness of such systems usually include periodic inspection and preventive maintenance. Our model incorporates these attributes; furthermore, it allows for imperfection in the inspection-repair process, e.g. brought about by human error.

The Model

A single equipment is subject to periodic inspections and preventive maintenance or repair actions. Let the time from the completion of a preventive maintenance period until the beginning of the next be I time units, and let the subsequent preventive maintenance period require R time units; both I and R will be taken to be fixed. Hence over a long period (say one year) the system presents itself as nominally "ready" a fraction of time equal to $I/(I + R)$, and down for inspection and maintenance, and hence unavailable, for a fraction of time $R/(I + R)$. Now admit the possibility that the system be additionally unavailable for one of two reasons: (a) at the end of an inspection-maintenance period the equipment is returned to active service in an inoperative condition, an event of probability δ $(0 \leq \delta \leq 1)$,

(b) at the end of an inspection-maintenance period the equipment is up, an event of probability $\bar{\delta} = 1-\delta$, but it fails before the next inspection, and is thus actually unavailable for the time following that failure until the next inspection-repair period begins. Let $F(t)$ be the distribution function, and $f(t)$ the density, of equipment failure time. If the inspection interval, I, is treated as a decision variable it is interesting to select its value so as to maximize long-run availability, or, equivalently, to minimize long-runs unavailability. In order to do so, first calculate the expected time unavailable during one cycle of length $I + R$:

$$\delta I + \bar{\delta} \int_0^I (I - t)\ f(t)dt + R\ ;$$

division by $I + R$ then gives the expected unavailable time per unit time as the latter depends upon I:

$$\bar{A}(I) = \frac{\delta I + \bar{\delta} \int_0^I (I - t)\ f(t)dt + R}{I + R}$$

$$= 1 + \bar{\delta} \left[\frac{-I + \int_0^I (I - t)\ f(t)dt}{I + R} \right] \qquad (7.1)$$

One may now choose I so as to minimize $\bar{A}$; differentiation shows that the optimum I must satisfy the equation

$$R = \frac{1}{1 - F(I)} \int_0^I tf(t)dt = \frac{IF(I) - \int_0^I F(t)dt}{1 - F(I)}\ ; \qquad (7.2)$$

since the denominator is a decreasing, and the numerator an increasing, function of I there is exactly one root of (7.2). Surprisingly at first glance, the optimum inspection interval, I_{opt}, does not depend upon δ, the probability of a failure during the inspection-preventive maintenance period. Of course the eventual system availability does depend upon this parameter; providing I_{opt} is chosen it turns out that

$$\bar{A}(I_{opt}) = 1 - \bar{\delta}[1 - F(I_{opt})]$$

Although (7.2) cannot usually be solved explicitly it turns out, in the case of exponential failures, to be

$$R = \frac{1}{\lambda}\left[e^{\lambda I} - (1 + \lambda I)\right] \approx \frac{\lambda I^2}{2}$$

when λ is small, so in this case

$$I_{opt} \simeq \sqrt{\frac{2R}{\lambda}}$$

and

$$\bar{A}(I_{opt}) \approx 1 - \bar{\delta} \exp(-\sqrt{2\lambda R}) ,$$

or, in terms of availability,

$$A(I_{opt}) \approx \bar{\delta} \exp(-\sqrt{2\lambda R}) .$$

REFERENCES

FELLER, W. (1957). An Introduction to Probability Theory and Its Applications. John Wiley and Sons, New York.

GAVER, D.P. (1972). "Point Process Problems in Reliability," in Stochastic Point Processes, ed. by P. A. W. Lewis, John Wiley and Sons, N.Y.

GAVER, D. P. (1978). "Simple Parameteric Hazard Models," paper in preparation.

GAVER, D. P. and CHU, B. (1977). "Estimating Equipment and System Availability by Use of the Jackknife," submitted for publication.

KARLIN, S. and TAYLOR, H. (1975). A First Course in Stochastic Processes, Academic Press.

KENDALL, M. G. (1947). The Advanced Theory of Statistics, Vol. I, Charles Griffin and Co., Ltd., London.

MILLER, R. P. (1964). "A Trustworthy Jackknife," Annals of Math. Statistics, 35, pp. 1594-1605.

MOSTELLER, F. and TUKEY, J. W. (1977). Data Analysis and Regression, Addison-Wesley Publ. Co., Reading, Mass.

PERRIN, C. S. (1975). "A Manning and Maintenance Effectiveness Model Applied to the Communication Division of a "Knox" Class Destroyer Escort," Naval Postgraduate School Master's Thesis.

TUKEY, J. W. (1958). "Bias and Confidence in Not-Quite Large Samples," Abstract in Ann. math. Statist. 59, p. 614.

QUALITATIVE EVALUATION OF THE M60A1 TANK CAMOUFLAGE BY OPERATIONAL IMAGERY INTERPRETERS

EDWARD R. EICHELMAN

AND

RONALD L. JOHNSON

US ARMY MOBILITY EQUIPMENT RESEARCH AND DEVELOPMENT COMMAND

FORT BELVOIR, VIRGINIA 22060

ABSTRACT. A continuing problem in the assessment of camouflage effectiveness has been the objective analysis of subjective data. This paper is concerned with such an evaluation for an M60A1 Tank. Thirty operational image interpreters analyzed the following camouflage prototypes: natural foliage, fender nets, two styles of gun barrel disrupters, and counter-shading. Each interpreter viewed aerial imagery of each condition. A forced choice of descending ratings was assigned. Mean ratings and associated variances were calculated. The scores were standardized, and the Z statistic was employed to determine significant differences. The effectiveness of foliage was significantly better, $\alpha = .01$, than counter-shading.

I. INTRODUCTION.

Up through World War II the development of camouflage involved a subjective, artistic approach rather than the scientific method now advocated. With the advent of more complex sensor systems the development of camouflage concepts has necessitated a more controlled approach based on stringently quantified data. The results of the analysis are then used as a data base

to identify the most promising camouflage concepts for further development. One such instance in which the U.S. Army is involved is the tactical camouflage of the M60A1 combat tank. The purpose of this study was to objectively evaluate the effectiveness of various prototype camouflage items for the M60A1 tank. It was accomplished through the use of operation image interpreters (II's).

II. DESIGN OF EXPERIMENT.

A. Targets. The test targets consisted of M60A1 tanks in the following conditions:

a. Pattern painted.

b. Pattern painted and natural foliage.

c. Pattern painted, countershading, and gun barrel disrupter (Type I).

d. Pattern painted, fender nets, and gun barrel disrupter (Type II).

These various conditions of camouflage will now be described in detail.

1. Pattern Paint.

The purpose of the camouflage paint patterns is to distort straight lines and edges of objects, alter perception of depths, and to reduce contrast with the surroundings and cause the object to blend with its background[1]. Camouflage paint patterns were developed by the U. S. Army Mobility Equipment Research and Development Command (MERADCOM). The pattern used in this test combines patches of the colors forest green, light green, sand, and black. It is the Summer U. S. and European verdant pattern.

2. Natural Foliage.

The natural foliage was inserted into specially placed brackets to disrupt the target's outline and distinct features. It was also intended to reduce the target to background contrast.

3. Countershading and Type I Gun Barrel Disrupter.

Countershading of the target consists of painting the normally dark or shadowed areas with light colors (e.g., white or gray) to reduce detection and identification by means of these visual contrast cues. The Type I gun barrel disrupter is an accordion like sleeve that slips over the gun barrel to break up the parallel straight edges as well as to distort its shadow.

4. Fender Nets and Type II Gun Barrel Disrupter.

The final camouflage condition evaluated, contained fender nets and a Type II gun barrel disrupter. Fender nets were designed to cover the visual cues of the tank's track system and lower portion of the hull. They consist of six foot long fiber glass rods supporting plastic garnish material from the Army's standard lightweight camouflage screening system (LWCSS). The Type II gun barrel disrupter is of an irregular fan shaped design which is attached along the top of the gun barrel.

B. Test Imagery.

The test imagery consists of a series of 4" X 5" color positives for each of the camouflage conditions. Scaled aerial photographs at 1:10,000 and 1:5,000 were taken of the front, back, top, and both sides of each target M60A1 tank. Additional ground level photographs were taken of each target for documentation. The target tanks were sited so that they were unobstructed by indigenous foliage.

C. Test Procedure.

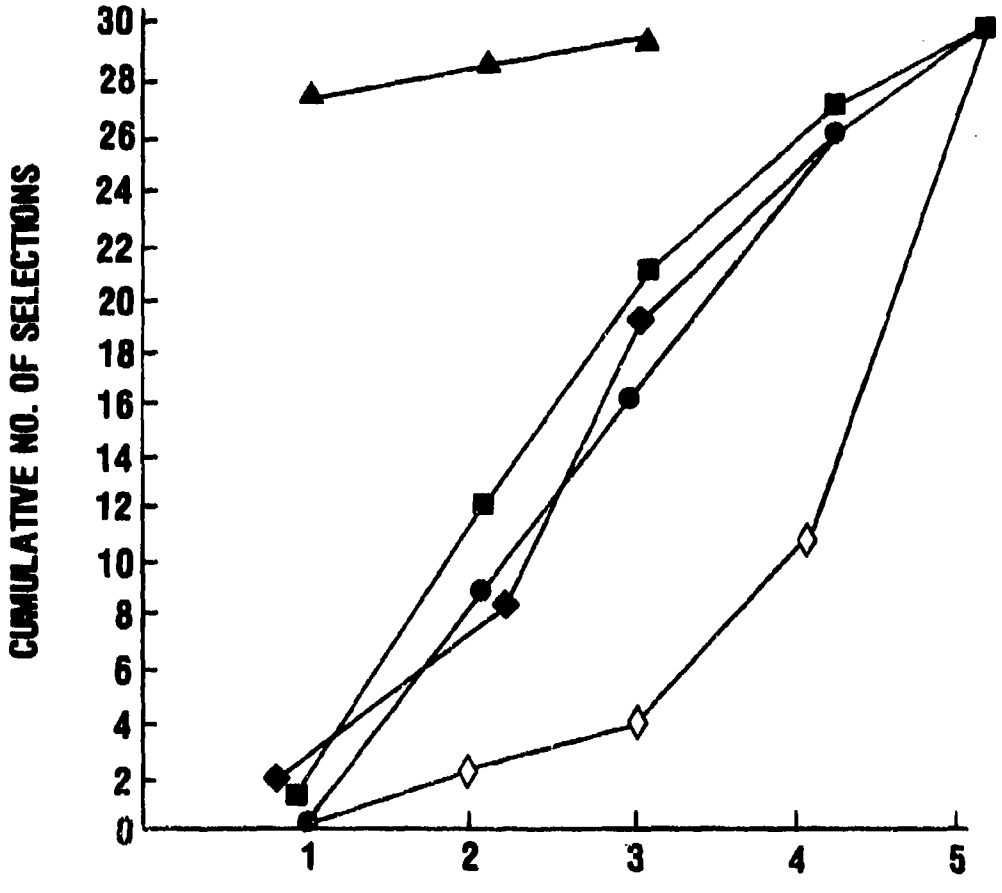
Thirty operational image interpreters (II's) participated in the camouflage evaluation. They were first shown the close-up, ground level pictures of the camouflaged tanks and given a brief description of the purpose of each type of camouflage. The pattern painted tank was defined as the base condition upon which the five types of camouflage were applied. They were than shown all of the 4" X 5" color positives of the camouflage conditions for evaluation. In order to provide objective results from this study, the II's were instructed to make a forced choice in analyzing the effectiveness of five types of camouflage. The ranking choices were as follows:

1. Most effective
2. Above average effectiveness
3. Average effectiveness
4. Below average effectiveness
5. Least effective

III. EXPERIMENTAL RESULTS.

The dependent variable of this test is the frequency with which each prototype camouflage was assigned a particular effectiveness value by the II's. The forced selection of effectiveness allowed the conversion of the subjective data into objective results. Figure I shows the cumulative totals of the forced selection of effectiveness. As an example, for countershading, the left end of the lower line with diamond points indicates zero choices as No. 1 (most effective); two choices as No. 2 (above average effectiveness); two more choices as No. 3 (average effectiveness) for a total of four; seven more as No. 4 (below average effectiveness) for a total

GRAPHIC RESULTS OF EVALUATION OF M-60A1 CAMOUFLAGE BY OPERATIONAL IMAGE INTERPRETERS



KEY

1—MOST EFFECTIVE
2—ABOVE AVERAGE EFFECTIVENESS
3—AVERAGE EFFECTIVENESS
4—BELOW AVERAGE EFFECTIVENESS
5—LEAST EFFECTIVE

▲ FOLIAGE WITH BRACKETS
■ FENDER NETS
● GUN BARREL DISRUPTER TYP I
◆ GUN BARREL DISRUPTER TYP II
◇ COUNTERSHADING

FIGURE 1

of eleven; and finally, nineteen more as No. 5 (least effective) for a total of thirty.

Table I is a numerical summary of the II data by camouflage effectiveness rating versus the type of camouflage.

The means and associated standard deviations were compared, using the Z statistic[2], to determine which means are statistically different from each other. The results are presented in Table II.

Any Z value greater than 2.576 indicates significance at an α of .01 shown by the values with asterisks.

IV. DISCUSSION.

The stated purpose of this study was to quantify the subjective evaluations of prototype camouflage for the M60A1 Tank. The problem faced in this study was one of obtaining objective data from facts that were subjective in origin. Four by five inch color positives were obtained of the M60A1 tank for four conditions of prototype camouflage. Photographs of the pattern painted tank were used as the base case. Thirty operational II's were shown all of the imagery. They used the forced choice rating technique to determine the effectiveness of the camouflage conditions on a five point scale, with one being most effective. The mean and standard deviation were determined for the frequency with which rating values were assigned to each condition of camouflage. The means and associated standard deviations were then subjected to the Z statistic to determine which condition of camouflage was significantly most or least effective. The resulting data was successfully used to determine the most promising candidates for further development.

TABLE I

CAMO EFFECT	WEIGHT	FOLIAGE	FENDER NETS	GUN BARREL DISRUPTER TYPE I	GUN BARREL DISRUPTER TYPE II	COUNTER SHADING
1	5	27	1	0	2	0
2	4	2	11	9	6	2
3	3	1	9	7	11	2
4	2	0	6	10	7	7
5	1	0	3	4	4	19
MEAN		4.87	3.03	2.70	2.83	1.57
STANDARD DEVIATION		.434	1.066	1.055	1.115	.898

TABLE II
Z OR STANDARD SCORES
FOR COMPARISON OF WEIGHTED MEAN CAMOUFLAGE
EFFECTIVENESS VALUES

	FOLIAGE	FENDER NETS	GUN BARREL DISRUPTER TYPE I	GUN BARREL DISRUPTER TYPE II	COUNTER SHADING
FOLIAGE					
FENDER NETS	8.72*				
GUN BARREL DISRUPTER TYPE I	15.20*	1.22			
GUN BARREL DISRUPTER TYPE II	10.54*	0.71	0.48		
COUNTER SHADING	18.13*	4.67*	4.48*	4.84*	

*SIGNIFICANT AT = .01

$Z_{.005} = 2.576$

$\alpha = .01$

$$Z = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2 + S_2^2}{N_1 + N_2}}}$$

$N = 30$

V. SUMMARY AND CONCLUSIONS.

Thirty operational II's evaluated tank camouflage effectiveness from 4" X 5" color positive aerial photographs of the camouflaged M60A1 Tanks. The means from a forced choice evaluation of the conditions of camouflage were objectively evaluated by use of the Z statistic. The data from Tables I and II, significantly ($\alpha = .01$), indicate that the use of natural foliage provides the best camouflage of those evaluated. The use of countershading has little or no camouflage effect or value. Fender nets and two types of gun barrel disrupters were significantly ($\alpha = .01$) better than the countershading, but significantly ($\alpha = .01$) inferior to foliage. Fender nets and the types of gun barrel disrupter did not differ significantly in camouflage effectiveness from each other. From the results of this study it was recommended that the use of foliage, fender nets and gun barrel disrupters, Types I and II, be subjected to additional development and testing. Countershading was not recommended. It is also noted that the use of forced choice rating can be very successful in an objective evaluation of data that is subjective in origin.

REFERENCES

1. AMC Circular No. 1-5, AMC Project Officer for M60A1 Tank Camouflage Pilot Program, 21 October 1975.
2. Walker, Helen M., and Lev, Joseph, Statistical Inference, Holt, Rinehart and Winston, New York, 1953.

DESIGN OF A FULL-SCALE TEST FOR U.S. ARMY HELICOPTER NAP-OF-THE-EARTH (NOE) COMMUNICATION SYSTEMS

Bernard V. Ricciardi
Communication and Sensor Division
U.S. Army Avionics Research and Development Activity
Fort Monmouth, New Jersey

Bruce C. Tupper and George H. Hagn
Telecommunications Sciences Center
SRI International
Menlo Park, California

ABSTRACT. Of particular interest to the Army is a reliable voice communication capability for helicopters that fly at Nap-of-the-Earth (NOE) altitudes. This flight regime at tree-top level or below is necessary for aircraft survival in the modern battlefield. At these altitudes, the present aircraft VHF/FM radio systems operate over only extremely short ranges and are essentially limited to line-of-sight (LOS) paths.

To quantitatively assess the performance and effectiveness of the nine candidate radio systems (both VHF/FM and HF/SSB) and communication methods, a large scale combined operational and engineering test was designed. The experiment design considered variables including range, altitude, terrain, time of day, frequency, and power that affect the radio channel (SNR). The tests were designed to determine how the performance of the non-LOS and LOS radio systems depended on these major variables. The test, conducted over a three-month period, involved over 100 personnel, and 1000 hours of flight testing, and utilized over 10,000 alpha-numeric (A-N) test messages to determine and evaluate the voice intelligibility of the radio systems.

This paper deals with a definition of the problem and development of measures of effectiveness (MOEs) to measure radio performance, the design of the experiment, and how the variables and dimensions of the test were treated. Although statistical principles were considered, a rigorous statistical design was not used; however, probability theory techniques were used for extension of the results to other terrains. Results are briefly discussed. Lessons learned from the tests are also summarized with recommendations given which could be applied to future operational tests of this nature.

1.0 INTRODUCTION: The Army is currently faced with a serious radio communication problem: communicating with the helicopter on the modern battlefield.

In order to survive on the modern battlefield aircraft must fly close to the surface of the earth in a Nap-of-the-Earth (NOE) region [1]. The NOE flight regime for helicopters is flying at extremely low altitudes, typically hover altitudes, at relatively low speeds below tree-top level in the battle area. Aircraft must fly at NOE altitudes to take maximum advantage of the terrain features for cover. Survivability and mission effectiveness in battle

area depend on how well the aircraft and crew can function under these strained conditions and how well communications can be maintained with the elements being supported. The problem of how to effectively communicate in the battle area while flying NOE resulted in the conduct of a full-scale operational field test of nine different communication systems. The main objective of the test was to compare and evaluate the communication effectiveness of the candidate radio systems under NOE conditions. The presently used tactical VHF/FM radio system was considered the baseline system for the tests.

Many variables existed for the NOE Communications test. Figure 1 shows the major test variables.

Variable	Condition	
Spatial:	Range Altitude	Terrain Siting
Time of Day:	Day Dawn	Night
Frequency Band/Modulation:		
	HF/SSB (2-8 MHz Below MUF) HF/SSB (8-30 MHz above MUF) VHF/FM (30-76 MHz)	
Power Output:	HF (40, 100, 200, 400W PEP) VHF (10, 40W)	
System Configuration (Links):	Air-Ground Ground-Air Air-Air	

FIGURE 1. TEST VARIABLES

These variables, and others, were considered in the design of the test to determine how communications range was affected with aircraft operating at various altitudes in various type terrain conditions. The tests described in this paper were supplemented by other engineering tests and by computer predictions of communications in operationally-significant areas such as Europe, the Mid-East, and Korea.

2.0 DESIGN OF THE TEST.

2.1 Measures of Effectiveness. To comparatively evaluate the performance of the candidate systems, two measures of effectiveness (MOEs) were developed.

2.1.1 Alpha-Numeric Test Messages. The first measure was a measure of communications effectiveness using randomly selected alpha-numeric

(A-N) characters sent through the radio channel. Communication effectiveness was defined as the percent of A-N characters correctly received, sent one way without repeats through the communication channel. This measure provided a quantitative comparison of each of the candidate radio systems as a function of the range and other test variables.

A test message containing an equal number of randomly selected letters and numbers was developed. This was called an A-N test message. The A-N test messages were formatted and transmitted as tactical spot reports by the tester. The tester determined that messages sent in this spot report format operationally resemble grid or target coordinates that helicopters routinely transmit over radio systems. Further, spot reports in this format sent one way through the channel without repeats are demanding on the communication channel. Finally, A-N messages in this format can be practically recorded in the helicopter by a test observer and graded at the end of the mission. Figure 2 shows a typical data recording sheet. A word consists of six randomly selected A-N characters. In this message characters and numbers are sent using the phonetic alphabet. These messages were copied down on answer sheets such as shown, graded and used as the primary measure of effectiveness for the tests.

Figure 2. Sample Data Sheet

DATE/TIME 1010/0930Z PAGE 3 OF 10

SYSTEM TEST CODE IFM1

RANGE (Km) 10 ALT NOE LEG: (IN) OUT

A/C NO 090 TAPE NO 7 HDG N

SPOT REPORT NO 201 RUN NO. 9

LINE	WORD					
A	9	7	ϕ	A	C	Z
B	B	E	I	3	5	9
C	O	L	S	X	Y	F
D	6	4	D	L	5	2
E	1	A	T	R	ϕ	O

RATE: TOO FAST TOO SLOW (OK)

READABILITY: 0 1 2 3 (4) 5

PERCENT CORRECT 27/30 A-N 90%

5 Perfectly Readable/No Perceptible Noise
4 Readable w/NOTICEABLE NOISE
3 Readable w/DIFFICULTY
2 Readable w/EXTREME DIFFICULTY
1 Unreadable/UNUSABLE
0 No Signal Heard

COMM CKT COMMENTS

ATMOSPHERIC NOISE NOTED

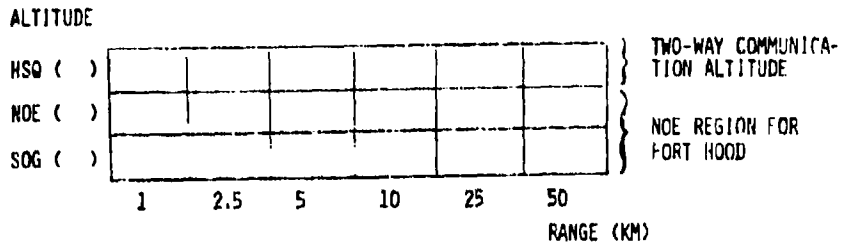
DATA TAKER BT TALKER BR

2.1.2 Height to Break Squelch. The second measure was the altitude required to establish two-way communications from the aircraft to the base station. In this case the aircraft would climb to whatever altitude was required to establish two-way communications to the base station. This provided an estimate of the vulnerability of the aircraft containing a candidate radio system to the enemy weapons threat. The units for the measure would be the height, in feet, above ground level (AGL) required to communicate above an NOE-situated site. This measure was made for the baseline system only and not for all of the candidates due to testing time limitations.

2.2 Test Variables. Many variables affected two-way helicopter communications (Figure 1). The principal variables were range, altitude, and terrain; next in importance was transmitter power used in the aircraft. Finally of importance was the link tested. Links (or modes) are air-to-ground (A-G), ground-to-air (G-A), and air-to-air (A-A). Performance over these links differs.

The range and altitude variables are shown in Figure 3 in the form of a range/height matrix. Ranges at which the communications equipment were tested were selected to include the failure range for the candidate systems, particularly the VHF/FM systems which operate under near LOS conditions.

FIGURE 3. RANGE-HEIGHT CELLS



2.2.1 Range. The range intervals at which the communication systems were tested were spaced logarithmically at operationally significant distances of 1, 2.5, 5, 10, 25, and 50 km. Actual ranges for the test differed slightly from these because of terrain and military reservation boundary limitations. Selection of ranges spaced at two to one multiples of distance results in excess of 10 dB incremental basic transmission loss for a groundwave signal between each site.

The test ranges were selected to identify the capabilities and limitations of each of two modes of transmission--groundwave mode and near vertical incidence skywave (NVIS) mode of propagation. VHF/FM radio systems operate in the groundwave mode of communication in which the launched signal generally follows the surface of the earth and is refracted or reflected by terrain irregularities along the path profile between the transmitter and the receiver. Signals in the VHF portion of the spectrum (30-76 MHz) are attenuated by both range and terrain. The test ranges of 1-10 km were selected prior to the tests to bracket the expected failure range of the VHF systems originating at a base station. The HF/SSB signals also propagate in groundwave mode but to longer ranges than their VHF/FM counterpart radio.

HF/SSB radios have the capability of operating both groundwave and in near vertical incidence (NVIS) mode. For NVIS mode the energy is directed from a horizontal radiator to the ionosphere and returned to the surface of the earth. Due to NVIS propagation, HF/SSB has the capability to operate at extended ranges independent of terrain effects. The 25 and 50 km points were selected to investigate the communication performance of HF/SSB radios in the NVIS mode.

2.2.2 Altitudes. The altitude intervals for the test were selected from an operational standpoint. Three altitudes were used:

- Skids on ground--this altitude defines the bottom of the NOE flight regime.
- NOE altitude--this altitude, approximately 3-ft AGL for Fort Hood terrain, represents the top of the NOE flight regime in the test.
- Height to break squelch altitude--this is the height above ground to which the aircraft must climb to establish two-way communications. This altitude is operationally significant in that the aircraft must climb to it in order to communicate to a remote base station. As the aircraft climbs above the NOE regime, its vulnerability to ground-based weapons increases.

As can be seen from Figure 3, the choices of six ranges and three altitudes resulted in a grid or matrix containing 18 cells. This matrix constituted the sampling grid.

2.3 Sampling Plan. The tester chose to use a factorial analysis for the analysis of the test data to relate the performance of the candidate

radio systems and test variables to the dependent variable, percent correct A-N score. A complete five-factor analysis of variance was planned. [2,3]. The five factors were radio system, range, time of day, altitude, and mode of transmission (A-A, A-G, G-A). A factorial analysis is generally used to determine the relationship among many test variables and the outcome (A-N score). To perform this analysis an assumption on the distribution of the data in required--that the data be normally distributed about the mean. The analysis of variance program run by the tester revealed significant interactions between the factors and also resulted in a large computed F-ratio for the candidate radio systems. On the basis of these results, the Newman-Keuls test was run to make pair-wise comparisons of the mean A-N score of the candidate radios and to determine significance. This test is also based on the Normal assumption.

The decision to perform an analysis of variance in this manner required iterative and equal sampling in each of the range-height cells for each of the conditions of the variables. This resulted in multiple sampling in each range-height cell to establish the required confidence levels. This approach is not recommended for future tests of radio systems in which the range characteristics of the radio systems can be estimated from propagation models.

A sequential sampling is more appropriate for a test program of this nature. Under this plan, samples would be taken at each of the range-altitude cells, only until the communication effectiveness, mean A-N test score in this case, could be estimated with a 95% confidence level. The number of samples required in each cell is dependent on the mean score and confidence level required. Sequential sampling is desirable to conserve expensive test resources and to redirect those resources (helicopters) to investigate other aspects of the NOE communication problem. A comparison of the two sampling approaches is shown in Figure 4.

FIGURE 4. COMPARISON OF EQUAL-OCCURRENCE AND SEQUENTIAL SAMPLING PLANS

EQUAL SAMPLES ANALYSIS OF VARIANCE (FACTORIAL)	SEQUENTIAL SAMPLING
1. REQUIRES NORMAL ASSUMPTION AND SAMPLES FROM POPULATION WITH SAME MEAN	1. DISTRIBUTION FREE
2. REQUIRES NO A PRIORI ASSUMPTIONS ON RADIOS	2. MAKES USE OF KNOWN PHYSICAL LAWS GOVERNING RADIOS
3. REQUIRES EQUAL SAMPLES IN ALL CELLS FOR FACTORIAL ANALYSIS OF VARIANCE	3. NOT REQUIRED
4. REQUIRES NORMAL DISTRIBUTION FOR CONFIDENCE STATEMENTS	4. CONFIDENCE STATEMENTS MADE ON CELLULAR BASIS, DISTRIBUTION FREE
5. NO THRESHOLD REQUIRED	5. REQUIRES THRESHOLD FOR DUAL HYPOTHESIS TESTING
6. LESS COST EFFECTIVE	6. MORE COST EFFECTIVE
NEWMAN KEULS--SIGNIFICANCE OF MEANS	
1. REQUIRES NORMAL ASSUMPTION FOR POPULATION DISTRIBUTION	(NOTE 4 ABOVE)

The sequential sampling approach was prepared but was not used during the tests. This approach implies the use of a channel utility estimator and is more operationally oriented; simply stated: Does the system work? To frame this objective, Does the channel work?, we must first define some quantitative measure of the term work. This can be done arbitrarily in percent of messages that can be correctly received in a particular test environment, but it must be decided on before the tests are begun. This can be achieved by prefield tests, such as screenroom tests run on radio systems using the test material, by experienced judgement, or by both methods. As an example, an A-N score of 80% may be a reasonable threshold between channel acceptability and unacceptability.

A three-level hypothesis testing procedure was proposed for sequential sampling:

- Take N samples of communication performance on the channel.
- Form an unbiased test statistic, based on the performance measured.
- Use this statistic to accept one of the following hypotehses:

  H1. The channel can support communications.

  H2. The channel cannot support communications.

  H3. Cannot be determined. More samples required.

The expected results of such a sampling plan are shown in Figure 5. Test thresholds and confidence levels required were determined a priori to the experiment. Channel quality is measured at the required confidence level and by using repetitive samples. In Figure 5, G indicates a good channel, B a bad channel, and no entry indicates more samples are required. It is proposed that once #1 and #2 has been accepted, then measurements under these test conditions will be terminated. In this mannner, experimental resources can be concentrated in areas where communication performance is at or near the critical value.

FIGURE 5. EXAMPLE OF SEQUENTIAL SAMPLING PLAN RESULTS

ALTITUDE (AGL)

G	G	G	G	G	?
G	G	?	?	?	B
G	?	?	?	B	B

RANGE (KM)

G = GOOD, MEETS H1
B = BAD, MEETS H2
? = CANNOT BE DETERMINED. MORE SAMPLES REQUIRED.

FIGURE 5

For this approach, confidence levels were determined using the binomial distribution based on independent Bernoulli trials.

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

where

P(x) = Probability that exactly x correct characteristics received in n independent trials

p = Expected probability character correctly received (preselected threshold or desired probability)

q = 1 - p

Suppose we wish to test the hypothesis that the channel is acceptable (p = 0.8), using ten transmitted characters (n = 10), of which three are correctly received (x = 3).

$$P(3) = \binom{10}{x} p^x q^{10-x}$$

$$= 7.865 \times 10^{-4}$$

Under these conditions, the probability of receiving 3 characters correctly, assuming that the true (desired) probability is 0.8, is approximately 10^{-3}, or 0.1%. Hence we can reject the hypothesis that the channel is acceptable with confidence, Q, where

$$Q = 1 - 7.864 \times 10^{-4}$$

$$= 99.92\%.$$

In informal correspondence, the test officer for FM-320 estimated that for mean A-N test scores used in the field with helicopters, 85% was acceptable without repeats, 75% acceptable with repeats, and less than 70% unacceptable [5].

2.4 Test Implementation. A detailed test plan was developed by the TRADOC Combined Arms Test Activity (TCATA) to implement the test at Fort Hood, Texas. This plan is extremely complex and is a tribute to the TCATA organization. The test involved six helicopters visiting the 18 range height cells in approximately 1 hour and 30 minutes--the time duration for their fuel load. At each cell the helicopter crew was required to send and receive an A-N message from a ground station or another aircraft. This was done in three, 2-hour intervals in each 24-hour period: at night, during dawn, and during the daytime hours. Over 10,000 A-N messages were transmitted, and received, and graded during the duration of the tests.

To handle the data generated by the large volume of messages, TCATA used a remote terminal, similar to a time-sharing terminal, to enter the A-N test

scores into a central computer. The computer was an Army-owned CDC-6500 computer located at Fort Leavenworth, Kansas, The test scores were entered at the end of each day. This type of data handling system is strongly recommended for any future tests having large amounts of data. The system has a number of distinct advantages:

- Mean A-N test scores for each of the candidate radio systems were computed and updated daily.
- Cumulative results were available in real-time to the test officer and others at TCATA interested in the progress of the tests.
- It permits ongoing analysis of the tests results on the basis of these results, and allows room for redirection of resources.

Figure 6 (extracted from the TCATA FM-320 Report) [2] shows an accumulative output, called a Table of Means for each candidate radio system tested. This table was prepared daily for use by the test officer to evaluate performance and to plan tests for the succeeding days.

FIGURE 6. PERCENT OF COMMUNICATION EFFECTIVENESS AT NAP-OF-THE-EARTH ALTITUDES AT ALL RANGES

System	Ranges (kilometers)						Mode		Altitude
	1.0	2.5	5.0	10.0	25.0	50.0	AG/GA	AA	NOE
AN/ARC-114 (BASELINE)	98	90	68	70	4	1	50	64	54
HF (400W)	98	98	95	96	91	83	95	92	94
HF (100W SF)	97	97	97	96	85	65	90	89	90
GROUND RETRANSMISSION	98	97	91	94	73	35	84	73	81
AIR RETRANSMISSION	98	90	80	93	58	46	80	77	79
SPECIAL RADIO	93	89	78	70	6	1	56	*	56
HF (200W)	98	97	97	98	79	59	90	84	88
HF (40W)	98	96	93	88	70	50	82	84	82
IMPROVED FM (40W)*	97	97	86	87	7	4	63	64	63

*As tested during FM-320. Further modifications (improvements) have since been made to this candidate.

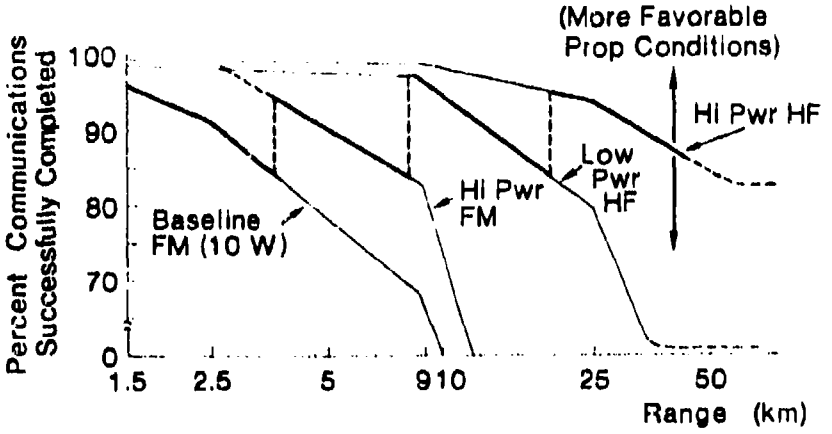
Source: TCATA FM-320 Report (Table 4-5).

3.0 RESULTS. Figure 7 is a plot of successful communications completed versus range for an A-N score equal to or greater than 80% for Fort Hood type of terrain. Success here is defined as an A-N score greater than or equal to 80%. The ordinate shows the percentage of the time that this score was equal to or exceeded. The data used to generate these curves were extracted from the TCATA FM-320 tests [2] for aircraft flying at two altitudes (skids-on-ground and a low hover), three time periods, and three communication modes (aircraft-to-ground, ground-to-aircraft, and aircraft-to-aircraft).

The curves show the advantage of an improved FM system over baseline and the improvement of a high power HF system (400W equivalent) over a low-power HF system (40W). For multipower radios, the minimum power setting should be used to achieve acceptable communication quality at the required range. Switching points for an A-N score of 85% are indicated on Figure 7. For additional and detailed analysis of this type, the reader is directed to reference [6].

Figure 7

TYPICAL COMMUNICATIONS SUCCESSFULLY COMPLETED VS RANGE FOR AN A-N SCORE ≥ 80% FOR FT HOOD TYPE TERRAIN (FM-320)



In summary, the following information was determined from the test:

A comparison of A-N scores for the nine systems at six test ranges.

The dependence of system performance on the test variables.

The relationship of the technical characteristics, such as power output, to communications performance.

Areas for improvement for the combined aircraft and ground station communication system.

4.0 LESSONS LEARNED. From the planning, conduct, and analysis of the tests, a number of lessons were learned that may be applicable to operational (and engineering) tests that the Army conducts in the future.

4.1 Measures of Effectiveness. The measure of effectiveness (MOE) selected to evaluate a communication system should be operationally significant and mission-related. A measure of effectiveness must have three characteristics. First, it must be measurable in the field. Second, it must be quantitative. Third, it must measure to what degree the objective is achieved. The MOE should have operational significance to decision makers.

4.2 Sampling Plan and Statistical Analysis. If the data for the test are to be analyzed statistically, it is recommended that the assumptions on the forms of the expected distribution of the data be carefully reviewed, and, if possible, be checked. For the FM-320 data, the tester assumed that the A-N test scores would be normally distributed about the mean. This did not prove to be the case. A sampling plan should be designed, written, and, if possible, tested before implementation of a full-scale test. The choice of an appropriate confidence level and the number of samples required under each set of variable conditions to achieve that level should be determined. The consequences of acquiring insufficent data (insufficient samples) should be investigated. Distribution free techniques should be used to estimate the required sample sizes for this type of test where the distribution form cannot be known in advance. Finally, the sampling plan should allow for test flexibility and redirection, if trends in the data so warrant.

4.3 Pretest Planning and Other Recommendations. The importance of pretest planning cannot be overstressed. It is important to review and exchange information among all test participant agencies and to change the design of the test if early results so warrant. Real-time data input and access are recommended for tests having large volumes of data. Finally, predictions or theoretical modeling should be accomplished before the start of the test and, if possible, be validated as part of the test procedure. This approach is mandatory for sequential testing. The test plan should obtain data for at least a spot check of any models which will later be used to extrapolate the test results to other situations.

5.0 CONCLUSIONS. The results of an operational/engineering test will be only as good as the planning inputs, the implementation of the test plan, and the analysis and reporting of the results.

In the fall of 1976, a large scale NOE communication test was performed by TCATA, which required 50-60 personnel, used 1,000 hours of helicopter flight time with 6 aircraft, and which used 10,000 alpha-numeric messages. This test was performed by TCATA with test inputs from the U.S. Army Avionics Research and Development Activity and U.S. Army Aviation Center to evaluate comparatively nine candidate radio communication systems.

The results of this test supplemented by additional analysis and computer predictions were a determining factor in the selection of a NOE radio

system for U.S. Army helicopters. The system selected will provide acceptable air-to-air, air-to-ground, and ground-to-air communications for helicopters operating in the NOE flight regime, and will represent a significant improvement over the present communications capability of Army helicopters.

REFERENCES

1. "Employment of Aviation Units in a High Threat Environment," U.S. Army Publication FM90-1.

2. LTC John M. Pinson, et al, "Nap-of-the-Earth Communications (NOE COMMO) System," Final Report RCS-ATCD-8, Headquarters, TRADOC Combined Army Test Activity, Fort Hood, Texas (March 1977).

3. "BMD-P Biomedical Computer Programs," Library of Congress Card No. 74-23702 (University of California Press, 1975).

4. B. J. Winer, Statistical Principles in Experimental Design (McGraw Hill, New York, New York, 1962).

5. Private correspondence from LTC J. M. Pinson, Test Officer, FM-320 (U.S. Army TCATA).

6. B. C. Tupper and G. H. Hagn, "Nap-of-the-Earth (NOE) Communications for U.S. Army Helicopters," Final Report, Contract DAAB-07-76-C-0868, SRI International, Menlo Park, California (in preparation).

TABLE LOOK-UP AND INTERPOLATION FOR A NORMAL RANDOM NUMBER GENERATOR, II

William L. Shepherd and John W. Starner, Jr.
Advanced Technology Office
Instrumentation Directorate
US Army White Sands Missile Range
White Sands Missile Range, New Mexico 88002

ABSTRACT. Results obtained since a paper of the same title was presented at the Twenty-Second Army Conference on Design of Experiment are described. An improved table look-up algorithm and more refined error norms are used. Comparison of the generator with several others is made.

1. INTRODUCTION. A paper with the same title was presented at the Twenty-Second Army Conference on Design of Experiment (Shepherd and Hynes [1]). We now present some results obtained since then. Some duplication will, of course, occur.

With

$$P(t) = \frac{1}{2} + \int_0^t \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv, \tag{1.1}$$

$$G(t) = P^{-1}(t),$$

and $\{u\}$ any output sequence for a uniform random number generator with density function equal to 1 over $[0,1]$ and 0 elsewhere, the sequence $\{G(u)\}$ can be thought of as the output sequence of an $n(0,1)$ random number generator (Abramowitz and Stegun [2], page 950).

A difficulty in using this idea is in the computation of $G(u)$. The earlier report described an interpolating quadratic spline which, once constructed, alleviates the difficulty and approximates $G(u)$ to within a prescribed accuracy. We now describe a somewhat improved procedure for constructing the spline. Blair, Edwards, and Johnson [3] furnished us with faster computation of the norm of the error, which in turn allowed a finer determination of the norm. At the same time, more compact storage of the coefficients was devised. Our experience with some uniform random number generators is presented, and the results of some statistical tests are given. The normal random number generator is compared with some others.

2. THE SPLINE APPROXIMATION FOR $G(t)$. From symmetry of $\{(t, G(t))\}$ about $(\frac{1}{2}, G(\frac{1}{2}))$, we need consider only $\frac{1}{2} \le t < 1$. First consider the knots

$$\frac{1}{2} = t_0 < t_1 < \ldots < t_{2N} < 1. \tag{2.1}$$

A continuously differentiable quadratic spline, x, with knots $\{t_i\}_{i=0}^{2N}$ can be represented by

$$x(t) = x(t_{2i+1}) + x'(t_{2i+1})(t-t_{2i+1}) + \tfrac{1}{2} x''(t_{2i+1}^-)(t-t_{2i+1})^2 \quad (2.2)$$

for $t_{2i} \le t \le t_{2i+1}$ and

$$x(t) = x(t_{2i+1}) + x'(t_{2i+1})(t-t_{2i+1}) + \tfrac{1}{2} x''(t_{2i+1}^+)(t-t_{2i+1})^2 \quad (2.3)$$

for $t_{2i+1} \le t \le t_{2i+2}$,

$i = 0, 1, \ldots, N-1$.

It can be shown that

$$x(t_{2i}) = G(t_{2i}),\ x'(t_{2i}) = G'(t_{2i}),\ \text{for } i = 0, 1, \ldots, N \quad (2.4)$$

if and only if

$$x'(t_{2i+1}) = \frac{2}{t_{2i+2}-t_{2i}} \left[G(t_{2i+2}) - G(t_{2i}) - \tfrac{1}{2}(t_{2i+2} - t_{2i+1})G'(t_{2i+2}) - \tfrac{1}{2}(t_{2i+1}-t_{2i})G'(t_{2i})\right], \quad (2.5)$$

$$x(t_{2i+1}) = G(t_{2i}) + \tfrac{1}{2}(t_{2i+1}-t_{2i})(G'(t_{2i}) + x'(t_{2i+1})), \quad (2.6)$$

$$x''(t_{2i+1}^-) = \frac{1}{t_{2i+1}-t_{2i}} (x'(t_{2i+1}) - G'(t_{2i})), \quad (2.7)$$

$$x''(t_{2i+1}^+) = \frac{1}{t_{2i+2}-t_{2i+1}} (G'(t_{2i+2}) - x'(t_{2i+1})). \quad (2.8)$$

This spline interpolates G(t) and G'(t) at every other knot. The computation of $G(t_{2i})$ can be done by the rational approximations of Blair, Edwards, and Johnson [3]. $G'(t_{2i})$ can be computed from

$$G'(t) = \frac{1}{P'(G(t))} = \sqrt{2\pi}\ e^{[G(t)]^2/2} \quad (2.9)$$

(2.5), ..., (2.8), in order, can then be used to compute the coefficients in (2.2), (2.3).

Since $G(1) = \infty$, we must choose $t_{2N} < 1$. Hence in extending x over $[t_{2N},1]$, we depart from interpolation and require that

$$x(t_{2N}) = G(t_{2N}),\ x'(t_{2N}) = G'(t_{2N}), \tag{2.10}$$

$$\int_{t_{2N}}^{t_{2N+1}} x(t)dt = \int_{t_{2N}}^{t_{2N+1}} G(t)dt := A_1, \tag{2.11}$$

$$\int_{t_{2N+1}}^{1} x(t)dt = \int_{t_{2N+1}}^{1} G(t)dt := A_2, \tag{2.12}$$

$$t_{2N+1} = \tfrac{1}{2}(1 + t_{2N}). \tag{2.13}$$

A_1 and A_2 can be evaluated by the formula

$$\int_a^b G(t)dt = \frac{1}{\sqrt{2\pi}} \left(e^{-[G(a)]^2/2} - e^{-[G(b)]^2/2}\right), \tag{2.14}$$

((2.14) can be obtained by the change of variables $t = P(u)$.)

With $\Delta := \frac{1}{2}(1-t_{2N})$, the conditions (2.10), ..., (2.13) are equivalent to

$$x''(t_{2N+1}^-) = \frac{6}{\Delta^3}\left(A_1 - G(t_{2N})\Delta - \tfrac{1}{2}G'(t_{2N})\Delta^2\right), \tag{2.15}$$

$$x'(t_{2N+1}) = G'(t_{2N}) + x''(t_{2N+1}^-)\Delta, \tag{2.16}$$

$$x(t_{2N+1}) = G(t_{2N}) + G'(t_{2N})\Delta + \tfrac{1}{2}x''(t_{2N+1}^-)\Delta^2, \tag{2.17}$$

$$x''(t_{2N+1}^+) = \frac{6}{\Delta^3}\left(A_2 - x(t_{2N+1})\Delta - \tfrac{1}{2}x'(t_{2N+1})\Delta^2\right). \tag{2.18}$$

Figure 1 illustrates $G(t)$ and $x(t)$ for $t_{2N} \le t < 1$. With this extension, $x(t)$ is a simple quadratic spline over $[\frac{1}{2}, 1]$. Table 1 gives the $4(N+1)$ coefficients corresponding to the odd numbered knots.

3. THE KNOT SEQUENCE AND SEARCH ALGORITHM. We now turn to the determination of a suitable set of knots. This set of knots must be chosen with a

Table 1. Center Knot Values and Coefficients for the Spline Approximation

i	t_{2i+1}	$x(t_{2i+1})$	$x'(t_{2i+1})$	$\frac{1}{2}x''(t_{2i+1}^-)$	$\frac{1}{2}x''(t_{2i+1}^+)$
0	.655000	.13828844	2.5220423	.14012709	.76637053
1	.655000	.39382947	2.7050517	1.0965414	1.9007992
2	.735000	.62797333	3.0423433	2.3748524	3.5848961
3	.795000	.82387042	3.5085989	4.3062559	6.0475303
4	.840000	.99443294	4.0958085	7.1208275	9.9256154
5	.875000	1.1503288	4.8417618	11.630958	15.879360
6	.900000	1.2815419	5.6846239	18.324066	23.668327
7	.920000	1.4050485	6.7007053	27.135737	37.354494
8	.935250	1.5160762	7.8989428	42.966455	52.183490
9	.945290	1.6008047	9.0084136	58.008280	72.988777
10	.955360	1.6992065	10.584978	82.930036	109.85763
11	.965460	1.8178920	13.013142	129.00044	185.42669
12	.973500	1.9348952	16.218389	217.36777	304.79406
13	.980000	2.0557069	20.523815	361.97811	532.58605
14	.985000	2.1700650	26.273978	638.66101	899.70241
15	.988500	2.2734119	33.060789	1062.6679	1486.5484
16	.991000	2.3656063	40.996771	1738.1733	2312.2766
17	.993000	2.4572319	51.011316	2694.9860	3901.1452
18	.994530	2.5446090	63.694708	4604.8235	5811.4999
19	.995536	2.6147976	76.246729	6613.0692	8795.6899
20	.996547	2.7013401	95.738746	10356.157	14967.435
21	.998500	2.9614552	175.59911	22099.942	213300.23

number of things in mind. The accuracy of our approximation depends on the knot spacing. The efficiency of our search algorithm in the table look-up depends on the exact placement of the knots, and the amount of storage depends on the number of knots used. These three considerations will be used simultaneously to obtain our knot sequence.

The accuracy to which we wish to approximate G(t) depends very much on the uniform generator used and the machine on which this algorithm is to be implemented. The most common (and most efficient) type of generator is the linear congruential type. Knuth [4], chapter 3, presents an excellent discussion of the linear congruential generator as well as some alternatives. It has been shown (Coveyou and MacPhereson [5] and Knuth [4]) that in a 35-bit word (as in the case of the UNIVAC 1108) using a linear congruential generator one can expect to have successive pairs of numbers independent only to an accuracy of about 10^{-4}. Successive K-tuples for K > 2 are independent for even smaller accuracies. It would be wasteful of effort to approximate G(t) to any greater accuracy for use with this uniform generator. It should be noted that the generator developed here is not suited for use in high resolution applications. If a greater accuracy is needed and a suitably accurate generator is obtained, a new knot sequence could be formed to make the spline sufficiently accurate.

The search algorithm determines, for any given t, a value of j so that $t_{2j} \leq t < t_{2j+2}$. Instead of using some binary search technique or a Fibonacci search, it was discovered that if we placed the knots carefully we could very simply compute an index from the value of t and then look up the value of j in a table using this index.

Let us choose the knots so that each even indexed knot is a multiple of .01 and also so that the maximum error over each interval $[t_{2j}, t_{2j+2}]$ is less than or equal to 10^{-4}. Further, we want each interval as long as possible to minimize the number of knots. This gives us the values listed in table 2.

Note that at .97 it is no longer possible to maintain an accuracy of 10^{-4} and a minimum spacing of .01. For any t in [.5, .97] let the index I_t be given

$$I_t = \lfloor 100t \rfloor - 49 \qquad (3.1)$$

This is very simple and fast to compute in FORTRAN. I_t is the index of the interval of length .01 in which t is to be found. Since all of the knots are on the boundaries of the intervals, we can look up in a table exactly which knot interval to which this index belongs.

For $t > .97$ we have a problem. The value .97 is not close enough to 1 to use the equal area criterion on this last interval, so we need more knots.

Table 2

J	t_{2j}	J	t_{2j}
0	.50	7	.91
1	.61	8	.93
2	.70	9	.94
3	.77	10	.95
4	.82	11	.96
5	.86	12	.97
6	.89		

Table 3

J	t_{2j}
13	.977
14	.983
15	.987
16	.990
17	.992
18	.994
19	.995
20	.996
21	.997

Table 4

I_t	$M1(I_t)$	I_t	$M1(I_t)$
1	0	25	2
2	0	26	2
3	0	27	2
4	0	28	3
5	0	29	3
6	0	30	3
7	0	31	3
8	0	32	3
9	0	33	4
10	0	34	4
11	0	35	4
12	1	36	4
13	1	37	5
14	1	38	5
15	1	39	5
16	1	40	6
17	1	41	6
18	1	42	7
19	1	43	7
20	1	44	8
21	2	45	9
22	2	46	10
23	2	47	11
24	2	48	12

Table 5

I_t	$M2(I_t)$
1	13
2	13
3	13
4	13
5	13
6	13
7	14
8	14
9	14
10	14
11	15
12	15
13	15
14	16
15	16
16	17
17	17
18	18
19	19
20	20
21	21
22	21
23	21

Starting at .97 we let the knots be multiples of .001. (See table 3.) We then use the index

$$I_t = \lfloor 1000t \rfloor - 976 \tag{3.2}$$

and look up the knot interval in a second index table. At the value .997 we can no longer maintain the accuracy 10^{-4} and the spacing .001. At this point, we use the equal area condition on the interval [.997, 1].

At the cost of extra storage and one further test, we could have formed a third sequence of knots starting at .997 with a spacing of .0001. We feel that the gain in accuracy near 1 does not justify the extra cost of 40 words of storage and one extra test.

We should mention here exactly what we mean by maximum error and how we compute the knots. To compute the maximum error over an interval, we compute the absolute difference between our approximant and an accurate rational approximation at 100 equally spaced points in the interval. (See Blair, Edwards, and Johnson [3].) The odd indexed knot t_{2j+1} is chosen inside the interval to an accuracy of .1% of the interval length to mimimize the maximum error over the interval. Starting with $t_0 = \frac{1}{2}$, for $j = 0,1,2,\ldots,n-1$ we compute t_{2j+1} and t_{2j+2} simultaneously to give the largest interval so that (1) the maximum error is minimized with respect to placement of the center knot, (2) the maximum error is less than or equal to 10^{-4}, and (3) the interval length is a multiple of .01.

The search algorithm is the following:

1. Input t (t is in [.5, 1])
2. If t > .97, skip to 5
3. $I_t = \lfloor 100t \rfloor - 49$
4. Return j = M1(I_t) (see table 4)
5. $I_t = \lfloor 1000t \rfloor - 976$
6. Return j = M2(I_t) (see table 5) (3.3)

4. A SPECIFIC GENERATOR AND STATISTICAL TESTING. In this section, we study a specific generator and present some empirical statistical tests. The tests are designed to study the distribution and serial correlation of the sequences generated by our algorithm. We choose a particular linear congruential uniform generator and use the tables presented here to form our generator. The uniform generator chosen was designed for a 35-bit integer word, and all of the tests were performed on a UNIVAC 1108 computing system.

The uniform generator used is of the form

$$U_{n+1} = (AU_n + C) \bmod m, \tag{4.1}$$

where $m = 2^{35}$ and A and C are chosen to give the uniform generator good statistical properties. The numbers U_n are the integers in the range $0 \le U_n \le 2^{35} - 1$. To obtain a value in (0, 1) simply divide U_n by 2^{35}.

The multiplier A is chosen to obtain good results in the spectral test. (See Knuth [4] and Coveyou and MacPhereson [5].) Coveyou and MacPhereson present several values of A and the results of the spectral test for each. We choose from their results the value A = 27214903917. Knuth presents the following criterion for choosing C. To minimize the serial pairwise correlation over the entire period, let

$$C \approx m\left(\frac{1}{2} \pm \frac{\sqrt{3}}{6}\right), \tag{4.2}$$

where $m = 2^{35}$ and C is odd. We choose the minus sign which gives the value C = 7261067085.

The first test performed on the normal generator is the Kolmogorov-Smirnov test. (See Knuth [4].) This test studies the distribution of sequences obtained from the generator. The empirical distributions of sequences of modest length (1000 numbers) are compared with the normal distribution. The maximum positive deviation (K^+) and maximum negative deviation (K^-) are determined for each sequence. The distribution of the values of K^+ and of K^- should be close to the Kolmogorov-Smirnov distribution. The deviations of these distributions from the Kolmogorov-Smirnov distribution are well within the confidence limits set forth by Knuth. Figure 2 shows these empirical distributions and the Kolmogorov-Smirnov distribution.

The second test is a measurement of the serial correlation for normally distributed sequences. We compute the following statistic for serial correlation

$$C_N = \frac{N(X_1X_2 + X_2X_3 + \ldots + X_NX_1) - \left(\sum_{j=1}^{N} X_j\right)^2}{N\sum_{j=1}^{N} X_j^2 - \left(\sum_{j=1}^{N} X_j\right)^2} \tag{4.3}$$

Anderson [6] has shown that for a truly random sequence of numbers the distribution of the serial correlation coefficients is for a large N asymptotically normal with mean $\frac{1}{N-1}$ and variance $\frac{N-2}{(N-1)^2}$. Figure 3 shows a comparison of the empirical distribution of the serial correlation coefficients of 50 sequences of 1000 numbers with the normal distribution with mean $\frac{1}{999}$ and variance $\frac{998}{999^2}$. The agreement is quite good.

5. COMPARISON WITH OTHER ALGORITHMS. We now compare our algorithm with two other popular algorithms in terms of speed, storage requirements, and ease of programming. There are many algorithms, and a discussion of most of the algorithms in use can be found in a paper by Ahrens and Dieter [7]. The two we choose here are probably the most commonly used.

One of the most popular algorithms is the polar algorithm. This is a modification by Marsaglia of the Box-Muller algorithm. The algorithm requires one floating divide, one square root, and one natural logarithm to generate two random numbers. It also requires approximately 2.5 uniform random numbers to generate two normal random numbers. The algorithm requires very little storage and is very easy to program. The difficulty with this algorithm is speed. The special function calls are very expensive.

Marsaglia, MacLaren, and Bray [8] present a faster algorithm (the rectangle-wedge-tail algorithm), which is based on the decomposition of the normal distribution into simple distributions. This algorithm is very fast, but requires much extra storage for tables. Further, to take full advantage of the speed of this algorithm, it should be programmed in machine language. There is no question that a machine language version of this algorithm is the fastest available; however, the difficulty of programming makes this algorithm somewhat inaccessable.

Our algorithm (the inverse distribution algorithm) requires some extra storage for tables. The amount required is, however, considerably less than the rectangle-wedge-tail algorithm. A FORTRAN implementation of our algorithm is also faster than a FORTRAN implementation of the rectangle-wedge-tail algorithm and is considerably faster than the polar method. Table 6 shows approximate times for the generation of one number with FORTRAN implementations of each of the algorithms and the approximate amount of extra storage required. The timings were made on the UNIVAC 1108 with the FORTRAN V compiler.

Table 6

Algorithm	Time	Storage
Polar	102 μsec	---
Rectangle-Wedge-Tail	82 μsec	707 words
Inverse Distribution	70 μsec	176 words

6. CONCLUSIONS. We have presented an algorithm for the generation of normally distributed random numbers. This algorithm is designed to be implemented in a high-level programming language such as FORTRAN. Compared with other good algorithms in FORTRAN implementations, our algorithm is the fastest and requires only a modest amount of storage. Because this algorithm is to be programmed in FORTRAN, it is portable. One must, of course, obtain a uniform generator that is designed for a given machine; however, the inverse distribution calculation is entirely machine independent. The inverse distribution approximation is accurate to 10^{-4}; however, the uniform numbers are at best independent to four places. A greater degree of accuracy is unnecessary and would materially add to the number of knots which effects both the efficiency and storage requirements. This method should be used in any application not requiring high resolution where ease of programming and speed are important and storage is not critical. This method can be used, with the appropriate table, to generate random sequences from any continuously differentiable distribution function.

REFERENCES

[1] Shepherd, W. L. and Hynes, J. N.; Proceedings of the Twenty-Second Conference on the Design of Experiments in Army Research, Development, and Testing; ARO Report 77-2; 1976; pp. 153-164.

[2] Abramowitz, M. and Stegun, I., "Handbook of Mathematical Functions," National Bureau of Standards AMS 55, 1964.

[3] Blair, J. M., Edwards, C. A., and Johnson, J. H.; Rational Chebyshev Approximations for the Inverse of the Error Function; Math. Comp. Volume 30; No. 136; 1976; pp. 827-830.

[4] Knuth, D., "The Art of Computer Programming, Vol. II: Seminumerical Algorithms," Addison Wesley, Reading, Massachusetts, 1969.

[5] Coveyou, R. R. and MacPhereson, R. D., Fourier Analysis of Uniform Random Number Generators, JACM, Volume 14, No. 1, 1967, pp. 100-119.

[6] Anderson, R. L., Distribution of the Serial Correlation Coefficient, Ann. Math. Stat., Volume 13, 1942, pp. 1-13.

[7] Ahrens, J. H. and Dieter, J., Computer Methods for Sampling from the Exponential and Normal Distributions, CACM, Volume 15, No. 10, 1972, pp. 873-882.

[8] Marsaglia, G., MacLaren, M. D., and Bray, T. A.; A Fast Procedure for Generating Normal Random Variables; CACM; Volume 7; No. 1; 1964, pp. 4-10.

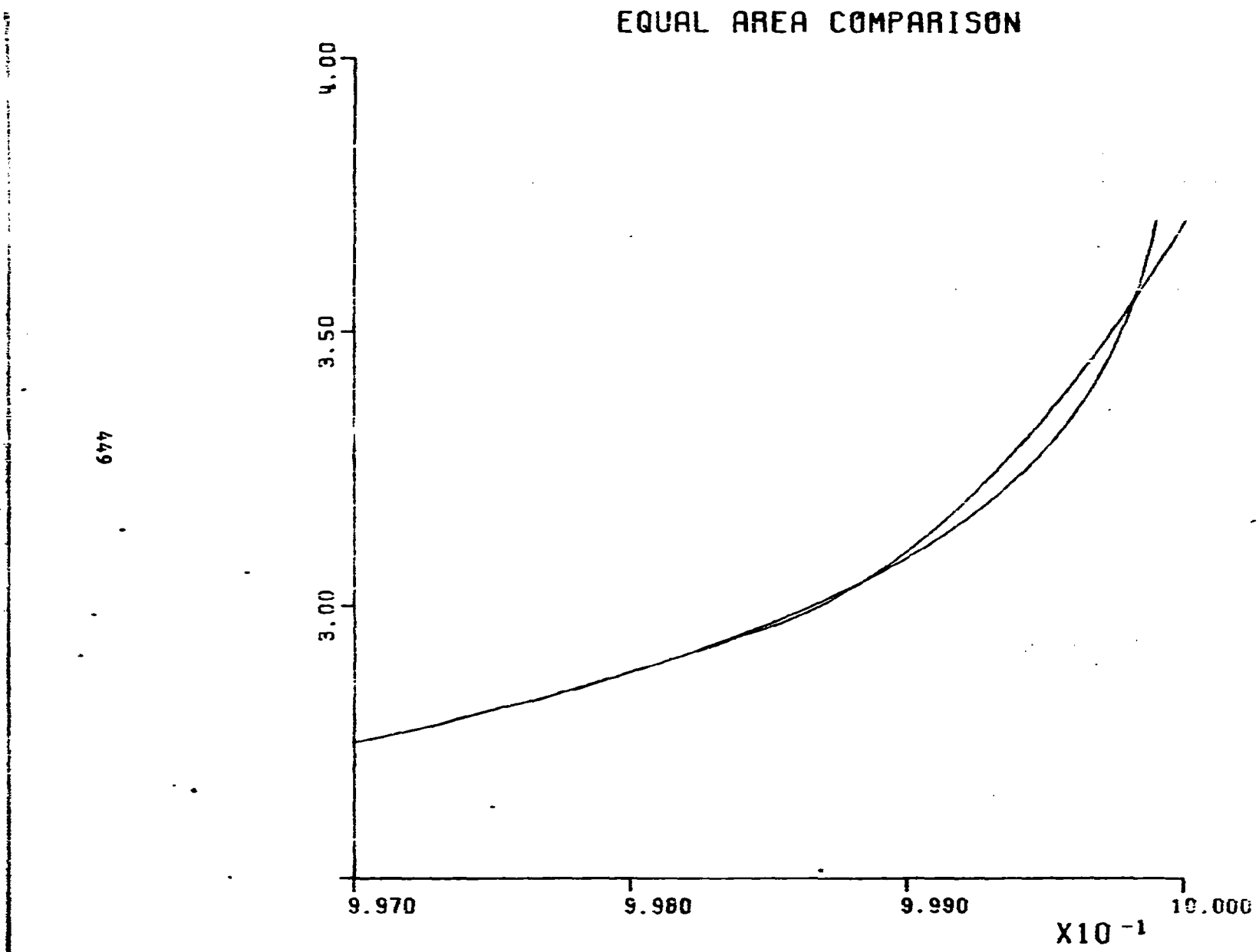


Figure 1

KOLMOGOROV-SMIRNOV TEST RESULTS

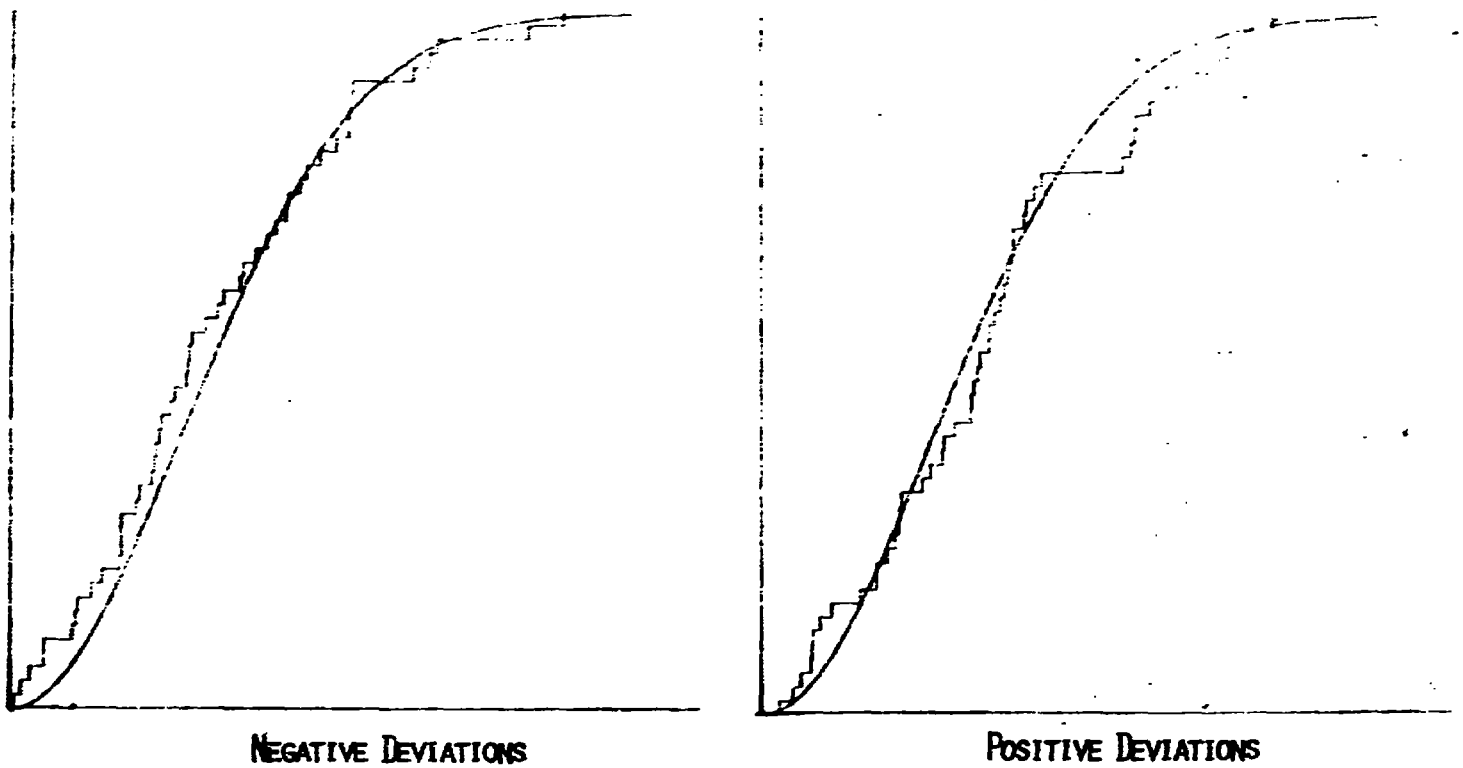


Figure 2

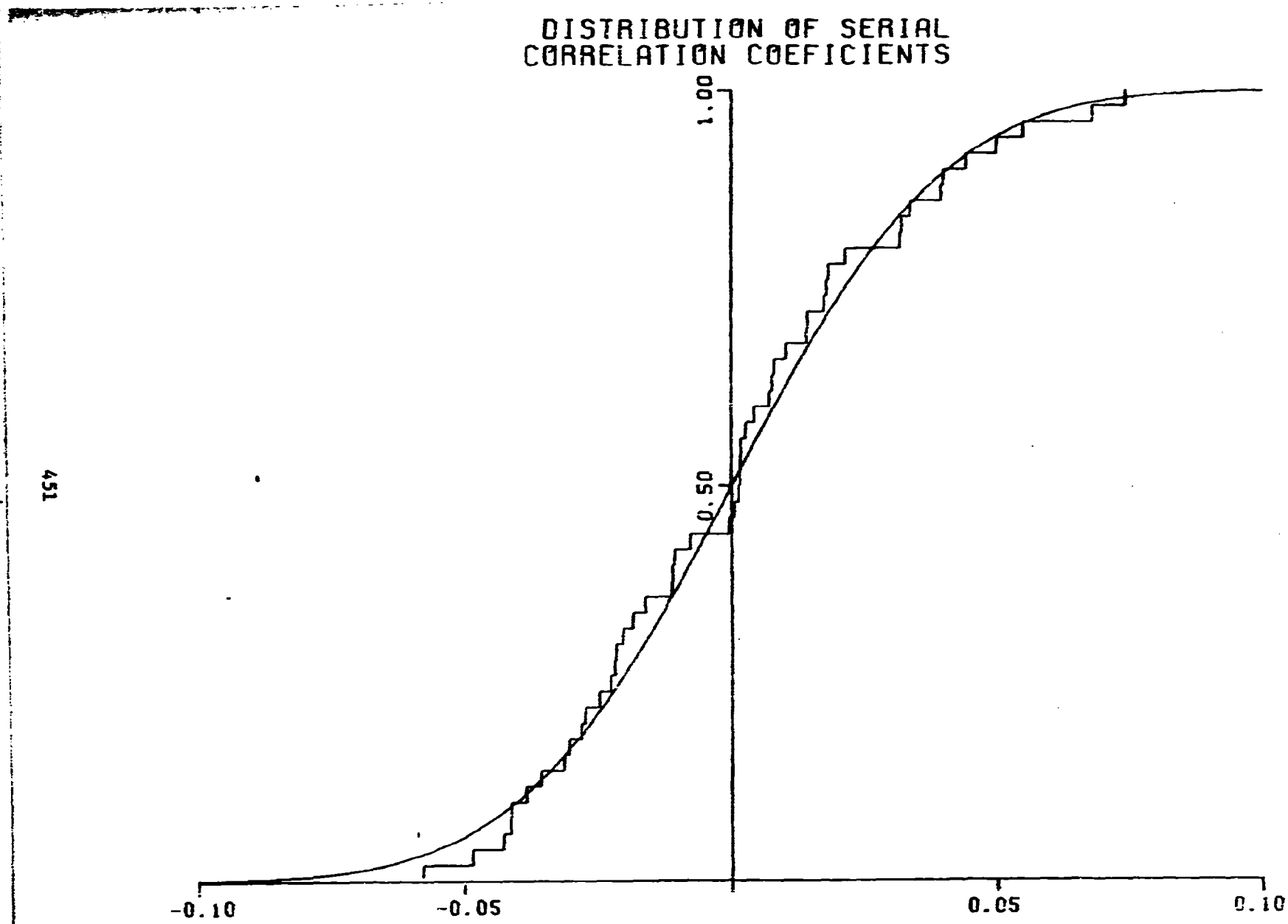

DISTRIBUTION OF SERIAL
CORRELATION COEFICIENTS
1.00
0.50
-0.10
-0.05
0.05
0.10

Figure 3

DIRECT DEGENERACY ATTAINMENT IN MARKOV CHAINS

Richard M. Brugger
Quality Evaluation Division
Product Assurance Directorate
US Army Armament Materiel Readiness Command
Rock Island, Illinois

ABSTRACT. Some procedures for solving for steady state probabilities are more complicated than necessary. This paper shows that by not introducing the equation reflecting that the sum of steady state probabilities is one into the matrix solution, the work becomes easier.

I. INTRODUCTION. This paper deals with the matter of determining steady state probability expressions for Markov chains. In particular, it deals with the matter of working with the set of equations from which the steady state probability expressions are derived.

As is well known, Markov chain methodology is often useful, and is sometimes the only methodology available, for dealing with certain types of problems related to such applications as determining sampling plan properties or analyzing the characteristics of a weapons system.

The motivation for this paper arose from a training course in which the author was enrolled. In this training course, a method of solution for the steady state probability expressions was presented which was much more complicated than the method which I had been using. Reviewing some of the more well-known textbooks that included material on Markov chains, it was noted that mathematical concepts of solution were presented, but generally no algorithms were provided to carry out these mathematical concepts. This paper, then, without benefit of references, will provide the algorithm from the training course and a simpler algorithm that the author has been using for some time. This simpler algorithm may not be well known, since, as mentioned, the better known textbooks on Markov chains tend to avoid detailed descriptions of algorithms.

Throughout the paper, ergodic chains only are considered.

II. THE LONGER METHOD. As an example, consider the chain represented by the matrix in Figure 1.

Figure 1
Transition Matrix

	S1	S2	S3
S1	p	q/2	q/2
S2	--	p	q
S3	p	--	q

In this chain, $p + q = 1$. Let $P(Sj)$ represent the steady state probability of state j. From the matrix, proceeding column by column, we can extract the following set of equations:

$$P(S1) = pP(S1) + pP(S3) \tag{1}$$

$$qP(S1) - pP(S3) = 0 \tag{2}$$

$$P(S2) = (q/2)P(S1) + pP(S2) \tag{3}$$

$$(1/2)P(S1) - P(S2) = 0 \tag{4}$$

$$P(S3) = (q/2)P(S1) + qP(S2) + qP(S3) \tag{5}$$

$$(q/2)P(S1) + qP(S2) - pP(S3) = 0 \tag{6}$$

Taking equations (2), (4), and (6) from above and taking into account that the sum of the steady state probabilities is equal to one, we have the following set of linear equations:

$$qP(S1) - pP(S3) = 0 \tag{2}$$

$$(1/2)P(S1) - P(S2) = 0 \tag{4}$$

$$(q/2)P(S1) + qP(S2) - pP(S3) = 0 \tag{6}$$

$$P(S1) + P(S2) + P(S3) = 1 \tag{7}$$

We shall see later that including equation (7) at this time was not wise.

We have a set of four equations in three unknowns, and we know that since the chain is ergodic, exactly two of the equations can be transformed into linearly dependent equations, thus producing a degeneracy which in effect reduces the set of equations to three linearly independent ones.

The process of working with these equations to attain this degeneracy can be done in a variety of ways. A standard approach is the so-called sweep out method, which we shall use here.

$$P(S1) - (p/q)P(S3) = 0 \tag{8}$$

$$P(S1) - 2P(S2) = 0 \tag{9}$$

$$P(S1) + 2P(S2) - 2(p/q)P(S3) = 0 \tag{10}$$

$$P(S1) + P(S2) + P(S3) = 1 \tag{11}$$

Subtracting (8) from (9), (10), and (11), we obtain:

$$P(S1) - (p/q)P(S3) = 0 \quad (12)$$

$$-2P(S2) + (p/q)P(S3) = 0 \quad (13)$$

$$2P(S2) - (p/q)P(S3) = 0 \quad (14)$$

$$P(S2) + (1 + (p/q))P(S3) = 1 \quad (15)$$

We see that equation (13) is simply minus one times equation (14), so we will discard equation (13).

Proceeding we obtain:

$$P(S1) - (p/q)P(S3) = 0 \quad (16)$$

$$P(S2) - (p/2q)P(S3) = 0 \quad (17)$$

$$P(S2) + (1 + (p/q))P(S3) = 0 \quad (18)$$

Continuing, we obtain:

$$P(S1) - (p/q)P(S3) = 0 \quad (19)$$

$$P(S2) - (p/2q)P(S3) = 0 \quad (20)$$

$$(1 + (p/q) + (p/2q))P(S3) = 1 \quad (21)$$

Solving for P(S3) in (21) and doing appropriate substitutions in (20) and (19), we finally obtain:

$$P(S3) = 2q/(2 + p) \quad (22)$$

$$P(S2) = p/(2 + p) \quad (23)$$

$$P(S1) = 2p/(2 + p) \quad (24)$$

It can be seen that even with a very simple example, a great deal of effort was expended in order to obtain a solution using this long method.

III. THE SHORTER METHOD. Refer again to Figure 1, the transition matrix for this example. We will work with the matrix differently at this time. First, we select the most complicated looking column of the matrix. This is column S3, since it contains an element in each row. We will then proceed to solve for each steady state probability in terms of P(S3).

We thus obtain:

$$P(S1) = pP(S1) + pP(S3) \tag{25}$$

$$P(S1) = (p/q)P(S3) \tag{26}$$

$$P(S2) = (1/2)q\ P(S1) + pP(S2) \tag{27}$$

$$P(S2) = (1/2)P(S1) = (p/2q)P(S3) \tag{28}$$

$$P(S3) = P(S3) \tag{29}$$

It is interesting to note that (29) permits us to disregard all of the elements in column S3. This is why we selected the most complicated looking column, because by so doing we eliminate more work.

Since the sum of the steady state probabilities equals one, and since

$$P(Sj) = \frac{a_j\ P(Sj)}{\sum_{i=1}^{3} a_i\ P(Si)}$$

$j = 1, 2, 3$, (where a_j represents the coefficient of P(S3) in (26), (28), and (29)) and since P(S3) cancels from each term, we can immediately write the solution as:

$$P(S1) = 2p/(2 + p) \tag{30}$$

$$P(S2) = p/(2 + p) \tag{31}$$

$$P(S3) = 2q/(2 + p) \tag{32}$$

As is obvious, this is much simpler than the other method.

THE CURSE OF THE EXPONENTIAL DISTRIBUTION IN RELIABILITY*

Leon H. Herbach
Polytechnic Institute of New York, 333 Jay st., Brooklyn, New York 11201

J. Arthur Greenwood
Oceanweather Inc., White Plains, New York

Saul B. Blumenthal
University of Kentucky, Lexington, Kentucky

The exponential is wrong
But works like a song.
Beware the Weibull:
It's incorrigible.—Anon.

All models are wrong.
Some work.—G. E. P. Box

ABSTRACT. The fact that failures follow the exponential distribution is almost universally accepted in reliability analysis. Two reasons are given for this assumption: (1) It is commonly assumed that electronic components do not wear out but are subject to random „shocks" which may cause failure. If these shocks form a Poisson process the underlying failure distribution *is* exponential. (2) Sufficiently complex equipment run for a sufficiently *long* time (failed components being replaced by good ones) will follow the exponential distribution. These reasons are investigated, especially the latter one. In many cases, equipment do not last *long enough* to reach the steady state alluded to in (2).

1. INTRODUCTION. The exponential distribution is used, almost exclusively, for the time between failures in reliability analysis. Even when it cannot be assumed that the failure distribution of a component is exponential, the exponential distribution is used for the time between failures of systems. The rationale for this is the belief that there is a theorem which states that for *large* systems the time between failures *is* exponentially distributed. Use of the exponential distribution simplifies the analysis considerably: it is well known that systems, whose failure law follows the exponential distribution, do not age; the exponential failure law is the only continuous distribution with this property. Since the analysis using any other failure law complicates the solution considerably, engineers are loth to give up use of the exponential. If retaining the exponential leads to incorrect conclusions, one might say that the reliability engineer is „being seduced by an easy solution" or is „cursed by the exponential distribution". The purpose of this paper is to state, somewhat colloquially but a little more precisely, the theorem

*Preparation of this paper was partially supported by the Office of Naval Research under Contract No. N00014-77-C-0601/NR042-377.

underlying the correct use of the exponential failure law for systems whose components fail according to another law, and to show the dangers when this theorem is not used correctly.

This paper is concerned with the *superimposed renewal process*, illustrated in Figure 1 for the case of $n = 5$ components connected in series. When any component fails, the system fails. We assume that a failed component is instantly replaced by a new one. The ×'s indicate times of failure for each component and the bottom line indicates the failures of the renewal process or system. One version of the exponential limit theorem [4] states that if one has a renewal process consisting of n components, with identical non-exponential failure laws, connected in series; then, for n greater than some n^* and t greater than some t^*, the times between failures of the system are indeed exponentially distributed. Intuitively the theorem states that for a sufficiently complex system, after some time t^* the components have been replaced at „random" times, and there is a random mix of ages of components. Thus the succeeding times of failure will occur at random--one of the postulates of a Poisson process, which implies that times between failures follow the exponential law.

We have investigated how large n^* and t^* must be for the limit theorem to yield a good approximation when the underlying component failure law is lognormal, gamma, or Weibull. For all those laws it appears that the dependence on n is not so crucial as the dependence on t; it is believed, however, that reliability engineers frequently ignore the dependence on t.

Actually the exponential limit theorem is more general than given above. Under certain conditions, the components need not all have the same failure distribution: in this case t^* would have to be larger yet, and the results given here would be even stronger.

2. RENEWAL DENSITY AND SYSTEM HAZARD. Although the mathematical details, which appear elsewhere [1, 2, 3], will not be repeated here, we will give some definitions, outline the techniques used, and present some cases to illustrate the results. Calculations are based on

$$h(t) = \text{renewal density of a component}$$

$$= f(t) + f(t)*f(t) + [f(t)]^{*3} + \ldots + [f(t)]^{*n} + \ldots,$$

where $[f(t)]^{*n}$ denotes the n-fold convolution of $f(t)$, i.e. the density of the distribution of the time to the nth failure of the component, measured from the initial time; and $f(t)$ is the failure density of a component. Thus $h(t)$ is the density of *all* failures for a specific component and $h(t)dt$ is the probability that, in the interval $(t, t+dt)$, the component either fails for the first time or fails for the second time if it was replaced prior to t or fails for the third time if it failed twice and was replaced prior to t, etc. It can be shown that $h(t) \to 1/\mu$ as $t \to \infty$, where μ is the mean time to failure of a component. Note that the renewal function

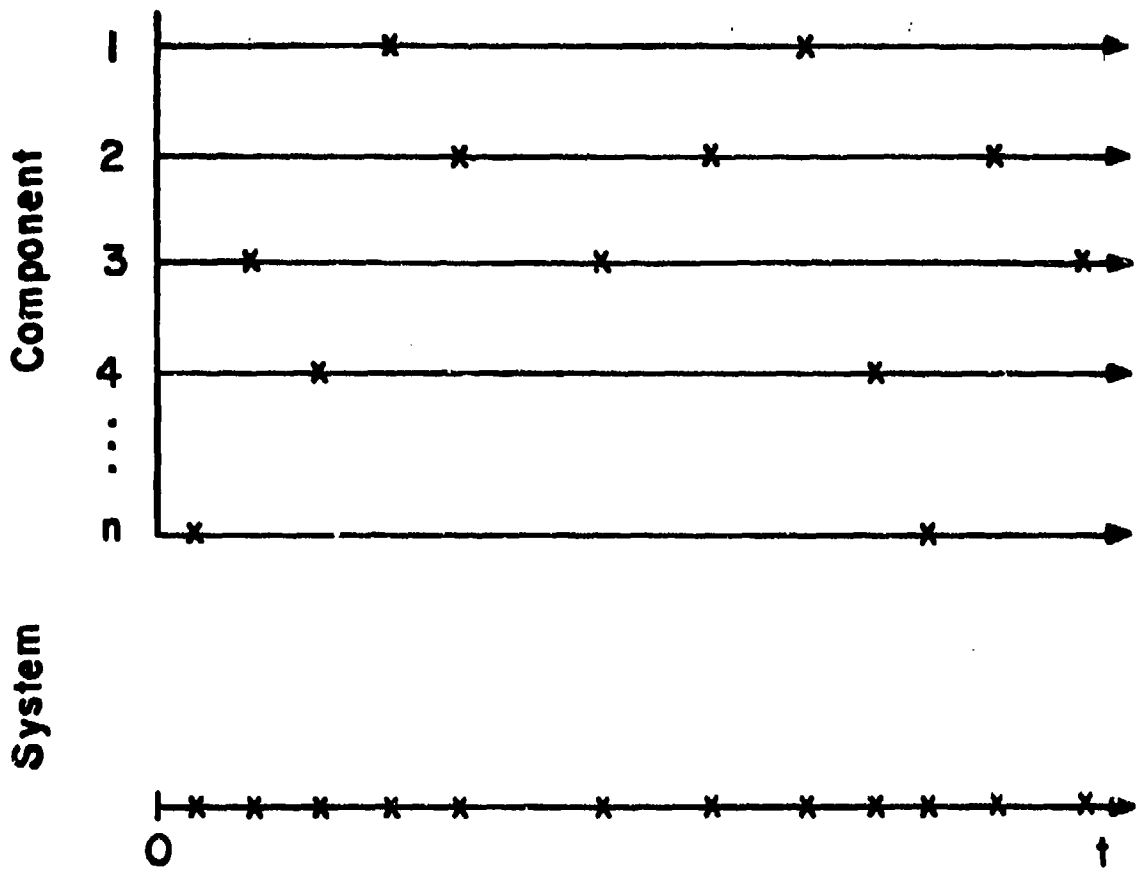


FIGURE 1. Failure times in a superimposed renewal process

$$H(t) = \int_0^t h(\tau)d\tau = \text{Expected number of failures up to time } t,$$

and that $H(t) \sim t/\mu$ - constant, where the constant reflects the fact that, for small t, $h(t)$ is typically less than $1/\mu$.

Let $h_s(t)$ be the system hazard so that $h_s(t)\Delta t$ is the probability that the system fails in the interval $(t, t+\Delta t)$, given that it was operating at time t. For $\Delta t << t$ the probability of more than one failure in the interval is negligible and $h(t)$ will be reasonably constant in the interval. These of course are the postulates of a Poisson process, and would suggest that some exponential limit will apply. In addition, $h(t)$ is an ensemble average over many components with different replacement histories. It is not an appropriate failure rate to use at time t for a component last replaced at some known time; in a system of n components connected in series, however, the summation of failures over components is a good approximation to this ensemble average for large n. Since we have assumed that failed components are instantly replaced, the mean number of system failures in the interval is rigorously n times the mean number of failures of a component. Thus $nh(t)$ is rigorously the system failure rate at time t, computed when the system is first put on test ($t = 0$). Because of the averaging over n components, $nh(t)$ is a good approximation to the system failure rate at time t, computed at time t, after we know the system history; and $h_s(t)\Delta t$ can also be considered as the mean number of system failures in $(t, t+\Delta t)$.

We are interested in

$$\mathcal{P} = \mathcal{P}(\omega;t,n) = \Pr\{\text{next failure occurs after } t+\omega \mid \text{present age is } t\}.$$

But, for large n, $\mathcal{P} \sim \mathcal{L}$, where

$$\mathcal{L} = \mathcal{L}(\omega;t,n) = \Pr\{\text{next failure occurs after } t+\omega \mid \text{failure occurred at } t\}.$$

Note that in $\mathcal{P}$ we have n components with unknown ages (although we do know the distribution of those ages), while in $\mathcal{L}$ we have $n-1$ components with unknown ages and 1 component which is new at time t. Thus the first moment of $\mathcal{L}$ is the mean time between failures. Define s, a dimensionless waiting time, by $\omega = \mu s/n$, where μ is the average time to failure of a component and μ/n is the average waiting time between system failures. Then it has been shown [3] that, neglecting terms of the order of n^{-3},

$$\mathcal{L}(\mu s/n;t,n) = \exp\{-\mu s h(t)\}\{1 - (\mu s)^2 L_1/2n - (\mu s)^3 L_2/24n^2\}, \qquad (1)$$

where

$$L_1 = L_1\{\mu s,\ h(t),\ h'(t)\} \text{ and } L_2 = L_2\{\mu s,\ h(t),\ h'(t),\ h''(t)\};$$

i.e. the „correction" terms depend on μs and the renewal density and its derivatives. This dependence is reasonable. For large ω (earlier in this section, when relating the system hazard to the Poisson process, ω was denoted Δt) the system hazard $h_s(t+\theta\omega)$, $0 < \theta < 1$, is not a constant; so that $h_s(t) \neq h_s(t+\omega)$. The mean number of failures in time ω is given by

$$\int_0^1 h_s(t+\theta w)\, w\, d\theta \ .$$

Using a Taylor expansion around t for the integrand will involve the derivatives of h.

Now, for n infinite, (1) becomes

$$\lim_{n\to\infty} \mathfrak{L}(\mu s/n; t, n) = \exp\{-\mu s h(t)\}, \tag{2}$$

and the waiting time is characterized by a non-homogeneous Poisson process. If, furthermore, $n \to \infty$, then $h(t) \to 1/\mu$ and we have

$$\lim_{t\to\infty} \lim_{n\to\infty} \mathfrak{L}(\mu s/n; t, n) = e^{-s}, \tag{3}$$

the limit theorem referred to in Section 1.

We shall present results based on (1) and (2) when the underlying failure distribution is gamma or Weibull. For the gamma we have

$$f(x) = x^{\alpha-1}\exp(-x/\theta)/\{\theta^{\alpha}\Gamma(\alpha)\},\ x > 0,\ \theta > 0,\ \alpha > 0; \tag{4}$$

$$\mu = \theta\alpha; \tag{5}$$

for the Weibull,

$$f(x) = px^{-1}(x/\theta)^{p}\exp\{-(x/\theta)^{p}\},\ x > 0,\ \theta > 0,\ p > 0; \tag{6}$$

$$\mu = \theta\Gamma(1+p^{-1}). \tag{7}$$

3. EXAMPLES. $\mathfrak{L}(w; t, n)$ is plotted as a function of t in Figures 2-9 for gamma and Weibull components. The smooth curve represents $n = \infty$, + represents $n = 64$ and × represents $n = 256$. Figures 2, 4, 5 appeared in [1]; Figures 3, 6, 7, in [3]; Figures 8, 9 were used in the oral presentation of [5] but did not appear in the Proceedings and have not been published previously.

In interpreting the gamma plots, Figures 2-7, several successive transformations from real time to coded time must be made. Start with T, the age of the system, and W, the waiting time, both in clock hours; so that we are concerned with failures in the interval $(T, T+W)$. Then transform: *(a)* Eliminate θ by computing $t = T/\theta$ and $w = W/\theta$. *(b)* The non-dimensional waiting time

$$s = nW/\mu = nW/(\theta\alpha) = nw/\alpha.$$

(c) The curves are indexed by e^{-s}, the double limit for n and T infinite, which is given equally spaced values from .05 to .95; thus

$$W = -\mu n^{-1}\log e^{-s}.$$

(d) Instead of t,

$$t/\alpha = T/\mu$$

was used in order to relate the plots to systems composed of elements having unit mean life regardless of α. To have used t would involve

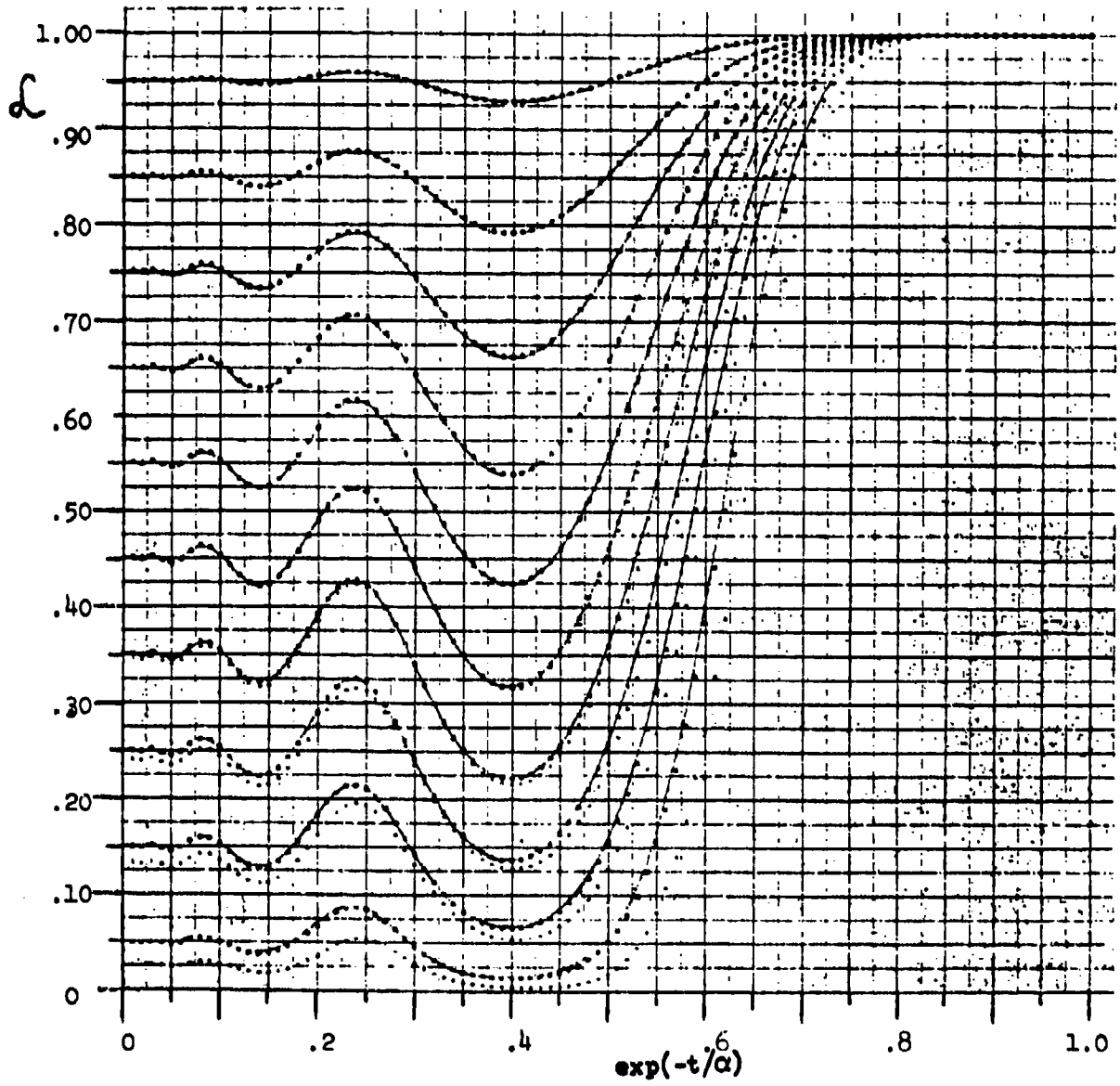


FIGURE 2. $\mathfrak{L}(\alpha s/n;\ t,\ n)$ for n gamma components: $\alpha = 12$

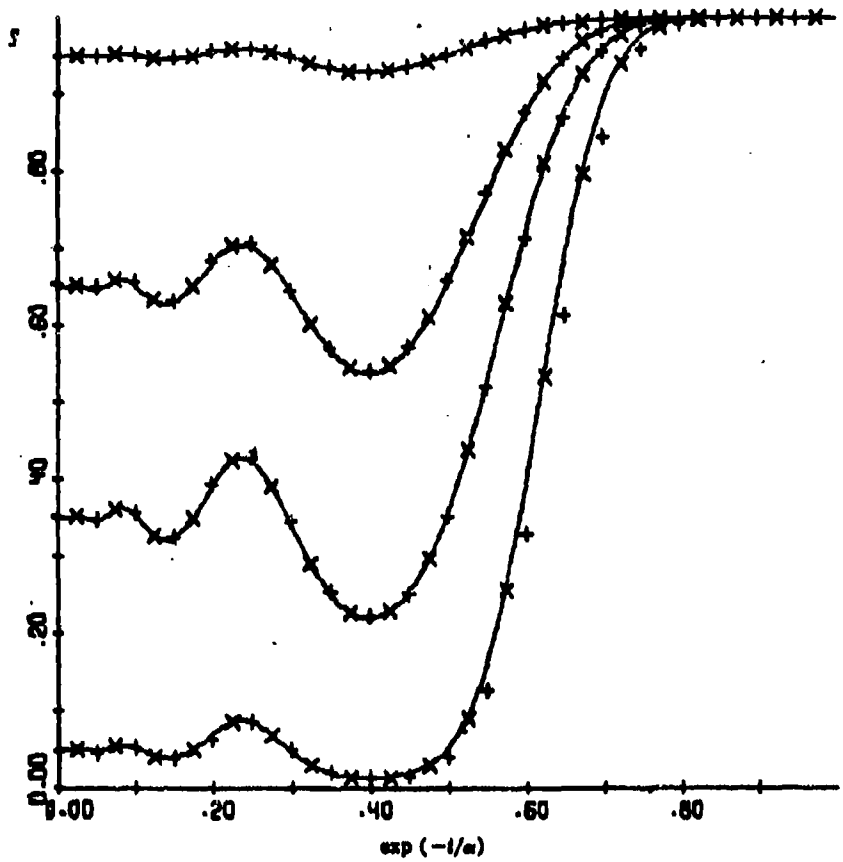


FIGURE 3. $f(\alpha s/n;\ t,\ n)$ for n gamma components: $\alpha = 12$

making plots for systems whose elements had different mean lives for different values of α and would make comparison of the results for different α more difficult, since both shape and mean life would be changing. *(8)* Finally, $\exp(-t/\alpha)$, rather than t/α, was taken as the argument, to „compress" the abscissa in the curves. This final normalization means that the gamma plots must be read from right to left: $t = 0$ and ∞ correspond to abscissas of 1 and 0 respectively. (The Weibull plots, Figures 8 and 9, read from left to right.)

The asymptotic probability e^{-s} ranges from 0.05 to 0.95 by steps of 0.10 in Figures 2, 4, 5 and by steps of 0.30 in Figures 3, 6, 7, 8, 9. Thus the top curve in Figure 2 corresponds to $s = \log(.95) \approx .05$; $\alpha = 12$, $\omega = \alpha s/n \approx .6/n$. Because ω depends on both α and s, each curve on any figure represents a different ω; the same ω, moreover, corresponds to different W as θ is varied.

To illustrate these somewhat confusing transformations that take W into s, consider a system with $n = 300$ components, $\alpha = 2$, and $\theta = 5000$ hours, so that $\mu = 10{,}000$ hours; and let the contemplated waiting time

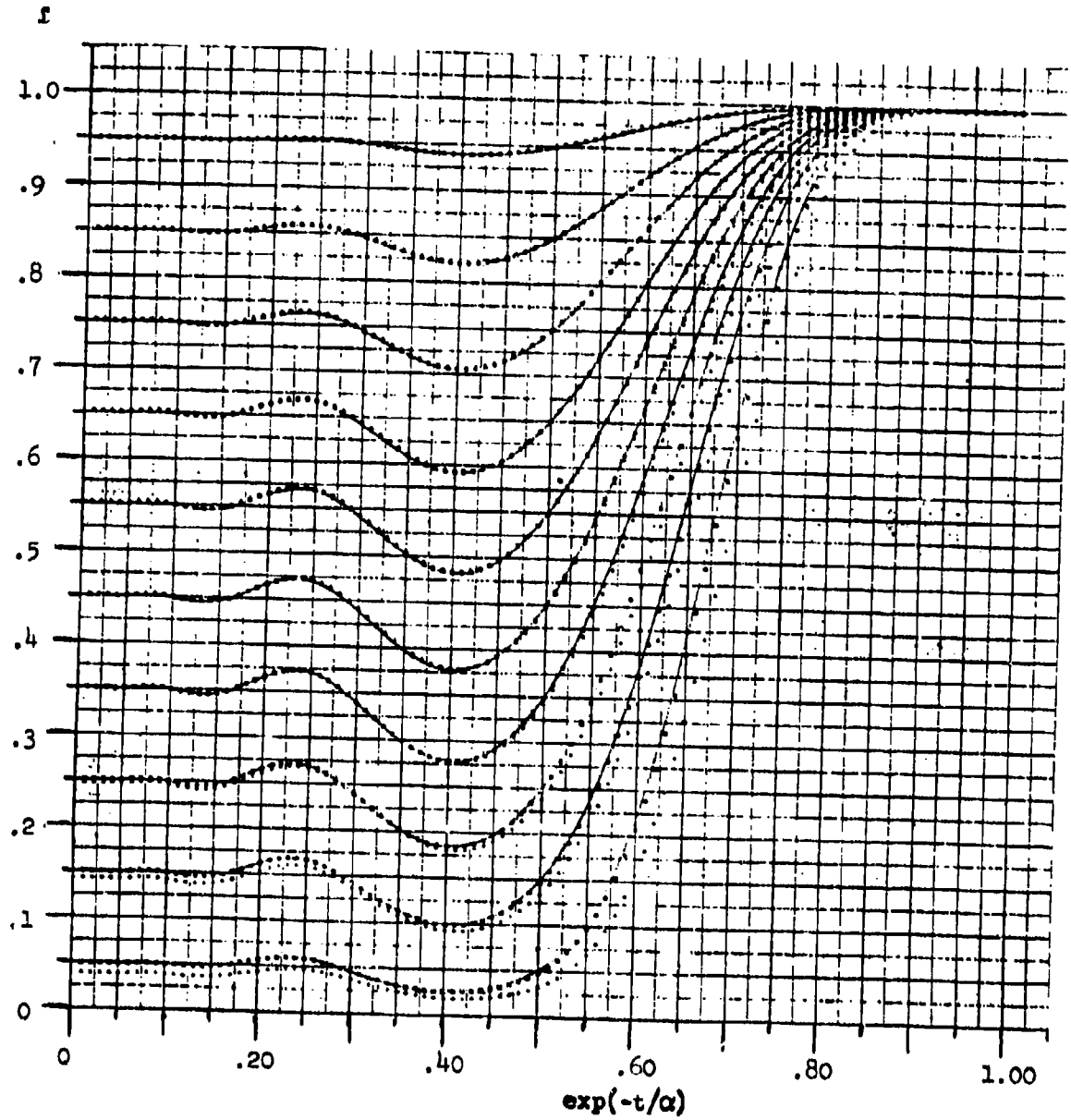


FIGURE 4. $\mathfrak{L}(\alpha s/n;\ t,\ n)$ for n gamma components: $\alpha = 8$

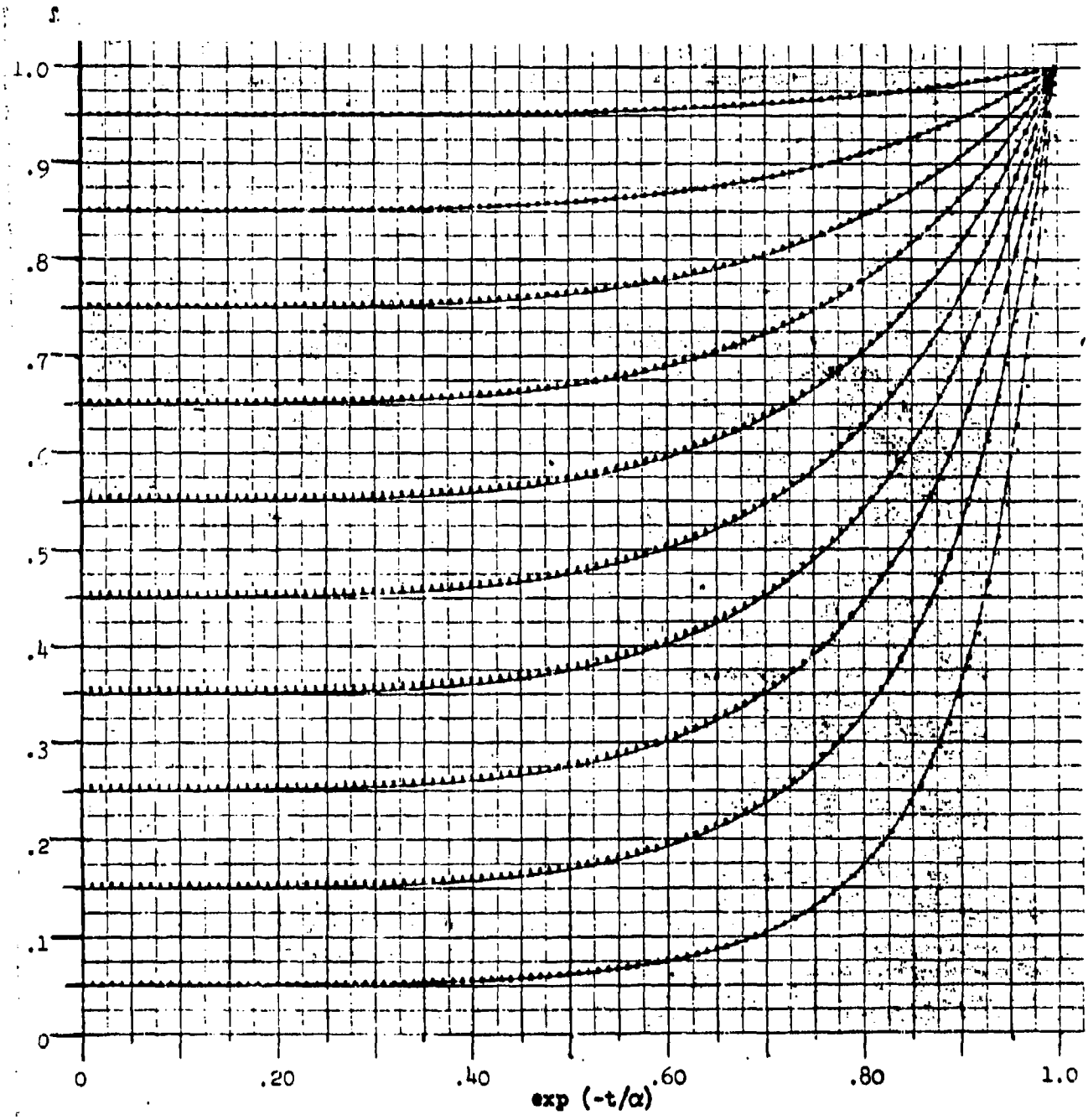


FIGURE 5. $\mathfrak{L}(\alpha s/n;\ t,\ n)$ for n gamma components: $\alpha = 2$

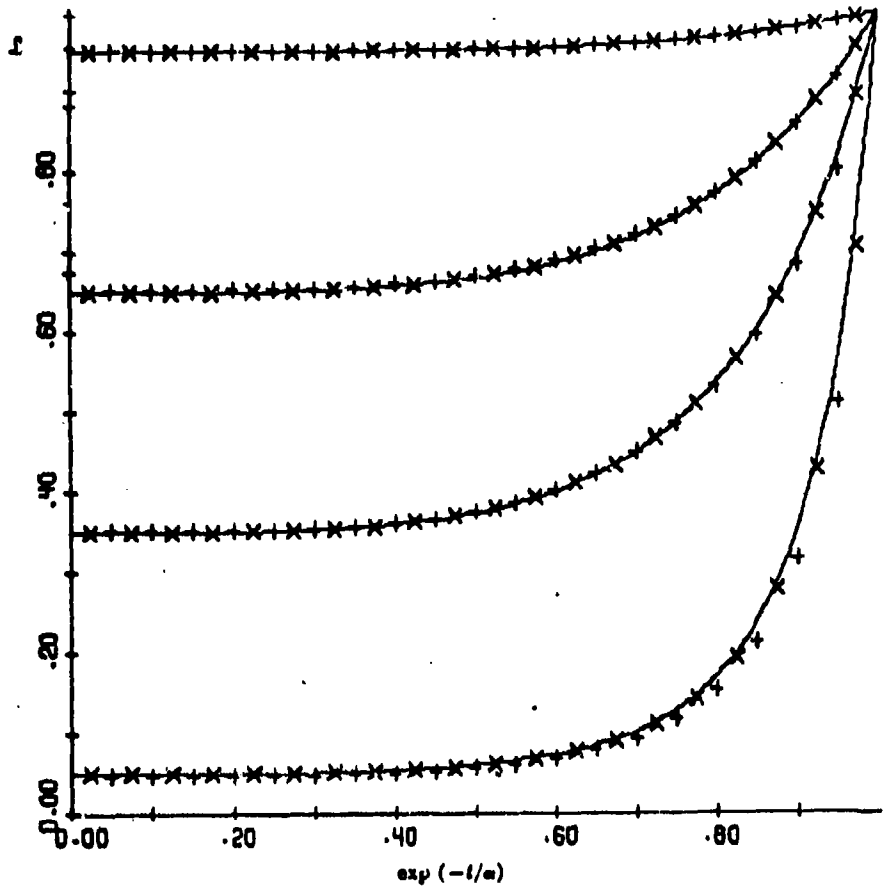


FIGURE 6. $\mathcal{L}(\alpha s/n;\ t,\ n)$ for n gamma components: $\alpha = 2$

ω = 100 hours. Then

$$s = nW/\mu = 300\times100/10{,}000 = 3,\ e^{-s} = 0.05.$$

Thus the time-equilibrium probability ($t = \infty$) that the system operates for at least 100 hours without a failure is 0.05.

As another example, suppose we desire to find the probability that a system of 100 components survives without failure for at least 24 hours when all of the components have the gamma distribution with $\alpha = 2$ and mean life 10,000 hours (θ = 5,000 hours). The system age is T = 10,000 hours. We have $t = 2$, $t/\alpha = 1$, $s = 100\times24/10{,}000 = 0.24$; so that

$$e^{-s} = 0.787,\ e^{-t/\alpha} = 0.368.$$

Interpolating in Figure 5, we find $\mathcal{L} \approx 0.792$. Alternatively one could show that $\mu h(t) = 0.984$ and use (2) to obtain

$$\mathcal{L} = e^{-.24\times.984} = e^{-.236} = 0.790.$$

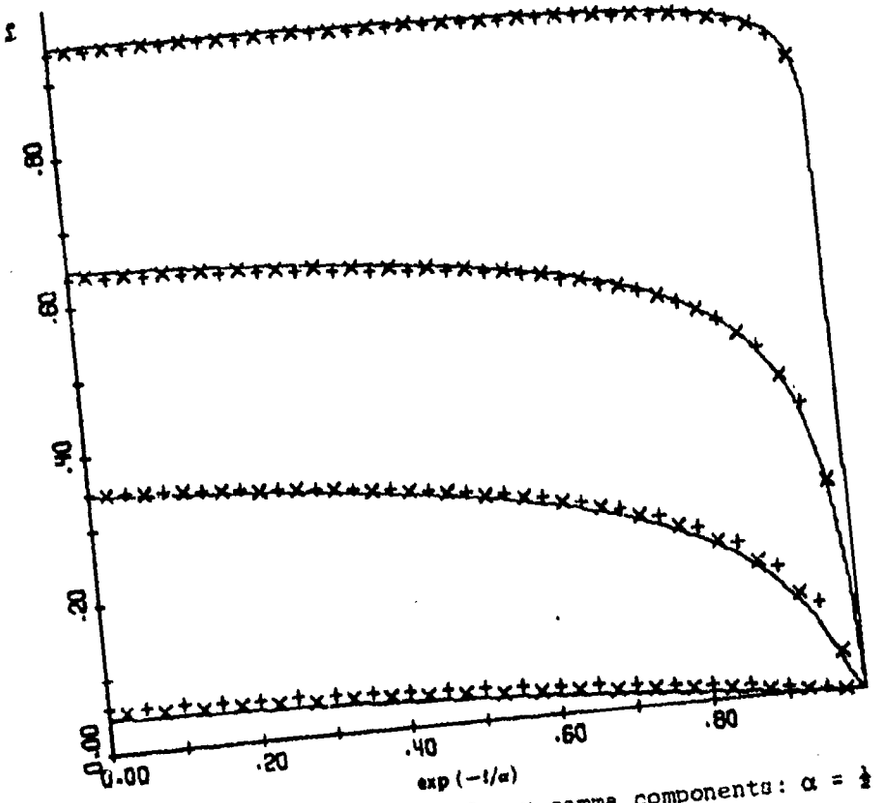


FIGURE 7. $f(\alpha s/n;\ t,\ n)$ for n gamma components: $\alpha = \frac{1}{2}$

It may be worth noting that when the time scale is changed (as from T to t), the same scale factor appears in the denominator in μ and in the numerator in h, so that the product $\mu h(t)$ is invariant.

Suppose the system, with size and age as above, consisted of gamma-distributed components with $\alpha = 4$ and mean life of $\mu = 10{,}000$ hours, so that $\theta = 2500$ hours; and we desire the survival probability for 24 hours as before. Then $t = 4$, $t/\alpha = 1$, $s = 0.24$, and the equilibrium probability is 0.787 as before. It can be shown that $\mu h(t) = 1.028$; and a calculation including terms in negative powers of n yielded $f = 0.782$.

At $T/\mu = t/\alpha = 1$, as in the last two examples, the time-dependent correction is only moderate; the next example will consider a system less well aged. Suppose the system consists of 100 components with $\alpha = 2$, $\mu = 10{,}000$ hours ($\theta = 5{,}000$ hours); $W = 24$ hours, and the system is $T = 2200$ hours old. Then $t = 2200/5000 = 0.44$, $t/\alpha = 0.22$, and $\exp(-t/\alpha) = 0.8$. Also, $s = 100\times24/10{,}000 = 0.24 = -\log(.787)$. Interpolation in Figure 5 at an abscissa of 0.8 yields $\mu h = .5904$ and

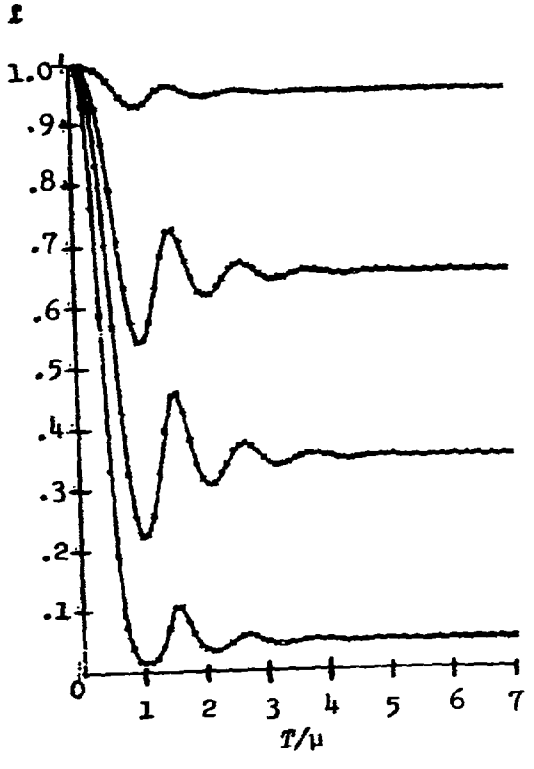


FIGURE 8. $p = 4.0$

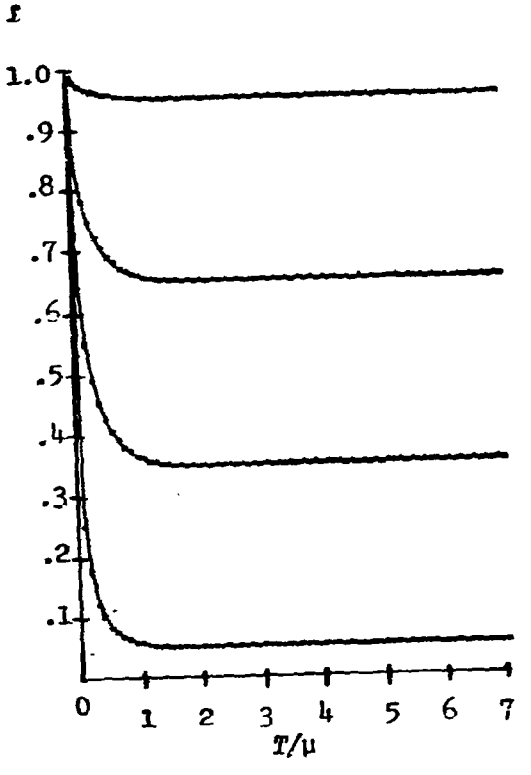


FIGURE 9. $p = 1.5$

$f\ (\mu s/n;\ T/\mu,\ n)$ for n Weibull components

$$\mathfrak{L} \approx e^{-.24\times.5904} = e^{-.141} = 0.868.$$

This is a somewhat larger survival probability than the time-equilibrium prediction would give. The difference is more striking if we consider the probability of surviving 240 hours so that $s = 2.4$;

$$e^{-s} = 0.091 \text{ and } \mathfrak{L} \approx e^{-\mu s h(t)} = e^{-1.4} = 0.247,$$

which is considerably larger than the equilibrium value, 0.091. The errors in ignoring system age are seen to be far greater for large waiting times than for small ones.

Several global conclusions can be drawn from these curves. The most important is that the effects of finite t are more important than the effects of finite n. This may be seen from the wide fluctuations of $\mathfrak{L}$ as t varies and the closeness* of ×'s and +'s to the smooth curve for $t = \infty$. The approach of $\mathfrak{L}$ to its limiting value for $\alpha = \frac{1}{2}$, as displayed in Figure 7, is monotonic increasing; this is because gamma components have decreasing hazard rates when $\alpha < 1$. Although we do not present the curve here, the same phenomenon has been seen for Weibull components with $p < 1$. As α (or p) gets larger there is a range of shape parameter for which the approach is monotonic decreasing, as shown in Figures 6, 7, 9. For still larger α or p the curve oscillates before damping in its approach to the equilibrium value; the larger α, the more oscillations are visible.

These oscillations were not expected, but they are genuine. Since hindsight is often 20/20, we now give an intuitive justification for the phenomenon. If the mean of the failure distribution of a component is large relative to its standard deviation (if the component has a small coefficient of variation) failures concentrated near the component mean life μ reduce the reliability, causing a relative minimum. After replacing the failed components, the reliability is increased, causing a maximum. But after an additional time μ the second generation of components will fail, causing a second maximum, etc. Thus we *expect* peaks to occur at values of T that are multiples of μ. The peaks get wider and shallower as T increases, until failures are essentially „random" and the exponential limit takes effect. This situation is illustrated in Figure 10. The upper set of curves represents $f(t)$ and its convolutions (time to second failure, time to third failure, etc.). The distribution of kth failures peaks at $t = k\mu$; its standard deviation is of the order of $\mu\sqrt{k}$ times the coefficient of variation of f. Thus the peaks do get wider and shallower as T increases. Another heuristic argument is illustrated by the lower curve in Figure 10, representing $h(t)$, the sum of the curves in the

*A comparison of the two curves for $\alpha = 2$, Figures 2 and 3, indicates that the approach for $n \to \infty$ is faster in Figure 3 than in Figure 2. Both curves represent computer plots. We had intended to include only Figure 3, but, having discovered the discrepancy, found it advisable to include both. Clearly one of the computer programs used was in error. The program is being rewritten; a correct tabulation and plot will be furnished on request.

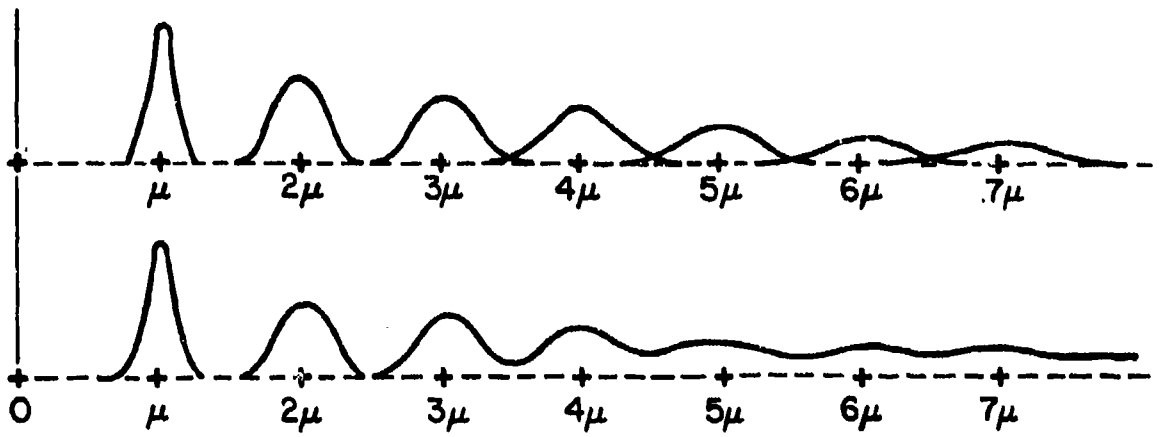


FIGURE 10. Schematic representation of $f^{*n}(t)$ (above) and $h(t)$ (below)

upper figure: it oscillates and then stabilizes to a constant value. But one observes from (1) that $\mathfrak{L}$ is essentially a monotonic function of $h(t)$ (L_1 and L_2 affect the size of the oscillation, but have little effect on its location) and one observes from (2) that the asymptotic $\mathfrak{L}$ for $n \to \infty$ *is* a monotonic function of $h(t)$, with sense reversed: the peaks of $h(t)$ are mirrored into the troughs of $\mathfrak{L}(t)$. It is well known that the coefficient of variation of the gamma and Weibull distributions decreases as α and p, respectively, increase.

The oscillations increase the value of T/μ needed before one can be sure that the deviation of $\mathfrak{L}$ from its limit is less than some specified value. For example, consider the curve of

$$e^{-\mu s h(t)} \quad \text{for } e^{-s} = 0.35$$

when $f(t)$ is a gamma density. Table 1 is obtained by finding on these curves the time beyond which the value of $\mathfrak{L}$ never deviates from 0.35 by more than 1% (i.e. 0.0035). Note that such a time as $T = 3.1\mu$ can be very large for highly reliable components. For example, if $\alpha = 12$, and $\theta = 1$ month, and $n = 256$, then on the average the system has 256 failures per year or one failure every 1.4 days. Yet the steady-state exponential limit is reached after 3.1 years! If $\alpha = 12$, and $\theta = 1$ year, and $n = 256$, then the system fails every 17 days; and the steady state is reached after 37 years! Do many systems last this long? if not, one should not be analyzing their reliability by means of the exponential assumption.

Table 2 illustrates how the mean life $\mu = \alpha\theta$ (for gamma components) enters the calculations. The first two lines were read from Figure 2. If $\theta = 15$ hours and $n = 256$, the MTBF of a component is 180 hours and there is a system failure every 42 minutes. If $\theta = 15$ years and $n = 256$, the MTBF of the system is 257 days; the last line of Table 2 indicates that steady state has not arrived after 165 years.

TABLE 1. Time for oscillations to die down as function of scale parameter

scale parameter	normalized time		coded time
α	t/α	$e^{-t/\alpha}$	$t = T/\theta$
$\frac{1}{2}$	3.0	.050	1.5
$\frac{3}{2}$	1.2	.301	1.8
2	1.2	.301	2.3
6	1.7	.183	10.3
12	3.1	.045	37.6

TABLE 2. Effect of scale parameter θ on reliability calculations: Poisson components, $\alpha = 12$

	$e^{-t/\alpha}$	0	.23	.40	.58
	$\mathfrak{L}$	.75	.795	.663	.90
	t/α		1.47	.92	.54
θ = 1 mo.	T =		17.6 mos.	11.0 mos.	6.5 mos.
15 hrs.			265 hrs.	165 hrs.	98 hrs.
15 yrs.			265 yrs.	165 yrs.	98 yrs.

REFERENCES

[1] Blumenthal, S.; Greenwood, J. A.; Herbach, L. 1969 Effect of system age on the distribution of waiting time between failures for serial systems. *Trans. 1969 Product Assurance Conference,* pp. 28-51. (IEEE CAT 69 C 54 PROD)

[2] ———; ———; ——— 1971 Superimposed non-stationary renewal processes. *J. Appl. Prob. 8:* 184-192.

[3] ———; ———; ——— 1973 The transient reliability behavior of series systems or superimposed renewal processes. *Technometrics 15:* 255-269.

[4] Drenick, R. F. 1960 The failure law of complex equipment. *J. SIAM 8:* 680-689.

[5] Herbach, L.; Greenwood, J. A.; Blumenthal, S. 1973 Transient minimum reliability for complex equipment whose components obey a Weibull failure law. *Bull. Internat. Statist. Inst. 45,* #1, 515-522.

APPLICATION OF TIME SERIES MODELS

George E. P. Box
University of Wisconsin
Madison, Wisconsin

1. The need for time series models

Statistical models with which the user is perhaps most familiar are of a form such that for the t'th of n observations

$$y_t = f(\underset{\sim}{x}_t, \underset{\sim}{\beta}) + u_t \qquad (1)$$

where y_t is the t'th observed value, $\underset{\sim}{x}_t$ is a vector of k independent (input) variables, & $\underset{\sim}{\beta}$ is a vector of p parameters. The error term u_t has zero mean & is often assumed to be distributed

i) normally,

ii) with constant variance σ^2 independent of t,

iii) independently of any other error u_s $(s \neq t)$.

Such models include those customarily associated with analysis of variance as well as with regression analysis. The practitioner is however frequently involved with data which occurs serially in time or space. Thus $y_1, y_2, \ldots, y_n$ might be successive observations of the positions of a missile observed every second, or of recruitment to the Army observed every month. For such data the errors are unlikely to be independent. A disturbance occurring at time t is likely to influence not only an observation made at time t but also subsequent observations at times t+1, t+2, etc. In such a case the errors u_t may be <u>serially correlated</u>.

Now statisticians have a great deal of experience with building models of the form of (1) and have available a battery of techniques which are appropriate when the assumptions mentioned above, (in particular the independence assumption) are true. Most notably maximum likelihood estimates of the parameters $\underset{\sim}{\beta}$ may then be obtained by use of the method of least squares (i.e. standard regression analysis).

It might therefore be asked whether serial correlation of errors will seriously invalidate these standard methods. Statisticians have traditionally seemed to worry most about the effects on non-normality rather than the effects of stochastic dependence of errors. It is therefore relevant to consider how badly the effects from violation of serial independence assumptions compare with those from non-normality.

Table 1 shows the result of sampling experiment (Box 1976, [1]) in which two samples of 10 observations from identical populations of the forms indicated were taken and subjected to a t-test (t) and a Mann-Whitney test (MW). The sampling was repeated 1,000 times and the number of results significant at the 5 percent point was recorded. Ideally, this number should be 50(that is, 5 percent of the total) but it has a standard deviation of about 7 because of sampling errors. More accurate results may be obtained by taking larger samples or by analytical procedures, however, since there is no practical difference between a significance level of say 4 percent and 7 percent, the present investigation suffices for illustration. Autocorrelation between adjacent values was introduced by generating observations so that ρ_1, the first serial correlation, had values of -0.4 and +0.4.

The frequences on the left are those obtained for a nonrandomized test. The frequencies on the right are obtained when the observations were randomly allocated to the two groups.

As is to be expected the significance level of the t-test is affected remarkably little by the drastic changes made in the marginal parent distribution--changes for which the "distribution-free"

Table 1.

Tests on Two Samples of Ten Observations having the same mean.

Frequency in 1,000 Trials of Significance at the 5 Percent Level Using the t-Test (t) and the Mann-Whitney Test

			With No Randomization			After Randomization		
	ρ_1	Test	Rectangular	Normal	Chi-square*	Rectangular	Normal	Chi-square*
Independent Observations	0.0	t	56	54	47	60	43	59
		MW	43	45	43	58	41	44
Autocorrelated Observations	-0.4	t	5	3	1	48	55	63
		MW	5	1	2	43	49	56
	0.4	t	125	105	114	59	58	54
		MW	110	96	101	46	53	43

* This parent chi-square distribution has four degrees of freedom and is thus highly skewed.

test provides insurance. Unfortunately, of course, both tests are equally impaired by error dependence unless randomization is introduced when they do about equally well. The point is, of course, that it is the act of randomization that is of major importance here not the introduction of the non-parametric test function.

In the situations we discuss in this paper, randomization is not possible and it is evident that in this case we face a serious problem if errors are serially correlated.

As a further illustration consider the following regression model used by Coen, Gomme & Kendall (1969), [16] to model quarterly data in which y_t is a stock market index, $x_{1,t-6}$ is a measure of U. K. car production lagged 6 periods & $x_{2,t-7}$ is a commodity index lagged 7 periods

$$y_t = \alpha + \beta_t + \beta_1 x_{1,t-6} + \beta_2 x_{2,t-7} + u_t$$

On the assumption of error independence, for which ordinary least squares is appropriate, estimates of β_1 and β_2 were calculated. These were 14.1 and -9.9 times their standard errors, indicating overwhelming significance. On this basis the authors of the paper believed that they could forecast future stock market prices. It was subsequently pointed out however (Box & Newbold 1971, [10]), that, as soon as proper provision was made for serial dependence in the errors, the apparent relationships disappeared.

2. ARIMA time series models

Models having their origins in the work of Yule [22] Slutsky [19] & Yaglom [21], which have been found to be of great practical value in representing serial dependence, employ stochastic difference equations of the form

$$\phi(B)u_t = \theta(B)a_t \qquad (2)$$

where u_t is the sequence to be modelled, B is the backshift operator such that $Bu_t = u_{t-1}$,

$\phi(B)$, called the autoregressive operator is such that

$$\phi(B) = 1-\phi_1 B-\phi_2 B^2,\ldots-\phi_p B^p$$

$\theta(B)$, called the moving average operator, is such that

$$\theta(B) = 1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q$$

and $\{a_t\}$is "white noise", that is, a source of independent random "shocks" roughly normally distributed about zero with constant variance σ_a^2.

The form of equation needed is often rather simple. Thus Fig. 1 shows a number of real time series together with the fitted stochastic models which have been found to represent them.

As illustrated by these examples, models for nonstationary time series may often be built by fitting a stationary model to a differenced series. Thus (see Figure 1(b)) the first difference $u_t-u_{t-1} = (1-B)u_t = \nabla u_t$ of the stock price series is represented by a stationary first order moving average model yielding the overall model

$$(1-B)u_t = (1-\theta B)a_t \text{ with } \theta = -.1$$

i.e.
$$u_t = u_{t-1} + a_t + .1a_{t-1}$$

Models of this kind have been used successfully to solve a wide variety of problems including

Forecasting future values of a series.

Smoothing Series (including seasonal adjustment of series).

Intervention Analysis (detecting & estimating effects of system changes buried in dependent noise).

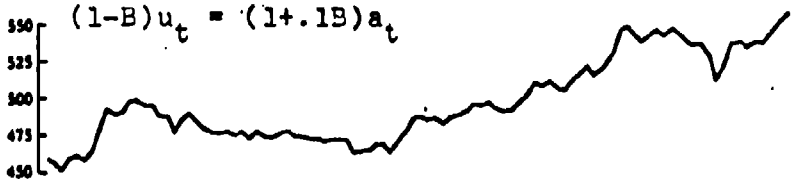
Control of Systems.

We now illustrate some of these applications with examples.

Fig. 1 $(1-.9B)u_t = 1.45+(1-.6B)a_t$

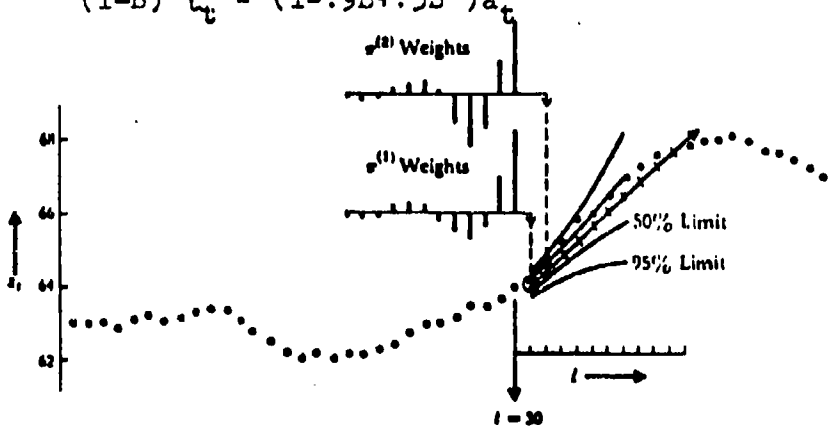
(a) Two-Hourly Concentration Readings: Chemical Process

$(1-B)u_t = (1+.1B)a_t$



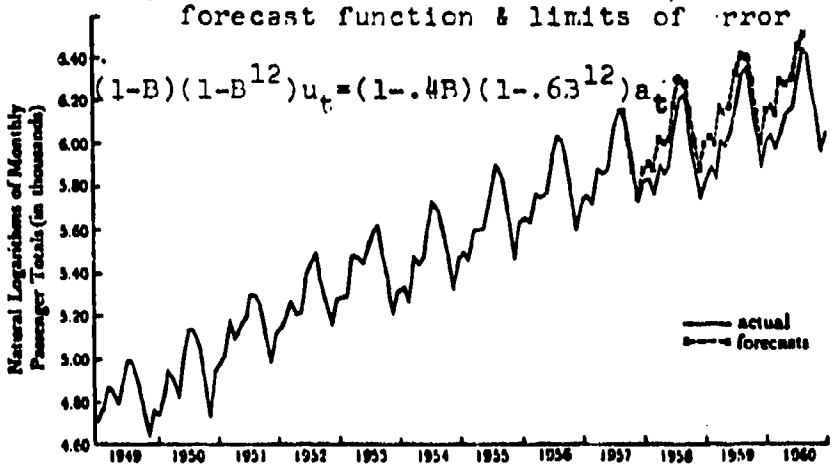
(b) Daily IBM Stock Prices

$(1-B)^2 u_t = (1-.9B+.5B^2)a_t$



(c) Series arising in a control problem with forecast function & limits of error

$(1-B)(1-B^{12})u_t=(1-.4B)(1-.63^{12})a_t$



(d) Seasonal series: logs of monthly passengers totals in international air travel. Forecasts for up to 36 months ahead all made from arbitrary origin July 1957.

3. Estimating future location of a missile

The following is taken from an MRC technical report (Box & Pallesen 1978, [11]) which describes the modelling of some missile data (made available by Mr. Paul Thrasher of the Quality Evaluation Division of White Sands, Missile Range). It shows how a stochastic difference equation model may be built & used to predict the future location of the missile. Details of the calculations will be found in the book by Box & Jenkins indicated here by B&J[8].

A model
$$\phi_p(B)\nabla^d z_t = \theta_q(B)a_t \qquad (3)$$
is said to be an ARIMA (Autoregressive Integrated Moving Average) model of order (p,d,q) if $\phi_p(B)$ & $\theta_q(B)$ are polynomials in B of degrees p & q respectively having zeros outside the unit circle.

3.1 Identification, Fitting and Checking of Model

The data series we are considering consists of 246 consecutive observations of the x-coordinate of a missile trajectory. The observations, z_t; $t = 1,2,\ldots,246$, were made with constant sampling interval and there are no missing or obviously aberrant values.

Modeling such a time series is conceived of as an iterative process involving three stages: identification, fitting and diagnostic checking. Identification is first performed along the lines of Chapter 6 in B&J [8]. Plotting the data z_t (Figure 2a) shows a smooth nonstationary series, whose autocorrelation function (Figure 2(b)) dies out extremely slowly. After differencing three times the series $\nabla^3 z_t$ appears stationary and its sample autocorrelation and partial autocorrelation function (Figures 2c and 2d) suggest that a reasonable model for $\nabla^3 z_t$ should include a few moving average parameters of low order. A clear identification is not possible

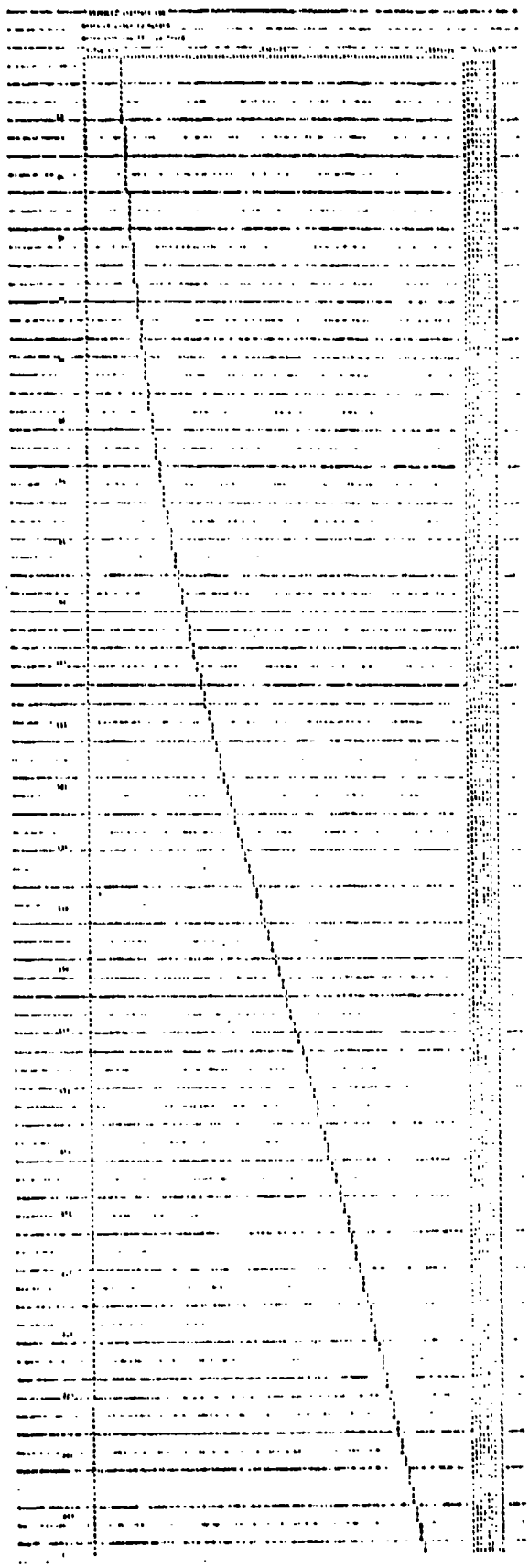


Figure 2a. Plot of data

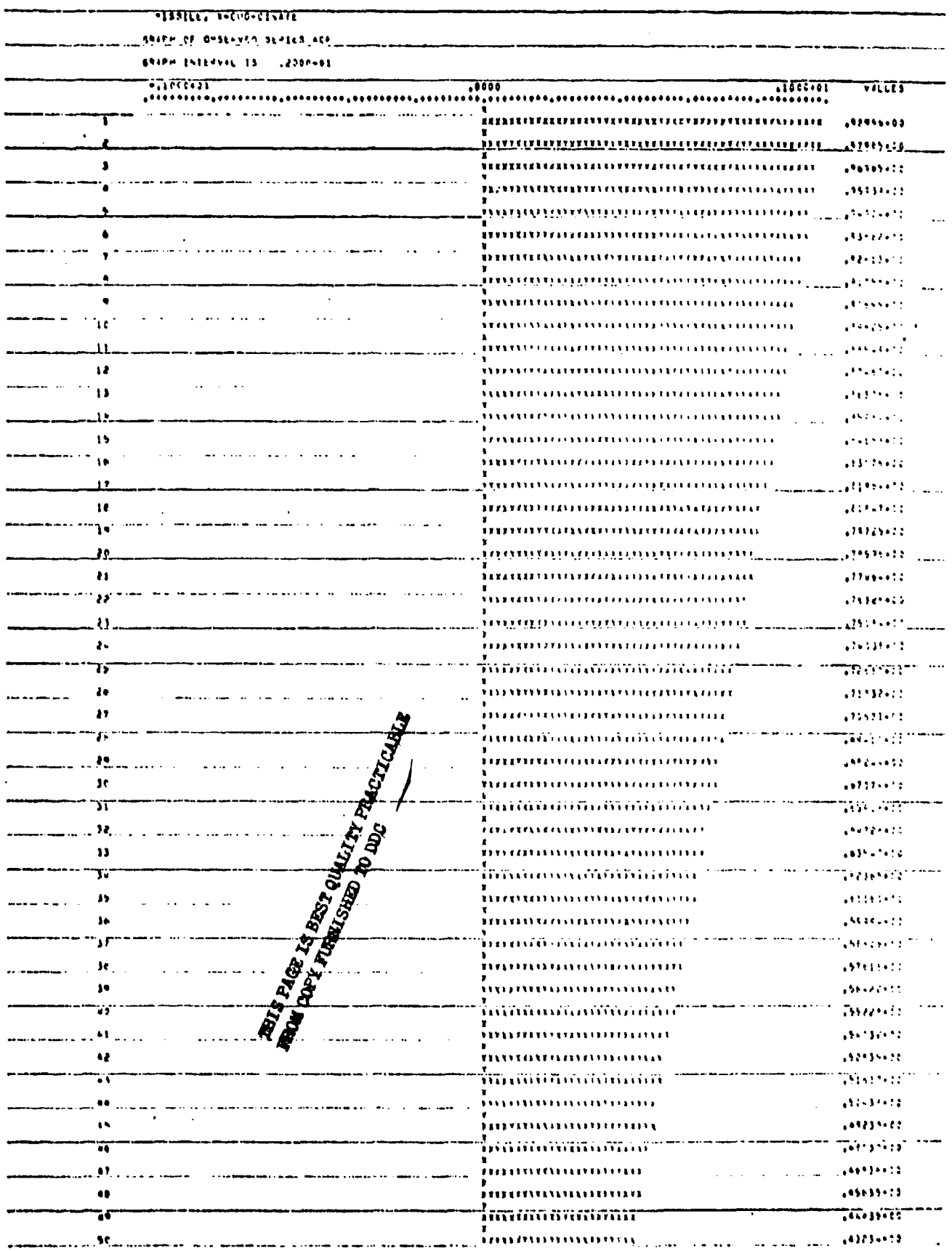


Figure 2b. Sample autocorrelation function of original series

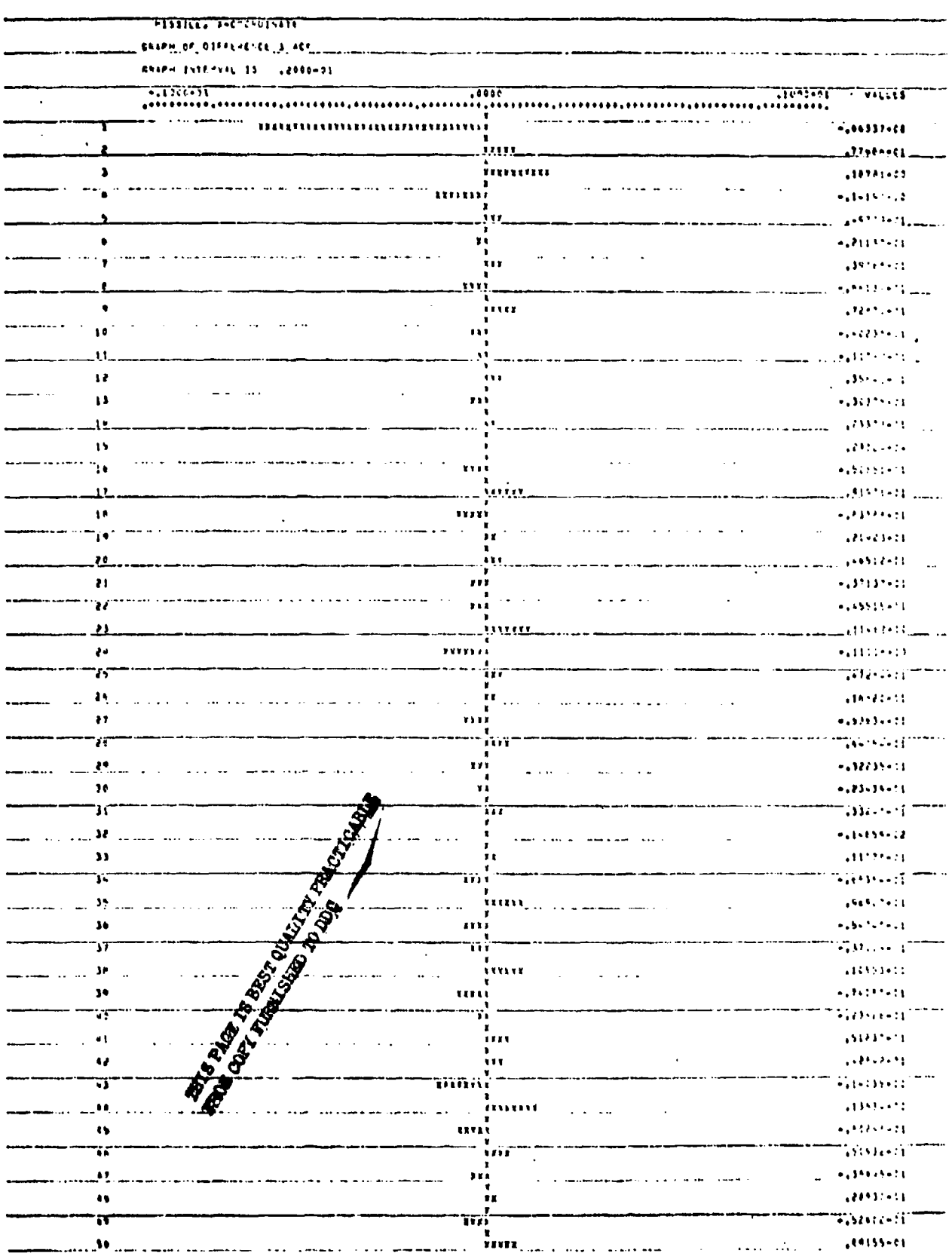


Figure 2c. Sample autocorrelation function for $\nabla^3 z_t$

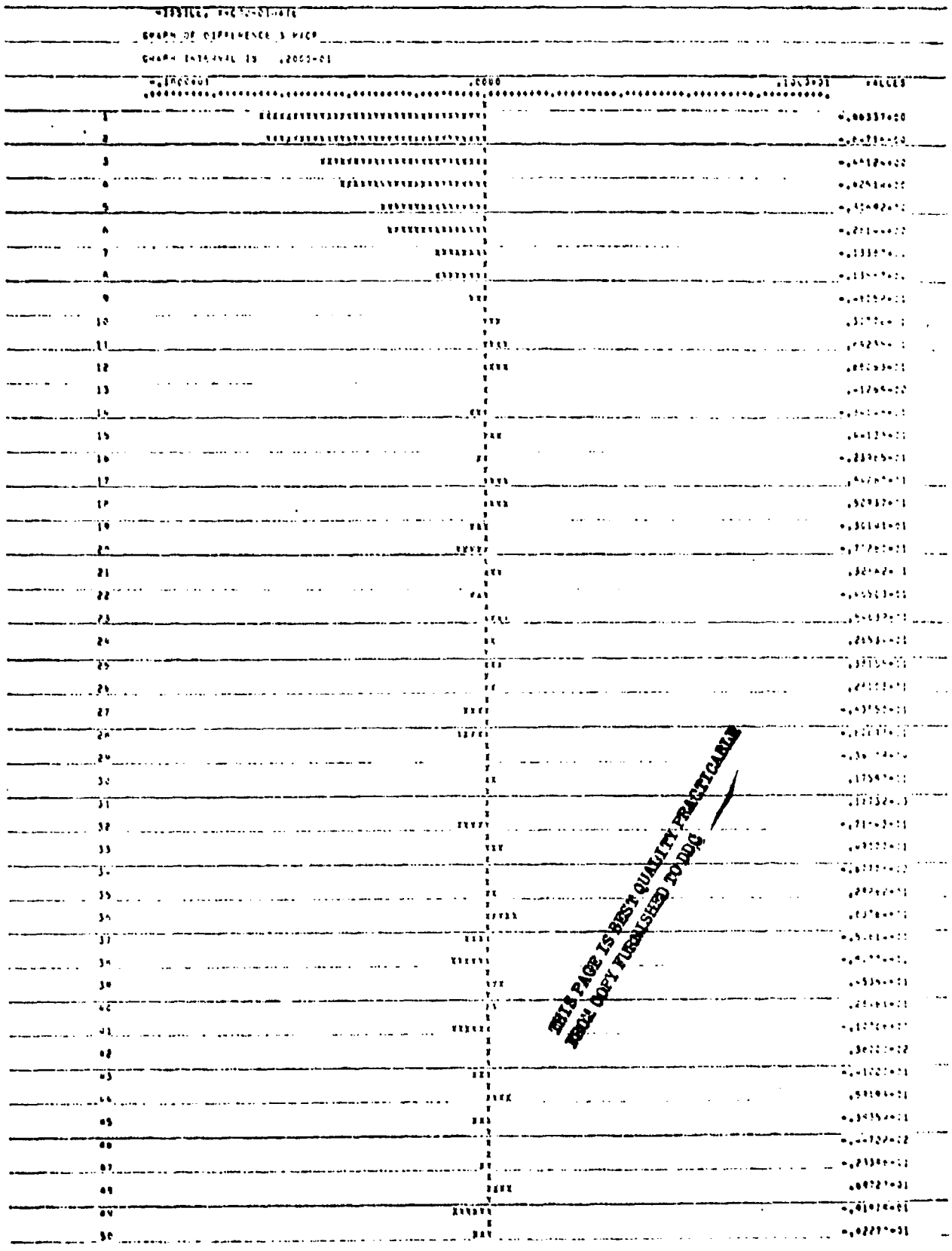


Figure 2d. Partial autocorrelation function for $\nabla^3 z_t$

at this point but a stochastic difference equation model of the form

$$\nabla^3 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3) a_t \qquad (4)$$

is considered worthy of being tentatively entertained.

Fitting this model by the method of Chapter 7 in B&J [8] gives the parameter estimates, residual sum of squares (RSS) and the residual mean square (RMS) listed in Table 2. If this model is adequate the RMS value provides an estimate of the variance, $\sigma_a^2 = E(a_t^2)$, which is the one step ahead forecast error variance.

Diagnostic checking (Chapter 8 in B&J [8]) involves examination of the residuals (the estimated a_t's) left after fitting this model to seek for departures from the "white noise" form. One way of doing this is to submit the residual $\{\hat{a}_t\}$ sequence to the identification procedure previously applied to $\nabla^3 z_t$. In fact the autocorrelation function of the residuals $\hat{a}_t$'s, Figure 3(a), suggests that while most of the dependence is being accounted for by the model, some significant low order autocorrelations remain, indicating some additional θ parameters are needed. Notice, that the diagnostic checking of the model (4) reveals model inadequacy and also identifies in which way the model should be modified.

After another cycle the model

$$\nabla^3 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4 - \theta_5 B^5) a_t \qquad (5)$$

is identified, and it fits the data very well, leaving residuals, Figure 3(b), which look like white noise. Figure 3(c) shows the sample autocorrelations of the residuals. This fitted model along with some other contenders are listed in Table 2. Additional models are fitted as a check that additional parameters would not

Table 2

Models fitted to Missile data

(x-coordinate)

(p,d,q)	Model	RSS	RMS (DF)
(0, 2, 2)	$\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2)a_t$ $\hat{\theta}_1 = .716 \begin{cases} .83 \\ .60 \end{cases}$ $\hat{\theta}_2 = -.517 \begin{cases} -.40 \\ -.63 \end{cases}$ (Moduli of roots: 1.39; 1.39 i.e. stable)	410.	1.69 (242)
(0, 2, 3)	$\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$ $\hat{\theta}_1 = .662 \begin{cases} .78 \\ .54 \end{cases}$ $\hat{\theta}_2 = -.114 \begin{cases} .03 \\ -.26 \end{cases}$ $\hat{\theta}_3 = -.425 \begin{cases} -.31 \\ -.54 \end{cases}$ (Moduli of roots: 1.14; 1.14; 1.83 i.e. stable)	348.	1.44 (241)
(0, 3, 3)	$\nabla^3 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$ $\hat{\theta}_1 = 1.731 \begin{cases} 1.75 \\ 1.72 \end{cases}$ $\hat{\theta}_2 = -.776 \begin{cases} -.76 \\ -.79 \end{cases}$ $\hat{\theta}_3 = -.104 \begin{cases} -.08 \\ -.13 \end{cases}$ (Moduli of roots: 1.014; 1.014; 9.39 i.e. stable)	247.	1.03 (240)

Table 2 Continued

(p,d,q)	Model	RSS	RMS (DF)
(0, 3, 4)	$\nabla^3 z_t = (1-\theta_1 B-\theta_2 B^2-\theta_3 B^3-\theta_4 B^4)a_t$ $\hat{\theta}_1 = 1.938 \begin{cases} 1.99 \\ 1.89 \end{cases}$ $\hat{\theta}_2 = -1.030 \begin{cases} -.95 \\ -1.11 \end{cases}$ $\hat{\theta}_3 = -.146 \begin{cases} -.02 \\ -.27 \end{cases}$ $\hat{\theta}_4 = .173 \begin{cases} .25 \\ .09 \end{cases}$ (Moduli of roots: 1.13; 1.13; 1.59; 2.86)	203.	.85 (239)
(0, 3, 5)	$\nabla^3 z_t = (1-\theta_1 B-\theta_2 B^2-\theta_3 B^3-\theta_4 B^4-\theta_5 B^5)$ $\hat{\theta}_1 = 2.078 \begin{cases} 2.11 \\ 2.04 \end{cases}$ $\hat{\theta}_2 = -1.291 \begin{cases} -1.21 \\ -1.37 \end{cases}$ $\hat{\theta}_3 = -.115 \begin{cases} .00 \\ -.23 \end{cases}$ $\hat{\theta}_4 = .395 \begin{cases} .51 \\ .28 \end{cases}$ $\hat{\theta}_5 = -.131 \begin{cases} -.04 \\ -.22 \end{cases}$ (Moduli of roots: 1.11; 1.11; 1.79; 1.79 1.93)	192.	.81 (238)
(0, 3, 6)	$\nabla^3 z_t = (1-\theta_1 B-\theta^2 B_2-\theta_3 B^3-\theta_4 B^4-\theta_5 B^5$ $-\theta_6 B^6)a_t$ (roots o. k.) $\hat{\theta}_6 \approx 0$	192.	.81 (237)

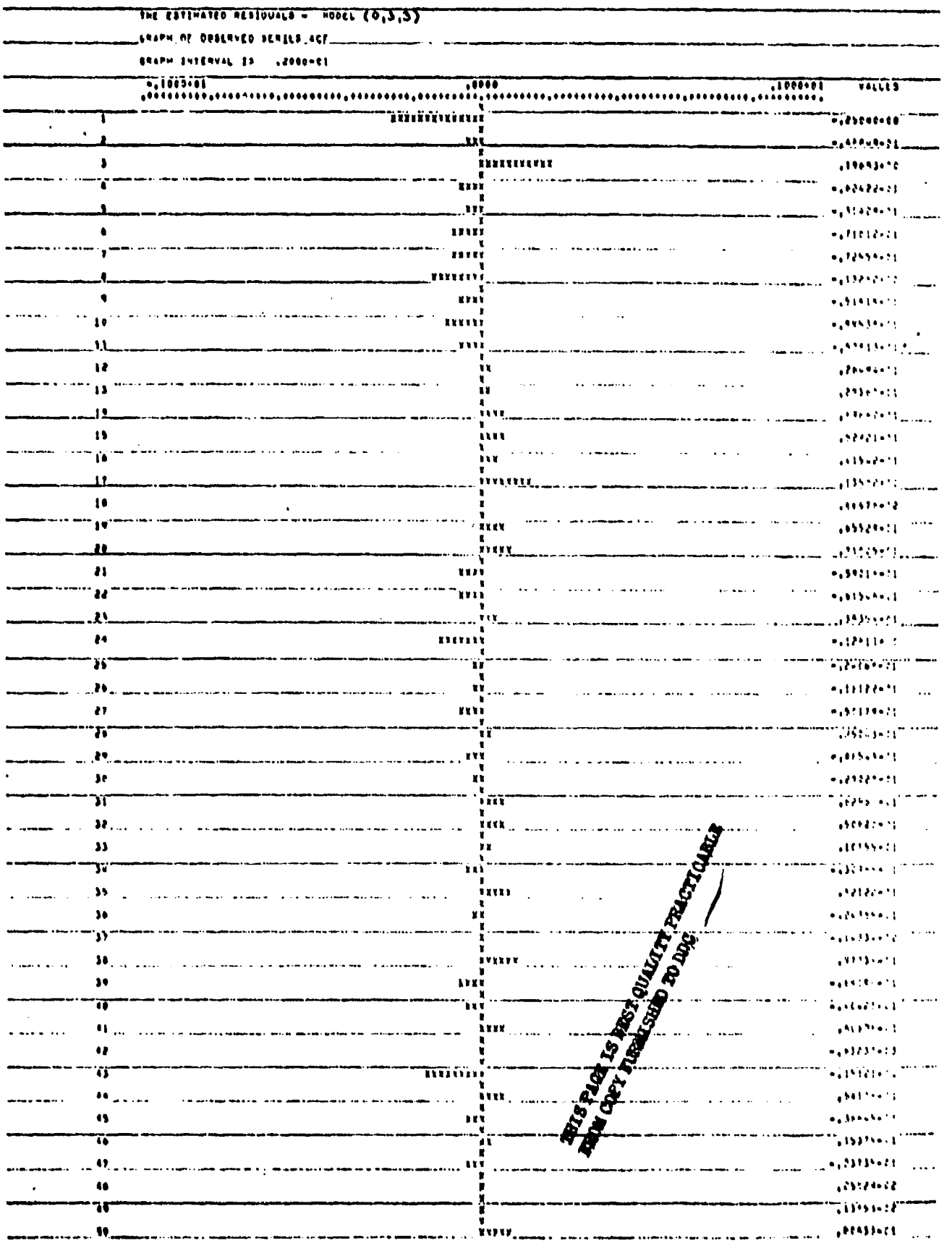


Figure 3a. Autocorrelation function of residuals from the (0,3,3) model

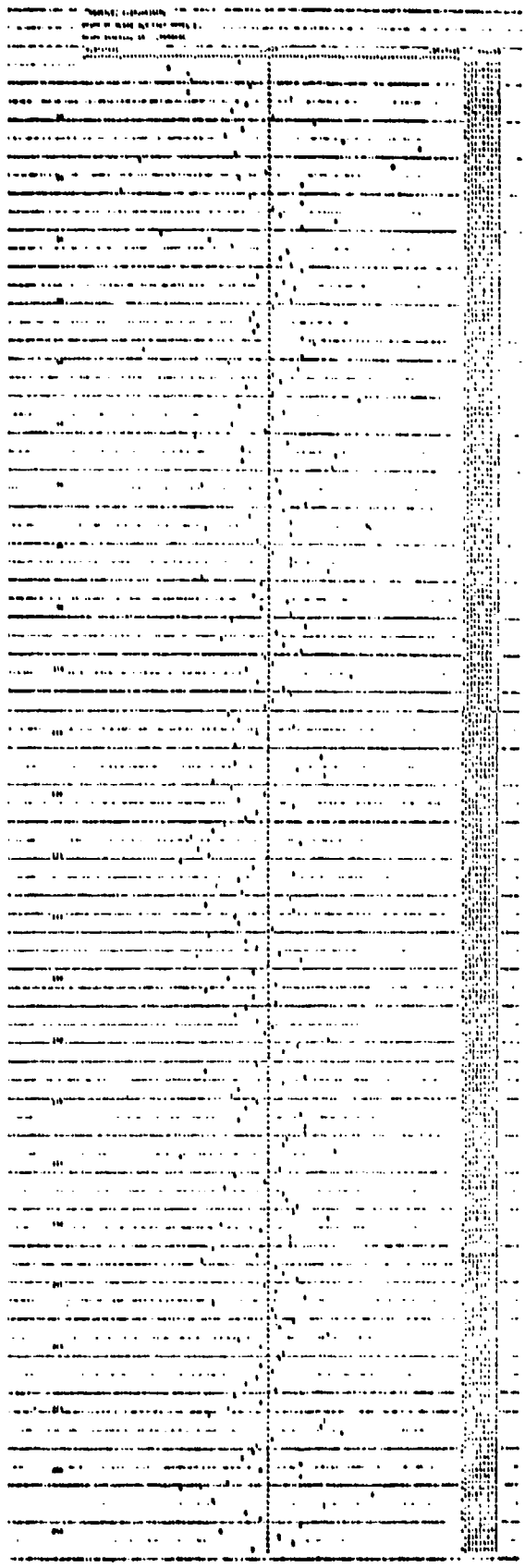


Figure 3b. Residuals from the (0,3,5) model

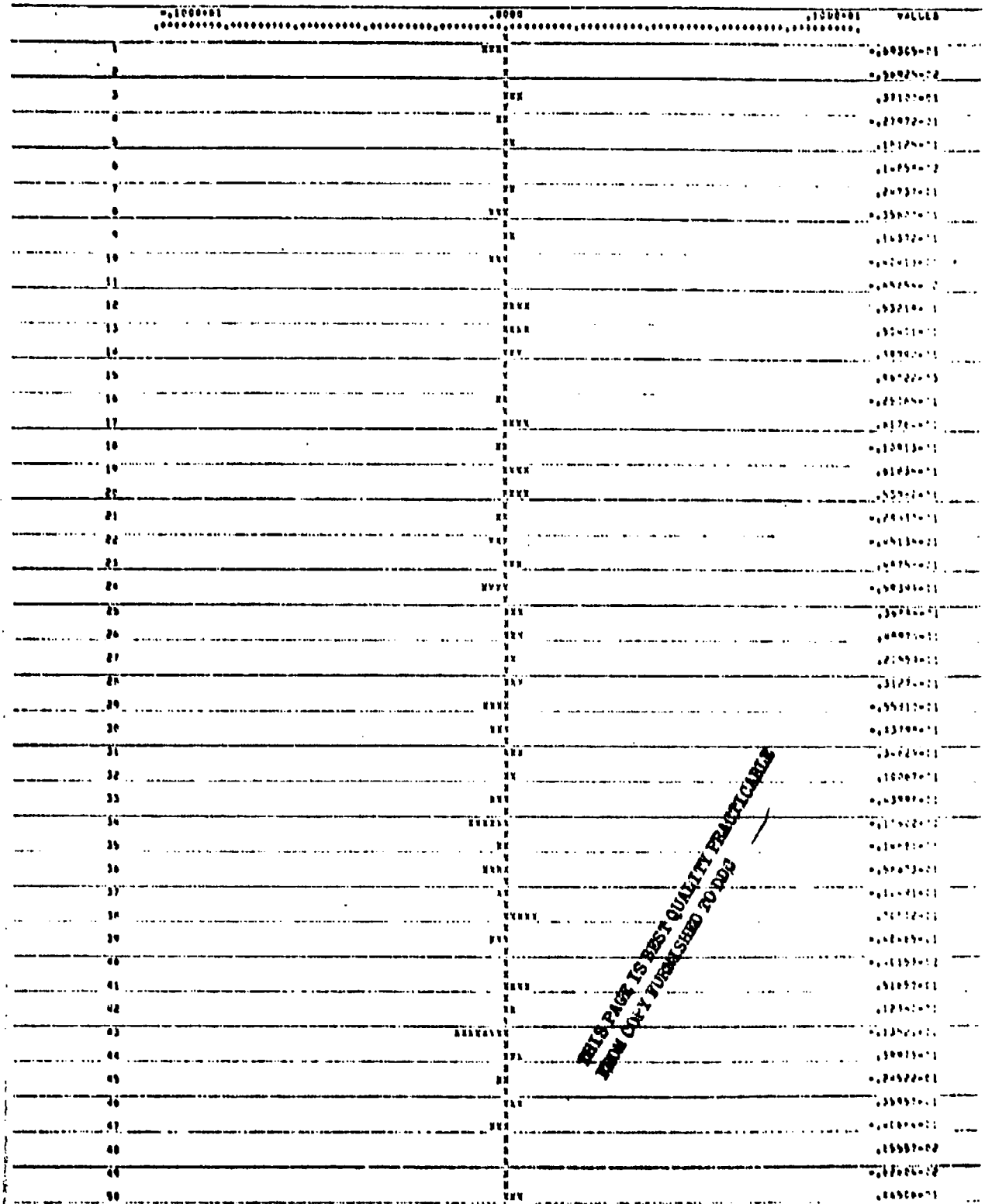


Figure 3c. Autocorrelation function of residuals from the (0,3,5) model

substantially improve matters (overfitting), and also to demonstrate that the chosen number of differencings is appropriate.

3.2 Checking zeros of $\theta(B)$

Regarding the operator

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4 - \theta_5 B^5 \qquad (6)$$

as a polynomial in B, it is shown in B&J [8] that a necessary requirement for a sensible model is that the zeroes of this polynomial be outside the unit circle (invertibility property).

It is important to check this and the moduli of the roots given in Table 2 indicate that the model is indeed invertible.

3.3. Forecasts

Accepting that the (0, 3, 5) model provides an adequate representation of the system (with the (0, 3, 4) model as a close runner-up) the forecasts produced are most easily calculated from the difference equation itself (see Chapter 5 of B&J [8]). From Equation (5) we find

$$z_t = 3z_{t-1} - 3z_{t-2} + z_{t-3} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_3 a_{t-3} - \theta_4 a_{t-4} - \theta_5 a_{t-5}. \qquad (7)$$

Then by taking conditional expectations of $z_{t+1}, z_{t+2}, \ldots, z_{t+\ell}$ at origin t (as described in B&J p. 130 [8]) the $1,2,3,\ldots,\ell,\ldots$ step ahead forecasts are:

$$\begin{aligned}
\hat{z}_t(1) &= 3z_t - 3z_{t-1} + z_{t-2} - \theta_1 a_t - \theta_2 a_{t-1} - \theta_3 a_{t-2} - \theta_4 a_{t-3} - \theta_5 a_{t-4} \\
\hat{z}_t(2) &= 3\hat{z}_t(1) - 3z_t + z_{t-1} - \theta_2 a_t - \theta_3 a_{t-1} - \theta_4 a_{t-2} - \theta_5 a_{t-3} \\
\hat{z}_t(3) &= 3\hat{z}_t(2) - 3\hat{z}_t(1) + z_t - \theta_3 a_t - \theta_4 a_{t-1} - \theta_5 a_{t-2} \\
\hat{z}_t(4) &= 3\hat{z}_t(3) - 3\hat{z}_t(2) + \hat{z}(1) - \theta_4 a_t - \theta_5 a_{t-1} \\
\hat{z}_t(5) &= 3\hat{z}_t(4) - 3\hat{z}_t(3) - \hat{z}(2) - \theta_5 a_t \\
\hat{z}_t(\ell) &= 3\hat{z}_t(\ell-1) - 3\hat{z}(\ell-2) - \hat{z}(\ell-3) \qquad \ell \geq 6
\end{aligned} \qquad (8)$$

Table 3

		FORECASTS		
Obs #	Actual value	Model (0, 3, 5)	Model (0, 3, 4)	Model (0, 3, 3)
201	13225.08	13224.78	13224.80	13224.99
202	13306.74	13305.80	13305.94	13306.46
203	13387.51	13386.70	13386.77	13387.51
204	13468.42	13467.20	13467.23	13468.14
205	13549.74	13547.33	13547.34	13548.34
206	13628.61	13627.10	13627.08	13628.12
207	13708.78	13706.49	13706.46	13707.48
208	13788.67	13785.52	13785.48	13786.40
209	13868.21	13864.18	13864.14	13864.91
210	13947.30	13942.47	13942.44	13942.99

In practice of course this is done automatically by the computer. For illustration, the forecasts produced by this model with an origin (for all forecasts) at t = 200 is shown in Figure 4. It will be noticed that the forecasts are in very close agreement with the actual values. Even the 10-step ahead forecast is hardly distinguishable from the actually observed value.

Table 3 lists the actual values and the forecasts numerically. The forecasts produced by the models (0, 3, 4) and (0, 3, 3) are also very good and they are included for comparison.

3.4. Error of Forecasts

In order to determine the error of the forecasts, it is helpful to write the model (3) in random shock form. Thus formally

$$z_t = \frac{\theta_q(B)}{\nabla^d \phi_p(B)} a_t = \psi(B) a_t \tag{9}$$

where

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots \tag{10}$$

And it is shown in B&J [8] p. 126-128 that the lead ℓ forecast error is

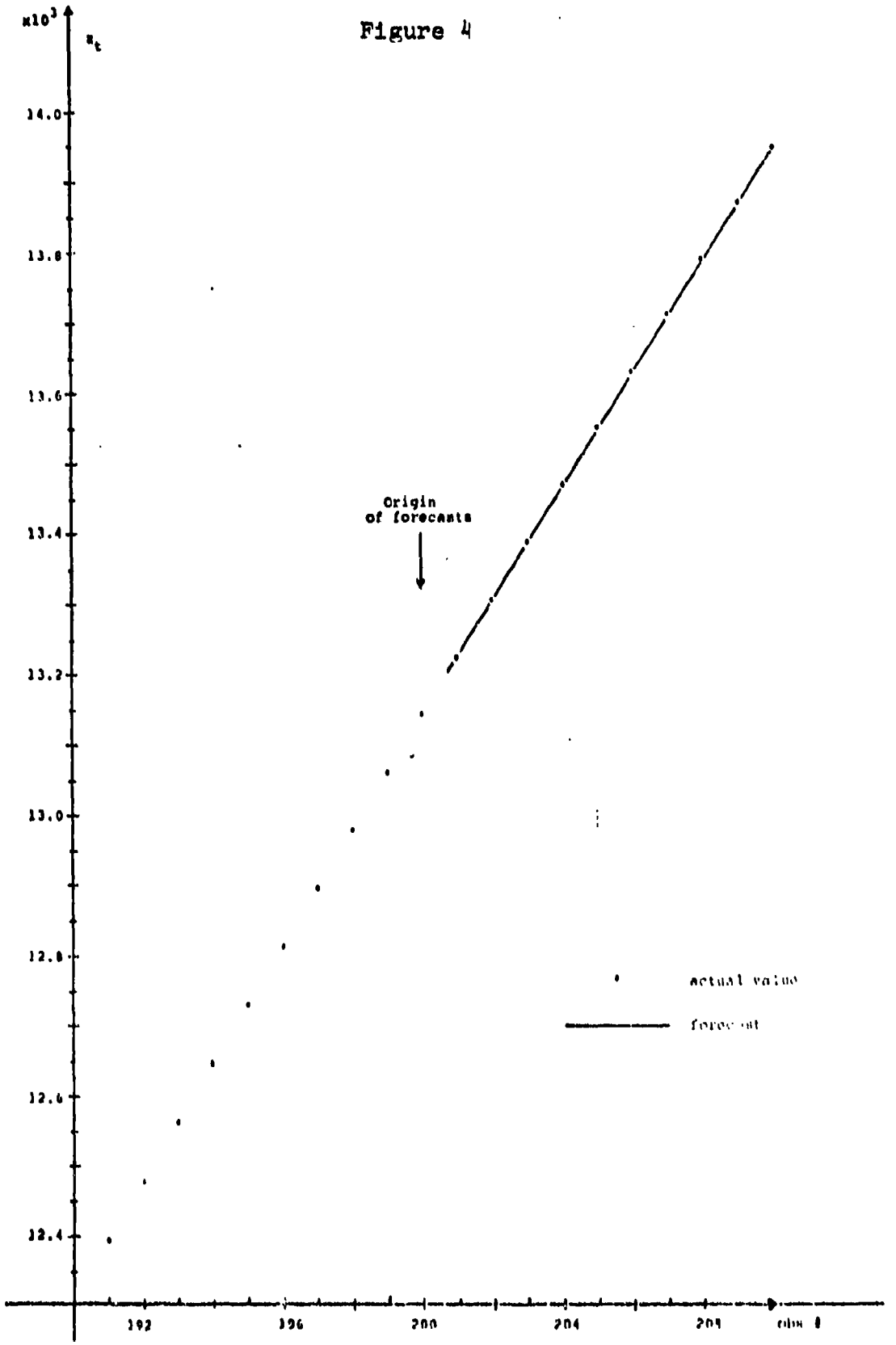
$$e_t(\ell) = z_{t+\ell} - \hat{z}_t(\ell) = a_{t+\ell} + \psi_1 a_{t+\ell-1} + \ldots + \psi_{\ell-1} a_{t+1} . \tag{11}$$

Whence the variance of the forecast error is

$$\begin{aligned} \text{var}[e_t(\ell)] &= E(z_{t+\ell} - \hat{z}_t(\ell))^2 \\ &= (1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{\ell-1}^2)\sigma_a^2 . \end{aligned} \tag{12}$$

For the fitted model (5) the ψ-weights are calculated by equating

Figure 4


x10^3
x_t
14.0
13.8
13.6
13.4
13.2
13.0
12.8
12.6
12.4
Origin
of forecasts
actual value
forecast
192
196
200
204
208
obs #

coefficients in (13), B&J [8] pp. 132-134.

$$(1 - 3B + 3B^2 - B^3)(1 + \psi_1 B + \psi_2 B^2 + \ldots) = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4 - \theta_5 B^5) \quad (13)$$

Specifically we find the ψ_j values given in Table 4. Using the estimated $\hat{\sigma}_a^2 = .81$ from Table 2, the variance of the forecast error is given for $\ell = 1,2,\ldots,10$. The last column in Table 4 lists ±2 standard errors, corresponding to approximately 95% probability intervals for the forecasts. We note, that these probability intervals are so narrow, that they cannot be distinguished from the forecasts themselves in a plot like Figure 4.

The above is all that is needed to compute forecasts and the standard deviations of forecast errors. What appears in the following sections is not necessary for calculation, but does illuminate the nature of the projection process.

3.5 Integral forms

As discussed in Chapter 4 pp. 103-114 of B&J [8], the equivalent integrated form of the model of Equation (5), is of some interest also. In this form the observations appear as a linear aggregates of past random shocks, their difference, sum, sum of sums, etc., plus a new random shock. Specifically the integrated model form

$$z_t = \lambda_{-2}\nabla a_{t-1} + \lambda_{-1}a_{t-1} + \lambda_0 S a_{t-1} + \lambda_1 S^2 a_{t-1} + \lambda_2 S^3 a_{t-1} + a_t \quad (14)$$

degenerates to different models from Table 2 when certain of the λ-coefficients are taken to be zero. Table 5 links models from Table 2 to their equivalent integrated forms, and lists estimated λ coefficients which can be calculated from the estimated θ's. Conversion formulas for the models under consideration are given in

Table 4

ψ-weights and forecast errors

j	ψ_j	l	$\mathrm{Var}[e_t(l)]$	Approx. 95% Probability Intervals
		1	.81	± 1.8
1	.922	2	1.38	± 2.3
2	1.057	3	1.81	± 2.7
3	1.520	4	3.74	± 3.9
4	1.916	5	5.95	± 4.9
5	2.376	6	9.15	± 6.0
6	2.900	7	13.62	± 7.4
7	3.488	8	19.70	± 8.9
8	4.140	9	27.77	± 10.5
9	4.856	10	38.20	± 12.4

Table 5

Integrated model forms

$$z_t = \lambda_{-2} \nabla a_{t-1} + \lambda_{-1} a_{t-1} + \lambda_0 S a_{t-1} + \lambda_1 S^2 a_{t-1} + \lambda_2 S^3 a_{t-1} + a_t$$

Model	λ_{-2}	λ_{-1}	λ_0	λ_1	λ_2	RSS
(0, 2, 2)	-	-	.483	.800	-	410.
(0, 2, 3)	-	.425	.035	.878	-	348.
(0, 3, 3)	-	-	1.104	.016	.149	247.
(0, 3, 4)	-	.173	.627	.197	.065	203.
(0, 3, 5)	.131	-.129	.716	.140	.064	192.

Table 6, but can more generally be found from equating coefficients in Equation 4.3.21, p.112 in B&J [8].

3.6. The eventual forecast function

One question of interest is what function is being selected for projecting the forecasts, i.e. what is the forecast function. It is shown in B&J [8] p. 139 that depending on the nature of the left hand operator, the model (3) could call for forecasts lying on an updating function that could consist of any combination of polynomials, exponentials and sine and cosine waves. What forecast function does the model imply for the present fitted (0, 3, 5) model?

The eventual forecast function for the (0, 3, 5) model satisfies the difference equation

$$\nabla^3 z_t(\ell) = 0 \tag{15}$$

which has as its solutions a polynomial in ℓ of 2^{nd} degree

$$\hat{z}_t(\ell) = b_0^{(t)} + b_1^{(t)}\ell + b_2^{(t)}\ell^2 \tag{16}$$

and applies for $\ell > q - p - d$ (i.e. $\ell > 2$).

In other words the model (0, 3, 5) implies, that the forecasted future values from some time origin t, will, except for slight deviations in the first two lead-times, follow a quadratic curve. (The (0, 3, 4) model which fits slightly less well implies that only one initial deviation occurs, while the (0, 3, 3) model implies that all forecasts lie on a quadratic curve).

Although the forecasts are best calculated directly from the difference equation as above it is enlightening to further consider their nature.

As the origin of forecasts is advanced the calculating process requires that coefficients b_0, b_1 and b_2 are sequentially updated.

Table 6

Conversion formulae, θ to λ

Model	Formulae
(0, 2, 2)	$\lambda_0 = 1 + \theta_2$ $\lambda_1 = 1 - \theta_1 - \theta_2$
(0, 2, 3)	$\lambda_{-1} = -\theta_3$ $\lambda_0 = 1 + \theta_2 + 2\theta_3$ $\lambda_1 = 1 - \theta_1 - \theta_2 - \theta_3$
(0, 3, 3)	$\lambda_0 = 1 - \theta_3$ $\lambda_1 = 1 + \theta_2 + 2\theta_3$ $\lambda_2 = 1 - \theta_1 - \theta_2 - \theta_3$
(0, 3, 4)	$\lambda_{-1} = \theta_4$ $\lambda_0 = 1 - \theta_3 - 3\theta_4$ $\lambda_1 = 1 + \theta_2 + 2\theta_3 + 3\theta_4$ $\lambda_2 = 1 - \theta_1 - \theta_2 - \theta_3 - \theta_4$
(0, 3, 5)	$\lambda_{-2} = -\theta_5$ $\lambda_{-1} = \theta_4 + 4\theta_5$ $\lambda_0 = 1 - \theta_3 - 3\theta_4 - 6\theta_5$ $\lambda_1 = 1 + \theta_2 + 2\theta_3 + 3\theta_4 + 4\theta_5$ $\lambda_2 = 1 - \theta_1 - \theta_2 - \theta_3 - \theta_5 - \theta_5$

For example the updating formulae for the (0, 3, 5) model can be found directly by relating (16) to the forecasting formula from the integrated model.

We find that the updating formulae derived below are

$$\begin{cases} b_0^{(t)} = b_0^{(t-1)} + b_1^{(t-1)} + b_2^{(t-1)} + \lambda_0 a_t \\ b_1^{(t)} = b_1^{(t-1)} + 2b_2^{(t-1)} + (\lambda_1 + \tfrac{1}{2}\lambda_2)a_t \\ b_2^{(t)} = b_2^{(t-1)} + \tfrac{1}{2}\lambda_2 a_t \end{cases} \tag{17}$$

Note that the first terms in the right of (17) simply allow for movement of the origin without changing the polynomial. The term involving the last random shock a_t appropriately updates the coefficient.

The updating formulae (17) are derived as follows. We have from Equation (14) that

$$z_{t+\ell} = \lambda_{-2}\nabla a_{t+\ell-1} + \lambda_{-1}a_{t+\ell-1} + \lambda_0 S a_{t+\ell-1} + \lambda_1 S^2 a_{t+\ell-1} + \lambda_2 S^3 a_{t+\ell-1} + a_{t+\ell} \tag{18}$$

Assuming $\ell > 2$ and taking expectations at origin t we find

$$\begin{aligned} \hat{z}_t(\ell) &= E(\lambda_0 S a_{t+\ell-1}) + E(\lambda_1 S^2 a_{t+\ell-1}) + E(\lambda_2 S^3 a_{t+\ell-1}) \\ &= (\lambda_0 S a_t) + (\lambda_1 S^2 a_{t-1} + \ell\lambda_1 S a_t) \\ &\quad + (\lambda_2 S^3 a_{t-2} + (\ell+1)\lambda_2 S^2 a_{t-1} + \frac{(\ell+1)\ell}{2}\lambda_2 S a_t) \\ &= (\lambda_0 S a_t + \lambda_1 S^2 a_{t-1} + \lambda_2 S^2 a_{t-1} + \lambda_2 S^3 a_{t-2}) \\ &\quad + \ell(\lambda_1 S a_t + \lambda_2 S^2 a_{t-1} + \tfrac{1}{2}\lambda_2 S a_t) \\ &\quad + \ell^2(\tfrac{1}{2}\lambda_2 S a_t) \end{aligned} \tag{19}$$

The coefficients $b_0^{(t)}$, $b_1^{(t)}$, $b_2^{(t)}$ in Equation (16) are now identified as

$$\begin{cases} b_0^{(t)} = \lambda_0 S a_t + \lambda_1 S^2 a_{t-1} + \lambda_2 S^2 a_{t-1} + \lambda_2 S^3 a_{t-2} \\ b_1^{(t)} = \lambda_1 S a_t + \lambda_2 S^2 a_{t-1} + \tfrac{1}{2}\lambda_2 S a_t \\ b_2^{(t)} = \tfrac{1}{2}\lambda_2 S a_t \end{cases} \tag{20}$$

Now it is seen that (17) can be rewritten as (20).

3.7 How are the data used in the forecast?

Still another way to interpret the forecasts is as a weighted sum of previous observations: Writing (5) as

$$\frac{\nabla^3}{\theta(B)} z_t = \pi(B) z_t = a_t \tag{21}$$

where

$$\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots \tag{22}$$

we find that

$$\hat{z}_t(\ell) = \pi_1 \hat{z}_t(\ell-1) + \pi_2 \hat{z}_t(\ell-2) + \ldots \tag{23}$$

where $\hat{z}_t(-h)$ is taken to mean z_{t-h} for $h = 0, 1, 2, \ldots$. The π-weights can be found by equating coefficients in the following identity after the θ-estimates have been substituted;

$$\nabla^3 = \theta(B)\,\pi(B)$$

$$(1 - 3B + 3B^2 - B^3) = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_4 B^4 - \theta_5 B^5)\cdot (1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - \ldots) \tag{24}$$

The π-weights (also denoted by $\pi^{(1)}$) are given in Figure 5: thus for example

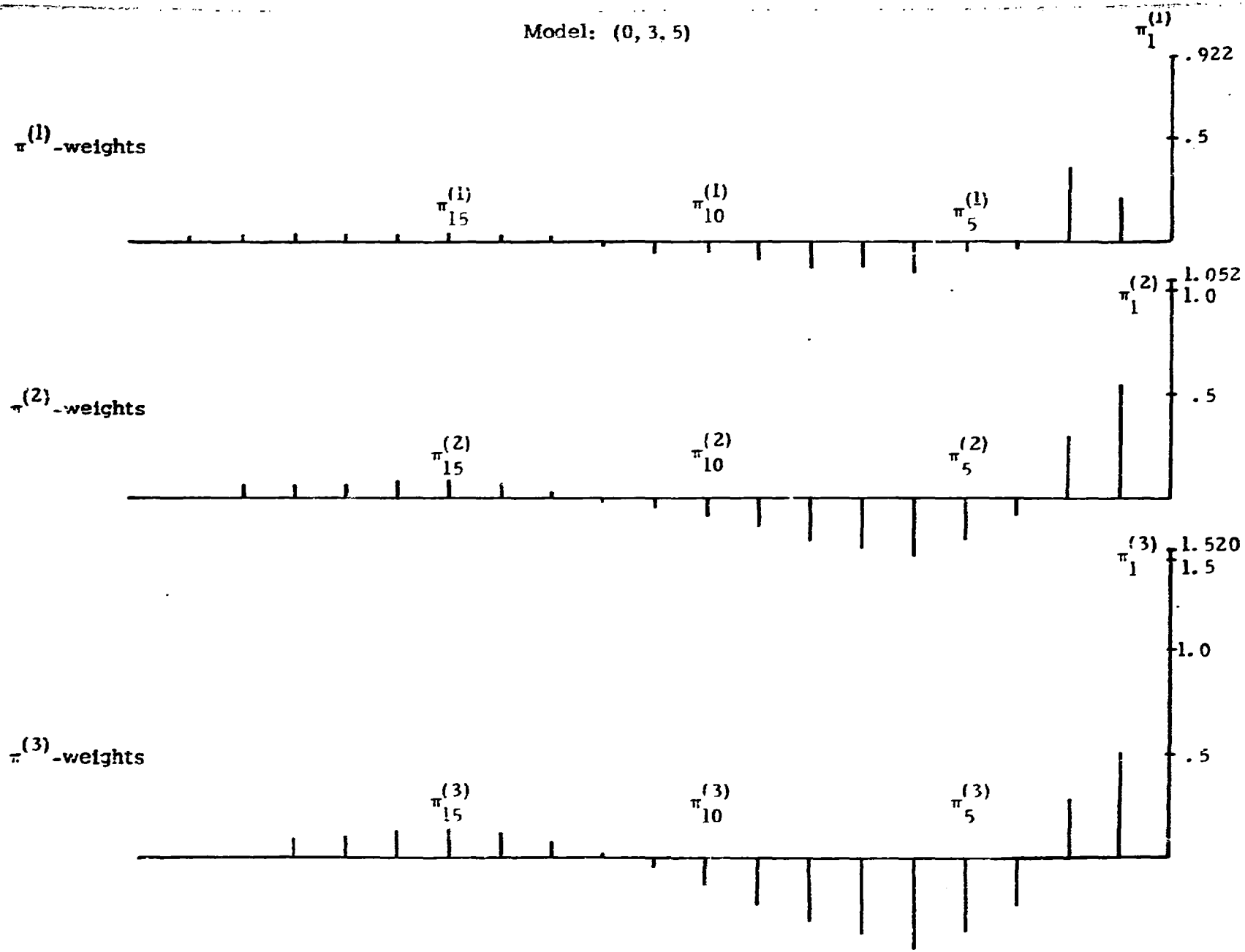


Figure 5. π-weights for the (0,3,5) model

$$z_t(1) = .922 \times z_t + .207 \times z_{t-1} + .355 \times z_{t-2} - .039 \times z_{t-3}$$
$$-.068 \times z_{t-4} - .171 \times z_{t-5} - .149 \times z_{t-6} - .143 \times z_{t-7}$$
$$-.107 \times z_{t-8} - .079 \times z_{t-9} - .046 \times z_{t-10} - .018 \times z_{t-11} \qquad (25)$$
$$+.008 \times z_{t-12} + .027 \times z_{t-13} + .041 \times z_{t-14} + .048 \times z_{t-15}$$
$$+.048 \times z_{t-16} + .044 \times z_{t-17} + .036 \times z_{t-18} + .026 \times z_{t-20} + \ldots$$

The two step ahead forecast can be found similarly by replacing z_t by $\hat{z}_t(1)$ and z_{t-j} by z_{t-j+1}, and so on for forecasts with higher lead times. However these forecasts may also be expressed directly as weighted sums of the observations z_t, z_{t-1}, z_{t-2},.... The weights $\pi^{(2)}$ and $\pi^{(3)}$ corresponding to the two and three step ahead forecasts respectively, are also shown in Figure 5. In the remainder of this paper a brief outline is presented of two other important applications of time series modelling.

4. Intervention Analysis

We frequently need to detect & estimate possible changes in the functioning of a system affected by known interventions.

For example Figure 6 shows monthly averages for ozone in parts per hundred millions (p.p.h.m) measured in downtown Los Angeles (Box & Tiao 1975 [13]). It is known that in January 1960 a law (rule 63) was put into effect whereby the amount of reactive hydrocarbons in gasoline sold throughout L.A. county was reduced. Can a change be detected at this point in the series? If so how large is it?

Furthermore modified engines were made compulsory for new cars introduced after 1966. Can any effect be detected which might plausibly be related to this intervention?

Standard statistical procedures will certainly be invalidated for examples of this kind because

(a) the noise a_t is highly dependent (& in this case seasonal).

(b) the effect of changes made may not be immediately felt but may have dynamic characteristics.

Difference equation models of the form

$$y_t = \frac{\omega(B)}{\delta(B)} + \frac{\theta(B)}{\phi(B)}a_t \qquad (26)$$

can take account of both difficulties.

For the Ozone data a model was developed of the form

$$y_t = \omega_1 x_{1t} + \frac{\omega_2 x_{2t}}{1-B^{12}} + \frac{\omega_3 x_{3t}}{1-B^{12}} + \frac{(1-\theta_1 B)(1-\theta_2 B^{12})}{1-B^{12}}a_t$$

In this expression x_{1t}, x_{2t}, & x_{3t} are indicator variables allowing for possible changes introduced by interventions.

$x_{1t} = \begin{cases} 0 & t < \text{Jan 60} \\ 1 & t \geq \text{Jan 60} \end{cases}$ Allows for step change of size ω_1 possibly associated with rule 63.

$x_{2t} = \begin{cases} 1 & \text{for summer months 66 onwards} \\ 0 & \text{for winter months 66 onwards} \end{cases}$ Produces a staircase function (step size ω_2) to represent possible effect of new car engines in summer conditions.

$x_{3t} = \begin{cases} 0 & \text{for summer months 66 onwards} \\ 1 & \text{for winter months 66 onwards} \end{cases}$ Produces a staircase function (step size ω_3) to represent possible effect of new car engines in winter.

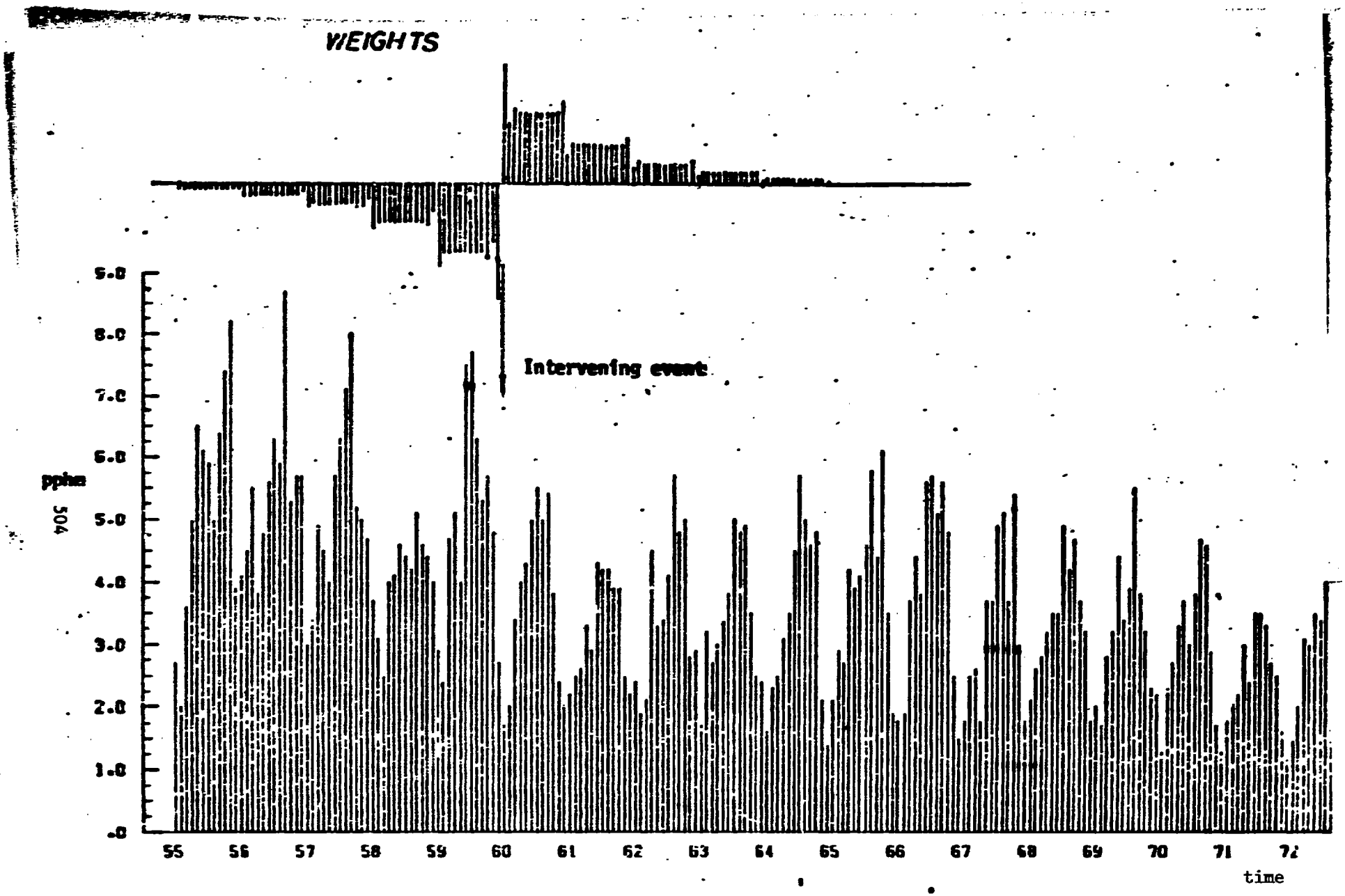


Figure 6. Monthly average of hourly readings of O_3 (pphm) at Downtown Los Angeles (1955-1972). (with the weight function for estimating the effect of intervening events in 1960)

Estimates were obtained as follows

$\hat{\omega}_1 = -1.09 \pm .10, \quad \hat{\omega}_2 = -.25 \pm .07 \quad \hat{\omega}_3 = -.07 \pm .06$

(with $\hat{\theta}_1 = -.24 \pm .03, \quad \hat{\theta}_2 = .55 \pm .06$)

This suggest that

(i) a step change of about -1.1 units occurred at about the time rule 63 was introduced.

(ii) that progressive changes of about -0.25 units per year occurred in the summer months after the new engines were introduced.

(iii) no detectable corresponding effect occurred in the winter.

Seasonal Adjustment

It frequently happens that time series such as inventories of equipment items, army recruitment etc. are highly seasonal. Changes are much more readily understood if appropriate seasonal adjustments are made. An empirical method for separating seasonal series into (i) a seasonal component (ii) a trend component (iii) an additional error component have been discussed by Julius Shiskin (1967), [18] & is presently used extensively, and is referred to as the X11 method. This method produces good results on the average. It is however unable to take account of the particualr properties of individual series. New research suggests that a model-based approach (Box, Hillmer & Tiao 1976), [2] can accomplish this. For example Figure 7 shows results obtained by the X11 method & by the model based method on a time series for unemployed males in the United States 20 & over.

Further research on stochastic difference equation models is currently undergoing vigorous development. In particular, research is being conducted into multivariate applications and to problems in control & the general identification of dynamic systems.

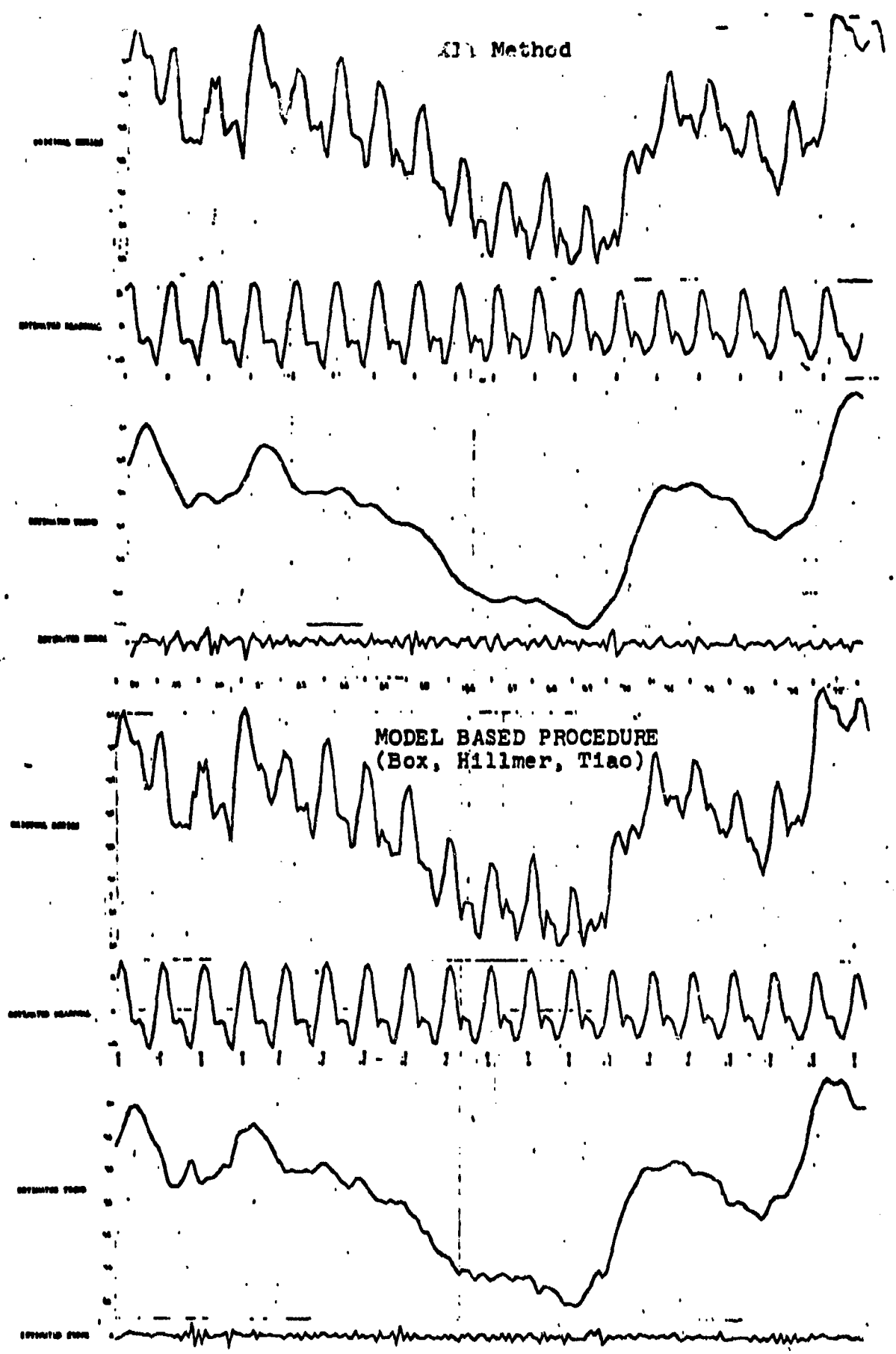


Figure 7. Separation of time series into components using X11 and Model-Based Procedure.

[1] Box, G.E.P. [1976]. Science and Statistics. Journal of the American Statistical Association, December 1976, Volume 71, No. 356 Application Section Pages 791-799.

[2] Box, G.E.P., Hillmer, S.C. and Tiao, G.C. [1976]. Analysis and Modelling of Seasonal Time Series. Tech. Report No. 465, Department of Statistics, University of Wisconsin, Madison.

[3] Box, G.E.P. and Jenkins, G.M. [1962]. Some statistical aspects of adaptive optimization and control. JRSS B, 24, 297.

[4] Box, G.E.P. and Jenkins, G.M. [1963]. Further contributions to adaptive quality control; simultaneous estimation of dynamics: non-zero costs. Bull. Intl. Stat. Inst. 34th Session, 943, Ottawa, Canada.

[5] Box G.E.P. and Jenkins, G.M. [1965]. Mathematical models for adaptive control and optimization, A.I. Ch.E. -I. Chem. E.Symp. Series, 4, 61.

[6] Box, G.E.P. And Jenkins, G.M. [1968]. Some recent advances in forecasting and control I. Applied Stat., 17, 91.

[7] Box, G.E.P. and Jenkins, G.M. [1969]. Discrete models for forecasting and control. Encyclopedia of Linguistics, Information and Control, Pergamon Press, 162.

[8] Box, G.E.P. and Jenkins, G.M. [1970]. Time Series Analysis: Forecasting and Control. Holden Day, San Francisco.

[9] Box, G.E.P., Jenkins, G.M. and Bacon, D.W. [1967]. Models for forecasting seasonal and non-seasonal time series. Advanced Seminar on Spectral Analysis of Time Series, ed. B. Harris, John Wiley, New York, 271.

[10] Box, G.E.P. and Newbold, P., [1971]. Some comments on a Paper of Coen, Gomme and Kendall. JRSS A, 134 229.

[11] Box, G.E.P. and Pallesen, L. [1978]. Adaptive Minimum Mean Square Error Forecast of Missile Trajectory using Stochastic Difference Equation Models. Technical Summary Report No. 1821, Mathematical Research Center, University of Wisconsin-Madison.

[12] Box, G.E.P. and Tiao, G.C. [1965]. A change in level of a non-stationary time series. Biometrika, 52, 181-192.

[13] Box, G.E.P. and Tiao, G.C. [1975]. Intervention Analysis with Applications to Economic and Environmental Problems, JASA, 70, 70-79.

[14] Cleveland, W.P. [1972]. Analysis and Forecasting of Seasonal Series. Unpublished Ph.D. thesis at University of Wisconsin.

[15] Cleveland, W.P. and Tiao, G.C. [1976]. Decomposition of Seasonal Time Series: A model for the Census X-11 Program. (To appear in JASA).

[16] Coen, P.G., Gomme E.D. and Kendall, M.G. [1969]. Lagged relationships in economic forecasting. J.R. Statist. Soc. A. 132, 133-152.

[17] Hillmer, S.H. [1976]. Time Series: Estimation, Smoothing and Seasonally Adjusting. Unpublished Ph.D. Thesis at University of Wisconsin.

[18] Shiskin, J., Young, A.H. and Musgrave, J.C. [1967]. The X-11 Variant of Census Method II Seasonal Adjustment Program, Technical Paper No. 15, Bureau of the Census, U.S. Department of Commerce.

[19] Slutsky, E. [1927]. The summation of random causes as the source of cyclic processes (Russian) Problems of Economic Conditions, 3, 1.

[20] Tiao, G.C., Box G.E.P. and Hamming, W.J. [1975]. Analysis of Los Angeles photochemical smog data: a statistical overview. J. Air Pollution Control Ass. 25, 260-268.

[21] Yaglom, A.M. [1955]. The correlation theory of processes whose n^{th} difference constitute a stationary process. Matem. Sb., 37, 141.

[22] Yule, G.U. [1927]. On the method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions, A, 226, 267.

ATTENDEES, DESIGN OF EXPERIMENT

CONFERENCE *

ALCOCER, Guillermo Mr.
USA White Sands Missile Range
WSMR, NM 88002

ANDERSEN, Gerald R. Dr.
USA Materiel Dev. & Readiness Comd
Alexandria, VA 22314

BAGGE, CARL F. Mr.
Agbabian Asso.
El Segundo, CA 90245

BALESTRI, R. J. Mr.
BDM
Albuquerque, New Mexico 87103

BATES, Carl B. Mr.
USA Concepts Analysis Agency
Bethesda, Maryland 20014

BECHHOFER, Robert Dr.
School of Opns Rsch & Industrial Engr
Cornell University
Ithaca, NY 14850

BOIRUN, Barclay H. Mr.
USA Avn Engineering Flight Actvy
Edwards AFB, CA 93523

BOX, George E. P. Dr.
Univ of Wisconsin-Madison
Math Rsch Center
Madison, Wisconsin 53703

BRATT, Howard M. Mr.
Army Applied Technology Lab
Ft Eustis, VA 23604

BRESLOW, Norman Dr.
University of Washington
Seattle, Washington 98101

BRUGGER, Richard M. Mr.
USA Armament Materiel Readiness Comd
Rock Island, Illinois 61200

---

*See end of this list for the names of the local attendees.

BRYANT, Wayne H.
NASA, Langley Research Center
Hampton, VA 23362

CARRILLO, Jesus E. Mr.
USA TRADOC Systems Analysis Activity
WSMR, New Mexico88002

CARROLL, Robert M.
Army Research Institute
Alexandria, VA 22314

CAVINESS, Jim Dr.
USA Administration Center
ATTN: DCO
Ft Benjamin Harrison, IN 46216

COLLINS, Jon D. Dr.
J. H. Wiggins Company
Redondo Beach, CA 90277

COPPOLA, Eugene E. Mr.
Watervliet Arsenal
Watervliet, NY 12189

CROW, Larry H. Dr.
USA Materiel Systems Analysis Activity
Aberdeen Proving Ground, MD 21005

CULP, David Dr.
BDM
Albuquerque, NM 87106

CULPEPPER, Gideon A. Mr.
USA WSMR
White Sands Missile Range, NM 88002

D'ACCARDI, Richard J. Mr.
USA Electronics Lab
Ft Monmouth, NJ 07703

DENTON, James Q. Mr.
National Science Foundation
Washington, DC 20550

DePRIEST, Douglas J. Dr.
Office Naval Research
Arlington, VA 22212

DYER, Danny Dr.
Univ of Texas at Arlington
Arlington, TX 76010

DZIWAK, Mr.
USA Armament Command
Dover, NJ 07801

EICHELMAN, Edward R. Mr.
USA Mobility Equip R&D Center
Ft Belvoir, VA 22060

EISENHART, Churchill Dr.
National Bureau of Standards
Washington, D.C. 20310

ERICKSON, Richard A. Mr.
Boeing Company
Seattle, Washington 98104

FEINGOLD, Harry Mr.
David Taylor Naval Ship R&D Center
Bethesda, MD 20014

FERRELL, Kenneth R. Mr.
USA Avn Engineering Flight Actvy
Edwards AFB, CA 93523

FOSTER, Walter D. Dr.
Armed Forces Inst of Pathology
Washington, DC 20310

GAVER, Donald P. Dr.
Naval Postgraduate School
Monterey, CA 93940

GREENLAND, C. Maxson Mr.
USA Rsch & Dev Comd
Chemical Systems Lab
Aberdeen Proving Ground, MD 21005

GRUBBS, Frank E. Dr.
USA Ballistic Research Lab
Aberdeen Proving Ground, MD 21005

HAGGSTROM, Gus W. Dr.
Rand Corporation
Santa Monica, CA 90406

HALL, Robert A. Mr.
Army Applied Technology Lab
Ft Eustis, VA 23604

HANNIGAN, Joseph F. Mr.
USA Engineer Topographic Labs
Ft Belvoir, VA 22060

HARRIS, Bernard Dr.
Univ of Wisconsin-Madison
Mathematics Research Center
Madison, WI 53703

HARRIS, J. L. Mr.
US Army Missile R&D Command
Redstone Arsenal, AL 35809

HARTLEY, H. O. Dr.
Dir, Inst of Statistics
Texas A&M Univ
College Station, TX 77840

HAYDEN, James S. Mr.
USA Avn Engineering Flight Actvy
Edwards AFB, CA 93523

HERBACH, Leon H. Dr.
Polytechnic Inst of New York
Brooklyn, NY 11201

HOPKINS, Alan Mr.
Letterman Army Inst of Rsch
Presidio of SF, CA 94129

JAYACHANDRAN, Toke Dr.
Naval Postgraduate School
Monterey, CA 93940

JOHNS, M. Vernon Dr.
Stanford Univ
Stanford, CA 94305

JOHNSON, Lyle Mr.
US Naval Surface Weapons Center
Dalgren, VA 22448

JONES, Daniel B. Dr.
Army Research Institute
Fort Monroe, VA 23351

KEATING, Jerome P. Mr.
Bell Helicopter Textron
Ft Worth, TX 76101

KHEIR, Naim A. Dr.
Univ of Alabama
Huntsville, AL 35807

KINEMOND, Gary A. Mr.
Sandia Labs
Albuquerque, NM 87115

KNACK, Fredrick H. CPT
USA TRADOC Sys Analysis Agency
White Sands Missile Range, NM 88002

KOWAL, Dennis M. CPT
USA Rsch Inst of Environmental Med
Natick, Mass 01760

KURKJIAN, Badrig Dr.
Mathematical Sciences Dept
Univ of Alabama
Huntsville, AL 35807

LAMBERSON, Leonard R. Dr.
Wayne State Univ
Detroit, Michigan 48202

LEACH, Paul J. MAJ
USA Medical R&D Command
Washington, D.C. 20314

LANK, Gerald W. Dr.
Mark Resources Incorporated
Marina Del Rey, CA 90306

LARSON, Harold Jr. Dr.
Naval Postgraduate School
Monterey, CA 93940

LAUNER, Robert L. Dr.
USA Research Office
Research Triangle Park, N.C. 27709

LEV, Ovadia Dr.
Merritt CASES Inc.
Redlands, CA 92373

MAIER, Milton H. Dr.
USA Rsch Inst for the Behavioral
& Social Sciences
Alexandria, VA 22314

McDONALD, Bruce J. Dr.
Office of Naval Research
American Embassy Tokyo
APO S.F. 96503

MELLINGER, J. J. Dr.
Army Research Institute
Alexandria, VA 22314

MILLER, Rupert Dr.
Dept of Statistics
Stanford Univ
Stanford, CA 94305

MOORE, J. Richard Dr.
USA Ballistic Rsch Lab
Aberdeen Proving Ground, MD 21005

MOORE, Richard L. Dr.
US ARRADCOM HQ
Dover, N. J. 07801

ORLEANS, Beatrice S. Miss
Naval Sea Systems Command
Washington, D.C. 20350

PALMER, Charlotte A. Ms
USA MILPERCEN, Rm 1N57
Alexandria, VA 22332

PASTRICK, Harold L. Dr.
US Army Missile R&D Command
Redstone Arsenal, AL 35809

PECK, Carl C. LTC
Letterman Army Inst of Rsch
Presidio of S.F., CA 94129

ROSE, Carol D. Mr.
USA Tank-Automotive R&D Command
Warren, MI 48090

ROSS, Edward W. Dr.
USA Natick R&D Command
Natick, MA 01776

RUSSELL, Carl T.
USA Opnl Test & Eval Command
Falls Church, VA 22041

SHEPHERD, William L. Mr.
USA White Sands Missile Range
WSMR, NM 88002

SIEGEL, Andrew F. Dr.
Univ of Wisconsin-Madison
Mathematics Rsch Center
Madison, WI 53703

SMITH, David M. Mr.
Lockheed-Georgia Company
Marietta, GA 30060

STARNER, John W. Dr.
USA White Sands Missile Range
WSMR, NM 88002

STEINHEISER, Frederick H. Jr., Dr.
US Army Research Institute
Alexandria, VA 22314

SUTHERLIN, Donald W. Dr.
USA Missile R&D Command
Redstone Arsenal, AL 35809

TANG, Douglas B. Dr.
Walter Reed Army Inst of Rsch
Washington, D.C. 20012

TAYLOR, Malcolm S. Dr.
USA Ballistic Rsch Lab
Aberdeen Proving Ground, MD 21005

THOMPSON, William A. Dr.
Univ of Missouri
Statistics Dept
Columbia, MO 65201

THRASHER, Paul H. Dr.
USA White Sands Missile Range
WSMR, NM 88002

TOMAINE, Robert L. Mr.
NASA, Langley Rsch Center
Hampton, VA 23369

TUPPER, Bruce Mr.
Stanford Research Institute
Stanford, CA 94305

WEINBERGER, M. A. Dr
Opns Research & Analysis Center
Ottawa, Canada

WILBURN, John B., Jr. Dr.
USA Electronic Proving Ground
Fort Huachuca, AZ 85613

WILLIS, Roger F. Mr.
USA TRADOC Systems Analysis Acty
WSMR, NM 88002

WITHERS, Lang P. Mr.
USA Opnl Test & Eval Command
Falls Church, VA 22041

WOLMAN, Agatha Mrs.
Federal Energy Admin
Washington, D.C. 20504

WOLMAN, William W. Dr.
US Dept of Trans, Fed Highway Admin
Washington, D.C. 20591

WONG, James T. Dr.
USA Rsch & Technology Labs
Moffett Field, CA 94035

WU, Chien-Fu Mr.
Dept of Statistics
Univ of Wisconsin-Madison
Madison, WI 53703

WYKOFF, Norman L. Dr.
USA Jefferson Proving Ground
Madison, N.J. 07940

YOUNG, Warren H., Jr. Dr.
NASA, Langley Rsch Ctr
Hampton, VA 23362

ZANE, Nancy Ms
Monterey Peninsula Unified School District
Monterey, CA 93940

LOCAL ATTENDEES

DESIGN OF EXPERIMENT CONFERENCE

DAVIS, Reed E., Jr. COL
Office of the Scientific Advisor
HQ CDEC
Ft Ord, CA 93941

STERNBERG, Jack J. Mr.
ARI
Presidio of Monterey, CA 93940

DOUGLAS, Robert MAJ
DCS Instrumentation
HQ USACDEC
Ft Ord, CA 93941

TILLER, Lawrence S. Mr.
BDM Corporation
Fort Ord, CA 93941

REHM, Walter MAJ
DCS Instrumentation
HQ USACDEC
Ft Ord, CA 93941

McKEARN, C. F. MAJ
DCS Plans
HQ USACDEC
Ft Ord, CA 93941

BANKS, James H. Dr.
Army Research Institute
Ft Ord, CA 93941

BROWN, Diane M. Ms
DCS Plans
HQ USACDEC
Ft Ord, CA 93941

BANKS, John E. Mr.
Office of the Scientific Advisor
HQ USACDEC
Ft Ord, CA 93941

LOVE, Gary
DCS Plans
HQ USACDEC
Ft Ord, CA 93941

BATESOLE, Robert D. Mr.
BDM
Ft Ord, CA 93941

BRYSON, Marion R. Dr.
Scientific Advisor
HQ USACDEC
Ft Ord, CA 93941

BARR, Brian CPT
DCS Plans
HQ USACDEC
Ft Ord, CA 93941

MALLIOS, William S.
BDM
Ft Ord, CA 93941

BROWN, David E. SP5
DCS Plans
HQ USACDEC
Ft Ord, CA 93941

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARO Report 78-2
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): PROCEEDINGS OF THE TWENTY-THIRD CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT, AND TESTING
5. TYPE OF REPORT & PERIOD COVERED: Interim Technical Report
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
8. CONTRACT OR GRANT NUMBER(s)

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 1978
13. NUMBER OF PAGES: 518

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): U. S. Army Research Office, P. O. Box 12211, Research Triangle Park, NC 27709
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. 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.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This is a technical report resulting from the Twenty-third 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.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

unbalanced experiments
measure of effectiveness
ratio data
sustained work in young women
least chi-square
polynomial splines
flight tests
simulations
analysis of variance
experimental designs
radars
confidence limits
programmable calculator
histogram and spectral analysis
rank analysis
under-identification
curve fitting
man-machine interface
spline functions
sensitivity evaluation
reliability
binomial data
jackknife
censored data
complex system availability
tank camouflage
markov chains
time series modelling

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified