ARMY RESEARCH LABORATORY

Proceedings of the Fifth Annual U.S. Army Conference on Applied Statistics, 19-21 October 1999

Barry A. Bodt
EDITOR

Hosted by:
UNITED STATES MILITARY ACADEMY

Cosponsored by:
U.S. ARMY RESEARCH LABORATORY
U.S. MILITARY ACADEMY
U.S. ARMY RESEARCH OFFICE
WALTER REED ARMY INSTITUTE OF RESEARCH
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
TRADOC ANALYSIS CENTER-WSMR

ARL-SR-110 July 2001

Approved for public release; distribution is unlimited.

20010828 155

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-SR-110 July 2001

Proceedings of the Fifth Annual U.S. Army Conference on Applied Statistics, 19-21 October 1999

Barry A. Bodt, EDITOR
Computational and Information Sciences Directorate, ARL

Hosted by:
United States Military Academy

Cosponsored by:
U.S. Army Research Laboratory
U.S. Military Academy
U.S. Army Research Office
Walter Reed Army Institute of Research
National Institute of Standards and Technology
TRADOC Analysis Center-WSMR

Approved for public release; distribution is unlimited.

Abstract

The fifth U.S. Army Conference on Applied Statistics was hosted by the United States Military Academy, West Point during 19–21 October 1999. The conference was cosponsored by the U.S. Army Research Laboratory (ARL), the U.S. Army Research Office (ARO), the United States Military Academy (USMA), the Training and Doctrine Command (TRADOC) Analysis Center-White Sands Missile Range, the Walter Reed Army Institute of Research (WRAIR), and the National Institute for Standards and Technology (NIST). The U.S. Army Conference on Applied Statistics is a forum for technical papers on new developments in statistical science and on the application of existing techniques to Army problems. This document is a compilation of available papers offered at the conference.

FOREWORD

The fifth U.S. Army Conference on Applied Statistics was hosted by the United States Military Academy (USMA), West Point during 19–21 October 1999. The conference was cosponsored by the U.S. Army Research Laboratory (ARL), the U.S. Army Research Office (ARO), USMA, the Training and Doctrine Command (TRADOC) Analysis Center - White Sands Missile Range, the Walter Reed Army Institute of Research (WRAIR), and the National Institute for Standards and Technology (NIST). The U.S. Army Conference on Applied Statistics is a forum for technical papers on new developments in statistical science and on the application of existing techniques to Army problems. The purpose of this conference is to promote the practice of statistics in the solution of these diverse Army problems.

The fifth conference was preceded by a short course, "Data Mining with Decision Trees," given by Professor Wei-Yin Loh of the University of Wisconsin-Madison. COL David C. Arney, Chair, Department of Mathematical Sciences, opened the conference. BG Fletcher M. Lamkin, Jr., Dean, Academic Board, offered welcoming remarks. Several distinguished speakers spoke during invited general sessions: Edward Wegman (keynote), George Mason University; David Banks, Bureau of Transportation Statistics; Daryl Pregibon, AT&T; Jayaram Sethuraman, Florida State University; Charles McCulloch, Cornell University; and Ingram Olkin, Stanford University. The first three of these talks were related to the data mining theme of the conference. An important moment in the conference was the awarding of the Army Wilks Medal to Edward Wegman of George Mason University. Professor Wegman was recognized "for fundamental work in mathematical and computational statistics, for conceptual innovation which has changed the direction and character of statistical research, for innovative problem solving in applied settings, and for continuing support of Army statisticians and scientists."

The Executive Board for the conference recognizes COL David C. Arney, USMA, for hosting the conference; LTC Philip Beaver, USMA, for handling all local details; Mr. David Webb, ARL, and Mr. Edmund Baur, ARL, for assisting with advertisement; Dr. Jock Grynovicki, ARL, for chairing the Army Wilks Award Committee, and Dr. Barry Bodt, ARL, for chairing the conference and serving as editor of the proceedings.

Executive Board		
Barry Bodt (ARL)	Robert Burge (WRAIR)	David Cruess (USUHS)
Paul Deason (TRAC-WSMR)	Eugene Dutoit (AIS)	Jock Grynovicki (ARL)
Robert Launer (ARO)	LTC Philip Beaver (USMA)	LTC David Olwell (NPS)
Carl Russell (JNTF)	Douglas Tang (WRAIR)	Deloras Testerman (YPG)
Mark Vangel (NIST)	David Webb (ARL)	Edward Wegman (GMU)

INTENTIONALLY LEFT BLANK.

Table of Contents*

* This Table of Contents contains only the papers that appear in the Proceedings.

FINAL PROGRAM
FIFTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

19-21 October, 1999

Hosted by the United States Military Academy
Cosponsored by:

U.S. Army Research Laboratory
U.S. Military Academy
U.S. Army Research Office
Walter Reed Army Institute of Research
National Institute of Standards and Technology
TRADOC Analysis Center-WSMR

Sunday, 17 October 1999

1300 - 1400 REGISTRATION (Highlands Room, Hotel Thayer)

1400 - 1800 TUTORIAL: (Highlands Room, Hotel Thayer)

DATA MINING WITH DECISION TREES
Wei-Yin Loh, University of Wisconsin, Madison

Monday, 18 October 1999

0800 - 1200 TUTORIAL: (Room 144, Thayer Hall)

1200 - 1330 Lunch

1330 - 1600 TUTORIAL

1800 - 2000 SOCIAL (Garden Terrace, Hotel Thayer)

Tuesday, 19 October 1999

0800 - 0845 REGISTRATION (Room 144, Thayer Hall)

0845 - 0915 CALL TO ORDER (Room 144, Thayer Hall)

Chairman: Barry Bodt, U.S. Army Research Laboratory

Host: COL David C. Arney Chair, Department of Mathematical Sciences, United States Military Academy

OPENING REMARKS: BG Fletcher M. Lamkin, Jr. Dean, Academic Board, United States Military Academy

0915 - 1145 GENERAL SESSION I (Room 144, Thayer Hall)

Chair: Robert Burge, Walter Reed Army Institute of Research

0915 - 1015 KEYNOTE ADDRESS

DATA MINING AND VISUALIZATION: SOME STRATEGIES
Edward Wegman, George Mason University

1015 - 1045 Break

1045 - 1145 REDUCTION IN PREDICTIVE ABILITY CAUSED BY DISCRETIZATION OF THE INDEPENDENT VARIABLE J
Jayaram Sethuraman, Florida State University

1145 - 1300 Lunch

1300 - 1430 CLINICAL SESSION I (Room 144, Thayer Hall)

Chair: Bernie Harris, University of Wisconsin, Madison

Panel:
Nozer Singpurwalla, George Washington University
Jim Thompson, Rice University

STATISTICAL METHODS TO VALIDATE THE NATIONAL MISSILE DEFENSE BATTLE MANAGEMENT COMMAND AND CONTROL SOFTWARE
Alyson Wilson, Los Alamos National Laboratory
Max Morris, Iowa State University

Tuesday, 19 October 1999 (Continued)

VALIDATION OF A LIVE-VIRTUAL FIELD EXPERIMENT USING CONSTRUCTIVE SIMULATION
Paul Deason, TRADOC Analysis Center, White Sands Missile Range

1430 - 1445 Break

1445 - 1545 GENERAL SESSION II (Room 144, Thayer Hall)

Chair: Robert Launer, U.S. Army Research Office

METHODS FOR COMBINING THE RESULTS OF INDEPENDENT STUDIES
Ingram Olkin, Stanford University

1545 - 1600 Break

1600 - 1730 CONTRIBUTED SESSION I (Room 144, Thayer Hall)

Chair: Deloris Testerman, Yuma Proving Ground

A METHODOLOGY FOR QUANTIFYING CRITICAL DECISION EVENTS DURING THE EXECUTION PHASE OF BATTLE
Jock Grynovicki, Kragg Kysor, and Michael Golden, U.S. Army Research Laboratory

FLEXIBILITY OF FRACTIONAL FACTORIALS IN FIELD TESTS
Carl T. Russell, Ballistic Missile Defense Organization

STATISTICS IN THE MILITARY OPERATIONS IN URBAN TERRAIN (MOUT) ADVANCED CONCEPTS TECHNOLOGY DEMONSTRATION (ACTD)
Eugene Dutoit, Dismounted Battlespace Battle Lab

1600 - 1730 CONTRIBUTED SESSION II (Room 444, Thayer Hall)

Chair: David Webb, U.S. Army Research Laboratory

METHODS TO ANALYZE THE EFFECT OF DECREASING DIGITAL IMAGE RESOLUTION ON TELEDERMATOLOGY DIAGNOSIS
Paul B. Hshieh, David Cruess, and Dennis A. Vidmar, Uniformed Services University of the Health Sciences

Tuesday, 19 October 1999 (Continued)

TVA USING A BAYESIAN APPROACH WITH NO PRIOR INFORMATION
Doug Frank, Indiana University, PA Ann E. M. Brodeen, U.S. Army Research Laboratory

MEMBERSHIP FUNCTIONS AND PROBABILITY MEASURES OF FUZZ SETS
Nozer Singpurwalla, George Washington University

1830 - WILKS AWARD BANQUET (Garden Terrace, Hotel Thayer)

1830 - 1930 SOCIAL (Cash Bar)

1930 - BANQUET

Wednesday, 20 October 1999

0800 - 0930 CLINICAL SESSION II (Room 144, Thayer Hall)

Chair: Jay Conover, Texas Tech University

Panel:
Edward Wegman, George Mason University
Charles McCulloch, Cornell University

TESTING FOR A DIFFERENCE BETWEEN TWO VARIANCE COMPONENTS OBTAINED FROM DEPENDENT DATA SETS
David W. Webb, U.S. Army Research Laboratory

OPTIMIZING A TARGET DETECTION ALGORITHM THROUGH USE OF A RECEIVER OPERATOR CURVE
Barry A. Bodt, Philip J. David, David B. Hillis, U.S. Army Research Laboratory

0930 - 0945 Break

0945 - 1200 GENERAL SESSION III (Room 144, Thayer Hall)

Chair: Doug Tang, Walter Reed Army Institute of Research

0945 - 1045 2001: A STATISTICAL ODYSSEY
Daryl Pregibon, AT&T

1045 - 1100 Break

1100 - 1200 AN INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS
Charles McCulloch, Cornell University

1200 - 1215 Break

1215 - 1315 CONTRIBUTED SESSION III (Room 144, Thayer Hall)

Chair: David Banks, Bureau of Transportation Statistics

A NEW MODEL FOR NONPARAMETRIC ACCELERATED LIFE TESTING
Miguel Arcones, Binghamton University; Paul Kvam, Georgia Institute of Technology; Frank Samaniego, University of California-Davis

Wednesday, 20 October 1999 (Continued)

1215 - 1315 CONTRIBUTED SESSION IV (Room 444, Thayer Hall)

Chair: David Cruess, Uniformed Services University of the Health Sciences

AN INTERVENTION ANALYSIS OF U.S. COCAINE PRICE AND PURITY
Samir Soneji, Columbia University; Robert Anthony and Arthur Fries, Institute for Defense Analyses

PROTECTING THE PUBLIC: APPLICATIONS OF STATISTICAL PROCESS CONTROL IN LAW ENFORCEMENT
LTC Dave Olwell and LCDR Rob Weitzman, Naval Postgraduate School

1315 - 1345 BREAK (Move to Boat Dock)

1345 - 1600 BOX LUNCH-HUDSON CRUISE

Thursday, 21 October

0815 - 0915 GENERAL SESSION IV (Room 144, Thayer Hall)

Chair: Paul Deason, TRADOC Analysis Center, White Sands Missile Range

MINING SUPERLARGE DATASETS
David Banks, Bureau of Transportation Statistics

0915 - 0930 Break

0930 - 1045 CONTRIBUTED SESSION V (Room 144, Thayer Hall)

Chair: Barry Bodt, U.S. Army Research Laboratory

SOME ASPECTS OF DEPENDENT FAILURES AND THEIR STATISTICAL ANALYSES
Bernie Harris, University of Wisconsin-Madison

DISRUPTIVE MODELS
Jim Thompson, Rice University

1045 -1100 BREAK

1100 - 1230 CONTRIBUTED SESSION VI (Room 144, Thayer Hall)

Chair: MAJ Philip Beaver, United States Military Academy

A SIMULATION PROCESS FOR DETERMINING THE LIFETIME OF THE M60 TANK TORSION BAR
Don Neal, SRRC

PROBLEMS OF CORRELATION IN THE PROBABILISTIC APPROACH TO COST ANALYSIS
Stephen A. Book, The Aerospace Corporation

PROTECTING THE FORCE: APPLICATIONS OF STATISTICAL PROCESS CONTROL IN BOSNIA
LTC Dave Olwell and CPT Paul Finken, Naval Postgraduate School

1230 ADJOURN

INTENTIONALLY LEFT BLANK.

Visual Data Mining

Edward J. Wegman
Center for Computational Statistics
George Mason University
Fairfax, VA 22030-4444

1. Introduction

This paper is intended to be illustrative of some of the visualization tools we have exploited in a data-mining-like framework. Huber (1992, 1994) developed a taxonomy of data set size which characterized data by increasing orders of magnitude. In earlier discussions, Wegman (1995), I have pointed out that somewhere around 10^6 to 10^7 bytes appears to be the practical limit for visualization of data while massive data sets easily venture into the range of 10^{12} bytes. The human eye has approximately 10^7 cones implying that visualizing one observation per cone would optimistically put the upper limit of visual resolution at about 10^7 observations. (I refer the reader to Wegman (1995) for more details on the limits of visualization and of computational feasibility.) Thus, data mining as such cannot successfully exploit visualization for truly massive data sets without some modification of the raw data. In Wegman (1999) I have suggested several approaches including binning and thinning to reduce the size of data sets making visual analysis more feasible.

With this caveat made, I would like to illustrate in this paper how several more or less standard statistical tasks can be carried out visually. Our basic approach involves a combination of three tools, parallel coordinates multidimensional displays, the *d*-dimensional grand tour, and saturation brushing. In combination these three tools are available in a down-loadable software called ExplorN (available at ftp://www.galaxy.gmu.edu/pub/software/) and also a commercial version called CrystalVision soon to be available.

Parallel Coordinates is a multidimensional visualization tool discussed by Inselberg (1985) and employed for data visualization by Wegman (1990). In order to represent a *d*-dimensional point, the basic idea is to draw *d* parallel axes labeling them according to the data variables. A point is then represented by locating the value of each variable (component) long its respective axis and then joining the resulting points by a broken line segment. Several such diagrams are found in this paper. A fuller discussion of the statistical and data analytic interpretations of parallel coordinate displays is given in Wegman (1990). In some sense a parallel coordinate display is a generalization of a two-dimensional scatterplot. We shall refer to "data clouds" even when strictly speaking the parallel coordinate display involves line segments.

The *d*-dimensional Grand Tour is a generalization of the 2-dimensional grand tour introduced by Asimov (1985). The basic idea of a grand tour is to look at a data cloud from all possible points of view. As implemented, the *d*-dimensional tour is a continuous geometric transformation of a *d*-dimensional coordinate systems such that all possible orientations of the coordinate axes are eventually achieved. The algorithm is described in Wegman (1991) and also in Wegman and Carr (1993). Coupled with the parallel coordinate display, these two techniques allow for an in-depth study of high dimensional data. Partial grand tours can be accomplished by holding one or more variables fixed. A grand tour is in some sense a generalization of a two-dimensional rotation although it is not a rotation in the conventional sense.

Saturation Brushing is a generalization of ordinary brushing. Ordinary brushing is accomplished by brushing a data cloud with a color for the purpose of visually isolating segments of the data. In data sets where there is considerable overplotting, ordinary brushing is potentially confusing in misleading, particularly where there is an animation such as rotation or grand tour. This is particularly the case because in many computer graphics algorithms, colors are drawn according to the z-depth, lowest z-depth points are drawn last. This can lead to apparently arbitrary changes of color and certainly gives no clue as to the amount of overplotting. In saturation

brushing, each point is assigned a highly desaturated color (nearly black) and when points are overplotted, their color saturations are added via the so-called α-channel. Thus heavily overplotted pixels have fully saturated colors whereas pixels with little overplotting remain nearly black. Saturation brushing is described by Wegman and Luo (1997) and is an effective method for dealing with large data sets. Coupled with parallel coordinates and the grand tour, these methods allow for an extremely effective visual approach to large, high-dimensional data.

2. Density Estimation and Rapid Data Editing

Saturation brushing can be used with either parallel coordinate displays or with ordinary scatterplots. In effect, by using saturation brushing with a single color (say white against a black background or gray against a white background) the brushed display becomes a density estimate. Local smoothing is not required except to the extent that the data is binned by the resolution of the screen. For ordinary high resolution displays this means there are approximately 1.3×10^6 bins or pixels. With the α-channel, this density estimation is accomplished with the same speed as the rendering of an ordinary two-dimensional image. Figure 2.1 shows the so-called Pollen data which consists of 3848 points. At first glance this data would simply seem to be five-dimensional with elliptical contours suggesting an uninteresting multivariate normal data set. However, when a desaturated parallel coordinate plot is examined in Figure 2.2 at least two interesting features can be observed. First a bright feature in the middle of the diagram suggests that there is a hidden structure. In addition the larger X-features in the display suggest that the overall elliptical clouds have a five-dimensional hole in the middle. The central structure can rapidly be isolated using a cutting tool. This is best done in the parallel coordinate display. The result is shown in Figure 2.3. The central structure shown in Figure 2.2 is actually 99 data points spelling the word EUREKA as illustrated in Figure 2.3. The five-dimensional hole mentioned above can be verified by doing a more conventional density estimate.

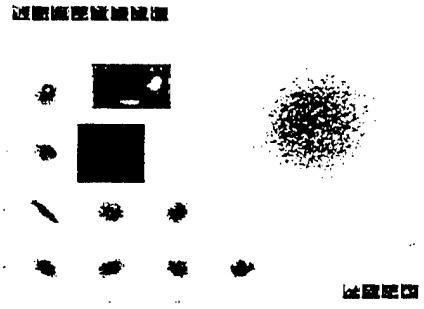


Figure 2.1 Scatterplot matrix of the Pollen data.

Figure 2.2 Desaturated parallel coordinate display of Pollen data.

3. Inverse Regression and Tree-Structured Decision Aids

The combination of parallel coordinates, partial grand tour and saturation brushing can be used in conjunction to achieve a form of inverse regression and a sort of tree structured decision aid. The data we use here to illustrate is 8-dimensional bank demographic data containing approximately 12,000 observations. One of the variables is profit, a variable we will treat as the dependent variable. After appropriate data editing for unknowns, we brush the profit variable with either red or green. Those observations with negative profit (loss) are brushed red, those with positive profit are brushed green. In an additive color system, red and green together make yellow. Thus regions of the covariates that show up as yellow are neutral with respect profit or loss. However, regions that are predominantly red reveal demographics that cause the bank to lose money while regions that are predominantly green represent the demographics of desirable customers. The partial grand tour is used to animate the demographic variables. Because the *d*-grand tour as we have implemented it forms orthogonal linear combinations of the demographic variables, we can visualize linear combinations of the demographic variables as a function of the dependent variable (profit). Thus we have a tool for inverse regression. Moreover by isolating linear combinations of demographic variables that are predominantly green or red, we can create a tree-structured decision rule for distinguishing desirable customers from undesirable

Figure 2.3 Central feature of the Pollen data, the word EUREKA.

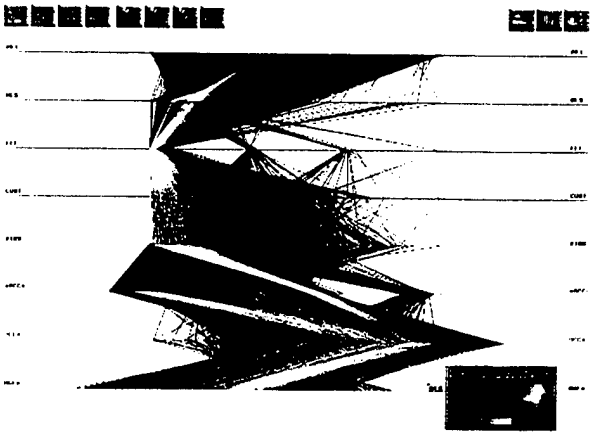


Figure 3.1 Bank data after editing, brushing profit variable with red and green, and a partial grand tour on the demographic variables

customers. Unfortunately full color graphics are not available in the printed form of this paper so the full impact of this technique cannot be seen here. However, a full version is available with the color graphics at our website, URL www.galaxy.gmu.edu/papers/datamining&visualization.html. The reader is encouraged to examine the sequence of images on our website.

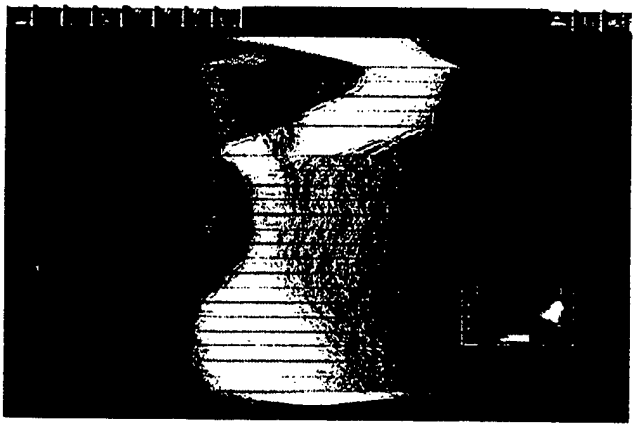


Figure 4.1 SALAD data in 13 dimensions showing three clusters.

4. Variable Selection and Dimensionality Reduction

Parallel coordinate displays coupled with saturation brushing can allow us to select variables visually for the purpose of discriminant analysis. This is illustrated with the so-called SALAD data, a thirteen-dimensional data set containing spectral response of some 10,000 plus chemical samples. The data is shown in Figure 4.1. The thirteen variables are intensity of spectral response in thirteen color bands with increasing wavelength. The variable on the bottom axis is a classification variable and is brushed with one of three colors red, blue, or green according to the class of chemicals. The idea is to look for one or more of the spectral bands which adequately discriminates the three classes of chemicals. Although less apparent in the black and white version, the additive color feature can again be exploited. Red + blue = magenta, red + green = yellow, and blue + green =cyan. Variable B10 separates blue and red, and in fact shows two distinct red clusters. Unfortunately, B10 does not discriminate red from green. However, in parallel coordinate displays, the slope of the line segments matters considerably. The slope of the green line segments between B9 and B10 is substantially different from the slope of the red line segments between the same axes. Thus the variable B10-B9, a surrogate for slope, will distinguish red from green. Thus only two variables, B9 and B10 are adequate to discriminate all three classes and in fact also discriminate the two subclasses of the red. This dimensionality reduction allows for real-time discrimination of chemical agents. Again we encourage the reader to view the color images at our website.

5. Clustering and Classification

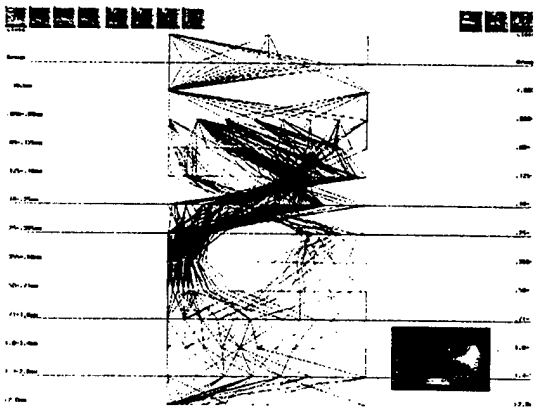


Figure 5.1 Oronsay sand particle distribution data for two distinct sites.

Our last example uses a combination of parallel coordinates and the grand tour to achieve a combination of clustering and classification. The data addressed here is a 12-dimensional dataset called the Oronsay sand data. This was treated more fully in Wilhelm et al. (1999). The basic idea is that samples of sand from various locations is taken and put through a series of sieves. The weight of sand isolated by each sieve is measured resulting in a distribution of particle sizes for each sand sample. The overall goal of the Oronsay sampling experiment was to classify Mesolithic sand samples as to whether they resembled contemporary beach sand or contemporary dune sand. Samples

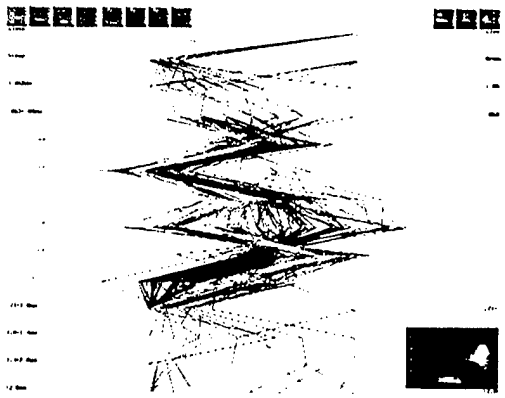


Figure 5.2 Oronsay sand data for so-called CC site after BUSH-TOUR strategy.

from two different locations in the Scottish Hebrides were taken and are illustrated in Figure 5.1. Although not recognized in the original analysis done by archeologists, Figure 5.1. shows that the two locations have distinctly different particle size distributions. Thus only the location with larger sample size (the sample colored black) was used in our analysis. Figure 5.2 is the result of following the BRUSH-TOUR strategy. The basic idea is to brush all visible clusters in the original orientation of the data with distinct colors. Then allow a grand tour rotation until new clusters show up. Brush the new clusters and repeat the BRUSH-TOUR until no new clusters are apparent. One can be confident of having isolated all clusters when the clusters identified by this technique move coherently under the grand tour rotation. In Figure 5.2, the beach sand is brushed with green, the dune sand is brushed with red and the unknown Mesolithic sand is brushed with black. The conclusion of our experiment was that although the Mesolithic sand more closely resembled contemporary dune sand it was still distinctly different and was really in a cluster by itself. The complete Oronsay sand analysis is available at our website http://www.galaxy.gmu.edu/papers/oronsay.html.

Acknowledgements

The preparation of this paper was supported by the United States National Science Foundation with a Group Infrastructure Grant DMS-9631351 and by the United States Army Research Office under Grant DAAG55-98-1-0404. I

References

Asimov, D., "The grand tour: a tool for viewing multidimensional data," ***SIAM J. Sci. Statist. Comput.***, 6, 128-143, 1985.

Huber, Peter J. (1992), "Issues in computational data analysis," in ***Computational Statistics***, Vol.2 (Y. Dodge and J. Whittaker, eds.), Physica Verlag, Heidelberg.

Huber, Peter J. (1994), "Huge data sets," in ***Compstat 1994: Proceedings***, (R. Dutter and W. Grossmann, eds.), Physica Verlag, Heidelberg.

Inselberg, A. (1985), "The plane with parallel coordinates," ***The Visual Computer***, 1, 69-91.

Wegman, E. (1990), "Hyperdimensional data analysis using parallel coordinates," ***Journal of the American Statistical Association***, 85, 664-675.

Wegman, E. (1991), "The grand tour in k-dimensions," ***Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface***, 127-136.

Wegman, E. (1995), "Huge data sets and the frontiers of computational feasibility," ***Journal of Computational and Graphical Statistics***, 4(4), 281-295.

Wegman E. and Carr, D. (1993), "Statistical graphics and visualization," in ***Handbook of Statistics 9: Computational Statistics***, (Rao, C. R., ed.), Amsterdam: North Holland, 857-958.

Wegman, E. and Luo, Q. (1997), "High dimensional clustering using parallel coordinates and the grand tour," ***Computing Science and Statistics***, 28, 352-360.

Wilhelm, A., Symanzik, J. and Wegman, E. (1999), "Visual clustering and classification: The Oronsay particle size data set revisited," ***Computational Statistics***, 14(1), 109-146.

INTENTIONALLY LEFT BLANK.

A Methodology for Quantifying Critical Decision Events During the Execution Phase of Battle

Dr. Jock Grynovicki
Mr. Kragg Kysor
Mr. Michael Golden

U.S. Army Research Laboratory
Human Research & Engineering Directorate

Introduction

The U.S. Army Research Laboratory (ARL) has undertaken a 5-year research program aimed at better understanding the distributed, non-linear decision-making process at the brigade level and above, as it is shaped by time, stress, team structure, staff experience, the environment and the introduction of digitization technology. Critical decision events were quantified using response data from key battle staff decision makers during the Crusader (howitzer) Concept Experimentation Program 3 (CEP 3) based on an ARL structured instrument called the "Decision Maker Self Report Profile (DMSRP)" (Golden, Grynovicki, & Kysor, 1999). The DMSRP is a data collection instrument designed to facilitate recording key data elements related to decisions made by commanders and staff officers during U.S. Army experiments and exercises. The DMSRP was not used during the planning phase but focused on the execution phase of battle.

The DMSRP was designed, in part, as the data collection complement to a cognitive engineering model of the decision-making process. This model's framework is based on a model known as the "Execution Decision Cycle," which was developed in early 1998 as a major component of a project titled "Cognitive Engineering of the Human-Computer interface for Army Battle Command Systems (ABCS)." The model incorporates recent theories of cognitive science and organizational psychology and presents a process depiction of how the commander engages in a variety of cognitive activities associated with executing combat operations. The "Execution Decision Cycle" model is described by Leedom, et al. (June 1998). The purpose of the present paper is to present a theoretically based approach to the analysis and classification of multivariate critical decision cognitive processes as recorded in the DMSRP instrument. Key findings from the multivariate analysis of the DMSRP application during the Crusader Concept Experimentation Program 3 (CEP 3) will be summarized.

Method

Procedures--Experiments to support the CEP 3 were conducted at Fort Hood, Texas, from 14 September to 16 October 1998. Some of these experiments consisted of a series of soldier-in-the-loop, interrelated simulation-supported studies designed to evaluate Crusader operational concepts. Participating battle staffs consisted of the Headquarters 3rd Brigade Combat Team (3 BCT), the battle staff and elements of a Field Artillery FA battalion. The CEP 3 supported battle staff training of mission-essential tasks including performance feedback for its after action

reviews (AARs). The training also helped participating units prepare for advanced collective training at the National Training Center (NTC). The participating battle staffs prepared operations orders (OPORDs) and the field artillery support plans (FASPs), based on approved Training and Doctrine Command (TRADOC) scenarios and division level orders.

In support of the CEP, the ARL assembled a data collection team from its Cognitive Engineering Research Program to systematically assess the battle command decision-making process during the execution phase of operations. Using a structured data collection instrument named the DMSRP, the ARL team observed and quantified critical elements of this process during the execution monitoring and adjustment of combat operations by the battalion command group. These observations were supplemented by in-depth interviews conducted with the battalion command group during daily "hot wash" AARs. The goals of this data collection were to identify (1) the types of information processes used by the command group, (2) the types of individual and collective mental activities and structures involved in the command group's "sense making" process, (3) the translation of the commander's intent and concept of operation into specific directives and battle adjustments, and (4) the methods that the commander used with his staff and battlefield operating system (BOS) to maintain proper focus of attention and awareness. In addition, an analysis of the insights from observations and the AARs was used to improve the DMSRP data collection instrument for future use.

Materials

Scenarios

Tactical scenarios, consisting of three separate phases were conducted three times each week for three consecutive weeks. The three phases consisted of (1) brigade movement to contact, (2) brigade defense in sector, and (3) division attack with the brigade offensive operations within the security zone. The three phases were conducted sequentially with one phase per day for a total of nine trials over the 3-week period.

Survey Questionnaire

The DMSRP was comprised of 15 major components and one section for providing additional comments in narrative form. Four of the major components had subsets which provided additional details for the data set. The major components, sub-sets, and the amplifying section appear in the DMSRP in the sequence shown in Table 1.

The purpose of each DMSRP component was to document the cognitive processes and associated explicit and tacit knowledge used by the decision maker at critical decision points in the mission. Additional goals were to collect critical decision data from decision makers located at different echelons (brigade, battalion, company) to develop insights regarding (1) the degree of mental model consistency or commonality among the key staff members and (2) the variability of the decision maker's information requirements over time, echelon, and battlefield situations.

Table 1

Outline of data item sequence in the DMSRP

Item	Description
1.	Location of the decision maker (TOC/Echelon)
2.	General critical decision event description
3.	Length of time for decision event
4.	Decision: Significant or Minor COA change
5.	Part or aspect of OPORD changed
6.	If significant COA change
	a. Principal Causes for COA change
7.	If minor COA change
	a. Principal Causes for COA change
8.	Process associated with decision making
	a. One immediately obvious response
	b. One option considered, mental simulation
	c. Multiple options considered, mental simulation
	d. Formal Option Generation by Staff (single/multiple)
	e. Manage the situation
9.	Triggering features or patterns in current situation
10.	Source of features or patterns
11.	Type of uncertainty experienced
12.	Uncertainty coping strategies
13.	Patterns of commander-staff interaction
14.	Decision maker's cognitive workload estimate
15.	Information processing activities (24 separate activities)

Analytical Methodology--A major cognitive decision-making program requirement was to develop a theoretical approach to the analysis and classification of the DMSRP-based multivariate critical decision-related cognitive processes. The data were collected during the execution phase of battle in a tactical operation center (TOC) under time pressure, uncertainty,

and complexity, in a constantly changing environment as depicted in Orasanu and Connolly (1993). The variables are subjective as are their units of measure and the values of these units. A key to our analysis approach was the use of multidimensional scaling to help visualize the cognitive processes associated with the human decision maker as a complex adaptive system. Psychologically, one can view the multidimensional space as a graphic depiction of the commander's mental model (Converse & Kahler, 1992). The vectors of the information element space reflect similarities and dissimilarities in the data. We analyzed the commanders' and battle staff members' multivariate critical decision event patterns and preferences using a weighted multidimensional scaling (WMDS) method and then confirmed these perceived patterns using discriminant analysis.

Weighted Multidimensional Scaling (WMDS)--In mapping the ARL cognitive model of command decision making, the DMSRP instrument attempts to chart the characteristics and concepts representing the cognitive processes associated with execution phase military decision making. The challenge of analyzing these complex processes was addressed through the use of multidimensional scaling techniques where objective scales of ranked order attributes, as reported by the subject decision makers, were constructed. These scales were subsequently mapped back to two-dimensional characteristics associated with the decision maker such as experience (i.e., "novice" or "expert"). The weighted Euclidean distance $D_{ijk} = [\sum w_{ka} (x_{ia} - x_{ja})^2]^{1/2}$, was used to account for individual differences in cognitive processes associated with critical event decision-making. The S-stress is used as a measure of fit S-stress = $[1/m \, \Sigma_k (\|E_k\| / \|T_k\|]^{1/2}$ in which $\|E\|$ is the sum of all squared elements of the error matrix defined as $\|E\| = \| T_k - D^2 \|$. S-stress can assume a value between 0 and 1. Smaller values represent a better quantification of the attributes.

Results

DMSRP data forms for 24 critical decision events were collected during the experiment. The following section provides a general overview of the analysis results for the data items comprising the DMSRP (see Table 1).

Decision Type, Time Window, and OPORD Changes

Regarding whether the decision was rated a significant or a minor change or adjustment of the current implemented course of action, 25% of the critical decisions were considered by the decision maker to be significant course of action (COA) changes while 75% were considered minor COA changes. The average time to make a decision was 39 minutes with a median time of 40 minutes for significant decisions and 12 minutes for minor decisions. The time range for significant and minor decisions did overlap. Most of the changes in the COA (54.5%) relative to the current operation order (OPORD) were related to "Friendly Scheme of Maneuvers" and "Fire Support." Overall, because there were two decision types (significant vs. minor COA change) and only a small sample size of data, no clear pattern in the type of change (significant vs. minor) could be identified.

Principal Causes for Adjustment in COA

This section of the DMSRP framed the decision by identifying the principal causes for the change or adjustment in the current implemented COA. It allowed the analyst to identify which elements of information available to the commander were critical in shaping his decision. We asked the decision maker to select the three principal causes for the decision to change the current COA and then rank order them with respect to each other. Using WMDS and discriminant analysis, we found significant differences between the expert decision makers and the novice decision makers, in the causes for the decision to change or adjust the course of action. The differences in the information elements' vector space reflected which items were considered important by each group. Specific clusters of information elements can be seen to represent concepts or information categories closely associated with the individual's experience level. Figure 1 depicts significant differences in the strategy and concepts used by the experts versus the novices (S-stress = 0.17, RSq = 0.92, Wilks' Lambda=0.71, sig.=0.032). The experienced decision makers (i.e., commanders) focused first on the change in the enemy's force projections, second on the blue force projections, and finally, on the tactics being implemented. The novice decision makers (i.e., commanders) focused on the blue force projection, the perception of the future plan, and the change in the reading or interpretation of the cues or patterns used for situational recognition.

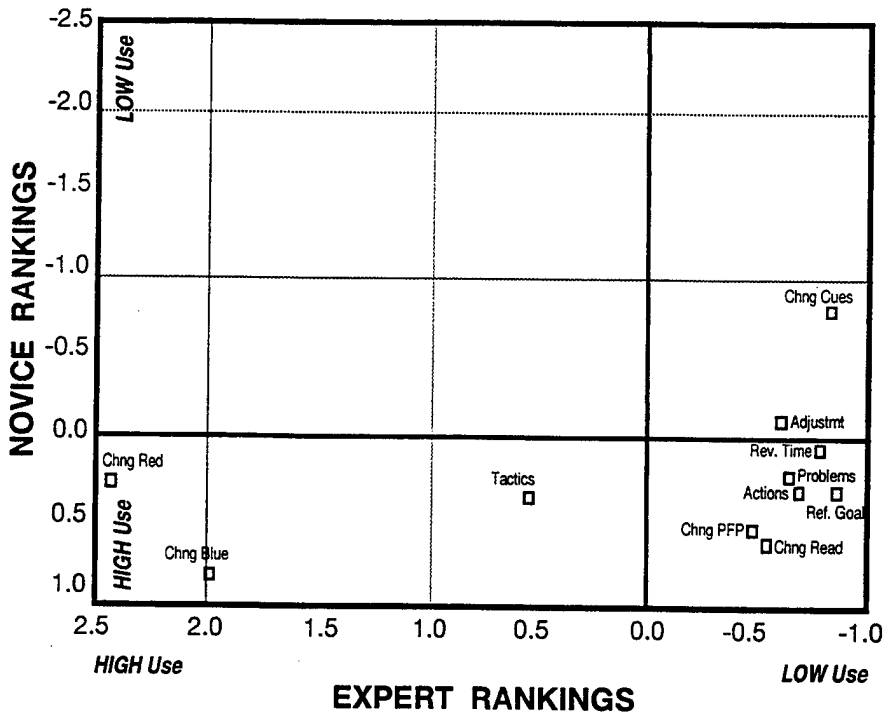


Figure 1. Principal causes for adjustments in COAs.

Process Associated With Decision Making

About half the decision-makers (52%), felt that the critical event situation suggested "one immediately obvious response, course of action change or adjustment." Forty-three percent of the respondents stated that "one response, course of action, or adjustment was not immediately obvious." Of these ten decision cases, 70% felt that multiple options were considered and evaluated sequentially using explanatory reasoning and storytelling based on a group oriented assessment. Finally, in only one case, the staff decided to manage the situation because of uncertainty, considered a formal option by directing the staff to generate new options, or used explanatory reasoning to consider one option.

Triggering Features or Patterns in the Current Situation and Principal Sources for Monitoring

The triggering features or patterns used by the commander and his staff to trigger one immediately obvious response varied by individual, echelon, and situation, with no quantifiable pattern. Triggers were cited such as weapon effectiveness, disposition of enemy and friendly forces, terrain, enemy tempo, and combat support. WDMS and discriminant analysis indicated that no significant difference was identified between the expert decision makers and novice decision makers regarding a difference in the principal sources of information (features or patterns) used to comprehend the current situation and trigger the decision event (S-stress=0.36, RSq=0.46, Wilks' Lambda=0.75, sig.=0.51). The expert decision makers primarily used reconnaissance and paper maps to comprehend the current situation. They also relied on the tactical net and information from the modular semi-automated forces simulation computer (MODSAF) to monitor the current situation. Other strategies were also used. The novice decision makers used numerous sources of information to help see the feature cues, indicators, or patterns in the current situation as seen in Figure 2.

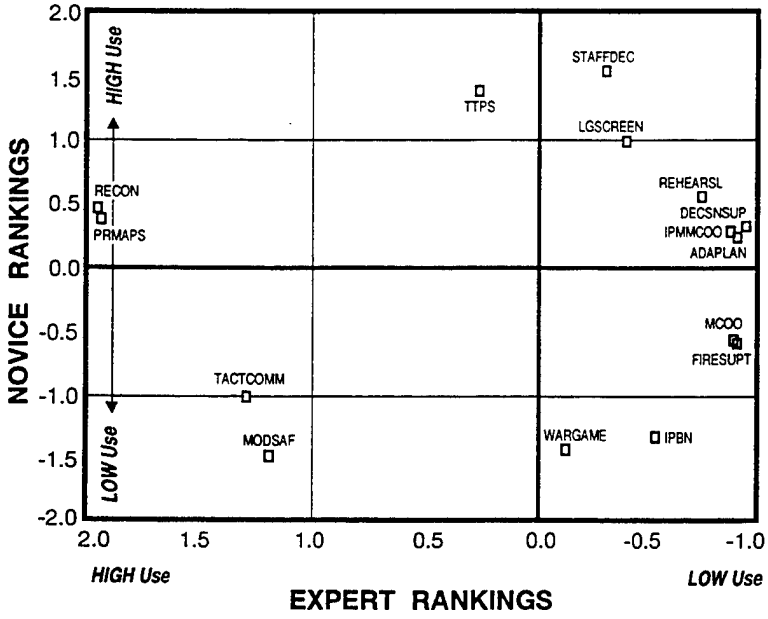


Figure 2. Multidimensional scaling of sources used by expert and novice decision makers to comprehend the current situation.

Types of Uncertainty Experienced and Coping Strategies Used

The majority (72%) of the staff experienced uncertainty because of incomplete information about the situation. Another 20% claimed to have an incomplete understanding of the situation even with complete information. Undifferentiated alternatives and confusion because of many meanings or interpretations accounted for the remaining reasons for uncertainty. To cope with uncertainty, the staff collected more information to reduce uncertainty (32%), made assumptions to deal with uncertainty (26%), weighted pros and cons (18%), formed understanding using plausible reasoning (11%), and forestalled (10%). No trend differences regarding the types of uncertainty experienced and coping strategies used could be determined between novice and experienced decision makers.

Patterns of Commander-Staff Interaction

The DSMRP results indicate that for this sample, the decision maker did not make decisions in isolation. Instead, the decision maker and his staff members performed as a well-formed team throughout the entire decision event a little more than a third (39%) of the time. Twenty-six percent of the time, the decision maker first set the general decision framework and then allowed the staff to complete the details. In the remaining cases (21%), the staff was hierarchically directed by the decision maker to provide specific input, and then the decision maker integrated his information to make the final decision.

Cognitive Workload Estimate

"Mental Demand" versus "Effort" Ratings. Most of the decision events were rated by the decision makers as "low" or "moderate" for mental demand (see Table 2). One interesting trend

Table 2

Mental Demand as a Function of
Military Experience

Novice=CPT-MAJ (18) Expert=LTC-COL (6)

	Experience	
Mental Demand	Novice	Expert
Very Low	5%	0%
Low	40%	0%
Moderate	30%	25%
High	25%	50%
Very High	0%	25%

was that the expert decision makers (i.e., colonels) rated the mental demand for the decisions they made as being "moderate," "high," or "very high." On the other hand, the novice decision makers (i.e., captains-majors) rated their mental demand at lower scale levels. However, while the expert commanders felt that their decisions were more mentally demanding, they regarded the process as involving less effort than did the novice commanders (see Table 3).

Table 3

Effort as a Function of Military Experience

Novice=CPT-MAJ (18) Expert=LTC-COL (6)

	Experience	
Effort	Novice	Expert
Very Low	10%	25%
Low	25%	25%
Moderate	35%	50%
High	25%	0%
Very High	5%	0%

Information Processing Activities

The utility of WMDS coupled with discriminant analysis was effective in quantifying the information processing activities conducted by the expert versus novice decision makers. WMDS indicated that a significant difference existed between these two groups in the type of information processing activities used (S-stress=0.15, RSq=0.96, Wilks' Lambda=0.64, sig.=0.029). The more expert decision makers (i.e., commanders) monitored specific objects or events that served as a qualitative indicator of a broader activity or trend within the battle space. They developed mission goals and priorities to achieve the desired battlefield end state. They interacted with their staff to reaffirm their decisions. In contrast, battlefield visualization, rule heuristics, and monitoring were the primary information processing activities used by the novice decision makers. As shown in Figure 3, the less experienced commanders focused on clarifying or prioritizing operational objectives, critical cue tracking, and relying on battle staff group assessment to monitor and collect information for assessment and decision making. They mentally applied doctrinal rules to interpret and evaluate a choice or option.

Summary

The DMSRP instrument along with the WMDS and discriminant analysis, has proved to be a useful tool in providing structured, diagnostic insights into the complex military decision making process.

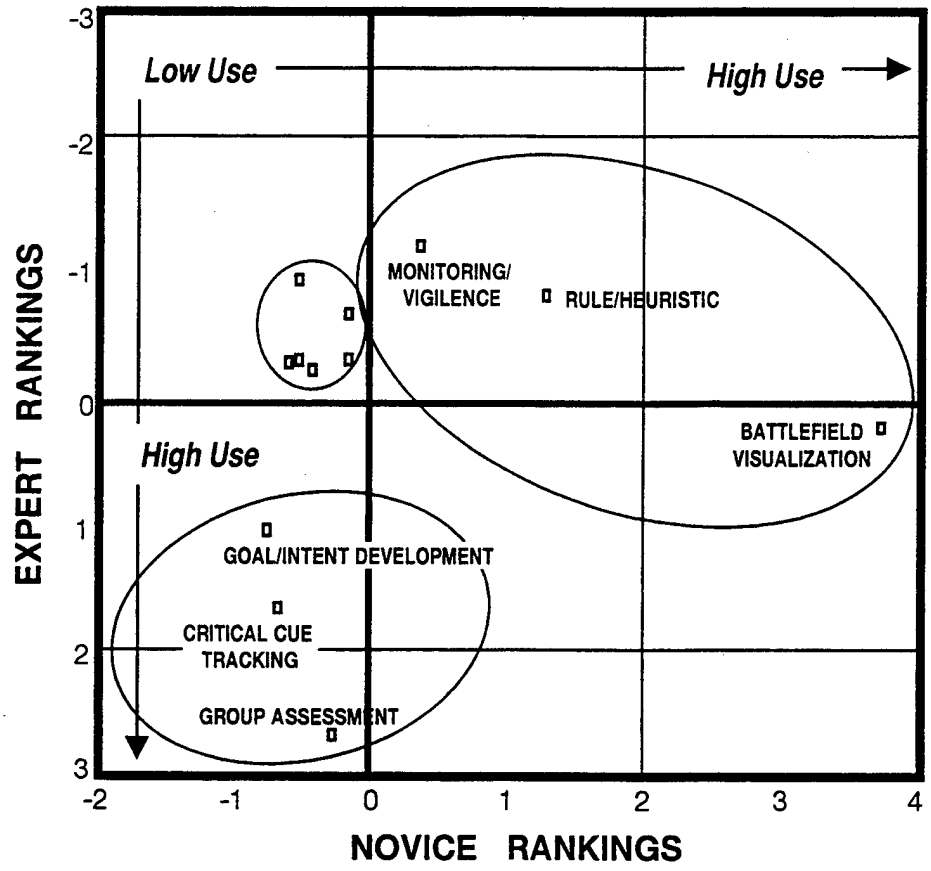


Figure 3. Information processing (IP) activities as a function of expert versus novice decision-maker rankings analyzed by a multidimensional scaling method.

References

(1) Converse, S.A. & Kahler, S.E. (1992). *Knowledge acquisition and the measurement of shared mental models.* (Contract #DAAL03-86-D-0001). Orlando, FL: Naval Air Warfare Center Training Systems Division.

(2) Golden, M., Cook, T., Grynovicki, J., & Kysor, K. (in press). *Decision Maker Self Report Profile.* ARL Technical Report.

(3) Leedom, D., Grynovicki, J., Pierce, L., Golden, M., Murphy, J., & Adelman, L. (1998). *Insights from Battle Command Reengineering II, CEP #1701: Cognitive Engineering of the Digital Battlefield*

(4) Orasanu, J. & Connolly, T. (1993). The reinvention of decision making. In G.A. Klein, J. Orasanu, R. Calderwood & C.E. Zsambok (Eds.), *Decision making in action: Models and methods,* 3-20. Norwood, NJ: ABLEX.

INTENTIONALLY LEFT BLANK.

FLEXIBILITY OF FRACTIONAL FACTORIALS IN FIELD TESTS

Carl T. Russell
Ballistic Missile Defense Organization
Joint National Test Facility
Schriever AFB, Colorado 80912

ABSTRACT

Fractional factorial design has sometimes been considered for use in field tests, but it has seldom been implemented. Typically, fractional factorials have been regarded as inflexible and fragile. Actually, traditional 2^{n-k} fractional factorial experimental design can produce implementable and robust field test designs that efficiently address multiple goals. This paper updates a paper presented in 1983[1] to reflect subsequent changes in technology, resource availability, and acquisition philosophy.

INTRODUCTION

Efficient experimental design offers several advantages in military field tests. The most frequently touted advantage of efficient design is resource reduction, which can be substantial. This advantage can be even greater, however, if it is viewed as exploiting available resources more efficiently. Experimental design technology developed over the past eighty years can be used to spread fixed resources efficiently over a wide variety of test conditions subject to complicated constraints. Moreover, modern computer hardware and software can now be used not only to visualize, analyze, and effectively portray the results of complicated experimental designs but also to assist in their development.

The designer of a military field test is typically presented with a very large number of factors possibly affecting test results and a continually changing test environment. The usual approach to field test design specifies that some factors be controlled, some be varied tactically, and the rest be uncontrolled (which frequently means ignored). Generally, the design specifies some minimally acceptable sample size for certain combinations of controlled conditions and gives quantitative guidance in terms of relative proportions of events for tactically varied factors. The test operator, however, must perform detailed test planning one trial at a time. In most instances, detailed test planning does not allow much tactical free play, but instead the *tactically varied factors are controlled by the test operator* within each trial. Moreover, the uncontrolled conditions usually end up being relatively constant throughout each trial, either through expedience of test conduct or because truly uncontrollable test factors tend to remain fairly constant over small regions of time or space. Thus the number of test factors actually controlled is frequently much larger than the number specified for control, and truly uncontrolled factors tend to vary from trial to trial rather than uniformly over the test.

This paper illustrates how sequences of fractional factorials can be used to manipulate a large number of test factors within a test framework that explicitly recognizes operational constraints and exploits the fact that truly uncontrollable test conditions tend to vary only from trial to trial. The author proposed such techniques in actual tests almost twenty years ago,[2,3] but he has only used them in a relatively simple case exploiting a sequence of Graeco-Latin Squares (a special fractional factorial). Floyd Hill's use of Graeco-Latin Squares in the Project STALK operational test (1953)[4] is another rare example. Three reasons that fractional factorials have not seen greater use in field testing are:

- Sequences of fractional factorials routinely require substantial changes in test conditions between trials. Unless those changes are carefully thought out, test conduct becomes extremely difficult and inefficient.
- Fractional factorials appear on the surface to be fragile—their basically sparse nature suggests that they would fall apart if observations are lost or corrupted. Even for designs in a small number of factors, this perception of fragility turns out to be incorrect.[5,6] For fractional factorial designs in a large number of factors the following argument suggests that the perceived fragility is overblown. One seldom looks at more than two or three factors together when analyzing data. Except for deliberate confounding, the nature of designs based on carefully constructed sequences of fractional factorials ensures that observations in any two or three factors are balanced in the remaining factors. Tables comparing two or three factors summarize a large number of observations in any cell, and those observations are generally balanced in the other factors. As more factors are formally considered in a fractional factorial, loss or corruption in relatively few observations becomes less likely to distort comparisons between any two or three factors.

- Demands for substantial free-play in field testing makes highly structured statistical design a very hard sell. With the current trend to make more use of Field Training Exercises (FTXs), the time may be right to put much more formal structure into any combined operational/developmental testing conducted.

The body of this paper consists of two examples. The first simple example illustrates the benefits obtainable from the fundamental ideas of experimental design. The second and primary example shows how those ideas can be extended in a systematic manner to field test design. Although this second example is constructed rather than actual, the constraints within which the example was developed are representative of those occurring in military field tests. The detailed guidance provided in the second example shows that a mathematically challenging experimental design can be specified in a practical manner that relieves the test operator of many technical scheduling problems without restricting the application of relevant military judgment or unduly reducing operational realism.

A SIMPLE ILLUSTRATIVE EXAMPLE OF EFFICIENT EXPERIMENTAL DESIGN

Consider the following situation. The test operator has four objects (A, B, C, D), each weighing 10 to 15 pounds (true but unknown weights of a, b, c, and d, respectively). The experimental goal is to estimate the weight of each object as accurately as possible. Four scales are available for weighings, each inaccurate by some constant amount over the range from 10 to 60 pounds. No more than four weighings can be made on any one scale. Thus any combination of the four objects can be weighed on each scale, but the weights obtained will vary by constant amounts depending on the scales used.

The simplest experimental approach would be to weigh each object on each scale (16 weighings total) and average the results. Table 1 summarizes the calculations and shows that all estimates are biased by an amount equal to the average error of scales. This bias cannot be removed.

TABLE 1. Estimation of Weights Using Simplest Approach

		Estimated Weights				
Object	True Weight	Scale 1	Scale 2	Scale 3	Scale 4	Average*
A	a	$a + s_1$	$a + s_2$	$a + s_3$	$a + s_4$	$a + \bar{s}$
B	b	$b + s_1$	$b + s_2$	$b + s_3$	$b + s_4$	$b + \bar{s}$
C	c	$c + s_1$	$c + s_2$	$c + s_3$	$c + s_4$	$c + \bar{s}$
D	d	$d + s_1$	$d + s_2$	$d + s_3$	$d + s_4$	$d + \bar{s}$

* Average denotes the arithmetic mean. In particular, $\bar{s} = (s_1 + s_2 + s_3 + s_4)/4$.

A more efficient experimental approach makes only two weighings per scale (8 weighings total) but obtains better estimates. Not only are a, b, c, and d estimated without scale bias, but each of the four scale errors is also estimated. The trick is to weigh combinations of objects on each scale. Table 2 summarizes the weighing scheme and the appropriate computations.

TABLE 2. Estimation of Weights Using Efficient Approach

Part 1: Measurements

Weighing	Description	Scale 1	Scale 2	Scale 3	Scale 4
First	Objects weighed	A	B	C	D
	True weight	a	b	c	d
	Estimated weight	$a+s_1$	$b+s_2$	$c+s_3$	$d+s_4$
Second	Objects weighed	B,C,D	A,C,D	A,B,D	A,B,C
	True weight	b+c+d	a+c+d	a+b+d	a+b+c
	Estimated weight	$b+c+d+s_1$	$a+c+d+s_2$	$a+b+d+s_3$	$a+b+c+s_4$
Combined	Difference	X1=a-b-c-d	X2=b-a-c-d	X3=c-a-b-d	X4=d-a-b-c
	Sum	$Y1=a+b+c+d+2s_1$	$Y1=a+b+c+d+2s_2$	$Y3=a+b+c+d+2s_3$	$Y4=a+b+c+d+2s_4$

Part 2: Final Estimates

a = (X1-X2-X3-X4)/4	s_1 = (Y1-T)/2
b = (X2-X1-X3-X4)/4	s_2 = (Y2-T)/2
c = (X3-X1-X2-X4)/4	s_3 = (Y3-T)/2
d = (X4-X1-X2-X3)/4	s_4 = (Y4-T)/2
T = a+b+c+d = -(X1+X2+X3+X4)/2	

The efficient procedure just discussed gets more useful results with 8 weighings than the simple procedure does with 16—the only catch is that there are no "degrees of freedom for error" left to permit standard statistical inference. If 16 weighings were really available, eight more weighings could be conducted to provide at least error estimates. This efficiency has another side. Within fixed resources, efficient design allows explicit consideration of more factors and conditions than a simplistic design. In this illustrative example, if resources had been available for only 8 weighings, the simple approach would force a choice between weighing only 2 objects (reducing scope) or using only two scales (reducing generality). Or two weighings per scale might be the maximum number possible or feasible. Military field tests are typically subject to both constraints: resources are limited, and only a limited number of observations (analogous to weighings) can be obtained in any single trial (analogous to scale). By carefully specifying combinations of factors to be tested together during any trial and cleverly stringing them together, efficient experimental design can be used to produce increases in scope (the number of factors examined) and generality (the number of conditions under which selected combinations of factors are tested). Since describing the effect of a wide variety of test conditions on total system performance should usually be the primary goal of a field test, efficient experimental design can contribute substantially to the attainment of that goal.

Fortunately, the illustrative example just given generalizes to tests of other sizes. In particular, with 64 scales and 8 weighings per scale, it is possible to estimate the weights of $64 \times 7 = 448$ objects as well as to estimate the bias for each of the 64 scales—thereby estimating 512 parameters with 512 observations. In practice, such a theoretical limit would not typically be attainable and would generally be undesirable if attained since some "degrees of freedom for error" are generally desirable in experimentation to permit statistical inference. In the next section, however, an example is constructed in which 8 performance observations (weighings) for a hypothetical tactical jammer are taken during each of 64 test periods (scales). The effects of 41 factors potentially influencing performance are estimated as well as 8 differences between trials and the mean (a few other parameters are also considered). This example illustrates both the flexibility inherent in fractional factorial design and the ability of fractional factorial design to provide a rich background for test execution.

AN EXAMPLE OF EFFICIENT EXPERIMENTAL DESIGN IN A HYPOTHETICAL FIELD TEST

A. UNDERLYING FRAMEWORK AND CONSTRAINTS

Suppose that a *hypothetical* ground-based tactical voice communications jammer capable of monitoring two frequencies and jamming one at a time is ready for operational/developmental test. This jammer is mounted on a vehicle, and the mission profile/operational mode summary calls for the jammer to operate against two tasked frequencies for a specified period of time from one position then to move a specified distance over certain types of terrain to a new position. Mission length is to vary between 30 and 60 minutes depending on the intensity of combat, with shorter missions corresponding to more intense combat. Roughly 50 percent of the missions are to be between 30 and 40 minutes long, 30 percent between 40 and 50 minutes long, and 20 percent between 50 and 60 minutes long. The jammer has software that supposedly prioritizes the two tasked frequencies and is to be tested against tactically representative voice communications over a variety of tactically deployed nets using various typical threat radios. Overall jammer performance is to be evaluated, and the effects of numerous factors potentially affecting jammer performance are to be quantified. In particular, effects of jammer to receiver distance, transmitter to receiver distance (link distance), threat radio power, and density of transmissions on each net are to be quantified in some operational sense. In addition, effects of net type (command, logistic, fire support, etc.) and echelon as well as jammer distance to the Forward Edge of Battle Area (FEBA—the notional boundary between friendly and hostile forces) are desired. Several possible synergistic effects are also of interest. Of special interest is possibly varying effectiveness of the prioritization software as the relative density of transmissions on the two tasked frequencies varies. Survivability of the jammer is to be addressed during testing in terms of the speed and accuracy with which jammer positions can be located. Logistic supportability is to be addressed throughout testing, hopefully in terms of the varying stresses applied to the jammer under the scenarios tested. Four jammers will be available for test.

B. USUAL APPROACH

At this stage, the test designer usually identifies factors and conditions for test, and an analyst is typically asked to divine the sample size necessary for "statistical significance." This determines the test length and the overall numbers of observations to be obtained. The test designer then writes guidance for the test operator in terms of general scenarios and matrices indicating the number of iterations to be conducted under each combination of test conditions. Finally, the test operator lays out and conducts the test in as efficient and operationally realistic a manner possible. Unfortunately, the testing problem described here is really too complicated for accurate sample sizing. Moreover, the illustrative example given earlier shows that method of test conduct can be more important than

sample size. Using efficient experimental design to indicate effective test methods, the analyst can clarify how much can be accomplished with any particular sample size. The following discussion indicates how this might occur.

C. DIRECT CONSEQUENCES OF TEST CONSTRAINTS

Within the underlying framework and constraints for the hypothetical jammer test, many options are available, both for test design and for test conduct. However, certain test features follow almost inevitably from the constraints described above.

The test would certainly be conducted as a series of trials with several trials per day. Since there are to be four jammers, each tasked against two frequencies, there is no reason to design a test involving more than eight nets. Although tasking more than one jammer against one net, examining nets consisting of more than one link each or using less than four jammers on some jamming trials are possible options, this example assumes that each jamming trial consists of all four jammers operating against 8 distinct nets consisting of one link each (each link/net would use a separate frequency throughout each trial). Then on any given trial, 16 *individuals*[†] must be positioned in 16 locations such that communications are possible between 8 pairs of locations (links) which exhaust the 16 locations. In any trial the two individuals on any link constitute a team and would have to use compatible radios. Although it might be possible to move radios and individuals to different positions between trials on the same day, such movement would make test conduct more difficult and less efficient (due to travel time and necessary rechecking for correct communications and configurations). Likewise, reconfiguring the links between trials would be practically infeasible, so teams would almost certainly remain on the same links throughout each day. It is reasonable to assume, however, that teams could use different frequencies on different trials during the same day and that teams and radios could be rotated between links from day to day. Links could be reconfigured from day to day, but it is difficult to lay out multiple links over which communication is actually possible, so it is desirable to reconfigure links only infrequently. Thus it is assumed that on each day a team consisting of two individuals with radios of one type are assigned to each link, that the individuals and radios are rotated between links from day to day, and that links are reconfigured only every week or two (if at all). Frequencies are rotated between links from trial to trial.

The situation involving rotation of jammers is somewhat different. For any trial, four jammer positions are necessary, and since each jammer must move between trials, each could move to a different position between trials. Since jammer survivability is an issue, surrogate threat direction finding would be necessary on some trials. If jammers merely rotated among four positions on a given day, then only one trial per day would be usable for survivability assessment. Thus it would be desirable to have sufficient number of jamming positions so that each could be used just once per day. Then survivability could be addressed, say, once a week for a whole day, and the jammers could rotate through the jamming positions in such a way that each position is used once per day and each jammer used each position once during the week.[‡] A new set of jammer positions would then be necessary only once per week.

In reality, a very large variety of link lengths and types is likely to be present on the battlefield. Since only eight will be tested at a time, however, selection must be made carefully. For the array of threat links to be realistic, it should contain links of different lengths at different distances from the FEBA. These distances represent communications at different echelon. If links are classified by echelon (say battalion to company (BC) and company to platoon (CP) and by length (short and long) then there are four possibilities: BC-short, BC-long, CP-short, and CP-long. If two links of each type are portrayed on each trial then the desired eight links would be present on each trial. Although other choices for classification of eight links are possible, the classification described is a likely choice. Similarly, a likely classification for jammer position is that a jammer can be located relatively close to the FEBA or relatively far from the FEBA, and with four jammers, it is reasonable to assume that on any trial, two jammers could be located relatively close to the FEBA and two located relatively far from the FEBA.

The actual content and density of messages presents a different problem. For each link and each trial a number of messages must be generated and given to teams for transmission. These messages must be scheduled so that they are sent during periods when jammers are actually operating. On any link during any trial, it would be desirable to send messages in each direction, preferably the same number of messages in each direction. During any one trial, it is necessary to portray different densities (i.e., different numbers of messages) on different links and to task jammers

---

[†] Throughout this paper, "*individuals*" refers to persons or automated programmable message generators. Since this paper describes a relatively short test with tactical voice (vice data) messages subject to degradation by jamming, it is doubtful that replacing persons with automated devices would be practical, but such an option should be considered in any actual test.

[‡] As developed later, "weeks" are four-day weeks; half of the first "week" is conducted, then the second "week" is conducted, finally the other half of the first "week" is conducted.

in such a way that some jammers are tested at high priority against denser links and some are tasked at low priority against denser links. Moreover, messages cannot be reused by the same teams. If messages were of 30 second average length, then an average of 12 messages per link (six minutes of transmission time per link) is a reasonable number of messages, but this requires 8×12=96 messages to be generated per trial; if 8 trials were conducted per day over, say, an 8-day test, then a total of 6,144 messages would be required. Since an obvious measure of jamming effectiveness is the percent of essential message elements (EMEs) received, analysis of each message is necessary to identify its EMEs and the number correctly received. The number of messages required for a test like this one, combined with processing for each message, suggests that some compromises with operational realism be made to permit reuse of messages. A specific version of such a compromise is discussed in the next subsection.

This subsection has indicated that the outlines of testing follow more or less directly from specified test constraints and the realities of test conduct. Each day 8 teams with specified radios are positioned on 8 links, and a number of trials are conducted with each of the 4 jammers tasked against 2 links on each trial. Frequency assignments rotate among links between trials and jammers change positions between trials so that no jamming position is used twice on the same day. Teams rotate among links from day to day and jammer positions are reused for several days. The threat array is laid out with links classified simply by length and echelon, and the threat array is changed rarely if at all. Some compromise with operational realism is made so that messages can be reused.

The next subsection indicates further more arbitrary choices dictating the method of test conduct.

D. FURTHER CHOICES AFFECTING TEST CONDUCT

Until now, little has been said about the number of trials per day. The specified average trial length is 40-45 minutes. Allowing an hour or so between trials for the jammers to conduct mobility exercise as well as time for administrative matters indicates that four jamming trials per day is reasonable and feasible. However, this choice leaves considerable dead time between trials for the threat array. In order to provide a baseline for jamming trials, it might be desirable (during a least a portion of the test) to exercise the threat array during these otherwise dead times by transmitting messages similar to those in jamming trials but with no jammers present. In what follows, such an approach is assumed, so each day consists of eight trials, four with jamming and four unjammed.

Reuse of messages was left open at the end of the last subsection. Such reuse not only reduces the number of distinct messages required (making test conduct and data reduction easier) but also improves overall test precision by reducing variation (a trick often used in constructive simulation). One way to solve the problem of message reuse is to prepare eight sets of messages each day, two of 6 messages each, four of 12 messages each, and two of 18 messages each. Each team member would send half the messages assigned to the team during any trial. Message sets would be rotated among teams from trial to trial in such a way that if a team transmitted an n-message set during a jamming trial, the same team would transmit another n-message set during the corresponding unjammed trial. On each jamming trial, jammers would be tasked so that one jammer was tasked against two 12-message sets, one jammer was tasked against one 12-message set and one 18-message set, one jammer was tasked against one 6-message set and one 12-message set, and one jammer was tasked against one 6-message set and one 18-message set. Each jammer would rotate through all four taskings during any day, and each message set would receive first priority twice and second priority twice during the jamming trials on any day. In other words, during any trial relative message densities in the ratios 1:1, 3:2, 2:1, and 3:1 would appear. Each jammer would see each ratio during each day, each ratio would appear twice with each priority assignment during each day, and each jammer would see each priority assignment at each ratio the same number of times during the test.

With the same number of messages transmitted during each trial, the only way to control combat intensity is to vary trial length. Fortunately, the correspondence of shorter trials to more intense combat is specified in the mission profile. One possible specification of trial lengths is one 30-minute trial, two 40-minute trials, and one 60-minute trial per day, which gives a distribution of trial lengths very similar to that specified in the mission profile. To conform to realistic scenarios, two possible arrangements of trial lengths are specified for jamming trials: 40-30-40-60 and 60-40-30-40. Each arrangement starts and ends with trials of relatively high intensity with lower intensity trials in between. Possible effect of trial ordering should be considered, and it would probably be worthwhile deliberately to vary starting times from day to day, starting trials on some days in early morning (say 0600) and starting trials on other days in early afternoon (say 1200).

In a truly operational setting, messages transmitted over any one net would be of similar character. That is, messages on command nets are different from messages on logistic nets. These in turn are different from messages on fire support nets, etc. In real operations, nets would be identified by the character of their messages, and tasking would be passed to the jammers based on assessment of each net's criticality at a given stage of battle. For a test of a

jammer divorced from a signal collection system, however, the realism of message traffic during short periods of time is much more important than the realism of message flow. Thus messages of several different types could be passed over each link during each test period. In particular, a convenient approach is to identify three types of messages on any day and generate all 8 message sets for that day in such a way that one third of the messages are of each type. If this were done—in particular if messages were scheduled so that each member of each team transmitted half the messages of each type during each trial—ultimate analysis could consider possibly varying effectiveness of jamming against various message types in addition to the other factors considered.

Possible distribution of threat radio types has not yet been mentioned. It is assumed that four relatively lower powered radios (each of the same type) as well as two each of two types of relatively high powered radios will be used during each trial. It is further assumed that each team will have two radios of the same type throughout each day but that radios will be rotated among links and teams in such a way that each radio type is used by each team the same number of times and each radio type is used on each link the same number of times.

During the course of the test, each jammer should be subjected to roughly the same stress as every other jammer. The design scheme sketched so far spreads tasking evenly over jammers. For a logistic supportability evaluation, however, it may be desirable to vary the stress from mobility exercises so that any large variations in logistic demand generated by varying test conditions can be assessed. One way to do this is to classify mobility runs according to roughness (relatively smooth or relatively rough) and length (relatively short or relatively long), and always run mobility trials before jamming trials. There are four combinations of conditions. If on each day each jammer ran mobility exercises under only one combination of conditions but if jammers rotated among conditions from day to day, then any large and consistent variation in logistic demand as stress varied might show up either in reliability failures or in degraded jammer performance (provided mobility exercises were always run before jamming trials).

The final consideration in the detailed design is the assumption that a typical array laid out on the ground is similar to that in Figure 1. If the array is laid out in such a way that it could be jammed from either side (sides I and II in Figure 1), then switching jammer positions from one side of the array to the other essentially creates a new array. The number of times a single array could be reused is effectively doubled if this approach is taken. Unfortunately, some possibly unacceptable degradation in operational realism occurs in this approach. The simple classification of links by length and echelon, if portrayed in a truly operationally realistic way, would tend to display longer links at higher echelons so that link length would be relative to echelon. Switching the array side from which jamming is done also switches echelon for each link. If long links at high echelon were longer than long links at low echelon when viewed from side I, long links at low echelon would be longer than long links at high echelon when viewed from side II. Thus if jamming from each side of the array were permitted, link length would have to be absolute rather than relative. This example assumes that the increased efficiency of test conduct obtained by switching sides is judged to outweigh the resulting degradation in operational realism.[†]

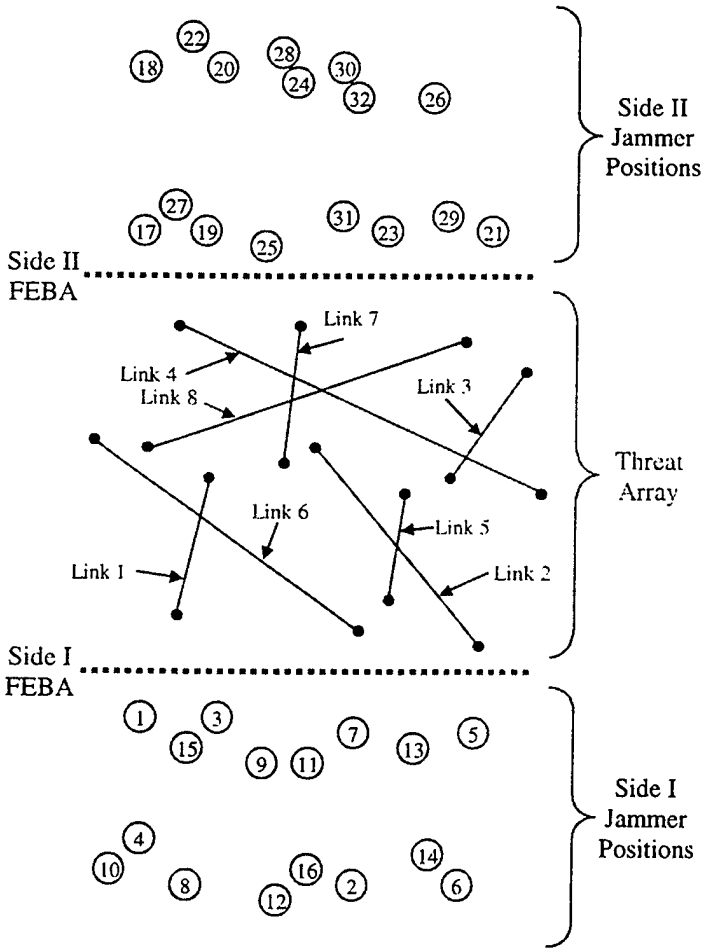


FIGURE 1. Notional Threat Array with Links and Jammer Postions Labelled

The design decisions and compromises discussed in this subsection were more arbitrary

[†] Types and density of messages would also vary operationally by echelon, an additional complication.

than those in subsection C. The decisions and compromises in both subsections, however, represent solutions to typical problems in field test design and conduct that must be solved either explicitly or tacitly in actual field tests. In fact, the problems in actual tests tend to be both more numerous and more complicated than those discussed here. As described in both subsections, the problems have been formulated and solved more analytically than is typical, because it has been assumed that an analyst was deeply involved in detailed planning and that the analyst's suggestions have been conscientiously considered when decisions and compromises were reached. Underlying the analyst's participation is the implicit promise that the analyst will produce unambiguous guidance which specifies a feasible way to perform all the various rotations of teams, jammers, radios, and scenario types outlined previously. ***The primary point of this paper is that solid analytical guidance can be given and that it can be developed systematically using the technology of efficient experimental design.*** When this is done, the analyst can explicitly describe any biases inherent in the test procedure and indicate what can be gained by further modification of test procedure or by conducting a longer test. The detailed guidance necessary for conducting the test as outlined is given in subsection E for an 8-day period (one complete rotation for a fixed threat array). Subsection G gives a systematic description of that guidance and indicates how guidance for any subsequent 8-day periods could be produced. If the decisions and compromises discussed in subsections C and D had been different, the guidance and technical details described in subsections E and G would also be different, but most alternate decisions and compromises could also have been incorporated into the design along the lines described.

E. DETAILED GUIDANCE FOR AN EIGHT DAY TEST

In order to provide guidance for an 8-day test in accordance with the outline developed in subsections C and D, three types of specifications are necessary:

- For each trial, start time and length must be specified.
- For each team on each trial, the following must be specified: the location of each team member, the radios used, and the message set used (with scheduling, half the messages given to each team member).
- For each jammer on each jamming trial, the jammer position and prioritized frequencies for jamming must be specified, as well as the type and length of course to be traversed before or after each trial.

Table 3 gives start time and length for eight days of trials conducted at the rate of eight per day (four jamming trials and four unjammed trials per day). Table 4 specifies location, radio type, message set, and frequency used by each team during each trial in Table 3. Table 5 specifies frequency used by each team during each trial in Table 3. Table 5 specifies position and prioritized frequencies used by each jammer during each jamming trial in Table 3 as well as the type and length of each course to be traversed before each jamming trial.

F. TEST REALISM

Provided the test operator lays out the threat array and jamming positions in an operationally realistic manner and provided the messages used are representative of threat communications, the jamming events portrayed in this test are operationally realistic even though they would never occur this way on the battlefield.

- Each message transmitted is presented in an operationally realistic manner to the jammer tasked against it.
- The variety of message presentations as a whole is representative of what might occur in combat.
- Each jammer is subjected to realistic combat stress.
- Jammer activity during any trial is realistic for survivability.

Such a test could do a very good job of characterizing operational jammer performance against representative threat messages. FTXs or live/virtual modeling and simulation must address bigger-picture issues such as tasking against multi-link nets, changes in tasking during jamming periods, and overall impact of jamming on threat effectiveness.

G. CONSTRUCTION OF THE DESIGN

This is a technical subsection which indicates how the fundamental concepts and notation of fractional factorial design were used to manipulate the many factors discussed in subsections B through E. The notation and theory is standard and will not be discussed here. Both John[7] and Montgomery[8] discuss the notation and theory.

The nature of the constraints in this hypothetical test dictated that it be laid out for the most part in powers of 2. This is the easiest possible case with which to work, and some constraints were forced into a power-of-two

TABLE 3. Start Time and Length (Minutes) for Each Trial

	Jamming Trials				Unjammed Trials		
Day	Trial Number	Mobility Start	Trial Start	Trial Length	Trial Number	Trial Start	Trial Length
1	1J1	1045	1200	60	1U1	1315	60
	1J2	1315	1430	40	1U2	1530	40
	1J3	1530	1645	30	1U3	1730	30
	1J4	1730	1845	40	1U4	1945	40
2‡	2J1	0600	0715	60	7U1	0600	60
	2J2	0830	0945	40	7U2	0830	40
	2J3	1045	1200	30	7U3	1045	30
	2J4	1245	1400	40	7U4	1245	40
3†	3J1	0445	0600	40	2U1	0700	40
	3J2	0700	0815	30	2U2	0900	30
	3J3	0900	1015	40	2U3	1115	40
	3J4	1115	1230	60	2U4	1345	60
4†‡	4J1	1145	1300	40	8U1	1200	40
	4J2	1400	1515	30	8U2	1400	30
	4J3	1600	1715	40	8U3	1600	40
	4J4	1815	1930	60	8U4	1815	60
5†	1J1	1045	1200	60	1U1	1315	60
	1J2	1315	1430	40	1U2	1530	40
	1J3	1530	1645	30	1U3	1730	30
	1J4	1730	1845	40	1U4	1945	40
6†‡	6J1	0600	0715	60	7U1	0600	60
	6J2	0830	0945	40	7U2	0830	40
	6J3	1045	1200	30	7U3	1045	30
	6J4	1245	1400	40	7U4	1245	40
7	7J1	0445	0600	40	2U1	0700	40
	7J2	0700	0815	30	2U2	0900	30
	7J3	0900	1015	40	2U3	1115	40
	7J4	1115	1230	60	2U4	1345	60
8‡	8J1	1145	1300	40	8U1	1200	40
	8J2	1400	1515	30	8U2	1400	30
	8J3	1600	1715	40	8U3	1600	40
	8J4	1815	1930	60	8U4	1815	60

†Threat array jammed from side II on these days, see Table 5.
‡Unjammed trials run first on these days.

TABLE 4. Team Specifications by Trial (Notes—Main Body of Table Follows)

* For each team the members are to be numbered as #1 and #2 and that the numbering remains the same throughout the test. On days marked with an asterisk, members numbered #1 are to be positioned nearest the FEBA; on the remaining days, members numbered #2 are to be positioned nearest the FEBA.

** Radio type codes are: L = low power radio, H1 = type 1 high power radio, H2 = type 2 high power radio.
Odd numbered links are short links, even numbered links are long links. Links 1, 2, 5, 6 are nearest side I; links 3, 4, 7, 8 are nearest side II (see Figure 1).
Echelon given for information only, link number determines position: BC = battalion-company, CP = company-platoon.

† Frequencies are numbered arbitrarily from 1 to 8. It is assumed that the same frequencies are used throughout the test and that the numbering remains the same throughout the test.

†† Message sets are numbered as follows:
1 = first 12-message set, 2 = second 12-message set, 3 = third 12-message set, 4 = fourth 12-message set,
5 = first 6-message set, 6 = second 6-message set, 7 = first 18-message set, 8 = second 18-message set.
A new message set is to be generated each day and used throughout the day. Each message set is to consist of three types of messages, one third of each type. For each trial separately, one half of the messages of each type in each set will be marked for transmission by team member #1, the other half will be marked for transmission by team member #2, and the marking will be done randomly. For each trial separately, each message set will be ordered randomly and randomly scheduled for transmission subject to suitable constraints. The analyst will aid in this randomization and scheduling.

‡ Only frequencies and message sets for jamming trials are explicitly listed. The same frequency is to be used on the corresponding unjammed trail. The message set to be used on the corresponding unjammed trial is determined as follows: if the message set on the jamming trial is number 2n (even) then use the message set numbered 2n-1 on the corresponding unjammed trial; if the message set on the jamming trial is number 2n-1 (odd) then use the message set numbered 2n on the corresponding unjammed trial.

TABLE 4. Team Specifications by Trial (Notes Precede Main Body of Table)

Part 1: Radio and Location by Day

Day	Specification**	Team -1-	-2-	-3-	-4-	-5-	-6-	-7-	-8-
1*	Radio	L	L	L	L	H1	H2	H1	H2
	Link	1	2	3	4	5	6	7	8
	Echelon	CP	CP	BC	BC	CP	CP	BC	BC
2	Radio	H1	H2	H1	H2	L	L	L	L
	Link	4	3	2	1	8	7	6	5
	Echelon	BC	BC	CP	CP	BC	BC	CP	CP
3	Radio	L	L	L	L	H2	H1	H2	H1
	Link	3	4	1	2	7	8	5	6
	Echelon	CP	CP	BC	BC	CP	CP	BC	BC
4*	Radio	H2	H1	H2	H1	L	L	L	L
	Link	2	1	4	3	6	5	8	7
	Echelon	BC	BC	CP	CP	BC	BC	CP	CP
5	Radio	L	L	L	L	H1	H2	H1	H2
	Link	6	5	8	7	2	1	4	3
	Echelon	BC	BC	CP	CP	BC	BC	CP	CP
6*	Radio	H1	H2	H1	H2	L	L	L	L
	Link	7	8	5	6	3	4	1	2
	Echelon	CP	CP	BC	BC	CP	CP	BC	BC
7*	Radio	L	L	L	L	H2	H1	H2	H1
	Link	8	7	6	5	4	3	2	1
	Echelon	BC	BC	CP	CP	BC	BC	CP	CP
8	Radio	H2	H1	H2	H1	L	L	L	L
	Link	5	6	7	8	1	2	3	4
	Echelon	CP	CP	BC	BC	CP	CP	BC	BC

Part 2: Frequency† and Message Sets†† by Trial (Frequency Number/Message Set Number)

Jamming Trial‡	Team -1-	-2-	-3-	-4-	-5-	-6-	-7-	-8-
1J1	1/4	2/1	3/7	4/2	5/6	6/3	7/5	8/8
1J2	2/2	1/7	4/1	3/4	6/8	5/5	8/3	7/6
1J3	3/8	4/5	1/3	2/6	7/2	8/7	5/1	6/4
1J4	4/6	3/3	2/5	1/8	8/4	7/1	6/7	5/2
2J1	6/1	5/4	8/2	7/7	2/3	1/6	4/8	3/5
2J2	5/7	6/2	7/4	8/1	1/5	2/8	3/6	4/3
2J3	8/5	7/8	6/6	5/3	4/7	3/2	2/4	1/1
2J4	7/3	8/6	5/8	6/5	3/1	4/4	1/2	2/7
3J1	5/7	6/2	7/4	8/1	1/5	2/8	3/6	4/3
3J2	6/1	5/4	8/2	7/7	2/3	1/6	4/8	3/5
3J3	7/3	8/6	5/8	6/5	3/1	4/4	1/2	2/7
3J4	8/5	7/8	6/6	5/3	4/7	3/2	2/4	1/1
4J1	2/2	1/7	4/1	3/4	6/8	5/5	8/3	7/6
4J2	1/4	2/1	3/7	4/2	5/6	6/3	7/5	8/8
4J3	4/6	3/3	2/5	1/8	8/4	7/1	6/7	5/2
4J4	3/8	4/5	1/3	2/6	7/2	8/7	5/1	6/4
5J1	2/2	1/7	4/1	3/4	6/8	5/5	8/3	7/6
5J2	1/4	2/1	3/7	4/2	5/6	6/3	7/5	8/8
5J3	4/6	3/6	2/5	1/8	8/4	7/1	6/7	5/2
5J4	3/8	4/5	1/3	2/6	7/2	8/7	5/1	6/4
6J1	5/7	6/2	7/4	8/1	1/5	2/8	3/6	4/3
6J2	6/1	5/4	8/2	7/7	2/3	1/6	4/8	3/5
6J3	7/3	8/6	5/8	6/5	3/1	4/4	1/2	2/7
6J4	8/5	7/8	6/6	5/3	4/7	3/2	2/4	1/1
7J1	6/1	5/4	8/2	7/7	2/3	1/6	4/8	3/5
7J2	5/7	6/2	7/4	8/1	1/5	2/8	3/6	4/3
7J3	8/5	7/8	6/6	5/3	4/7	3/2	2/4	1/1
7J4	7/3	8/6	5/8	6/5	3/1	4/4	1/2	2/7
8J1	1/4	2/1	3/7	4/2	5/6	6/3	7/5	8/8
8J2	2/2	1/7	4/1	3/4	6/8	5/5	8/3	7/6
8J3	3/8	4/5	1/3	2/6	7/2	8/7	5/1	6/4
8J4	4/6	3/6	2/5	1/8	8/4	7/1	6/7	5/2

Notes to TABLE 4 appear on the previous page.

TABLE 5. Jammer Specifications by Trial

*Part 1: Position and Frequency Jammed by Priority (L/F1/F2)**

Trial	Jammer Number -1-	-2-	-3-	-4-
1J1	13/4/6	14/5/3	15/1/7	16/8/2
1J2	6/1/7	5/8/2	8/4/6	7/5/3
1J3	4/3/5	3/6/4	2/2/8	1/7/1
1J4	11/2/8	12/7/1	9/3/5	10/6/4
2J1	7/2/8	8/7/1	5/3/5	6/6/4
2J2	16/3/5	15/6/4	14/2/8	13/7/1
2J3	10/1/7	9/8/2	12/4/6	11/5/3
2J4	1/4/6	2/5/3	3/1/7	4/8/2
3J1	25/8/2	26/1/7	27/5/3	28/4/6
3J2	18/5/3	17/4/6	20/8/2	19/1/7
3J3	24/7/1	23/2/8	22/6/4	21/3/5
3J4	31/6/4	32/3/5	29/7/1	30/2/8
4J1	19/6/4	20/3/5	17/7/1	18/2/8
4J2	28/7/1	27/2/8	26/6/4	25/3/5
4J3	30/5/3	29/4/6	32/8/2	31/1/7
4J4	21/8/2	22/1/7	23/5/3	24/4/6
5J1	29/5/3	30/4/6	31/8/2	32/1/7
5J2	22/8/2	21/1/7	24/5/3	23/4/6
5J3	20/6/4	19/3/5	18/7/1	17/2/8
5J4	27/7/1	28/2/8	25/6/4	26/3/5
6J1	23/7/1	24/2/8	21/6/4	22/3/5
6J2	32/6/4	31/3/5	30/7/1	29/2/8
6J3	26/8/2	25/1/7	28/5/3	27/4/6
6J4	17/5/3	18/4/6	19/8/2	20/1/7
7J1	9/1/7	10/8/2	11/4/6	12/5/3
7J2	2/4/6	1/5/3	4/1/7	3/8/2
7J3	8/2/8	7/7/1	6/3/5	5/6/4
7J4	15/3/5	16/6/4	13/2/8	14/7/1
8J1	3/3/5	4/6/4	1/2/8	2/7/1
8J2	12/2/8	11/7/1	10/3/5	9/6/4
8J3	14/4/6	13/5/3	16/1/7	15/8/2
8J4	5/1/7	6/8/2	7/4/6	8/5/3

Part 2: Type and Length of Course to be Traversed Before Each Jamming Trial

Day	Jammer Number -1-	-2-	-3-	-4-
1	short, smooth	short, rough	long, smooth	long, rough
2	long, rough	long, smooth	short, rough	short, smooth
3	short, smooth	short, rough	long, smooth	long, rough
4	long, rough	long, smooth	short, rough	short, smooth
5	long, smooth	long, rough	short, smooth	short, rough
6	short, rough	short, smooth	long, rough	long, smooth
7	long, smooth	long, rough	short, smooth	short, rough
8	short, rough	short, smooth	long, rough	long, smooth

* Position numbers are given in Figure1, and frequencies are numbered as in Table 4. F1 is the first priority frequency, F2 is the second priority frequency.

framework to make the problem more tractable. The application of Graeco-Latin squares to the real jammer test mentioned in the Introduction was based on powers of 3, which was harder to do in a systematic manner.[9] On the other hand, designing in powers of two can be quite flexible, as this example has shown. Experiments with n factors, each at 2 levels are 2^n designs, and they can be run (or partially run) in sequences of 2^{n-k} fractional factorials constructed using the notation and theory of finite abelian groups.

The factors and levels used in this example are tabulated in Tables 6 and 7. The overall design was constructed by starting with a 2^3 full factorial in factors A-C, adding factor D by confounding it with the ABC interaction to give the 2^{4-1} design defined by I = +ABCD, then adding factors E, F, and G by confounding them with the two-factor interactions AB, AC, and BC, to give the 2^{7-4} fractional factorial design consisting of the points *efg, abe, acf, bcg, adg, bdf, cde,* and *abcdefg*.[†] The defining contrast for this fraction is I = ABFG = ACEG = ABCD = EFG, and the

[†] This fraction is closely related to the simple example which uses I=-ABCD and blocking factors AB, AC, and AD. Now A-D characterize physical conditions relevant to effect of jamming on the link while E-G characterize jammer tasking conditions.

design is a saturated design of resolution III (see John[7] section 8.4 or Montgomery[8] section 11-5; John[7] gives a method due to Margolin[10] for deriving defining contrasts from a basic set of letters). By repeating the design in 16 trials (blocks) generated by the defining contrasts I = ±ABFG = ±ACEG = ±ABCD = ±EFG, all 128 effects in the full factorial design become estimable, except that the effects in the defining contrast subgroup are confounded with blocks. Two new factors corresponding to"Side II" and unjammed trials (A_o and F_o in Table 6) were added by quadrupling the number of trials (running the 16 trials already discussed at each of the four combinations of $\pm A_o$ and $\pm F_o$). Table 8 gives the relationship of defining contrast signs to trial numbers. Formally, the blocks actually used have defining contrasts generated by:

$$I = \pm F_o = \pm A_o = \pm ABFG = \pm ACEG = \pm ABCD = \pm EFG.$$

In this example, the seven two-factor interactions AB, AC, AD, BD, CD, and EF and the four three-factor interactions ABC, ABD, ACD, and BCD were thought likely to be relatively important but all other interactions were thought likely to be relatively unimportant. Thus (apart from block effects) of the 512 estimable effects obtainable from running 64 trials, only 21 (mean, 9 main effects, 7 two-factor interactions, and 4 three-factor interactions) appeared likely to be important. New factors were added by confounding them with what were anticipated to be relatively unimportant interactions between the nine factors A-G, A_o, and F_o. In total 32 factors were examined. Including various interpretable interactions of interest, 62 distinct effects were explicitly considered. (Some of the 31 effects which characterized the 32 jammer positions were not fully examined and are not included in this total; many of those were partially confounded with other effects of possible interest.) ***Exactly how new factors were added is indicated in Table 9, which gives the alias chains in factors A-G with the new factors set equal to appropriate interactions. The horizontal lines in Table 9 divide the table into the effects aliased in trials with the mean, A, B, E, C, F, G, and D, respectively. The top block is the defining contrast subgroup for the entire 2^{7-3} fraction, the other blocks are cosets. The individual rows of the table indicate how those alias chains break for trials within a day (each row represents an alias chain once four jamming or unjammed trials in a day are complete). The columns indicate breaks in alias chains once an entire four-day week of trials is complete.***[†] The full table of alias chains with A_o and F_o incorporated is too large to include, but it can be obtained from Table 9 by writing after each row of Table 9 the same row multiplied by F_o and writing to the right of this new table the same table multiplied by A_o. A partial table of the alias chains for the mean (I) with all 9 factors is given for illustration in Table 10 (this is the defining contrast subgroup for the entire design). The following equation summarizes how new factors were added:

$$\begin{aligned} I &= CFGH = BDEJ = A_oEGK = A_oACEFL = BCDEGM = A_oBN = BEFO = ABCEFP = CEFQ \\ &= AEGR = AEFGS = ABDEFT = A_oABCFGU = BFV = CDFW = A_oCDFGX = BDFY \\ &= A_oACEGZ = A_oB_oABFG = C_oABCD = D_oEFG = A_oE_oBCEF = G_oACEG. \end{aligned}$$

Since this hypothetical example is formally constructed, it is easy to discuss test expansions or reductions. Eight days of testing were explicitly considered using 8 trials per day with 8 separate links per trial, so essentially[‡] $2^9 = 512$ observations were planned, and the overall design represents a 2^{32-23} (1/8,388,608) fraction of a 2^{32} design (4,294,967,296 points) run in 64 blocks of size eight. If another eight days of testing were desired, the design could simply be repeated (preferably with a physically modified target array and new jammer positions). A more desirable approach would not only modify link and jammer positions but also change the signs of some of the added factors. Ignoring changes in sign for the special contrasts involving the pure position indicators N, O, X, and Y (used as a device to rotate jammer positions within a day but having no general meaning outside a particular physical set up) and L and M (which specify message rotation), sign changes could generate a test $2^{17} = 131{,}072$ times as long as the present one (and lasting almost 3,000 years). Thus, the design given here can easily be extended to a test of any desired length. However, even the hypothetical eight-day test discussed here would likely allow quite detailed and accurate examination of most performance factors of interest. In earlier days of extensive stand-alone operational and developmental testing, logistics considerations would probably have driven a longer test, in which case the eight-day extension in which the signs of BFV and CDFW are changed might be particularly appropriate. Continuation of unjammed trials beyond the first 4 days of test could be hard to justify. Then unjammed trials could be discontinued after the first four days and messages could be used for two days by assigning message sets as

[†] The breaks by row and column occur because EFG and ABCD change sign within day but ACEG and ABFG only change sign from day-to-day, see Table 8.

[‡] Actually, since the 2 ends of each link could be considered separately in analysis and an average of 6 messages would be sent to each receiver during each trial, there are 512×12 = 6,144 observations total for analysis.

TABLE 6. Factors and Levels Used in This Example

Factor	Description	Levels
A	Link length	(1)=short; *a*=long. See N, O below.
B	Echelon	(1)=CP (company-platoon); *b*=BC (battalion-company)
C	Radio power	(1)=low; *c*=high
D	Jammer distance from FEBA	(1)=short; *d*=long. See Table 7.
E	Jamming priority	(1)=priority 2; *e*=priority 1
F	Relative message density	(1)=low; *f*=high
G	Jammer position indicator	Relative left/right indicator, used with D to ensure 4 positions used per trial. See Table 7.
H,J	Jammer number	(1)=#1; *h*=#2; *j*=#3; *hj*=#4
K	Radio power refinement	(1)=low; *k*=low; *c*=high type 1; *ck*=high type 2
L,M	Message set number	(1)=#2; *l*=#1; *fm*=#4; *fl*=#3; *m*=#6; *lm*=#5; *f*=#8; *flm*=#7; f_o=#1; $f_o l$=#2; $f_o fm$=#3; $f_o fl$=#4; $f_o m$=#5; $f_o lm$=#6; $f_o f$=#7; $f_o flm$=#8. See note †† to Table 4.
N,O	Physical link location	(1)=#1; *a*=#2; *n*=#3; *an*=#4; *o*=#5; *ao*=#6; *no*=#7; *ano*=#8. See Figure 1 and note * to Table 4.
P,Q,R	Teams	(1)=#1; *p*=#2; *q*=#3; *pq*=#4; *r*=#5; *pr*=#6; *qr*=#7; *pqr*=#8
S,T,U	Frequencies	(1)=#1; *s*=#2; *t*=#3; *st*=#4; *u*=#5; *su*=#6; *tu*=#7; *stu*=#8
V	Roughness of mobility	(1)=smooth; *v*=rough
W	Length of mobility	(1)=short; *w*=long
X,Y	Jammer position indicator	Used with G and D to ensure 16 positions per day. See Table 7.
Z	Team member position	(1)=no asterisk; *z*=asterisk. See note * to Table 4.
A_o	Side of threat array jammed	(1)=Side II; a_o=Side I. See Table 7.
B_o	Scenario intensity order	(1)=40/30/40/60; b_o=60/40/30/40
C_o,D_o	Trial order within day	(1)=fourth trial; c_o=second trial; d_o=third trial; $c_o d_o$=first trial
E_o	Start time on day	(1)=0600; e_o=1200
F_o	Jammer use in trial	(1)=unjammed; f_o=jamming
G_o	Jammed trial order	(1)=unjammed trial first; g_o=jamming trial first

TABLE 7. Coding of Levels for Jammer Position (Involves Factors D, F, X, Y, and A_o from Table 6)

	Position Number from Figure 1*		
Side II Level	Side II	Side I	Side I Level
(1)	17	1	a_o
d	18	2	$a_o d$
g	19	3	$a_o g$
dg	20	4	$a_o dg$
x	21	5	$a_o x$
dx	22	6	$a_o dx$
gx	23	7	$a_o gx$
dgx	24	8	$a_o dgx$
y	25	9	$a_o y$
dy	26	10	$a_o dy$
gy	27	11	$a_o gy$
dgy	28	12	$a_o dgy$
xy	29	13	$a_o xy$
dxy	30	14	$a_o dxy$
gxy	31	15	$a_o gxy$
dgxy	32	16	$a_o dgxy$

*Odd numbers denote jammer positions nearest the FEBA (low level of Factor D).

TABLE 8. Relationship of Defining Contrast Signs to Trial Numbers

Sign of Defining Contrasts Excluding F_0					Trial Number	
A_0	ABFG	ACEG	ABCD	EFG	With $+F_0$	With $-F_0$
+	+	+	+	+	1J1	1U1
+	+	+	+	-	1J2	1U2
+	+	+	-	+	1J3	1U3
+	+	+	-	-	1J4	1U4
+	+	-	+	+	2J1	2U1
+	+	-	+	-	2J2	2U2
+	+	-	-	+	2J3	2U3
+	+	-	-	-	2J4	2U4
-	+	+	+	+	3J1	3U1
-	+	+	+	-	3J2	3U2
-	+	+	-	+	3J3	3U3
-	+	+	-	-	3J4	3U4
-	+	-	+	+	4J1	4U1
-	+	-	+	-	4J2	4U2
-	+	-	-	+	4J3	4U3
-	+	-	-	-	4J4	4U4
-	-	+	+	+	5J1	5U1
-	-	+	+	-	5J2	5U2
-	-	+	-	+	5J3	5U3
-	-	+	-	-	5J4	5U4
-	-	-	+	+	6J1	6U1
-	-	-	+	-	6J2	6U2
-	-	-	-	+	6J3	6U3
-	-	-	-	-	6J4	6U4
+	-	+	+	+	7J1	7U1
+	-	+	+	-	7J2	7U2
+	-	+	-	+	7J3	7U3
+	-	+	-	-	7J4	7U4
+	-	-	+	+	8J1	8U1
+	-	-	+	-	8J2	8U2
+	-	-	-	+	8J3	8U3
+	-	-	-	-	8J4	8U4

specified in Table 4 for jamming trials on odd-numbered days beyond 4 (days with number 2k+1, k>1) and relabeling message sets (by interchanging message sets 2n-1 and 2n for n = 1, 2, 3, 4) on even numbered days beyond 5. Reduction of test days from eight to four would prohibit rotating all jammers through all positions and all teams through all links and would reduce the ratio of observations to effects from about 16:1 to less than 10:1 (some additional confounding would also result). Reductions in test length below four days would introduce substantial confounding of effects.

The display in Table 9 proved to be very useful as a working tool for constructing this design. It is a group multiplication table indicating by its rows and columns how the alias chains (corresponding to elements in various quotient groups) decompose as trials (blocks) are conducted. For instance, physical location (effects A, N, O) and team numbers (effects P, Q, R) must stay with the same link lengths, echelons, and radios throughout each day. Thus P and A, Q and N and B, and C and O and R, respectively, must be on the same rows of Table 9. Similarly, V and W (which define stress on the jammers due to mobility exercises) must be on the same rows as H and J (which determine jammer number) to ensure that each jammer is repeatedly subjected to the same stress throughout any one day. On the other hand, H and J must be in the same alias chains as D and G (which determine relative short/long and left/right jammer position) on any trial, but H and J must be on different rows from D and G to ensure that jammers rotate from trial to trial within each day; X and Y (aliased with I and therefore fixed for any trial) further define jammer rotation by ensuring that four different jammer positions are used in each trial throughout any day. Similar considerations influence confounding of other effects. ***Some fiddling is required to insert effects in this way, since even with Table 9 as a guide to enable the analyst to keep track of previous work as modifications are made, each modification must be checked to make sure it has the intended impact.*** Insertion of X and Y so that they rotated jammers from trial-to-trial in such a way that jammers saw teams, links, frequencies, and density ratios—as well as jamming positions—proved especially difficult. Luckily, improvements in computer software since the original paper[1] was written make such checks quite easy. In fact, without such software, attempting construction of designs as elaborate as the one in this example is extremely error prone. The original paper is rife

TABLE 9. Alias Chains for Main Effects A-G Showing How New Effects Were Added (See Text)
(**Bold** Effects Considered Relatively Important)

I	**ABFG=A_oB_o**	**ACEG=G_o=A_oZ**	**BCEF=A_oE_o**
ABCD=C_o	**CDFG=A_oX**	**BDEG=B_oC_o**	ADEF
EFG=D_o	ABE	ACF	**BCG=A_oXY**
ABCDEFG=C_oD_o	CDE	**BDF=Y**	ADG
A	BFG	**CEG=CK**	**ABCEF=P**
BCD=VW	**ACDFG=HJ**	ABDEG	DEF
AEFG=S	BE	CF	ABCG
BCDEFG=A_OFM	ACDE	ABDF	**DG**
B=A_ON	AFG	ABCEG	**CEF=Q**
ACD	BCDFG	DEG	**ABDEF=T**
BEFG	AE	ABCF	CG
ACDEFG	BCDE	DF	**ABDG=A_oFLM**
AB=A_OAN	**FG**	**BCEG=PQ**	**ACEF=A_oL**
CD	ABCDFG	ADEG	BDEF
ABEFG	**E**	BCF	ACG
CDEFG	ABCDE	ADF	**BDG=ST**
C	**ABCFG=A_oU**	**AEG=R**	**BEF=O**
ABD	DFG	**BCDEG=M**	ACDEF
CEFG	ABCE	AF	BG
ABDEFG	DE	BCDF	ACDG
AC	**BCFG=PR**	**EG=A_OK**	**ABEF=AO**
BD	ADFG	ABCDEG	CDEF
ACEFG	**BCE=A_oSU**	**F**	ABG
BDEFG	ADE	ABCDF	CDG
BC	**ACFG=QR**	ABEG	**EF=NO**
AD	BDFG	**CDEG=A_oTU**	ABCDEF
BCEFG	**ACE=A_oFL**	ABF	**G**
ADEFG	**BDE=J**	**CDF=W**	ABCDG
ABC	CFG	**BEG=PQR**	**AEF=A_oANO**
D	**ABDFG=A_oLM**	ACDEG	BCDEF
ABCEFG=H	CE	**BF=V**	AG
DEFG	ABDE	**ACDF=A_oSTU**	BCDG

Rows show breaks in alias chains once trials on a day are complete.
Columns show breaks in alias chains once four days of trials are complete.

TABLE 10. Alias Chain for I with Factors A-G, A_o, and F_o Considered (Defining Contrast Subgroup)

I	ABFG	ACEG=G_o	BCEF	A_o	A_oABFG=B_o	A_oACEG=Z	A_oBCEF=E_o
F_o	F_oABFG	F_oACEG	F_oBCEF	A_oF_o	A_oF_oABFG	A_oF_oACEG	A_oF_oBCEF
ABCD=C_o	CDFG	BDEG=B_oC_o	ADEF	A_oABCD	A_oCDFG=X	A_oBDEG	A_oADEF
F_oABCD	F_oCDFG	F_oBDEG	F_oADEF	A_oF_oABCD	A_oF_oCDFG	A_oF_oBDEG	A_oF_oADEF
EFG=D_o	ABE	ACF	BCG	A_oEFG	A_oABE	A_oACF	A_oBCG=XY
F_oEFG	F_oABE	F_oACF	F_oBCG	A_oF_oEFG	A_oF_oABE	A_oF_oACF	A_oF_oBCG
ABCDEFG=C_oD_o	CDE	BDF=Y	ADG	A_oABCDEFG	A_oCDE	A_oBDF	A_oADG
F_oABCDEFG	F_oCDE	F_oBDF	F_oADG	A_oF_oABCDEFG	A_oF_oCDE	A_oF_oBDF	A_oF_oADG

with errors, and initial manual attempts to correct the errors introduced as many new errors as it corrected. Fortunately it is easy to automate calculation of the entire design using existing statistical software. A spreadsheet can also be used but not so conveniently. JMP® [11] was used to implement the calculations in Figure 2 using indicator functions, ceiling functions, and modulo arithmetic. Automation reduces the time needed to insert and check a new factor from hours to minutes and substantially reduces errors.

This subsection has sketched many of the detailed technical considerations that can be used to design efficient field tests. The methods are not elementary, but they are within the capabilities of a well-trained analyst equipped with suitable statistical software. Moreover, the design resulting from such a process suggests a multitude of possible analyses that are feasible using readily available statistical software. On the other hand, these methods are well beyond the capabilities of a typical test operator, which implies that much more active participation by analysts in detailed test planning is necessary to apply these methods in practice.

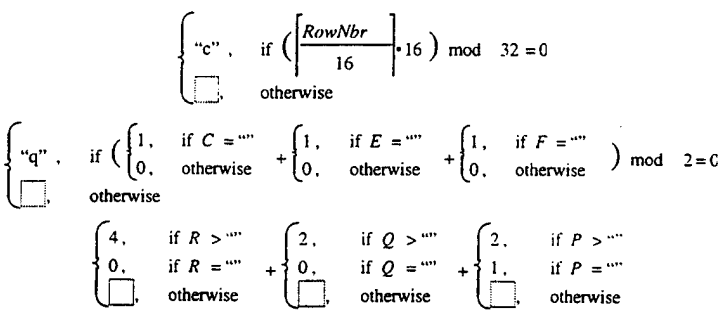


Figure 2. JMP® Formulas for C, Q, and Teams

SUMMARY

This paper has shown through one illustrative example and one hypothetical example how formal experimental design techniques could be used to design executable, robust, and realistic field tests. It shows that a sequence of fractional factorials can be used to spread fixed resources efficiently over a wide variety of test conditions subject to complicated constraints, yielding a resulting design that can be described by a few test matrices. It exploits the fact that many compromises with reality are necessary in order to conduct a field test and that the test operator typically specifies many conditions that would be uncontrolled in actual combat. When compromises with reality must be made or normally uncontrolled factors must be specified in order to have an executable field test, an analyst using techniques of efficient experimental design can suggest methods for specifying test conditions likely to yield marked improvement in the scope, generality, and clarity of test results without making the test impractical to conduct.

Until recently, demands for free play in operational testing have made formal statistical design of field tests a very hard sell. Now the trend is away from performing lots of formal developmental and operational testing and towards more use of FTXs in materiel acquisition—together with modeling and simulation in the simulation based acquisition paradigm. This trend presents an opportunity to add more structure to the formal combined developmental and operational testing part of the materiel acquisition process, relying on training exercises to make up for any lost free play. It may not be possible in most real test planning to push formal statistical design as far as it was pushed in the hypothetical example of this paper. However, the potential for improved efficiency, precision, and accuracy from a conscientious formal experimental design effort certainly makes such increased formal design efforts worthwhile.

REFERENCES

1. Russell, C.T. "Using Efficient Experimental Design to Accommodate a Wider Variety of Test Conditions within Resource Constraints." Proceedings of the Twenty-Second Annual US Army Operations Research Symposium, vol. 1, p. 4-354, 1983.
2. Russell, C.T. "The Potential Utility of Crossing a Fractional Factorial with a Full Factorial in the Design of Field Tests." Raleigh: US Army Research Office Report No 81-2, Proceedings of the Twenty-Sixth Conference on Design of Experiments, p. 335, 1981.
3. Russell, C.T. "Selling a Complicated Experimental Design to the Field Test Operator." Raleigh: US Army Research Office Report No 82-2, Proceedings of the Twenty-Seventh Conference on Design of Experiments, p. 293, 1982.
4. Hill, F.I. "Recollections of First Years of CDEC." Aberdeen Proving Ground: Army Research Laboratory, ARL-SR-59, Proceedings of the Second Annual US Army Conference on Applied Statistics, p. 119, 1997.
5. Russell, C.T. "The Perversity of Missing Points in the 2^4 Design." Raleigh: US Army Research Office Report No 83-2, Proceedings of the Twenty-Eighth Conference on Design of Experiments, p. 499, 1983.
6. Russell, C.T. "A Second Look at the Perversity of Missing Points in the 2^4 Design." Raleigh: US Army Research Office Report No 87-2, Proceedings of the Thirty-Second Conference on Design of Experiments, p. 351, 1987.
7. John, P.W.M. Statistical Design and Analysis of Experiments, First Edition. New York: Macmillan, 1971.
8. Montgomery, D.C. Design and Analysis of Experiments, Third Edition. New York: Wiley, 1991.
9. Bose, R.C. "On the Application of the Properties of Galois Fields to the Problem of the Construction of Hyper-Graeco-Latin Squares." Sankya, vol 3, p. 323, 1938.
10. Margolin, B.H. "Systematic Methods of Analyzing 2^m3^n Factorial Experiments with Applications." Technometrics, vol. 9, p. 245, 1967.
11. Sall, J. et. al. JMP® User's Guide, Version 3 of JMP, Cary: SAS Institute Inc., 1995.

INTENTIONALLY LEFT BLANK.

STATISTICS IN THE MILITARY OPERATIONS IN URBAN TERRAIN (MOUT) ADVANCED CONCEPTS TECHNOLOGY DEMONSTRATION (ACTD)

Eugene Dutoit
Dismounted Battlespace Battle Lab
Fort Benning, GA 31907

ABSTRACT

This paper will present an overview of the MOUT ACTD to include the purpose and scope. This is a joint effort involving the Army and Marines with experimentation taking place at both Fort Benning GA and Camp Lejeune NC. About *forty* technical MOUT deficiencies were identified and technical solutions were proposed to meet these deficiencies. Field experiments had to be designed to examine each of the proposed technical solutions. In several cases, simulation experiments were also conducted to evaluate the proposed solution in a force-on-force context. The Army Dismounted Battlespace Battle Lab, the Marine Warfighting Lab, AMSAA, TRAC, OPTEC, TEXCOM, HRED and ARI were all involved in the total evaluation effort. This paper will show, for a *few selected technologies*, examples of the experimental design and statistical analysis procedure for both the ground and simulation experiments.

INTRODUCTION

Many military analysts regard military operations in urban terrain (MOUT) as the most likely battlefield in the twenty-first century. It will be a complex and resource intensive scenario were potential adversaries may have military parity with US forces. The Russian way of conducting MOUT, as seen in Chechnya, is to fight without regard for collateral damage or non-combatants. Essentially, this is a rubble and then clean up philosophy. This is not an optional policy for the United States. One of the purposes of the MOUT ACTD is to provide technological dominance in MOUT for both Soldiers and Marines. This is to be done by examining technological candidates that will improve command/control/communication and computers, engagement capability, force mobility and overall force protection.

GENERAL EXPERIMENTAL OVERVIEW

As of the first quarter of FY00, there were twelve field experiments conducted in support of the MOUT ACTD. The Army conducted six separate experiments at Fort Benning and the Marines conducted four separate experiments at Camp Lejeune. In addition to these separate service experiments, there were two joint service experiments conducted at Camp Lejeune and Fort Benning respectively. Experimentation started in the second quarter of FY98 (Army experiment 1) and was completed at the end of the fourth quarter FY99 (the second joint experiment at Fort Benning). Both Camp Lejeune and Fort Benning have MOUT villages constructed on post complete with one, two and three story buildings, streets, and underground sewer systems. These villages are used for MOUT training and experimentation. The Fort Benning village (the McKenna MOUT site) is instrumented with video capture, playback, and storage and retrieval capabilities. SIMLAS and separate sensor arrays working together can send data to the operational test-visualization system (OT-VIS) in the form of a two dimensional array. An after action review theater has been constructed so that simulated combat operations data can be sent for reviewing the operations in real time. This facility can also record important information that is available for playback and analysis.

The general scheme of data collection included measures of system / technology performance (MOP) which were obtained through side experiments conducted on each technology. Qualitative data were obtained through the use of questionnaires. Subject matter experts were also asked to provide information to scoring conferences. Combat measures of effectiveness (MOE) were obtained primarily from experiment controller assessments and video capture and replay. A data authentication group was formed to screen

---

Approved for public release; distribution unlimited.

for logical errors or missing data. The data authentication group (DAG) was made up of representatives from the Army, Marines, the Technical Project Office and the Army Systems Analysis Activity (AMSAA).

A comprehensive listing of the MOUT technology types that were investigated in the overall ACTD are presented in the tables below.

Table 1 (Requirements investigated in the four Marine experiments).

Man portable shield.	Non explosive breach.
Clearly identify friendly forces.	Get on top of buildings.
Stun grenade.	Powered optics.
See through wall sensor.	Countersniper detection.
Lightweight common target designator	See in building at all times.

Table 2 (Requirements investigated in the six Army experiments).

Remote marking.	Joint protection.
Protective mask.	Improved obscurants.
Blunt training munition.	Non line of sight radio.
Medical evacuation.	Improved intelligence collection system.
Improved sling.	Blow man-sized hole.
Point munitions.	Personnel protection kit.
Hearing protection.	Frangible ammunition.
Personnel restraints.	Rapid mapping and virtual mission planner.

GENERAL SCHEME FOR EACH LIVE EXPERIMENT

1. Perform new equipment training (NET). This will insure that the Soldiers and Marines can use the new experimental technologies so that they get a fair evaluation in all of the experiments.

2. Design / verify the design of the technology side experiment. Consideration is given to insure that all users (subject Soldiers and Marines) see all technologies (including the base case) in some random order.

3. Execute the side experiment for each requirement. This is a technology / user experiment with a focus on measures of effectiveness (MOP). The base case technologies are included.

4. Analysis of these data result in the selection of the best technologies to be used in other experiments and in force-on-force simulations. Sometimes the base case is best.

5. Execute the tactical experiments for each requirement technology (base case too). This is a technology / user experiment with a focus on the measures of effectiveness (MOE).

6. Execute collective tactical experiments where several technologies are used together as a suite. The base case force is also considered.

AN EXAMPLE OF THE RESULTS OF A LIVE FIELD EXPERIMENT

This section will present an example overview of the live field experiment that evaluated the performance of four technical candidates to fill the requirements of the medical evacuation device (see Table 2 above). It is important to point out that this was not the only live field experiment performed to evaluate the four candidates.

MEDICAL EVACUATION REQUIREMENT

Experiment: To determine the average time required for unpacking each alternative stretcher.

Method:

a. Eight teams were formed by randomly assigning two people per team (thus forming teams of equal ability).
b. Each of the eight teams used each of the four alternative stretchers, but not in the same order.
c. Each team "unpacked" each alternative four times. Times (in seconds) for unpacking the stretchers were recorded.
d. Consistent start and stop criteria were established for all trials.

Results:

1. Analysis across all four unpackings.

Stretcher Alternative	Average time to unpack
Alternative 1	13.71 sec
Alternative 2	66.87 sec
Alternative 3	92.53 sec
Alternative 4	43.06 sec

P = .000

Post Hoc (Tukey/LSD) indicated that the unpacking times for Alternative 1 < Alternatives 2,3,4.

2. Analysis considering unpacking trials (1,2,3,4). This was done to determine if there was any practice effect.

Stretcher alt	Unpack 1	Unpack 2	Unpack 3	Unpack 4	Comments
Alt 1	18.4	12.7	12.1	11.6	Sig linear contrast (.038)
Alt 2	97.2	60.9	56.1	53.2	Sig linear & quad contrast (.000)
Alt 3	164.1	77.7	62.3	66.0	Sig linear & quad contrast (.002 & .036)
Alt 4	52.6	44.3	40.2	35.2	Sig linear contrast (.000)

Comment: All four stretcher alternatives showed a significant negative trend (indicating learning across trials) in the time required to "unpack."

3. Analysis considering the fourth trial.

Stretcher alternative	Average unpacking time for the fourth trial
Alternative 1	11.6 sec
Alternative 2	53.2 sec
Alternative 3	66.0 sec
Alternative 4	35.2 sec

P = .000

Post hoc analysis still indicated that for the 4th trial, the unpacking time for Alternative 1 < Alternatives 2, 3, 4.
Note: This was also confirmed by the nonparametric Kruskal-Wallis test.

4. Operational Insights.

Overall, alternative 1 requires less time to unpack. This is true across all four trials and for the fourth trial itself. All four alternatives showed significant learning effects across trials. However, the results obtained from the first analysis (averaged across all four trials) is probably more representative of the expected unpacking times under operational conditions when multiple practice trials would not take place.

GENERAL SCHEME FOR EACH SIMULATION

1. Determine the base case technology and the alternative(s) based on the results of a live fire field experiment.

2. Establish the force-on-force combat scenario (with guidance from the proponent).

3. Start the input data authentication process with oversight from AMSAA.

4. Select the appropriate force-on-force model: JCATS for inside of the city and buildings
JANUS for outside of the city

5. Both the independent evaluators (TRADOC Analysis Center,TRAC-White Sands Missile Range) and the AMSAA are present during the gaming development, the *simulation experimental design*, and for the "production" simulations.

6. Perform data analysis of the simulation results. This is a joint effort between the (Dismounted Battlespace Battle Lab, AMSAA and TRAC.

7. The independent evaluators (TRAC and AMSAA) write the final simulation report.

AN EXAMPLE OF THE RESULTS OF A COMBAT SIMULATION

This section will present the results of a live field experiment to determine which optic alternative(s) provides the best probability of hit and the quickest engagement time (Table 1). The best optical alternative data are then used in the JCATS combat model to determine if the better alternative optics provide the small unit force with increased lethality and survivability in a simulated combat environment. As pointed out in the previous example, this is only one of several experiments that were conducted to evaluate the optical alternatives.

OPTICS REQUIREMENT

Experiment: To determine which optic alternative(s) provides the best probability of hit and the quickest engagement time.

Method (A Special Shoot House was constructed at Camp Lejeune, NC):

a. Twenty Marine riflemen were assigned to the experiment. Each rifleman (*Rifle = M16*) used each of three optic/sighting alternatives. The order of use of optic alternative was "randomized."
b. Each Marine was given six rounds, one round to engage each target twice. There were three targets placed at 26 feet, 39 feet, 68 feet respectively.
c. Each rifleman placed on the same ready position line.
d. An electronic timer "beeped" to start the engagement of the first target. The timer recorded time from the beep to the report of the rifle. The shooter did this twice for each target at each range.
e. The engagement times and number of hits was recorded.
f. This was done in day and night conditions.

Results:

1. Average probabilities of hit (across both lighting conditions and three target distances) are given below:
a. Alternative 1; the base case (iron sights in the day/PAQ4 laser pointer for the night) = .88.
b. Alternative 2 = .85
c. Alternative 3 = .87

There was no statistically significant difference between these probabilities of hit.

2. Probability of hit (finer detail)

	Target at 26 ft	Target at 39 ft	Target at 68 ft
Alt 1 (Base) Day	.85	.88	.73
Alt 1 (Base) Night	.98	.93	.93
Alt 2 Day	.98	1.00	.90
Alt 2 Night	.73	.80	.70
Alt 3 Day	1.00	.98	.90
Alt 3 Night	.71	.82	.76

NOTE: Alternatives 2 and 3 provided significantly better probabilities of hit in the daytime but the significantly better probability of hit associated with alternative 1 (base case) used with the PAQ4 at night offsets the daytime advantage of alternatives 1 and 2. <u>The base case is preferred at night</u>.

3. Average Engagement Times (Seconds)
a. Day firings.
(1) Alt 1(base) = 1.51 sec
(2) Alt 2 = 1.47 sec
(3) Alt 3 = 1.86 sec
Both Alt 1 and Alt 2 are significantly less (.05) than Alt 3.

b. Night firings.
(1) Alt 1 (base with PAQ4) = 1.83 sec
(2) Alt 2 = 2.80 sec
(3) Alt 3 = 2.83 sec
Alt 1 is significantly less (.05) than Alt 2 and Alt 3. <u>The base case is preferred at night.</u>

APPLYING THESE RESULTS IN FORCE ON FORCE SIMULATION (FOR INSIGHTS).

1. This was a joint effort between the Army Battle Lab, USMC, TRAC-WSMR and AMSAA.

2. Purpose: Examine the contribution of the alternative optics to small unit lethality and survivability in a simulated combat environment.

3. Model: Joint Conflict and Tactical Simulation (JCATS). This combat model provides high-resolution in urban terrain (in buildings, rooms, and halls) and detailed modeling of small unit tactics.

4. Data authenticity:

a. The Data Authentication Group examined the data from the experiment described above and determined they were sound enough to be used for analysis.

b. AMSAA examined the data from the experiment and determined that the data were suitable for the intended simulation.

5. Case Descriptions

a. Daytime

Base case = M16 with base case (iron sights)
Alternative = Average of Alt 2 and Alt 3 in probability of hit and engagement times.

b. Night

Base case = M16 with PAQ4 laser pointer and Night Vision Goggles.
Alternative = Average of Alt 2 and Alt 3 in probability of hit and engagement times.

6. Abbreviated Scenario Description.

A Marine infantry squad (13-man squad) has to enter and secure a building. The mission ends when 7 Marines or the entire enemy force is killed (This criterion is based on Marine input).

7. Measures of Effectiveness (MOE).

a. Total Threat Losses.
b. Total Friendly Losses.
c. Loss Exchange Ratio (LER). Red losses / Blue Losses

8. The Average Results For Five Iterations of Each Case

Optic/Light conditions	Red Losses	Blue Losses	LER
Base, Day	4.0	4.8	.85
Alt, Day	3.0	6.4	.57
Base, Night	4.0	4.4	.96
Alt, Night	3.6	7.0	.54

9. Wilcoxon Signed Ranks Test Results:

a. Blue Losses between Base and Alternative at Night (results are in favor of the base case); P=.031
b. Resultant LER between Base and Alternative at Night (results are in favor of the base case); P=.031

10. Bottom line

For no MOE did the alternative case force perform better (numerically and statistically) than the base case.

OTHER STATISTICAL WORK IN THE MOUT ACTD

HUMAN RESEARCH AND ENGINEERING DIRECTORATE (HRED)

HRED used statistical design and analyses procedures extensively in the evaluation of human factors implications of all the technologies that were examined in the MOUT ACTD. HRED constructed *all* of the questionnaires and rating scales. These questionnaires were instrumental in the selection of the technologies that were used in the tactical and collective field experiments.

ARMY EVALUATION COMMAND (AEC)

The AEC conducted a separate and independent analysis and evaluation of the experiments conducted in the ACTD. Of particular note was the application of correspondence analysis that combined the results of the Soldiers and Marines rating profiles of the technical candidates (obtained from the questionnaires developed by HRED cited above) into clusters that indicated similarity of ratings.

A FINAL WORD

This paper represents the tip of the statistical and evaluation iceberg associated with the MOUT ACTD. At the present time a comprehensive database is being developed to catalog all of the data generated during the twelve experiments. If any reader of this report would like to know some of the details about these experiments and the resultant data, I will be available to discuss their questions. I can be reached as follows:

Gene Dutoit
(706) 545-5844/7000
DSN: 835-5844/7000
dutoite@benning.army.mil

INTENTIONALLY LEFT BLANK.

Methods to Analyze the Effect of Decreasing Digital Image Resolution on Teledermatology Diagnosis

Paul B. Hshieh, Ph. D.
David Cruess, Ph.D.
Dennis A. Vidmar, MD.
Uniformed Services University of the Health Sciences

ABSTRACT

Teledermatologists concern what level of digital image resolution is good enough to interpret a well taken clinical image. The purpose of this paper is to use different statistical methods to analyze the data and compare their advantages and disadvantages. A total 180 different dermatology slides were divided equally into 3 logical competitor sets (LCSs). Four specific diagnoses (include " other") were in each LCS with three difficult levels. In each slide, 3 digital images of different resolutions were created. The diagnosis with its associate confidence response were obtained from eight staff dermatologists' decision to each LCS. Four statistical methods were used to analyze this data. There are multiple-choice Receiver Operating Characteristic (ROC) curves, logistic regression model (binary response, ignore the level of confidence), cumulative logistic regression (ordinal responses) and baseline-category logits model (nominal response). We consider that the regression model approach has better interpretation than the ROC curve when fits the data appropriately.

INTRODUCTION

Teledermatology consultation is most valuable when applied over great distances under the best of circumstances. Several investigators[1] have shown that good quality digital images can usually substitute for conventional film based clinical photography. Images are transmitted more slowly because the image is not compressed that much (2:1 or 3:1). This requires using a very high-bandwidth line to send the image, which is expensive and not available everywhere. What level of digital image resolution is good enough for a dermatologist to interpret a well taken clinical image? A study to compare the fidelity characteristics of three levels of digital image resolution, Low (600 dpi, 720X500 pixels), Medium (1200 dpi, 950X650 pixels) and High (2000 dpi, 1490X1000 pixels) have been performed by Vidmar & others[2]. The purpose of this paper is to analysis the same data with different statistical methods and comparing their results. We will present four statistical methods (include ROC) to analyze this data, interpret the results, and state the advantages and disadvantages of those methods. **First method**, compare the level of digital image resolutions with multiple-choice Receiver Operating Characteristic (ROC) curves. **Second**, fit the logistic regression model when only consider the readers diagnosis responses and disregard their confidences. **Third**, consider the levels of multiple-choice, described as above, as ordinal responses and fit the cumulative logistic regression models. **Fourth**, consider the levels of multiple-choice as nominal responses and fit the baseline-category logits model. We consider that the regression model approach has better interpretation than the ROC curve when fits the data appropriately.

DESIGN

A total 180 different dermatology slides were obtained from Dr. Douglas Perednia[3] of Oregon Health Science University. These slides were divided equally into 3 logical competitor sets, (LCSs) pigmented lesions (LCS-A), tan or flesh-colored papules (LCS-B), and papulosquamous lesions (LCS-C) each with 60 slides. Within each LCS, there were three specific diagnoses and one non-specific "other" diagnosis. There were 15 distinct slides of each the four possible diagnoses in each LCS. These 15 slides were divided into 3 "easy', 5 "moderately difficult", and 7 "difficult" cases. From each one of the 180 original slides, 3 digital images of different resolutions were created: Low (600 dpi, 720X500 pixels), Medium (1200 dpi, 950X650 pixels) and High (2000 dpi, 1490X1000 pixels). A

white cardboard frame was placed around the edge of the centrally placed image before each session. As a result, the viewer had no cues to determine the resolution of the desktop. Eight staff dermatologists were asked to decide which one of the four diagnoses of each LCS was the correct diagnosis, along with their level of confidence in that diagnosis (very certain, fairly certain, just guessing).

STATISTICAL METHODS

RECEIVER OPERATING CHARACTER (ROC) CURVE TESTING METHOD:

A ROC curve for binary response is a graph that plots the sensitivity verse 1-specificity at different cut-off points. The sensitivity is defined as the probability of a test result is positive when the disease is present, and the specificity is the probability that a test result is negative when the disease is not present. For the ROC curve testing method to multiple-Choice, previous articles by Swets and others have shown that diagnostic subgroups and test images with varying degrees of diagnostic difficulty can be used to tailor ROC analyses to more complex situations. A standard ROC curve to multiple-choice outcomes simply plots the true-positive and false-positive fractions that result from six decision thresholds. The six decision thresholds were defined by the combination of the reader diagnoses (correct or incorrect), any P and N, and confidences (very certain, fairly certain and just guessing), say C, F, and G; that is PC, PF, PG, NG, NF, NC. We also consider six decision thresholds are in ranking order as NC<NF<NG<PG<PF<PC. ROC curves correspond to three level of resolutions were generated by level of difficulties, level of Logical Competitor Sets (LCSs) and combination of them. The area under the ROC curve and Standard error of area are calculated by using a statistical software, MedCalc. The formation of ROC curve for low resolution with easy level for LCS-A, is explained as follow (the other ROC curves will be created by the same way).

Step 1. For each viewer, we create the viewer decision table with a specific diagnosis, MM (row) by decision levels (column) then fill the frequencies in the cells. And the other specific diagnoses (NV,SK, OP) will do the same way as well.

Step 2. Sum all specific diagnoses for each viewer.

Step 3, Sum all viewers, all specific diagnoses, for Low resolution, Easy level of Difficulty, Pigmented lesions (LCS-A).

Step 4, Create a normalized version table as below, and use it to form the ROC curve.

Table: Easy level of difficult , Pigmented lesions (LCS-A), Low resolution

Cumulative Version						
T-DX	**Viewer Decision**					
	PC	**PF**	**PG**	**NG**	**NF**	**NC**
ALL DD+	47	75	79	85	93	96
ALL DD-	3	11	17	41	141	288

Normalized Version						
T-DX	**Viewer Decision**					
	PC	**PF**	**PG**	**NG**	**NF**	**NC**
ALL DD+	0.49	0.781	0.823	0.885	0.969	1
ALL DD-	0.01	0.038	0.059	0.142	0.490	1

Finally, the results indicated that that there is no significant difference among the level of resolutions.

LOGISTIC REGRESSION (LOGIT) MODEL:

Ignore the levels of certainty, and consider the correct and incorrect diagnostic as a binary outcome, then fit the regression model by using SAS " PROC LOGISTIC ". The logistic regression model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots + \beta_k X_{ki}$$, where $i = 1, 2, \ldots\ldots, n$ individuals, π_i probability of correct diagnostic, and $X_1, \ldots\ldots X_k$ are corresponding to covariates, level of difficulties, level of digital image resolutions, readers, and level of LCSs and possible interaction terms.

The deviance, Pearson chi-square and Hosmer and Lemeshow Goodness-of-Fit test all indicate the model fit the data reasonable well and the result also indicate that there is no significant difference among the level of resolutions.

MULTINOMINAL LOGISTIC REGRESSION (MLR) MODEL:

Consider the levels of multiple outcomes as nominal and fit the baseline-category logit model use the j=6 as the baseline. The model can be written as $\log\left(\frac{\pi_j}{\pi_6}\right) = \alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + ... + \beta_{jk}X_k$.

where j=1,2...5. In this model there are two restrictions must be imposed to estimate the equations. One, the probabilities are nonnegative; the other, sum of probabilities must equals to 1.

Under two restrictions, the probabilities can be solved as

$$\pi_1 = \frac{\alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + ... + \beta_{1k}X_k}{1 + \sum_{j=1}^{5}\alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + ... + \beta_{jk}X_k}, \quad \pi_2 = \frac{\alpha_2 + \beta_{211}X_1 + \beta_{22}X_2 + ... + \beta_{21k}X_k}{1 + \sum_{j=1}^{5}\alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + ... + \beta_{jk}X_k}$$

. . .

$$\pi_6 = \frac{1}{1 + \sum_{j=1}^{5}\alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + ... + \beta_{jk}X_k}$$

To estimate the parameters, we obtain the log likelihood equation, then maximize it with numerical method. The SAS procedure, **PROC CATMOD**, will be employed to fit the model. The interpretation will be similar to logistic regression model but modified. The likelihood ratio test is a goodness of fit test, the p-value is 0.1196 indicates the MLR model fits data reasonable well, and there is no significant difference among the level of resolutions.

CUMULATIVE LOGISTIC REGRESSION(CLR) MODEL:

In the MLR model, we consider the multiple-choice responses are unordered categories. When the categories are ordered, it would not be incorrect to simply ignore the ordering and estimate a MLR model. For this experiment, consider the levels of multiple-choice as an order categories, that is NC<NF<NG<PG<PF<PC.
The cumulative logistic regression models is written as

$$\log\left(\frac{p(y \le j)}{1 - p(y \le j)}\right) = \alpha_j + \beta_1 X_1 + ... + \beta_k X_k ,$$

where, there are k independent variables and j =1,2..5 are correspond to order level of (NC<NF<NG<PG<PF) with $p(y \le 1) \le p(y \le 2) \le p(y \le 3) \le p(y \le 4) \le p(y \le 5)$, and $p(y \le 6) = 1$, the PC level is correspond to 6. The CLR model imposes restriction, proportional odd assumption, on the data. The assumption must be tested before accept the model. The interpretation of CLR model is similar to the logistic regression model. The SAS procedure, PROC LOGISTIC, is widely used to fit the CLR model. The p-value of the Score test for the proportional odds assumption is less than 0.0001, which indicate the proportion assumption does not hold. Therefore no further discussion is need.

CONCLUSION

Except Cumulative Logistic Regression (CLR) model does not fit the data, all methods have the same conclusion. For large data set like this, the process to construct the ROC curve is time consuming and should be handled with carefully. It is good and easy for a clinician to see the ROC curve and judge the difference through the eye vision, however, it is impossible to control the confounding when more explanatory variables are involved. We consider that the regression model approach has better interpretation than the ROC curve when the model fits the data appropriately.

REFERENCE

1. Sneiderman CA, Cookson JP and Hood AF. A comparison of video and photographic display of skin lesions for morphology and recognition. J Biocommun. 1991;18:22-25.
2. Vidmar Dennis A., Cruess David , Hshieh Paul, Dolocek Quentin, Pak Hon, Gwynn Marjorie, Mather Mary, Montemorano Andrew, Powers James, Sperling Leonard: The effect of decreasing digital image resolution on teledermatology diagnosis, 1999. (submitted)
3. Perednia, Doulgas A., Gaines, John A., Burtruille, Tony W: Comparison of the clinical informativeness of photographs and digital imaging media with multiple-choice receiver operating characteristic analysis. Arch Dermatol 1995;131:292-297.

Testing for a Difference Between Two Variance Components Obtained from Dependent Data Sets (Clinical Paper)

David W. Webb
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD

Statement of Problem

Over the last 15 years, the U.S. Army research community has tested several proposals (some implemented, some not) intended to reduce the component of total system error attributable to gun tubes. This error component, also known as tube-to-tube dispersion and denoted by σ_T^2, has been shown to play an important role in the delivery accuracy of the M1 series main battle tank. Most of the proposals have been evaluated through live-fire testing, making use of principles of experimental design to control extraneous error sources and statistical inference to draw scientifically valid conclusions.

The most common experimental design used to evaluate treatment effects on σ_T^2 is a nested design (hereafter denoted Design I; see Table 1) whereby $2a$ gun tubes are divided evenly into a control group and a treatment group. Each gun tube fires two or more "occasions"[1] of rounds. To ease computations required of the analysis, Design I is balanced, with each tube firing the same number of occasions and each occasion comprised of the same number of rounds. The response for each round fired is an ordered pair (x,y) representing the horizontal and vertical coordinates of the point of impact on a target. Although the response is bivariate, it has been historically shown that for direct-fire ammunition,

[1] An occasion is loosely defined as a group of rounds fired consecutively and under as near identical conditions as possible. The need for this definition arose over the years as test engineers discovered, for example, that a group of 10 rounds fired consecutively in the morning and a second group of 10 rounds fired from the same vehicle that afternoon often have distinctly different centers of impact.

the horizontal and vertical coordinates are independent. As a result, typical analyses treat the data as independent univariate responses.

Control Tubes				Treatment Tubes		
Tube	Occasion	Rounds		Tube	Occasion	Rounds
1	1	$(x,y)_{1,1,1} \Lambda (x,y)_{1,1,n}$		a+1	ab+1	$(x,y)_{a+1,ab+1,1} \Lambda (x,y)_{a+1,ab+1,n}$
	2	$(x,y)_{1,2,1} \Lambda (x,y)_{1,2,n}$			ab+2	$(x,y)_{a+1,ab+2,1} \Lambda (x,y)_{a+1,ab+2,n}$
	M	M			M	M
	b	$(x,y)_{1,b,1} \Lambda (x,y)_{1,b,n}$			ab+b	$(x,y)_{a+1,ab+b,1} \Lambda (x,y)_{a+1,ab+b,n}$
2	b+1	$(x,y)_{2,b+1,1} \Lambda (x,y)_{2,b+1,n}$		a+2	ab+b+1	$(x,y)_{a+2,ab+b+1,1} \Lambda (x,y)_{a+2,ab+b+1,n}$
	b+2	$(x,y)_{2,b+2,1} \Lambda (x,y)_{2,b+2,n}$			ab+b+2	$(x,y)_{a+2,ab+b+2,1} \Lambda (x,y)_{a+2,ab+b+2,n}$
	M	M			M	M
	2b	$(x,y)_{2,2b,1} \Lambda (x,y)_{2,2b,n}$			ab+2b	$(x,y)_{a+2,ab+2b,1} \Lambda (x,y)_{a+2,ab+2b,n}$
M	M	M		M	M	M
a	(a-1)b+1	$(x,y)_{a,(a-1)b+1,1} \Lambda (x,y)_{a,(a-1)b+1,n}$		2a	(2a-1)b+1	$(x,y)_{2a,(2a-1)b+1,1} \Lambda (x,y)_{2a,(2a-1)b+1,n}$
	(a-1)b+2	$(x,y)_{a,(a-1)b+2,1} \Lambda (x,y)_{a,(a-1)b+2,n}$			(2a-1)b+2	$(x,y)_{2a,(2a-1)b+2,1} \Lambda (x,y)_{2a,(2a-1)b+2,n}$
	M	M			M	M
	ab	$(x,y)_{a,ab,1} \Lambda (x,y)_{a,ab,n}$			2ab	$(x,y)_{2a,2ab,1} \Lambda (x,y)_{2a,2ab,n}$

TABLE 1. Design I frequently used in live-fire testing of gun tubes, using two independent groups of a gun tubes.

Because the control and treatment groups are disjoint, their data are independent. For a given group of tubes, a mathematical model for either coordinate of the impact location is given by the equation:

$$Y_{jkm} = FZ + T_j + O_{k(j)} + R_{m(jk)},$$

where

Y_{jkm} = the impact for the m^{th} projectile fired during the k^{th} occasion from the j^{th} tube;

FZ = the fleet zero, or the overall population mean;

T_j = the effect of the j^{th} tube for $j = 1, 2, \ldots, a$;

$O_{k(j)}$ = the effect of the k^{th} occasion for $k = 1, 2, \ldots, b$; and

$R_{m(jk)}$ = the experimental error, for $k = 1, 2, \ldots, n$.

The terms T_i, $O_{j(i)}$, and $R_{k(ij)}$ are assumed to have normal distributions with mean zero and variances σ_T^2, σ_O^2, and σ_R^2, respectively.

Although several hypotheses may be tested from the data collected under Design I, the one of primary concern to the research engineers is $H_O: \sigma_{T_C}^2 \leq \sigma_{T_U}^2$ versus $H_A: \sigma_{T_C}^2 > \sigma_{T_U}^2$, where the subscripts refer to either the control (C) or treatment (T) group. An exact test of this hypothesis using generalized p-values was originally derived by Zhou and Mathew (1994). The generalized P-value for this hypothesis is given by $P = \Pr(\tau \geq 1)$ where $\tau = \dfrac{ss_{T_U}/\chi^2_{a-1} + ss_{O_C}/\chi^2_{a(b-1)}}{ss_{T_C}/\chi^2_{a-1} + ss_{O_U}/\chi^2_{a(b-1)}}$. In this expression, the SS terms are sums of squares taken from the independent analyses of variance for each tube type and the χ^2 terms are chi-square random variables with degrees of freedom indicated by the subscript. The generalized P-value is estimated by repeatedly generating random values of τ.

Now suppose a another design, denoted hereafter as Design II, calls for a single set of gun tubes to fire before treatment and then again after treatment (see Table 2). The independence that existed between the two halves of Design I is therefore lost in Design II. In addition, Design II is in part characteristic of a factorial design because of the fact that each tube is used in the before- and after-treatment phases of the study.

Similar to Design I, the primary hypotheses of interest are $H_O: \sigma_{T_{before}}^2 \leq \sigma_{T_{after}}^2$ versus $H_A: \sigma_{T_{before}}^2 > \sigma_{T_{after}}^2$, however it is unclear to me how to properly perform this test. I have considered several strategies, none of which seem to test the desired hypothesis. For example, a mathematical model for either coordinate of the impact location may be given by the equation:

$$Y_{ijkm} = FZ + P_i + T_j + PT_{ij} + O_{k(ij)} + R_{m(ijk)},$$

where

Y_{ijkm} = the impact for the m^{th} projectile fired during the k^{th} occasion from the j^{th} tube under the i^{th} treatment;

P_i = the effect of the i^{th} treatment for $i = 1$ (pre-), 2 (post-); and

PT_{ij} = the interaction of the i^{th} treatment and the j^{th} tube;

Before Treatment				After Treatment		
Tube	Occasion	Rounds		Tube	Occasion	Rounds
1	1	$(x,y)_{1,1,1} \Lambda (x,y)_{1,1,n}$		1	ab+1	$(x,y)_{1,ab+1,1} \Lambda (x,y)_{1,ab+1,n}$
	2	$(x,y)_{1,2,1} \Lambda (x,y)_{1,2,n}$			ab+2	$(x,y)_{1,ab+2,1} \Lambda (x,y)_{1,ab+2,n}$
	M	M			M	M
	b	$(x,y)_{1,b,1} \Lambda (x,y)_{1,b,n}$			ab+b	$(x,y)_{1,ab+b,1} \Lambda (x,y)_{1,ab+b,n}$
2	b+1	$(x,y)_{2,b+1,1} \Lambda (x,y)_{2,b+1,n}$		2	ab+b+1	$(x,y)_{2,ab+b+1,1} \Lambda (x,y)_{2,ab+b+1,n}$
	b+2	$(x,y)_{2,b+2,1} \Lambda (x,y)_{2,b+2,n}$			ab+b+2	$(x,y)_{2,ab+b+2,1} \Lambda (x,y)_{2,ab+b+2,n}$
	M	M			M	M
	2b	$(x,y)_{2,2b,1} \Lambda (x,y)_{2,2b,n}$			ab+2b	$(x,y)_{2,ab+2b,1} \Lambda (x,y)_{2,ab+2b,n}$
M	M	M		M	M	M
a	(a-1)b+1	$(x,y)_{a,(a-1)b+1,1} \Lambda (x,y)_{a,(a-1)b+1,n}$		a	(2a-1)b+1	$(x,y)_{a,(2a-1)b+1,1} \Lambda (x,y)_{a,(2a-1)b+1,n}$
	(a-1)b+2	$(x,y)_{a,(a-1)b+2,1} \Lambda (x,y)_{a,(a-1)b+2,n}$			(2a-1)b+2	$(x,y)_{a,(2a-1)b+2,1} \Lambda (x,y)_{a,(2a-1)b+2,n}$
	M	M			M	M
	ab	$(x,y)_{a,ab,1} \Lambda (x,y)_{a,ab,n}$			2ab	$(x,y)_{a,2ab,1} \Lambda (x,y)_{a,2ab,n}$

TABLE 2. Design II proposed for live-fire testing of gun tubes, using the same group of a gun tubes before and after treatment.

The term P_i is a fixed effect so that $P_1 = -P_2$, while PT_{ij} is assumed to have normal distribution with mean zero and variance σ^2_{PT}. All others terms are as defined in Design I. However, the usual prescribed F-tests for each source of variation, including the interaction PT, do not test the desired hypothesis.

Another possible strategy is a multivariate analysis offered by Mathew (1999). In this analysis a bivariate model is used to capture the dependent structure of the test design. The model is given by the equation

$$\begin{pmatrix} Y_{1jkm} \\ Y_{2jkm} \end{pmatrix} = \begin{pmatrix} FZ_1 \\ FZ_2 \end{pmatrix} + \begin{pmatrix} T_{1j} \\ T_{2j} \end{pmatrix} + \begin{pmatrix} O_{1k(j)} \\ O_{2k(j)} \end{pmatrix} + \begin{pmatrix} R_{1m(jk)} \\ R_{2m(jk)} \end{pmatrix},$$

where a "1" in the subscripts signifies "pre-treatment", and "2" signifies "post-treatment". Each term is the multivariate analog of the terms in the univariate model of Design I. If we assume that $\begin{pmatrix} T_{1j} \\ T_{2j} \end{pmatrix}$ is distributed bivariate normal with variance $\begin{pmatrix} \sigma^2_{T_1} & \rho \\ \rho & \sigma^2_{T_2} \end{pmatrix}$, then perhaps a direct test of the desired hypothesis exists. Although I think this may be a viable strategy, I have yet to investigate it in more detail.

In the absence of a better, proven methodology, I continue to use generalized P-values under the mistaken assumption of independent groups of gun tubes to offer preliminary sample size recommendations and power estimates for the engineers involved in this study. I would appreciate comments and/or suggestions from the panel on the aforementioned or other analytical approaches.

References

Mathew, T. Private communications at University of Maryland Baltimore County, Baltimore, MD, September 1999.

Zhou, L. and T. Mathew. "Some Tests for Variance Components using Generalized P-Values." *Technometrics*, Vol. 36, pg. 394-402, 1994.

INTENTIONALLY LEFT BLANK.

On Combining Information from Ordered Experiments

Miguel A. Arcones

Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902, USA

Paul H. Kvam

Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Francisco J. Samaniego

Division of Statistics, University of California, Davis, CA 95616, USA

April 11, 2000

Abstract

For any two random variables X and Y with distributions F and G defined on $[0, \infty)$, X is said to stochastically precede Y if $P(X \leq Y) \geq 1/2$. The applicability of stochastic precedence in reliability modeling and in life and stress testing is discussed. The problem of estimating the underlying distribution(s) of experimental data under the assumption that they obey a stochastic precedence (sp) constraint is examined in detail. Two estimation approaches are used to construct estimators that conform to the sp constraint, and each is noted to lead to a consistent estimator of the underlying distribution. The asymptotic behavior of each of the estimators is described. Evidence is presented, both analytically and via simulation, which demonstrates that the new estimators outperform the corresponding empirical distribution functions (edfs). An application involving life testing experiments in the context of developmental and operational testing in the DoD acquisitions process is treated in the final section. All theoretical results are stated without proof; proofs will be published elsewhere.

---

*Approved for public release; distribution is unlimited.

1 Introduction

The notion that one random variable tends to be larger than another is one that can be quantified in many different ways. Among the best known stochastic relationships in the literature are stochastic ordering, uniform stochastic (or hazard rate) ordering and likelihood ratio ordering, denoted here by $X \leq_{st} Y$, $X \leq_{hr} Y$ and $X \leq_{lr} Y$, respectively. (When $X \sim F$ and $Y \sim G$, we will use the inequality $F \leq G$ as interchangeable with $X \leq Y$.) Definitions and a comprehensive discussion of these and other orderings can be found in the recent monograph by Shaked and Shanthikumar (1994).

The weakest of the orderings mentioned above, namely, $X \leq_{st} Y$, is still too strong an assumption in many problems in which one is inclined to believe that the X population is somehow smaller than the Y population. If F and G are the cumulative distribution functions (cdfs) of X and Y respectively, the stochastic ordering assumption prescribes uniform domination (i.e. $F(t) \geq G(t)$ for all t) of one distribution by the other. While that domination may well hold over an important part of the range of the relevant variables, it may be known to fail over another part of the range (due to infant mortality or planned obsolescence, for example), or may simply be unknown or unknowable over the entire range. Furthermore, stochastic ordering often fails when comparing distributions from different parametric families, and may be quite an unmanageable concept when the two cdfs of interest are not available in closed form. For these reasons, one might wish to entertain the possibility of alternative formulations of the relationship between two random variables.

With this motivation, we introduce the following stochastic relationship as a way of comparing distributions:

Definition: Let X and Y be independent random variables with distributions F and G, respectively. Then the variable X is said to stochastically precede the variable Y if $P(X \leq Y) \geq 1/2$. This relationship will be denoted by $X \leq_{sp} Y$ or, equivalently, by $F \leq_{sp} G$.

Suppose X and Y are independent, continuous random variables, with $X \sim F$ and $Y \sim G$. It is then easily seen that stochastic ordering implies stochastic precedence. If $X \leq_{st} Y$, then $P(X \leq Y) = \int_X (1 - G(x)) dF(x) \geq \int_X (1 - F(x)) dF(x) = 1/2$. It is thus apparent that stochastic precedence is a less restrictive assumption on the relationship between two random variables than stochastic ordering.

The assumption $X \leq_{sp} Y$ is equivalent to the assertion that the median of the variable $Y - X$ is greater than or equal to zero. The relationship is thus seen to be different from,

but of the same ilk as, the more familiar restriction $E(Y - X) \geq 0$. Statistical inference under the latter restriction and its natural generalizations has been studied extensively, and constitutes an important part of the field of order-restricted inference (see Robertson, Wright and Dykstra (1988)).

Interest in the probability $P(X \leq Y)$, where X and Y are independent random variables, has a fairly long history. Birnbaum (1956), for example, considered the problem of the estimating $P(X \leq Y)$ on the basis of two independent samples, advocating a scalar multiple of the Mann-Whitney statistic for this purpose and deriving one-sided confidence intervals based on his estimator. Church and Harris (1970) studied a particular parametric version of this problem arising in reliability, pointing out the relevance of this probability in the modeling of stress-strength relationships. When Y represents the stress placed on a component under test and X represents its (breaking) strength, then $P(X \leq Y)$ is simply the probability that the component fails. Johnson (1988) provides a comprehensive review of work on modeling and inference related to stress-strength testing. While the probability $P(X \leq Y)$ has received a good deal of attention, its utility in ordering the variables X and Y, as in the relationship $X \leq_{sp} Y$ defined above, has not heretofore been carefully explored.

Our main interest here is in nonparametric estimation of the underlying distribution function F when it is known that F obeys a stochastic precedence constraint. Specifically, under the assumption that $F \leq_{sp} G$, we will consider both one and two-sample estimation problems. In the one sample case, the dominating distribution G is treated as known. Our goal is to develop estimators of F (and of G, when appropriate) which obey the postulated *sp* constraint. If the edf satisfies the *sp* constraint, then it will serve as a suitable estimator. The real challenge, of course, is to develop a good estimator in the more typical circumstance in which the edf violates the constraint.

In section 2, we construct an estimator of F that satisfies the stochastic precedence constraint by *shrinking* the sample (generated from F) until the constraint holds. An alternative estimator for F, based on *shifting* the data rather than shrinking it is treated in section 3. Both estimators can be shown to be consistent; their asymptotic behavior will be identified. No proofs are provided here. For proofs, the interested reader is referred to Arcones, M.A., Kvam, P.H. and Samaniego, F.J. (1999). Antecedents for the work presented here include Grenander's (1956) and Marshall and Proschan's (1965) studies on estimating a distribution with monotone failure rate, Boyles and Samaniego's (1984) study on estimating a survival curve under the "new better than used" constraint, Brunk et al. (1966) and Dykstra (1982) on estimation under a stochastic ordering constraint, and Rojo and Samaniego (1991, 1993), Mukerjee (1996) and Arcones and Samaniego (1999)

on estimation under a uniform stochastic ordering constraint. In the final section, the two proposed estimators are applied to simulated data typical of the data available from developmental and operational testing in the context of DoD acquisitions programs.

2 Estimation via Data Rescaling

Let F and G be two continuous distributions on the positive real line, and assume that $F \leq_{sp} G$. If the empirical distribution functions (edfs) based on respective samples from F and G fail to meet the *sp* constraint, we seek to modify one or both edfs until the constraint is achieved. In this section, we do this by minimally rescaling the observations to achieve the *sp* constraint. We'll assume that F and G are continuous with support in $(0, \infty)$. Here, we treat both the one-sample case, where G is assumed known, and the two-sample case, where samples are available from both populations, using the "rescaling" strategy.

We consider the one-sample sase first. Let us assume that G is known and that a random sample $X_1, ..., X_n$ is available from F. The results we derive here hold somewhat more generally than for estimation under a stochastic precedence constraint. We will obtain a consistent estimator of F under the assumption that F satisfies the constraint $E[\phi(X)] \leq 0$, where $\phi(.)$ is an arbitrary non-decreasing function on $[0, \infty)$. When $\phi(x) = G(x-) - 1/2$, the inequality $F \leq_{sp} G$ is equivalent to $E[\phi(X)] \leq 0$. Define

$$\theta_n = \sup\{t \geq 0 : n^{-1} \sum_{j=1}^{n} \phi(tX_j) \leq 0\}. \tag{2.1}$$

We have that $n^{-1} \sum_{j=1}^{n} \phi(\theta_n X_j -) \leq 0 \leq n^{-1} \sum_{j=1}^{n} \phi(\theta_n X_j +)$. Let $\lambda_n = \min(\theta_n, 1)$, and define our estimator of F as a function of λ_n:

$$\hat{F}_1(x) = n^{-1} \sum_{j=1}^{n} I(\lambda_n X_j \leq x). \tag{2.2}$$

By construction, $\hat{F}_1$ stochastically precedes G. The statistic λ_n is the scale factor used to shrink the data. That is, when $\int \phi(x) dF_n(x) > 0$, we multiply the set $X_1, ..., X_n$ by λ_n, with $0 < \lambda_n < 1$, so that $\int \phi(x) d\hat{F}_1(x) = 0$. It is well known that $\{n^{1/2}(F_n(x) - F(x)) : x \in \mathbb{R}\}$ converges weakly to $\{W(F(x)) : x \in \mathbb{R}\}$, where $\{W(u) : 0 \leq u \leq 1\}$ is a Brownian bridge. If $E[\phi(X)] < 0$, then by the law of large numbers $\Pr\{n^{-1} \sum_{j=1}^{n} \phi(X_j) \leq 0\} \to 1$, which implies $\Pr\{\hat{F}_1(x) = F(x), \text{ for each } x\} \to 1$. This fact implies that when the stochastic precedence is strict, (that is, $P(X \leq Y) > 1/2$), $\hat{F}_1$ has the same asymptotic limit as F_n. We record this result as

Theorem 1. If $E[\phi(X)] < 0$, then

$$\{n^{1/2}(\hat{F}_1(x) - F(x)) : x \in I\!R\} \xrightarrow{w} \{W(F(x)) : x \in I\!R\}.\blacksquare$$

If $P(X \leq Y) = 1/2$, the *sp*-constraint can more strongly affect the asymptotic variance of $\hat{F}_1$. The difference is seen in part (ii) and (iii) of the theorem below.

Theorem 2. Let $U_n = n^{-1/2}\sum_{j=1}^{n}(\phi(X_j) - E[\phi(X_j)])$. If $E[\phi^2(X)] < \infty$, then U_n converges in distribution to $U \sim N(0, \mathrm{Var}(\phi(X)))$. In addition, $\{n^{1/2}(F_n(x) - F(x)) : x \in I\!R\}$ and U_n converge jointly to $\{W(F(x)) : x \in I\!R\}$ and U with covariance function $\mathrm{Cov}(W(F(x)), U) = \mathrm{Cov}(I(X \leq x), \phi(X))$. Define $\zeta(t) = E[\phi(tX)]$. If $E[\phi(X)] = 0$ and $\zeta'(1)$ exists and is positive, then

(i) $n^{1/2}(\theta_n - 1) + (\zeta'(1))^{-1}U_n \xrightarrow{\Pr} 0$ and $n^{1/2}(\lambda_n - 1) + (\zeta'(1))^{-1}U_n^{+} \xrightarrow{\Pr} 0$.

(ii) If $\lim_{h\to 1+}(|F(hx) - F(x) - x(h-1)F'(x)|)/(h-1) = 0$, then

$$n^{1/2}(\hat{F}_1(x) - F(x)) \xrightarrow{\mathrm{d}} W(F(x)) + xF'(x)(\zeta'(1))^{-1}U^{+}.$$

(iii) If $\lim_{h\to 1+}\sup_{x\geq 0}(|F(hx) - F(x) - x(h-1)F'(x)|)/(h-1) = 0$, and

$\sup_{x\geq 0} xF'(x) < \infty$, then

$$\{n^{1/2}(\hat{F}_1(x) - F(x)) : x \geq 0\} \xrightarrow{\mathrm{w}} \{W(F(x)) + xF'(x)(\zeta'(1))^{-1}U^{+} : x \geq 0\}. \quad \blacksquare$$

Let's consider the difference in performance of $\hat{F}_1$ and F_n in two specific examples. If X and Y have a Uniform$(0,1)$ distribution, then $\mathrm{MSE}(\hat{F}_1(x)) = x(1-x) + x^2(x - 5/6)$. It follows that this MSE is smaller than $\mathrm{MSE}(F_n(x))$ for $x < 5/6$. The integrated mean squared error of $\hat{F}_1$, defined as $\mathrm{IMSE}(\hat{F}_1) = \int \mathrm{MSE}(\hat{F}_1(x))dF(x) = 5/36$, is slightly smaller than that of F_n, $1/6$. If X and Y are distributed as exponential with mean $\mu = 1$, then $MSE(\hat{F}_1) = (1 - e^{-x})e^{-x} + 2xe^{-2x}(x/3 - 1 + e^{-x})$. In this case, $MSE(\hat{F}_1) \leq MSE(F_n)$ for all $x < 2.8214$, which is approximately $x_{0.94}$, the 0.94 quantile of F. The integrated mean squared error of $\hat{F}_1$ is 77/648=0.1183, again less than that of F_n. These examples show that the available improvement with $\hat{F}_1$ depend on the underlying distributions of F and G. While uniform improvement over F_n cannot be guaranteed, we see from the above that $\hat{F}_1$ can offer improvement upon the MSE of F_n over a large portion of the effective support set of the distribution F as well as according to the global measure IMSE.

The two-sample case is, as might be expected, somewhat more complex. Let us now consider the estimation of F in the case in which G is also unknown. We assume that an independent random sample $Y_1, ..., Y_m$ from G is available, along with the original sample $X_1, ..., X_n$ from F. In this case, we have to estimate two distribution functions simultaneously. Let F_n and let G_m be the empirical distributions based on $X_1, ..., X_n$ and on $Y_1, ..., Y_m$, respectively.

Let $H(t) = P(Y < tX)$, and define $\hat{H}_n(t) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} I(Y_i < tX_j)$. Analogous to (2.1), define

$$\hat{\theta}_n = \sup\{t \geq 0 : \hat{H}_n(t) \leq 1/2\}, \tag{2.3}$$

and let $\lambda_n = \min(\theta_n, 1)$. Given $0 \leq t \leq 1$, we define the two–sample estimators of F and G to be $\hat{F}_{1,t}(x) = n^{-1} \sum_{j=1}^{n} I(\lambda_n^{1-t} X_j \leq x)$ and $\hat{G}_{1,t}(x) = m^{-1} \sum_{j=1}^{m} I(\lambda_n^{-t} Y_j \leq x)$, respectively. Observe that $\hat{H}_n(t) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} I(\lambda_n^{-t} Y_j < \lambda_n^{1-t} X_j)$. Hence, for each $0 \leq t \leq 1$, $\hat{F}_{1,t}(x) \leq_{sp} \hat{G}_{1,t}(x)$. At $t = 0$, we achieve the *sp* constraint by rescaling only the sample from F. At $t = 1$, only the sample from G is rescaled, and for values $t \in (0, 1)$, both samples are simultaneously rescaled.

When the stochastic precedence between F and G is strict, the large sample behavior of these two estimators is described in the following result, stated without proof.

Theorem 3. If $H(1) < 1/2$ and $m, n \to \infty$, then, for each $0 \leq t \leq 1$,

$$\{n^{1/2}(\hat{F}_{1,t}(x) - F(x)) : x \in I\!R\} \xrightarrow{w} \{W_1(F(x)) : x \in I\!R\}$$

and

$$\{m^{1/2}(\hat{G}_{1,t}(x) - G(x)) : x \in I\!R\} \xrightarrow{w} \{W_2(G(x)) : x \in I\!R\}.$$

The behavior of $\hat{F}_{1,t}$ and $\hat{G}_{1,t}$when $F =_{sp} G$ must be described by considering a collection of subcases. In general, the limiting distribution is non Gaussian, but can be identified explicitly. The theorem below is an example of results of this type. Detailed statements and proofs may be found in Arcones, Kvam and Samaniego (1999).

Theorem 4. Suppose that $H(1) = 1/2$, $H'(1) > 0$, $m, n \to \infty$, $\mathrm{Var}(G(X-)) > 0$ and $\mathrm{Var}(F(Y)) > 0$. If

$$\lim_{h \to 1+} |F(hx) - F(x) - x(h-1)F'(x)|/(h-1) = 0,$$

$$\lim_{h \to 1+} |G(hx) - G(x) - x(h-1)G'(x)|/(h-1) = 0$$

and $n/m \to c$ for some $0 \leq c < \infty$, then

$$n^{1/2}(\hat{F}_{1,t}(x) - F(x))$$
$$\xrightarrow{d} W_1(F(x)) + (1-t)(\mathrm{Var}(G(X-)) + c\mathrm{Var}(F(Y)))^{1/2} x F'(x)(H'(1))^{-1}U^+$$

and

$$n^{1/2}(\hat{G}_{1,t}(x) - G(x))$$
$$\xrightarrow{d} c^{1/2}W_2(G(x)) - t(\mathrm{Var}(G(X-)) + c\mathrm{Var}(F(Y)))^{1/2} x G'(x)(H'(1))^{-1}U^+.$$

3 Estimation via Data Translation

In this section, we present an alternative estimator for F (denoted by $\hat{F}_2$) based on transforming the data with a *location* rather than a scale change to achieve stochastic precedence. If needed, the data $X_1, ..., X_n$ are minimally shifted by some constant amount to the left until the edf based on the shifted data stochastically precedes G. In the two sample case where G is also unknown, we simultaneously shift the data $Y_1, ..., Y_m$ (from G) by a constant to the *right* until the *sp*-constraint holds.

While the methods employed in the present section can be applied to problems involving positive random variables (on which we focused in Section 3), they also apply more broadly. Here, we assume only that $F \leq_{sp} G$, with F and G being continuous cdfs on the real line. We treat one- and two-sample problems below.

Focusing first on the one-sample case, let us assume that G is known and that a random sample $X_1, ..., X_n$ is available from F. As before, the results we derive here hold somewhat more generally than for estimation under a stochastic precedence constraint. We will obtain a consistent estimator of F under the assumption that F satisfies the constraint $E[\phi(X)] \leq 0$, where $\phi(.)$ is an arbitrary non-decreasing function on $[0, \infty)$. As previously mentioned, when $\phi(x) = G(x^-) - 1/2$, the inequality $F \leq_{sp} G$ is equivalent to $E[\phi(X)] \leq 0$. Define

$$\theta_n = \sup\{t \in I\!R : n^{-1}\sum_{j=1}^{n} \phi(t + X_j) \leq 0\}. \tag{3.4}$$

We have that $n^{-1}\sum_{j=1}^{n} \phi(\theta_n + X_j^-) \leq 0 \leq n^{-1}\sum_{j=1}^{n} \phi(\theta_n + X_j+)$. Let $\lambda_n = \min(\theta_n, 0)$, and define our estimator of F as a function of λ_n:

$$\hat{F}_2(x) = n^{-1}\sum_{j=1}^{n} I(\lambda_n + X_j \leq x). \tag{3.5}$$

By shifting the data an amount λ_n, we have $\hat{F}_2$ stochastically preceding G. The location-shift statistic λ_n is analogous to the scale-shift statistic λ_n from Section 3.1. The properties of the estimators are similar, as well, but they are not identical in any case of interest. This fact is made clear in the theorems below.

Theorem 5 below states that if stochastic precedence is strict, $\hat{F}_2$ has the same asymptotic limit as F_n. Theorem 6 examines the asymptotic limit of $\hat{F}_2$ in the case that stochastic precedence is not strict.

Theorem 5. If $E[\phi(X)] < 0$, then

$$\{n^{1/2}(\hat{F}_2(x) - F(x)) : x \in I\!R\} \xrightarrow{w} \{W(F(x)) : x \in I\!R\}.$$

Theorem 6. Define $\zeta(t) = E[\phi(t + X)]$. If $E[\phi(X)] = 0$ and $\zeta'(0) > 0$, then

(i) $n^{1/2}\theta_n + (\zeta'(0))^{-1}U_n \xrightarrow{\Pr} 0$ and $n^{1/2}\lambda_n + (\zeta'(0))^{-1}U_n^+ \xrightarrow{\Pr} 0$.

(ii) If $\lim_{h\to 1+}(|F(x+h) - F(x) - hF'(x)|)/h = 0$, then

$$n^{1/2}(\hat{F}_2(x) - F(x)) \xrightarrow{\mathrm{d}} W(F(x)) + F'(x)(\zeta'(1))^{-1}U^+.$$

(iii) If $\lim_{h\to 1+} \sup_{x\geq 0}(|F(x+h) - F(x) - hF'(x)|)/h = 0$, and $\sup_{x\geq 0} F'(x) < \infty$, then

$$\{n^{1/2}(\hat{F}_2(x) - F_n(x)) : x \geq 0\} \xrightarrow{\mathrm{w}} \{W(F(x)) + F'(x)(\zeta'(1))^{-1}U^+ : x \in I\!R\}.$$

The difference between $\hat{F}_1$ and $\hat{F}_2$ is best appreciated through an example. The two estimators, and the edf F_n, all based on samples of size ten, are graphed in Figure 1 against the true Weibull distribution with shape parameter α=2 and scale parameter β=1, so that $F(x) = 1 - exp(-x^2)$, $x > 0$. For illustration, we chose $G = F$. The edf, graphed as a solid-line step function in Figure 1, clearly violates the constraint of stochastic precedence for this sample. The scale transformation estimator, $\hat{F}_1$, is graphed as a dashed-line step function while $\hat{F}_2$, is graphed as a dotted-line step function. $\hat{F}_1$ is shifted left of F_n by multiplying all the observed data by 0.8253 to ensure $\hat{F}_1$ stochastically precedes G. $\hat{F}_2$ is shifted to the left by subtracting 0.1732 from each observation. The disagreement between $\hat{F}_1$ and $\hat{F}_2$ is most dramatic at values of x for which $F(x)$ is close to 1.

The asymptotic MSE of the estimator $\hat{F}_2$ neither dominates or is dominated by the asymptotic MSE for the scale-translation estimator. For example, if X and Y have a

Uniform$(0,1)$ distribution, we have from Section 2 that $\mathrm{MSE}(\hat{F}_1(x)) = x(1-x) + x^2(x - (5/6))$. For the location-translation estimator, $\mathrm{MSE}(\hat{F}_2(x)) = x(1-x)/2 + 1/24$. The integrated MSE for $\hat{F}_2 = 1/8$, which is slightly smaller than the IMSE for $\hat{F}_1$. On the other hand, when X and Y have identical exponential distributions, the IMSE for $\hat{F}_2$ is slightly larger than that of $\hat{F}_1$. In the case $\mu = 1$, then $\mathrm{MSE}(\hat{F}_1(x)) = e^{-x}(1-e^{-x}) + xe^{-2x}(2e^{-x} - 2 + x/6)$ and $\mathrm{MSE}(\hat{F}_2(x)) = e^{-x}(1-e^{-x}) + e^{-2x}(e^{-x} - 5/6)$. Here, $\mathrm{IMSE}(\hat{F}_1) = 0.1183 < \mathrm{IMSE}(\hat{F}_1) = 0.1389$.

Turning to the two-sample case, let us assume that an independent random sample $Y_1, ..., Y_m$ from G is available, along with the original sample $X_1, ..., X_n$ from F. We proceed similarly to Section 2, but, in this case, we need not assume that the r.v.s are nonnegative. Let $H(t) = P(Y < t + X)$, and define $\hat{H}(t) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} I(Y_j < t + X_j)$. Analogous to (2.3), define

$$\hat{\theta}_{n,m} = \sup\{t \in I\!R : H(t) \leq 1/2\}, \tag{3.6}$$

and again define $\lambda_{n,m} = \min(\theta_{n,m}, 0)$. We define the two-sample estimator of F and G to be $\hat{F}_{2,t}(x) = n^{-1} \sum_{i=1}^{n} I((1-t)\lambda_{n,m} + X_i \leq x)$ and $\hat{G}_{2,t}(x) = m^{-1} \sum_{j=1}^{m} I(-t\lambda_{n,m} + Y_j \leq x)$. Note that, by definition, $\hat{F}_{2,t}(x) \leq_{sp} \hat{G}_{2,t}(x)$. At $t = 0$, only data from F are shifted (to the left), and at $t = 1$, only data from G are shifted (to the right). For values of $t \in (0,1)$, both samples are shifted. Examples of the large sample behavior of $\hat{F}_{2,t}$ are given in the two results below.

Theorem 7. If $H(0) < 1/2$, and $m, n \to \infty$, then

$$\{n^{1/2}(\hat{F}_{2,t}(x) - F(x)) : x \in I\!R\} \xrightarrow{w} \{W(F(x)) : x \in I\!R\}.$$

Theorem 8. Suppose that $H(0) = 1/2$, $H'(0) > 0$, $m, n \to \infty$, $\mathrm{Var}(G(X-)) > 0$ and $\mathrm{Var}(F(Y)) > 0$.

If

$$\lim_{h \to 0+} \sup_{x \in I\!R} |F(x+h) - F(x) - hF'(x)|/h = 0,$$

$$\lim_{h \to 0+} \sup_{x \in I\!R} |G(x+h) - G(x) - xG'(x)|/h = 0,$$

$\sup_{x \in I\!R} F'(x) < \infty$, $\sup_{x \in I\!R} G'(x) < \infty$, and $n/m \to \infty$, then

$$\{m^{1/2}(\hat{F}_{1,t}(x) - F(x)) : x \in I\!R\} \xrightarrow{\mathrm{w}} \{(1-t)(\mathrm{Var}(F(Y)))^{1/2} xF'(x)(H'(1))^{-1}U^+ : x \in I\!R\}$$

and

$$\{m^{1/2}(\hat{G}_{1,t}(x) - G(x)) \xrightarrow{\mathrm{d}} W_2(G(x)) + t(\mathrm{Var}(F(Y)))^{1/2} xG'(x)(H'(1))^{-1}U^+ : x \in I\!R\}.$$

4 Discussion

The approach we have taken to the estimation of F, given $F \leq_{sp} G$, is unabashedly ad hoc. It is an approach that has considerable intuitive appeal when F and G are continuous. Both estimation methods we have considered can be applied to failure time (i.e., non-negative) data, though we consider this to be the natural domain of applicability of $\hat{F}_1$. For measurement (i.e., real-valued) data, $\hat{F}_2$, based on a change in location, is clearly the more suitable. Since most applications in reliability involve positive random variables, both approaches constitute new and useful techniques for analyzing reliability data when stochastic precedence is a reasonable assumption. Because there is no universally accepted approach to constrained nonparametric estimation, it is common to seek to exploit the specific structure of the constrained class one is working with. Our approach has been to transform the data in a minimal way so that the empirical distribution of the transformed data will satisfy the *sp* constraint. Our theoretical results show that this approach is efficacious, yielding estimators which satisfy the assumed constraint for any sample size and which inherit many of the good properties of F_n (and G_m) asymptotically.

We devote the remainder of this section to a discussion of an application that arises in typical military acquisitions programs. Specifically, let us consider an estimation problem involving data from developmental and operational testing. The standard protocols leading to the production and development of a new military system involve two separate testing phases. While system prototypes are under development, testing is carried out for the purpose of studying the system as well as to allow for defect detection and possible system improvement. Data from the later stages of the DT phase might be viewed as controlled performance testing to determine if the system is ready for independent inspection and ultimate procurement. The second testing phase is carried out by an independent agent whose role it is to estimate system performance under anticipated use conditions and to make recommendations regarding system production. It is often found that DT data, which is typically obtained under carefully controlled conditions, presents a somewhat more optimistic picture of system quality than does the OT data. This suggests that a natural ordering relationship exists between DT and OT data. Simulated samples of size $m = 20$ and $n = 10$ from DT and OT experiments are displayed in the table below. These data were generated independently from distributions F_{OT} and G_{DT} which satisfy the constraint $F_{OT} <_{sp} G_{DT}$.

DT DATA		OT DATA
28.9377	89.8388	44.6414
14.0814	3.4894	25.5603
14.7829	12.4447	14.0053
1.1059	31.5321	25.2872
0.6971	16.8332	4.5816
1.7558	10.3292	16.9442
44.1989	0.9279	5.2248
30.9623	38.9988	21.9833
14.3511	34.8822	5.5573
4.3106	19.0173	34.1653

Table 1. Simulated DT and OT Data

Under the assumption that $F_{OT} \leq_{sp} G_{DT}$, and given the data

$$X_1, ..., X_{10} \overset{iid}{\sim} F_{OT}$$

and

$$Y_1, ..., Y_{20} \overset{iid}{\sim} G_{DT}$$

as displayed in Table 1, we have computed the estimators $\hat{F}_1$ and $\hat{F}_2$ of F_{OT} using the two-sample versions of the data rescaling and the data translation strategies with the parameter t set equal to zero. Specifically, the two estimators we have computed are specified in the equations below

If

$$\frac{1}{mn}\sum_i \sum_j I(X_j \leq Y_i) < \frac{1}{2},$$

let

$$\hat{F}_1(x) = \frac{1}{n}\sum_j I(\lambda_1 X_j \leq x),$$

where

$$\lambda_1 = \sup\left\{t : \frac{1}{mn}\sum_i \sum_j I(tX_j \leq Y_i) \geq \frac{1}{2}\right\}$$

If

$$\frac{1}{mn}\sum_i\sum_j I(X_j \le Y_i) < \frac{1}{2},$$

let

$$\hat{F}_2(x) = \frac{1}{n}\sum_j I(X_j - \lambda_2 \le x),$$

where

$$\lambda_2 = \inf\left\{t : \frac{1}{mn}\sum_i\sum_j I(X_j - t \le Y_i) \ge \frac{1}{2}\right\}.$$

Figure 2 displays the estimator $\hat{F}_1$, plotted against the empirical distributions F_n and G_m. Figure 3 shows $\hat{F}_2$ plotted against F_n and G_m. In both instances, it can be seen that the constrained estimators $\hat{F}_1$ and $\hat{F}_2$ are adjusted to conform with the *sp* constraint, each being situated to the left of F_n. Since the actual data were generated from two exponential distributions which actually satisfy the *sp* constraint, it is both natural and, as our theoretical results show, efficatious, to "correct" the empirical distribution F_n of the OT data so that it (minimally) satisfies the *sp* constraint relative to G_m.

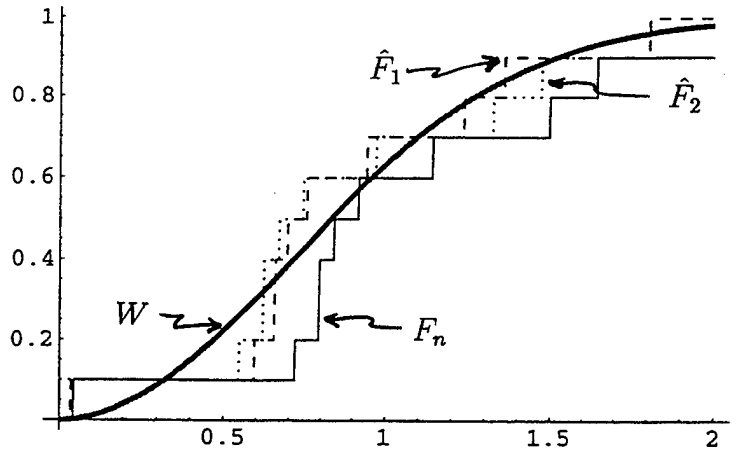


Figure 1: Weibull Distribution Function (gray line) with estimators F_n (solid line), $\hat{F}_1$ (dashed line), and $\hat{F}_2$ (Dotted line) under the constraint of stochastic precedence.

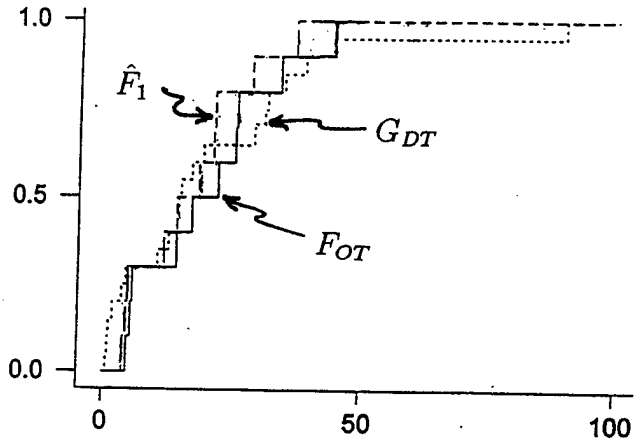


Figure 2: The sp estimator and the edfs F_n and G_n based on shrunken data.

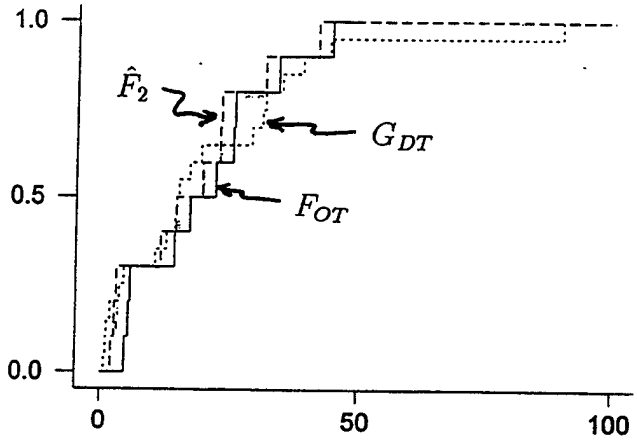


Figure 3: The sp estimator and the edfs F_n and G_n based on translated data.

References

[1] Arcones, M.A. and Samaniego, F.J. (1999) "On the asymptotic distribution theory of a class of consistent estimators of a distribution satisfying a uniform stochastic ordering constraint", *Annals of Statistics*, to appear.

[2] Arcones, M.A., Kvam, P.H., and Samaniego, F.J. (1999)"Nonparametric Estimation of a Distribution Function Subject to a Stochastic Precedence Constraint." Technical Report No. 360, Division of Statistics, University of California, Davis.

[3] Birnbaum, Z. W. (1965) "On a use of the Mann-Whitney statistic". *Proceedings of the Third Berkeley Symposium on Probability and Statistics*, Vol. 1, University of California Press, 13-17.

[4] Boyles, R.A. and Samaniego, F. J. (1984) "Estimating a survival curve when new is better than used", *Operations Research*, 32, 732-40.

[5] Brunk, D.R., Franck, W.E., Hanson, D.L. and Hogg, R.V. (1966) "Maximum likelihood estimation of two distributions of two stochastically ordered random variables", *Journal of the American Statistical Association*, 61, 1067-1080.

[6] Church, J. D. and Harris, B. (1970) "The estimation of reliability from stress-strength relationships", *Technometrics*, 12, 49-54.

[7] Dykstra, R. L. (1982) "Maximum likelihood estimation of the survival functions of stochastically ordered random variables", *Journal of the American Statistical Association*, 77, no. 379, 621 – 627.

[8] Grenander, U. (1956) "On the theory of mortality measurement, part 2", *Scand. Akt.*, 39, 125-53.

[9] Johnson, R. A. (1988) "Stress-strength models for reliability". In *Handbook of Statistics, Vol. 7: Quality Control and Reliability*, Editors: P.R. Krishniah and C. R. Rao, 27-54. Elsevier, New York.

[10] Koroljuk, V. S. and Borovskich, Y. V. (1994) *Theory of U–statistics.* Kluwer Academic Publishers. Dordrecht.

[11] Marshall, A.W. and Proschan, F. (1965) "Maximum likelihood estimation for distributions with monotone failure rate", *Annals of Mathematical Statistics*, 36, 69-77.

[12] Mukerjee, H. (1996) "Estimation of survival functions under uniform stochastic ordering", *Journal of the American Statistical Association*, 91, 1684-89.

[13] Robertson, T., Wright, F.T. and Dykstra, R.L. (1988) *Order Restricted Statistical Inference*, New York: John Wiley and Sons.

[14] Rojo, J. and Samaniego, F. J. (1991), "On nonparametric maximum likelihood estimation of a distribution uniformly stochastically smaller than a standard", *Statistics and Probability Letters*, 11, 267 – 271.

[15] Rojo, J. and Samaniego, F. J. (1993), "On estimating a survival curve subject to a uniform stochastic ordering constraint", *Journal of the American Statistical Association*, 88, no. 422, 566 – 572.

[16] Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and their Applications*, Academic Press, Inc., Boston.

INTENTIONALLY LEFT BLANK.

TIME SERIES INTERVENTION ANALYSES OF U.S. COCAINE PRICES

Samir Soneji, Robert Anthony, Arthur Fries, and Barry Crane
Institute for Defense Analyses (IDA)
1801 N. Beauregard Street
Alexandria, VA 22311
ssoneji@ida.org

ABSTRACT

Previous work at IDA showed that major counter-cocaine interdiction operations in the South American source zone preceded significant price increases of cocaine on U.S. streets. This paper presents an Autoregressive Integrated Moving Average (ARIMA) model of this relationship. The U.S. Drug Enforcement Agency (DEA) has collected and maintained extensive data on the domestic narcotics market, including price, purity, and weight of cocaine purchases by undercover agents, via its System to Retrieve Information from Drug Evidence (STRIDE). We present time series intervention analyses of cocaine price series derived from the STRIDE database and characterize the effects of specific source-zone interdiction operations occurring in South America since 1991. Specifically, we obtain analytical representations of times between interdiction events and associated increases in the price series, the corresponding magnitudes of increase, and the follow-on relaxation times for the market to return to lower prices.

INTRODUCTION

Previous work at IDA established that the street prices of cocaine increase by the month-to-month fluctuations following each major source-zone interdiction operation in the Andean growing countries of South America.[1] Additionally, street purity and all available indicators of usage were demonstrated to be anticorrelated with the street price increases, i.e. as price increases usage decreases. This paper extends that earlier effort to model the effects of major source-zone interdiction operations on street prices. It shows that every significant price increase excursion above a stable floor price can be associated with such an operation.

For nearly 20 years, the U.S. DEA has utilized undercover agents to make cocaine purchases. The DEA determines the price, purity, and weight of the purchased cocaine, and maintain these data and other related information in its STRIDE database. While these purchases by DEA agents were not designed to be a random sample, they have been performed in a consistent enough fashion to extract a meaningful time series of relative changes in street price.[1] As before, our analysis begins by normalizing the STRIDE transaction price for cocaine with adjustments for both purity and weight to produce a unit price:

$$normalized\ unit\ price = \frac{100 \times price\ (\$)}{purity\ (\%) \times weight\ (grams)}. \qquad (1)$$

Because the distribution of normalized unit prices exhibits an extraordinarily heavy right tail, we create a stable street price index that is the *median* of many successive transactions. Whereas IDA formerly took medians of each successive 100 transactions, we take medians for each successive month to accommodate standard time series analyses. There are more than 100 transactions in nearly every month. Hereafter, we use the shorthand "street price" to denote "monthly median normalized unit price."

As shown in Figure 1, cocaine street prices in the U.S. were highest in the early-to-mid 1980s when the cocaine epidemic peaked. Prices declined dramatically as additional illicit businesses competed to supply this drug of choice, finally leveling off at roughly $65 per pure gram in 1989. That year, as part of the Bush "War on Drugs," (WOD) the U.S. military and other government agencies, in conjunction with several Central and South American governments, initiated a complex combination of interdiction operations to curb the production and flow of cocaine. Interdiction activities included eradication of crops, arrests of cartel leaders, counter transport across the Caribbean, and heavy seizures; all contributed to the 1989-90 upward surge in cocaine prices. For several years thereafter, interdictions were executed in the

transit zone (Central America and the Caribbean) and in the source zone (the Andean nations of South America).

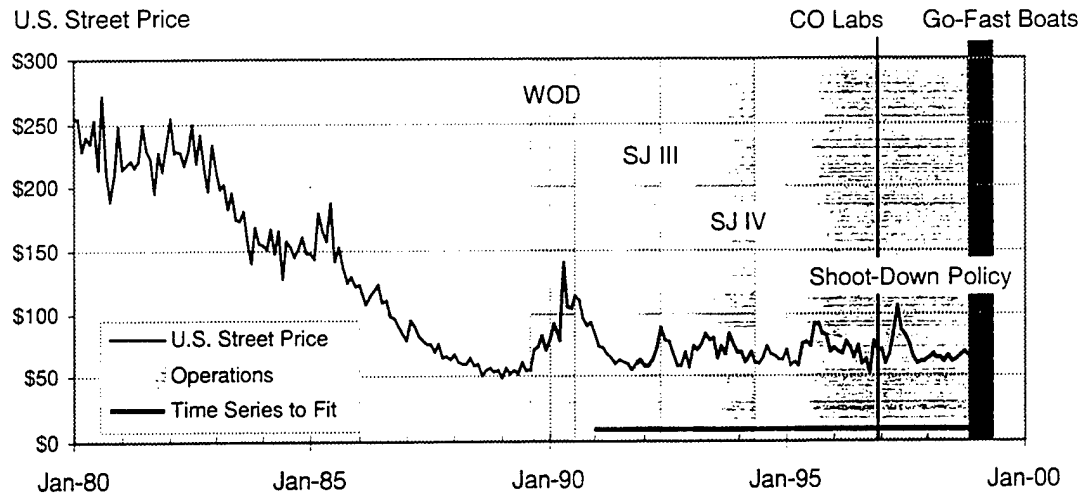


Figure 1. U.S. Street Prices and Interdiction Operations over the Time-Series Analysis Interval

Our analysis focuses on source-zone interdiction operations beyond early 1991 after prices stabilized from the *War on Drugs*. We selected major operations that logically could have disrupted the flow of cocaine by inflicting significant damage on a large portion of the cocaine supply. These interdictions were identified from an exhaustive search and examination of several sources including the U.S. Southern Command, the U.S. embassy in Peru, and other organizations in the Andean countries.[2] We came to focus on the essential elements of the transportation chain and central processing laboratories.

Figure 2 shows the operations we selected in more detail. For example, Operations *Support Justice III (SJ III)*, from November 1991 through April 1992, and *SJ IV*, from January 1993 through April 1994, interdicted trafficker flights of coca base from Peru to Colombia. During Operation *SJ III*, the Peruvian Air Force (FAP) fired upon some trafficker aircraft, but the U.S. ended support soon after the FAP accidentally engaged a U.S. C-130 killing one airman. Operation SJ IV began with much more restrictive rules of engagement, essentially precluding lethal interdiction. Although the interdictions did not stop many trafficker flights, it strongly depressed coca base prices in Peru. After Congressional and Presidential review, the U.S. agreed to resume intelligence, detection, and monitoring support to the FAP under new rules of engagement. Before firing upon suspect aircraft, the FAP were to adequately signal suspected traffickers to land for inspection. In Peru, "shoot-down policy" enforcement began in March 1995, and immediately, most trafficker flights were deterred and coca base prices fell to the degree farmers abandoned their fields. This policy is still in force, and air trafficking has not recovered.

In December 1996 and January 1997, the Colombian Government attacked and destroyed a major cocaine laboratory complex that was estimated to produce about one-third of all of Colombia's cocaine (CO Labs). This event occurred after the publication of the previous IDA report, and shows up in the street price on Figure 2 as a distinct bump in mid-1997. Finally, in November 1998, a joint force engaged the go-fast boat lanes in the western Caribbean – striking the Yucatan bases first and soon afterwards along the Colombian coast. Follow-up operations continue, and these appear as significant price rises near the end of the depicted series.

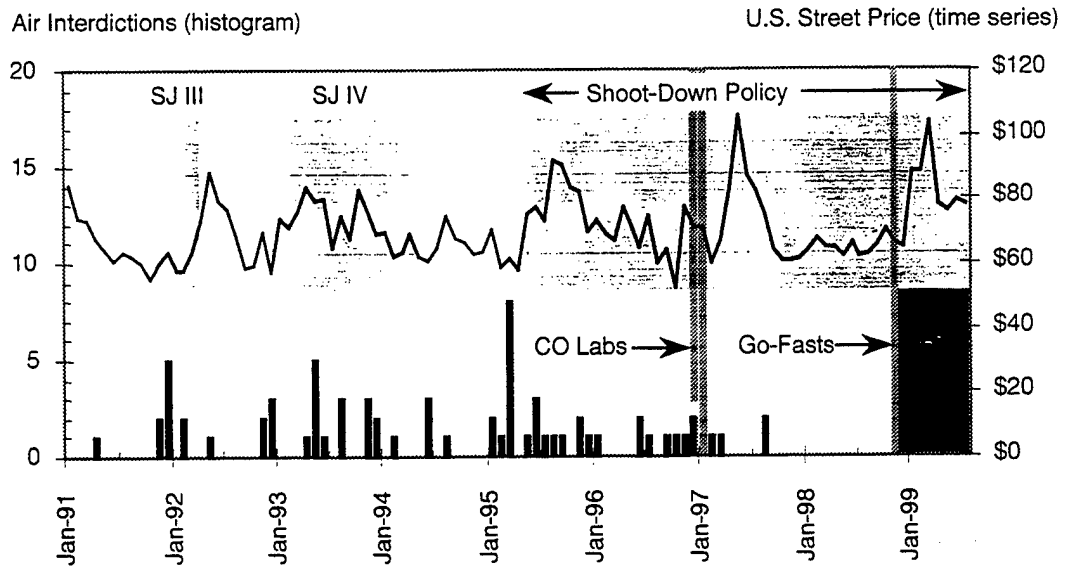


Figure 2. U.S. Street Price, Major Interdiction Operations, and Operational Indicators

We analyze the price series consisting of 112 months from January 1991 through April 1999, spanning the three air interdiction operations in Peru, the Colombian laboratory attacks, and go-fast boat interdictions. We also examine the lag time from the initiation of the 1998 attacks on the go-fast lanes until their impact was felt on U.S. streets. Although the actual relationship between source-zone interdiction operations and effects on street prices might result from numerous factors (e.g., high pilot fees during an operation, complex mechanisms causing price rise, or possible long-tailed relaxation), the limited time period of analyzable data can only support parsimonious forms of modeling. We, therefore, adopt the Autoregressive Integrated Moving Average (ARIMA) process to establish a stationary baseline model and the simplest adequate model formulated in terms of linear operators. These linear operators will represent each of the events shown in Figure 2 as interventions. In this formulation, the ARIMA fit to these interventions will represent a delayed, abrupt price rise followed by a decaying relaxation to the price floor at a street price of $55 to $65.

The remainder of this paper is organized as follows. First we apply ARIMA time series methodologies to derive a "baseline" model without modeling the interdiction activities. We then construct simple intervention expressions to model the specific interdiction events. These intervention expressions are introduced one-at-a-time, i.e., separately for each category of interdiction class relative to the baseline model. Next, we obtain a dynamic regression model formulation that incorporates all statistically significant intervention terms. To establish the impact of the "shoot-down policy," we characterize the differences between effects attributable to interdiction events occurring before and after implementation of the shoot-down policy. Finally, we show the significance of each term within the model and estimate the degree to which source-zone interdiction explains the major features of the street price series.

BASELINE ARIMA TIME SERIES MODEL

Standard ARIMA analyses rely on the assumption one can transform the subject time series into a stationary process whose mean and variance are constant through time, and whose autocorrelations have been eliminated to within the limits of reasonable statistical tests. In order to achieve this requisite first order stationarity, we took differences of the natural logarithms of the price series and corrected for month-to-month autocorrelations. This baseline ARIMA model estimates what the price series would be given all of its *past history*, but does not include intervention events for the interdictions in the *forecasted* month. Our baseline ARIMA(0,1,1) model (exponentially weighted moving average model) obtained for the price series (y_t) is:

$$\hat{z}_t = \ln(y_t) - \ln(y_{t-1}) = (1-\hat{\theta}B)\hat{a}_t = (1 - \underset{(0.08,0.48)}{0.28}\, B)\hat{a}_t\,, \tag{2}$$

where $\hat{\theta} = 0.28$ and the 95% confidence interval for θ is (0.08,0.48). This meets the invertibility condition $|\theta| < 1$ and leads to a simple expression for the one-step ahead forecast:

$$\ln(\hat{y}_{t+1}) = \sum_{i=0}^{\infty}[\hat{\theta}^i \times (1-\hat{\theta}) \times \ln(y_{t-i})] \quad or \quad \hat{y}_{t+1} = e^{\sum_{i=0}^{\infty}[\hat{\theta}^i \times (1-\hat{\theta}) \times \ln(y_{t-i})]}\,, \tag{3}$$

where the weights corresponding to the immediately preceding points rapidly diminish: 0.72, 0.20, 0.06, ... The predicted value for time t_{n+1} under this ARIMA (0,1,1) model is strongly related to the actual value at time t_n and to a much lesser degree to values at earlier times because the autocorrelations are weak. For our problem, all of the intervention processes quench exponentially with infinitesimal impact after at most 6 months.

Figure 3 shows that the forecasted baseline series consistently appears shifted one month to the right when compared to the observed series, a direct result of the exponentially weighted moving average. The ARIMA baseline model, therefore, underestimates the rises to the peaks and overestimates the declines from the peaks because it considers trends in the past history and not immediate impacts from exogenous factors. Our full model will incorporate intervention terms arising from specific interdiction events, and these will be compared to the baseline model using the formal order selection Akaike information criterion (AIC).

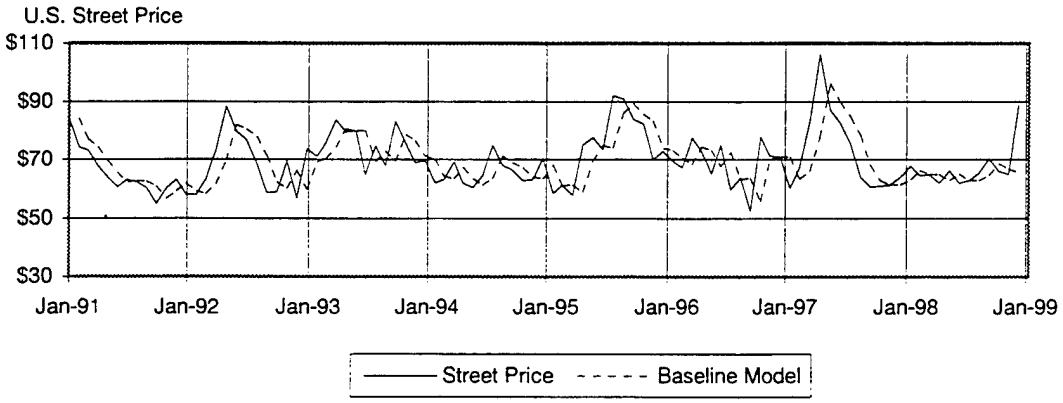


Figure 3. Baseline Model ARIMA (0,1,1) One-Month Forecast

CHARACTERIZATIONS OF SPECIFIC INTERDICTION EVENTS

The framework we use for evaluating hypotheses based on interdiction events employs both the linear transfer function (LTF) and Box-Jenkins Sample Cross Correlation methodologies.[3-5] The LTF method calls for a multi-lag model to obtain the general response pattern and examines the relationship among the estimated weights through maximum likelihood estimation. The Box-Jenkins method examines the cross correlation between time series at different lags and is used to confirm the LTF results.

From observation of Figure 2, we see that, following some time after every interdiction event, street prices first rise and then decay toward the pre-interdiction price. This pattern of damage and recovery can be modeled via the LTF method through first positive and then negative weights. The general form of the dynamic regression model with damage and recovery is then:

$$z_t = C + \frac{\omega_{dam}(B)}{\delta_{dam}(B)} B^{b_{dam}} x_t + \frac{\omega_{rec}(B)}{\delta_{rec}(B)} B^{b_{rec}} x_t + N_t\,.$$

Here, C is a constant term, and N_t is a stochastic disturbance described by an ARIMA process. The numerator $\omega(B)$ represents a pattern of interventions or initial recovery rates, $\delta(B)$ represents the decay process, and B^b captures the respective delay times, where recovery follows damage. X_t is a binary deterministic variable for the pulse interventions that takes the value 1 for a month with an interdiction event and 0 for all others.[3]

For example, the simplest adequate characteristic response for attacks on the Colombian cocaine laboratory complex is:

$$\hat{z}_t = \underset{(0.09,0.37)}{0.23} B^4 x_{lab,t} + \frac{\underset{(-0.20,0.00)}{-0.10}}{1 - \underset{(0.15,1.07)}{0.61}} B^6 x_{lab,t} + (1 - \underset{(0.20,0.56)}{0.38})\hat{a}_t. \quad (4)$$

There is an abrupt increase of 0.23 on the natural-log scale ($e^{0.23} - 1 = 26$ percent in the untransformed true price scale) 4 months after the December 1996 lab events and similarly for the January 1997 events. Starting the two months later, there is an exponentially decaying price decrease with an initial value of –0.10 and a decay constant of 61 percent, corresponding to a nominal 2 months relaxation time. Relative to the general form, the constant term, C, is not significantly different from zero in this and all interdiction classes. All of the other intervention classes share the same pattern of abrupt damage ($\omega(B)=\omega_0$ and $\delta(B)=1$) and gradual recovery ($\omega(B)=\omega_0$ and $\delta(B)=1-\delta_1 B$, $0<\delta_1<1$). The two exceptions are the medium intensity air interdictions before and after the *Shoot Down Policy*. They have no significant exponential decay in the recovery process in the full model.

Rather than choosing broad operational period, e.g., all the months in *SJ III*, we modeled *SJ III* with characteristic response terms similar to (4), but triggering the interventions on those months with air interdictions. We further subdivided the months by classes of interdiction intensity – low, medium, and high. Low intensity months were defined to be those with a single air interdiction. Medium intensity months had two to three interdictions, while high intensity months had four or more.

The *Go Fast Boat Attacks* operation lasted several months in the Caribbean Sea and focused on the transportation capabilities of drug traffickers. We chose a single month (November 1998) to represent the initiations of the damage component of the campaign. We only examined the delay and initial magnitude, however, because the ongoing operation is too complex to model in a parsimonious manner.

THE COMPLETE MODEL

Table 1, below, shows the estimated damage and recovery parameters for each interdiction class fit independently from the other interdiction classes. Estimated 95 percent confidence intervals in the natural log first order difference space accompany each parameter. Many of these are wide due to large standard errors obtained. Note that none of the low intensity events (months with only a single plane interdicted) had a statistical or visual effect on predicting the peaks in the price series. In the majority of classes, the estimated damage and recovery are statistically significant. The current model is not able to adequately capture the weak recovery process for the 9-event *Medium Intensity, Before Shoot Down* class of interdictions.

One of the most compelling pieces of evidence for a causal connection between source-zone interdiction events and price series upward excursions is the consistent 5 month lag time for all five of the independently fit air interdiction classes representing operations in Peru. Since we examined an interval of 6 months for the signature of an up month followed by a down, the chance of all 5 yielding 5 month lags at random is only $(1/6)^5 < 0.0002$. If the associations between source-zone events and street price excursions had been spurious (e.g., the next randomly occurring street price excursion after the event), then we would not expect any consistent value for the lag time. Similarly, the lag time from the Colombian laboratory attacks was 4 months, 1 month closer to the U.S. streets than the Peruvian interdictions – again consistent with the causal hypothesis. Finally, the go-fast events occurring even closer to the United States fit to a lag time of only 2 months.

Table 1. Estimated Damage and Recovery Parameters by Interdiction Class

		Damage		Recovery		
Interdiction Class	Number of Occurrences	Lag	Estimated Impact (95% CI)	Lag	Estimated Initial Recovery (95% CI)	Estimated Exponential Decay Constant (95% CI)
Support Justice III, IV, and Between Operations, Medium Intensity	9	5	0.04 (−0.04,0.12)	6	−0.04 (−0.12,0.04)	−0.06 (−2.10,1.98) deleted because of instability
Support Justice III, High Intensity	1	5	0.26 (0.04,0.48)	6	−0.13 (−0.31,0.05)	0.62 (−0.08,1.30)
Support Justice IV, High Intensity	1	5	0.19 (−0.03,0.41)	6	−0.11 (−0.11,0.31)	0.60 (−0.48,1.52)
After Shoot Down Policy, Medium Intensity	5	5	0.16 (0.00,0.32)	6	−0.08 (0.00,0.16)	0.53 (0.00,1.03)
After Shoot Down Policy, High Intensity	1	5	0.23 (0.01,0.55)	6	−0.07 (−0.07,0.21)	0.80 (0.26,1.34)
Lab Attacks	2	4	0.23 (0.09,0.37)	6	−0.10 (−0.20,0.00)	0.61 (0.15,1.07)
Go Fast Boat Attacks	1	2	0.37 (0.11,0.70)	not modeled	not modeled	not modeled

DYNAMIC REGRESSION MODEL

All of the interdiction classes individually add to the predictive ability of the combined model, except for the two low intensity classes. The AIC value is significantly reduced from −135 for the baseline model to −177 for the final dynamic model indicating a stronger fit. In the estimated LTF model, most estimated parameters were significant at the 95% level. Those parameters that failed to meet individual significance remained in the model to maintain zero mean for the residuals. The only exceptions were two parameters (decays on *Before and After Shoot Down Mediums*) that were deleted due to instability and strong multicollinearity effects. There is significant interaction between the first *After Shoot Down Medium* and the *After Shoot Down High* events because of the interference caused by the slow decay of the high intensity event. The estimated LTF for the final model takes the form:

$$
\begin{aligned}
\hat{z}_t = {} & +0.08B^5 x_{\textit{before shoot down policy medium},t} - 0.07B^6 x_{\textit{before shoot down policy medium},t} \\
& + 0.33B^5 x_{\textit{SJIII high},t} - \frac{0.16}{1-0.54B} B^6 x_{\textit{SJIII high},t} \\
& + 0.18B^5 x_{\textit{SJIV high},t} - \frac{0.13}{1-0.60B} B^6 x_{\textit{SJIV high},t} \\
& + 0.16B^5 x_{\textit{after shoot down policy medium},t} - 0.05B^6 x_{\textit{after shoot down policy medium},t} \\
& + 0.29B^5 x_{\textit{after shoot down policy high},t} - \frac{0.10}{1-0.87B} B^6 x_{\textit{after shoot down policy high},t} \\
& + 0.16B^4 x_{\textit{lab},t} - \frac{0.09}{1-0.67B} B^6 x_{\textit{lab},t} \\
& + 0.31B^2 x_{\textit{boat},t} + (1-0.53B)\hat{a}_t .
\end{aligned}
\tag{7}
$$

Figure 4 shows the improved ability of our final full dynamic regression model to predict all of the peaks in street price (numbered peaks). Rather than present a separate figure for each estimated response function, we display the final fitted curve incorporating *all* of the modeled responses. For example,

comparing the baseline in Figure 3 with the model in Figure 4 reveals how much more faithfully the model traces the price rise and fall following the Colombian laboratory attacks.

There is little difference between the parameters for individual classes (Table 1) and the combined model shown above because the most parsimonious model was chosen to reduce the multicollinearity. Most intervention terms apply only locally about one specific interdiction operation, i.e., intervention terms do not overlap – with two notable exceptions. The high and medium intensity events in March and June 1995 interact with each other and the resulting cumulative response exhibits a double spike (c) in Figure 4. In the second, the damage coefficient for the laboratory attacks drops from 0.23 to 0.16 in the full model due to an interaction with a medium intensity interdiction month.

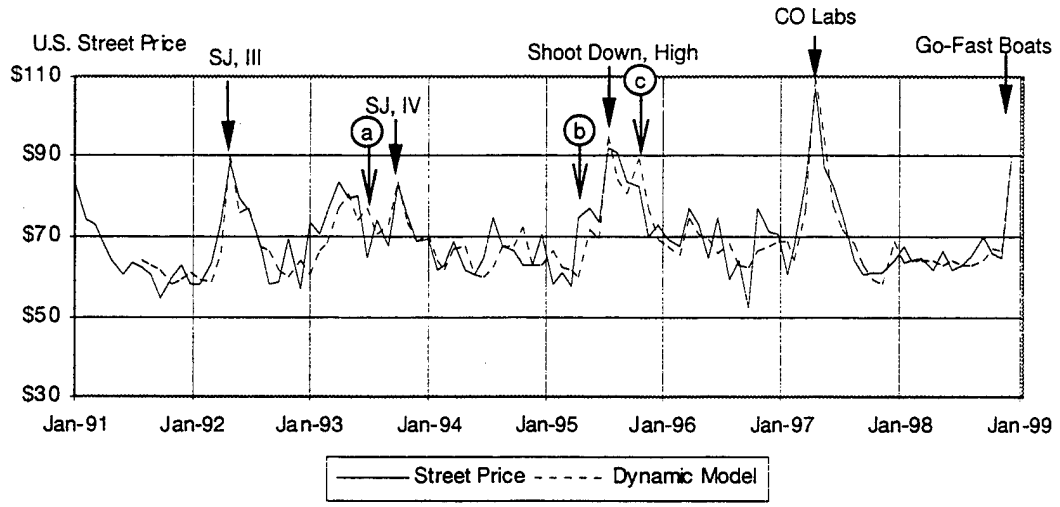


Figure 4. Full Dynamic Regression Model One-Month Future Forecast

The dynamic model also significantly reduces the residuals between the forecasted and observed prices. (Figure 5) First note that both series of residuals appear stationary, and this is born out by our tests of fixed mean, freedom from autocorrelation, and apparently stable variance. The power of our model to reduce all of the strong deviations from the zero mean is also evident in Figure 5. The baseline had 16 excursions outside the ±0.18 range, while the model had only 4 such excursions, most of which were minor by comparison.

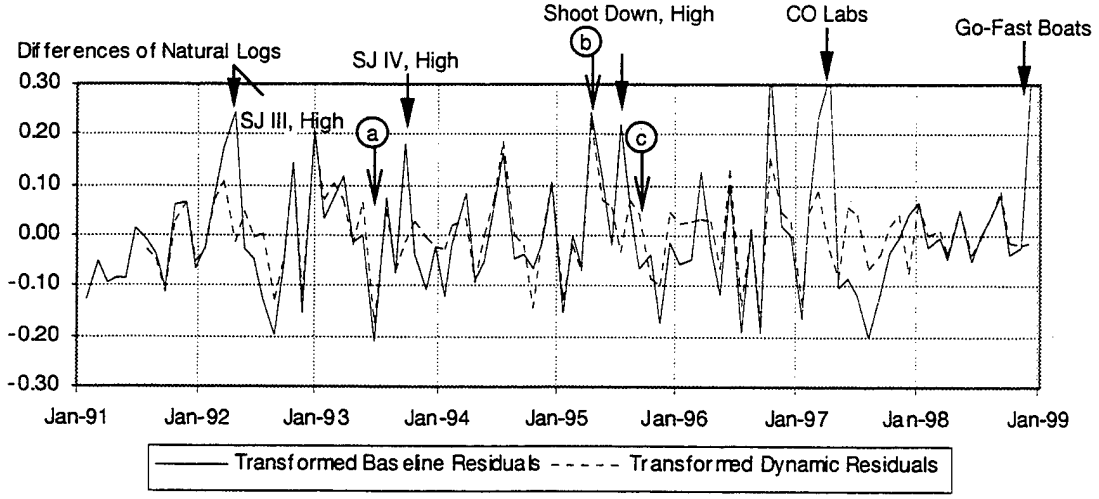


Figure 5. Baseline versus Dynamic Regression Residuals

We can explain two of the excursions as artificialities of our choice of a parsimonious ARIMA representation of the actual interdiction operations in the source-zone countries. At (a), SJ IV began without any medium or high-intensity events although there were two medium events just before SJ IV; therefore, without an "event" representing the intimidation of simply beginning an operation, the model dipped with the data. The second is at (b) just before implementation of the shoot-down policy in Peru but after it was implemented in Colombia. A medium event in Peru two months before implementation was not weighted sufficiently as a Pre-policy event to bring the model forecast up to the actual data.

Below we analyze observed responses associated with high- and medium-level interdiction events, separately for the period prior to the implementation of the shoot-down policy and for the period subsequent to its implementation. After the implementation, the U.S. Government provided detection and monitoring support to the Peruvian interdictor aircraft that were authorized to use lethal force, and these operations were sustained indefinitely. This was decreased the air trafficker flights from Peru to Colombia by over 80 percent.

PRE-SHOOT-DOWN POLICY

For the pre-shoot-down policy period, interdiction effects were moderate for the medium intensity (two to three planes) events, but quite pronounced for the high intensity (four or more planes) events. Thus, there appears to be a threshold somewhere between our "medium" and "high" intensity designations that determines whether there will be a significant effect on U.S. streets. Figure 6 compares the characteristic response of the four pre-policy air interdiction classes. Note that the flat profile for the low-intensity events confirms they are completely insignificant.

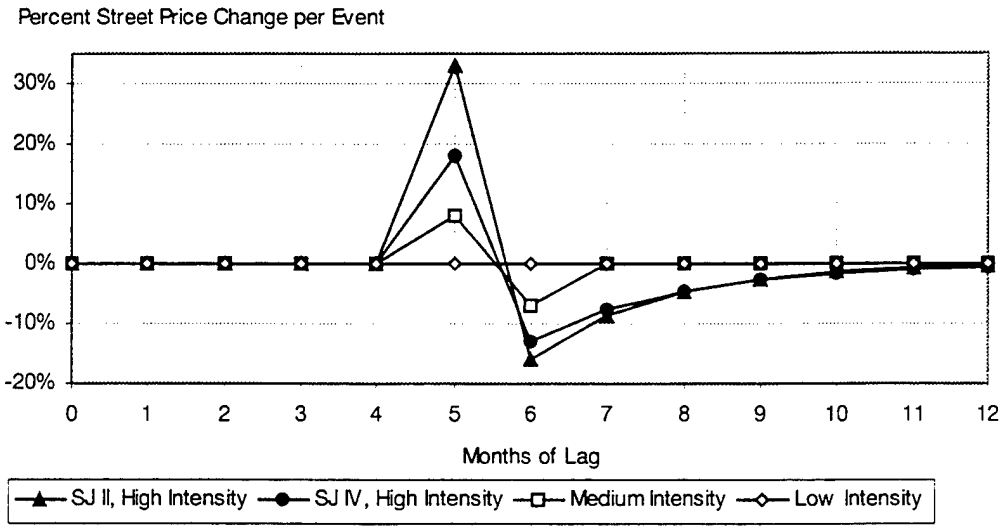


Figure 6. Price Impacts of Pre-Shoot-Down Policy Air Interdiction Classes

The December 1991 (*SJ III*) high intensity lethal event had a strong price effect, 33 percent increase, 5 months after the interdictions. The next month, an exponentially decaying decrease starting with 16 percent recovery and a decay rate with 2 month relaxation time. The same intensity class event in May 1993 (*SJ IV*), but with non-lethal consequences, led to a slightly weaker 18 percent price increase followed by a 13 percent recovery the next month and a similar 2 month relaxation-time exponential decay. The medium intensity events, by contrast, had almost negligible impact on the price series, a modest 8 percent increase 5 months after the interdiction. Medium intensity events had no observable recovery relaxation response while low intensity had no visual or statistically significant effect.

POST-SHOOT-DOWN POLICY

For the post-shoot-down period, the street price impact increases with the air interdiction event intensity. The Table 1 and full model coefficients as well as Figure 7 show the price impacts of the post-shoot-down interdiction classes. After the shoot-down policy, the 5 medium intensity events had two-times the impact as the mediums before the policy. The high-intensity interdiction classes with lethal threat, *SJ III* and the post-*Shoot-Down Policy* operations, had over 60% greater impact as the non-lethal high-intensity interdiction class *SJ IV*. And the recovery from the high-intensity event after the policy implementation was much slower, 4.5 months versus 2 months before. Note that the low intensity class again did not exhibit a statistical or visual impact on street prices.

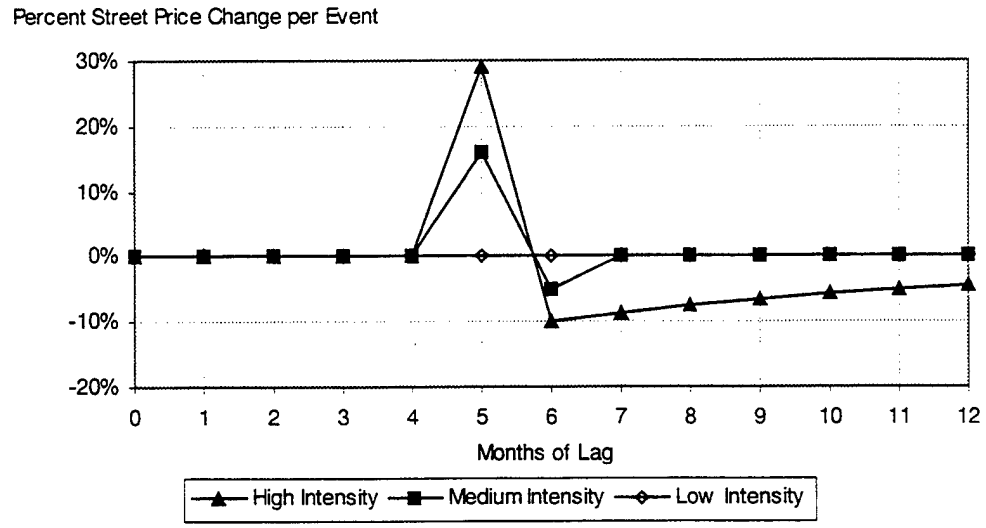


Figure 7. Price Impacts of Post-Shoot-Down Policy Air Interdiction Classes

SUMMARY AND DISCUSSION

A parsimonious ARIMA process representing the impacts of source- and transit-zone counter-cocaine interdiction operations can model the significant upward price movements of the U.S. cocaine street price time series from 1991 to early 1999. We obtained stationary residuals for both the baseline and full models. This modeling provides new operationally useful information, e.g., more accurate estimates for the time lags along the distribution chain from source country to the U.S. streets and approximations to the traffickers' recovery time.

We have demonstrated that the U.S. street price index developed at IDA provides a useful statistic for analyzing the impact of source-zone operations on the cocaine market at the street level in the United States. This finding may also provide the basis for analyzing regional differences within the United States given sufficient sample size. During the analysis period from 1991 to early 1999, the street prices continually return to a floor in the range of $55 to $65. We conjecture that this reflects the minimum price traffickers in the distribution chain require as compensation for the risks of being in this illicit business.

In combination, several findings from our ARIMA modeling results give compelling evidence that source-zone interdictions were responsible for the significant excursions above a stable baseline for U.S. cocaine street prices. First, all seven independently fit and statistically significant event classes yielded completely consistent lag time estimates. All 5 interdiction classes of events in Peru led corresponding price rises in the U.S. by 5 months (probability less than 0.0002 by chance), the laboratory attacks in Colombia led price rises by 4 months, and the attacks on go-fast lanes in the western Caribbean led price rises by 2 months.

Second, the quantitative intensities associated with each interdiction class logically followed the progression of operational intensities. For air interdictions in Peru, high intensity months (with four or more interdictions) had larger fitted impact coefficients than medium intensity months (with 2 or 3 interdictions). Low intensity months (with only one interdiction) did not have statistically significant fitted coefficients. Months in which trafficker pilots might be shot down if they did not land for inspection had larger fitted coefficients than months with non-lethal consequences for non-compliance. In fact, lethal medium-intensity interdictions after the implementation of the shoot-down policy had five-times the impact as medium intensity events before the policy. Finally, the post-policy interdiction events had much slower trafficker recovery times: 4.5 months versus 2 months for high-intensity events and 2 months versus none at all for medium-intensity events.

Third, our model explains *all* significant peaks in the street price that occurred between 1991 to early 1999 in terms of transit- and source-zone interdiction operations. For example, the first event, *Support Justice III*, followed an 8-month period of stable prices. Also, there was a 15-month period of stable street prices between the end of the affects of the attacks on Colombian cocaine labs and the next attack, those on go-fast lanes in the western Caribbean. If there were other reasons for significant price excursions, these quiescent periods offered ample opportunity for them to exhibit their effects.

Areas for potential future research include ARIMA univariate and multivariate modeling of cocaine street purity time series because decrements of purity persisted after the shoot-down policy and dropped even more after the recent go-fast attacks. There is evidence for complex seasonal components in the residuals from our model, which might have operational utility. Finally, multivariate analyses might provide an integrated and more sensitive measure of the impact on the illicit cocaine business by integrating several independent indicators of cocaine use described in IDA's previous report.[1]

ACKNOWLEDGEMENTS

Portions of the authors' research were conducted at the Institute for Defense Analyses (IDA) under tasks sponsored by the U.S. Office of the Secretary of Defense (OSD), Assistant Secretary of Defense/Special Operations and Low Intensity Conflict and the Office of Drug Enforcement Policy and Support. In addition, the IDA Summer Intern Program funded part of Samir Soneji's work.

The views expressed in this paper are solely those of the authors. No official endorsement by IDA, OSD or the cited OSD offices is intended or should be inferred.

REFERENCES

1. Crane B., Rivolo R., and Comfort G. "An Empirical Examination of Counterdrug Interdiction Program Effectiveness," Paper P-3219, Institute for Defense Analyses, Alexandria, Virginia, January 1997.

2. Anthony R., Crane B., and Hanson S. "Deterrence Effects and Peru's Force-Down/Shoot-Down Policy," Paper P-3472, Institute for Defense Analyses, Alexandria, Virginia, February 2000.

3. Pankratz, A. Forecasting with Dynamic Regression Models, John Wiley and Sons, New York, New York, 1991.

4. Bowerman B. and O'Connell R. Forecasting and Time Series: An Applied Approach, Duxbery Press, Belmont, California, 1993.

5. Enders, W. Applied Econometric Time Series, John Wiley and Sons, New York, New York, 1995.

PROBLEMS OF CORRELATION IN THE PROBABILISTIC APPROACH TO COST ANALYSIS

Stephen A. Book
The Aerospace Corporation
El Segundo, CA 90245

ABSTRACT

Since the probabilistic approach to cost analysis became *de rigueur* over the past decade in responding to DoD-issued requests for proposals (RFPs), it has been realized that correlation among element cost distributions contributes significantly to the total cost's uncertainty. Unfortunately for cost analysts, correlation turned out to be a more sophisticated mathematical concept than appeared at first glance. Except in the case of uniquely malleable normal distributions, the correlation coefficient ρ cannot range freely over the range $-1.00 \leq \rho \leq 1.00$ irrespective of the way marginal program-element cost distributions intertwine to form their multivariate distribution. Therefore a cost analyst cannot be sure that his or her *a priori* choice of numerical values for inter-element correlation coefficients can actually be attained in practice. One well-known case involves E.J. Gumbel's "first bivariate exponential distribution" (1960), in which the correlation between the two marginal distributions is restricted to the range $-0.40365... \leq \rho \leq 0.00$. In the case of a perennial favorite of cost analysts, the correlation between two lognormal random variables associated with correlated standard normal variables cannot be less than $-0.367879... = (e^{-1}-1)/(e-1)$. Further complicating the situation is the impossibility of expressing the total-cost probability distribution in closed analytic form and the resultant need to model it by Monte Carlo simulation of the sum of pairwise correlated random variables. As it happens, generating the required sequences of correlated random numbers in support of this objective is a nontrivial problem. In addition, availability of several commercial software products that purport to circumvent the problem by providing sequences of rank-correlated random numbers has turned out to be a distraction standing in the way of a concerted effort to solve the problem properly. The present report describes the state of the art in modeling total-cost probability distributions as statistical sums of correlated program-element cost distributions.

THE COST-RISK IMPERATIVE

While it is agreed that the "best" estimate of system cost should accompany a program proposal up the DoD funding chain, no agreement exists as to what "best" means: is it the "most likely" cost, the 50^{th}-percentile cost, the "average" cost? The typical "accounting" approach is to estimate the most likely cost of each program element and then to sum ("roll up") those estimates. Preponderance of high-end risks over low-end uncertainties (especially in technical development programs), though, causes the roll-up estimate to underestimate actual total cost by a wide margin, inducing cost overruns *attributable purely to the mathematics* of the procedure. For a justification of this fact, see Reference 2. DoD's solution to this problem is to require the cost of <u>every</u> element to be modeled as a random variable, with probability distributions defined for all elements [Reference 1, page 2-5], correlations or functional relationships between them estimated, and the distributions summed statistically by Monte Carlo sampling. The result is a probability distribution of total system cost, from which meaningful estimates of the median, 70th percentile, and other relevant quantities can be determined.

"Cost-risk analysis" is the term used by cost analysts to refer to any procedure by which cost estimates are provided in the form of random variables rather than deterministic numbers. The name derives from the fact that, for any deterministic cost estimate, there is some degree of risk that the program will be unable to be completed and meet its stated objective at that particular funding level. Cost-risk analysis recognizes that there is a probability of success associated with each fixed cost estimate, a fact consistent with the representation of program cost as a random variable.

Representation of program cost as a random variable for the purpose of appropriately modeling uncertainty carries with it the assumption that understanding the degree of uncertainty in an estimate is a necessary aspect of cost analysis. Perhaps the most universal statistical descriptor of a random variable's degree of uncertainty is its standard deviation (or equivalently its square, the variance). It is this fact that makes correlation between program-element costs a critical factor in the estimation of total program cost.

Suppose there are n separate program elements that comprise the total program and that their costs are, respectively, the random variables $X_1, X_2, \ldots, X_n$. Total program cost is then the sum of those random variables, namely

$$S_n = \sum_{k=1}^{n} X_k \ . \tag{1}$$

Uncertainty in total program cost is measured by its variance:

$$Var(S_n) = \sum_{k=1}^{n} \sigma_k^2 + 2\sum_{j=2}^{n} \sum_{i=1}^{j-1} \rho_{ij}\sigma_i\sigma_j \tag{2}$$

where $\{\rho_{ij}: 1 \le i \le n, 1 \le j \le n, i \ne j\}$ are the pairwise correlations between pairs (X_i, X_j) of program-element costs, and $\{\sigma_i: 1 \le i \le n\}$ are the standard deviations of the costs $\{X_i: 1 \le i \le n\}$, respectively.

WHY INTER-ELEMENT COST CORRELATIONS ARE NOT ZERO

Consider the case of a space system. At the highest level of the "work-breakdown structure" (WBS), which is simply the list of all program elements that have costs associated with them, there are just three elements: space-resident satellites, a launch system, and a ground-based control and data-analysis system. The respective costs X_S, X_L, and X_G may be positively correlated for several reasons, among which are (1) an increase in sizes, weights, and numbers of satellites to be orbited induces an increase in launch costs, either through the number of launches required or capability of the individual launch vehicles, and (2) an increase in number and data-gathering capability of the satellites forces an increase in ground-operations costs, either through the complexity of the tasking and control system software or the number and size of ground-station facilities. On the other hand, X_S, X_L, and X_G may very well be negatively correlated for different reasons, one of which is that reducing the complexity of on-board satellite software and communications hardware may tend to increase ground costs by complicating the ground software, while decreasing launch costs due to reduction in size of on-orbit hardware.

Further down into the WBS, costs are more highly correlated because they correspond to specific items that physically occupy adjacent locations within the satellites or ground stations or to software packages that operate specific pieces of hardware.

WHY CORRELATION MATTERS IN COST ANALYSIS

For a WBS comprised of n program elements having costs, the variance of total-program cost S_n is as given in Equation (2). If program-element costs were in fact all pairwise uncorrelated, namely all $\rho_{ij} = 0$, the variance of total-program cost would be

$$Var(S_n) = \sum_{k=1}^{n} \sigma_k^2 \tag{3}$$

At the other extreme, if all inter-element correlations were equal to +1

$$Var(S_n) = \sum_{k=1}^{n} \sigma_k^2 + 2\sum_{j=2}^{n} \sum_{i=1}^{j-1} \sigma_i\sigma_j = \left(\sum_{k=1}^{n} \sigma_k\right)^2 \tag{4}$$

Because "ignoring" the issue of correlation when estimating total-program cost is tantamount to setting all $\rho_{ij} = 0$, the following order relationships imply that such a tactic will lead to underestimating actual program cost uncertainty:

$$\sum_{k=1}^{n}\sigma_k^2 \le \sum_{k=1}^{n}\sigma_k^2 + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1}\sigma_i\sigma_j \le \left(\sum_{k=1}^{n}\sigma_k\right)^2. \tag{5}$$

To understand the significance of the error due to ignoring correlation, it is important to get an intuitive "feel" for the magnitude of the impact of inter-element correlation on total-program cost uncertainty. To do this, let's look at a simple, yet instructive, scenario. Suppose each $\sigma_k = \sigma$, a constant, for $1 \le k \le n$, and $\rho_{ij} = \rho$ for $1 \le i \le n$, $1 \le j \le n$. Then

$$Var(S_n) = n\sigma^2 + n(n-1)\rho\sigma^2 = n\sigma^2[1+(n-1)\rho] \tag{6}$$

At its two extremes over the interval $0 \le \rho \le 1$, the total-cost variance is

$$\begin{aligned} Var(S_n) &= n\sigma^2 \quad when\ \rho = 0 \\ &= n^2\sigma^2 \quad when\ \rho = 1 \end{aligned} \tag{7}$$

As can be seen, $Var(S_n)$ spans a wide range of possible values, the high-end value equal to n times the low-end value. This means that the larger number of elements in the WBS, the greater the underestimation of the total-cost variance attributable to ignoring correlation.

A more precise way of expressing the impact of correlation on total-cost uncertainty is to calculate two ratios: (1) $100u\%$, the percentage underestimation of the total-cost standard deviation that is attributable to setting all correlations zero when, in fact, they are equal to $\rho > 0$, and (2) $100v\%$, the percentage overestimation of the total-cost standard deviation that ensues when all correlations are set to 1 instead of to their correct value ρ. Expressions for u and v are as follows:

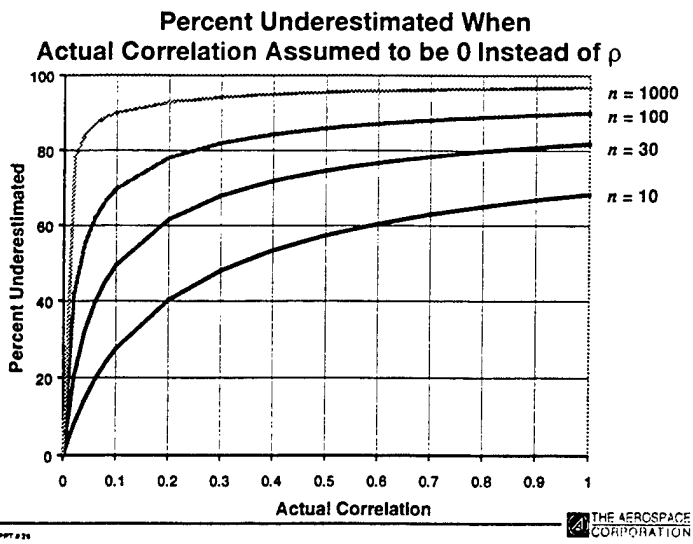
$$u = \frac{\sqrt{n\sigma^2[1+(n-1)\rho]} - \sqrt{n\sigma^2}}{\sqrt{n\sigma^2[1+(n-1)\rho]}} = 1 - \frac{1}{\sqrt{1+(n-1)\rho}} \tag{8}$$

and

$$v = \frac{\sqrt{n^2\sigma^2} - \sqrt{n\sigma^2[1+(n-1)\rho]}}{\sqrt{n\sigma^2[1+(n-1)\rho]}} = \sqrt{\frac{n}{1+(n-1)\rho}} - 1 \tag{9}$$

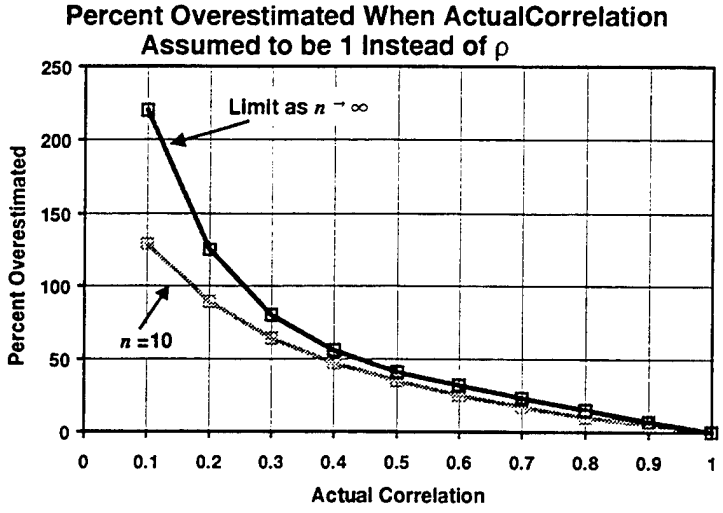
The percentages $100u\%$ and $100v\%$ are graphed for various values of ρ and n in Figures 1 and 2, respectively. In looking at these graphs, it is important to remember that nothing can ever be underestimated by more than 100%, but the possible extent of overestimation is unbounded. A glance at these graphs provides convincing evidence that correlation matters in establishing the extent of uncertainty in a cost estimate.

Figure 1
Maximum Possible Underestimation of Total-Cost Sigma



THE AEROSPACE CORPORATION

Figure 2
Maximum Possible Overestimation of Total-Cost Sigma



THE AEROSPACE CORPORATION

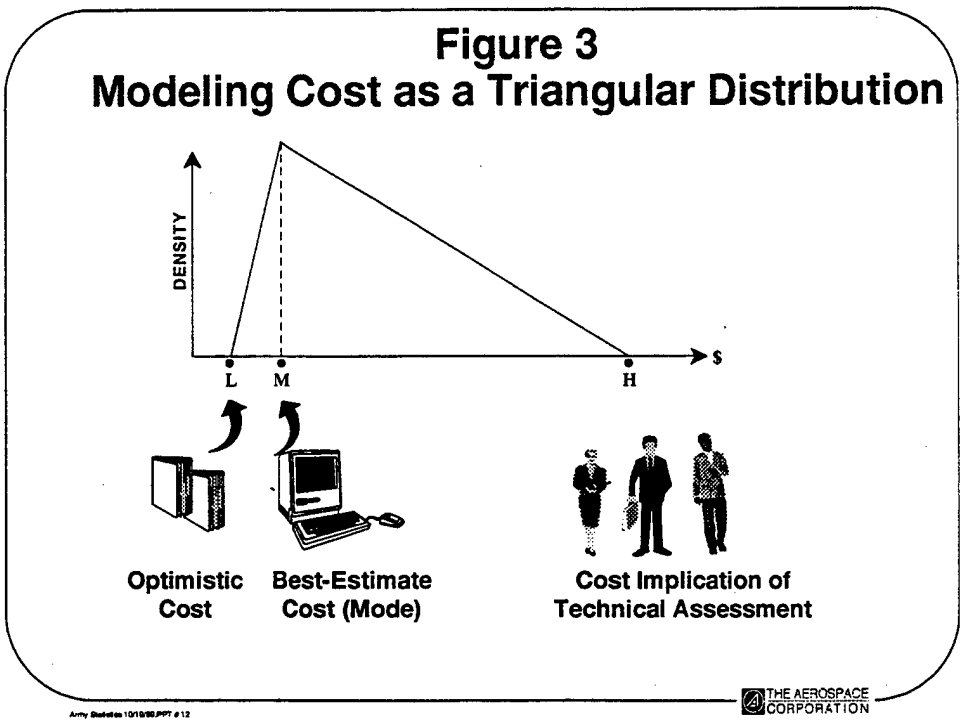
CHOOSING NONZERO NUMERICAL VALUES FOR CORRELATION

Having established that correlation between pairs of WBS-element costs won't go away if we simply ignore it, we face the task of assigning nonzero numerical values to the inter-element correlations. There are a number of ways of selecting nonzero numerical values that do a reasonable job of modeling the actual impact of one element's cost upon another's, but that process is not the subject of this report. We

assume that this task has already been done, i.e., we already have an $n \times n$ matrix of numbers $\{\rho_{ij}: 1 \le i \le n, 1 \le j \le n, -1 \le \rho_{ij} \le 1\}$, that, in our opinion, adequately represent the inter-element correlations.

At this point several very deep mathematical difficulties of several different kinds arise. To set the stage for discussing these mathematical difficulties, let's explain what's involved in the mathematics underlying this whole process. We have a sequence of n random variables, each of which represents the cost of a WBS element. These random variables are not independent, but rather they are linked according to one or more dependency mechanisms, of which a correlation matrix is probably the simplest kind. (Whether or not a correlation matrix is the "correct" dependency mechanism is not the subject of this report — here we are assuming that it is.) We view the distributions of the n WBS-element costs as the marginal distributions of a multivariate probability distribution that models the joint behavior of the n costs.

Suppose, for example, that the n WBS-element costs $X_1, X_2, \ldots, X_n$ are all normally distributed with means $\{\mu_i: 1 \le i \le n\}$, variances $\{\sigma_i^2 : 1 \le i \le n\}$, and a self-consistent set of correlations $\{\rho_{ij}: 1 \le i \le n, 1 \le j \le n\}$. Then the joint probability distribution is uniquely defined and, of course, multivariate normal. Our work would be done at this point were it not for the unfortunate fact that the normal distribution is not a good model for element costs, especially for elements of high-technology state-of-the-art defense systems. A more appropriate model is a skewed distribution, such as the triangular distribution of Figure 3. The triangular distribution is a three-parameter distribution defined by its lower bound (the most optimistic estimate of the cost, usually supplied by the contractor in his proposal), its mode (the most likely cost, usually provided by a government cost-estimating team), and its upper bound (the "worst-case" cost, typically derived from information generated by a government technical-risk assessment team). In the case of defense systems, many of which "push" the state of the art, the preponderance of probability lies above



the mode. Other common skewed distributions that serve as better models for cost elements than does the normal are the lognormal and the exponential.

Necessity to model costs as non-normal probability distributions raises some apparently very deep mathematical issues. In all cases, even those involving normal distributions, the preliminary question of whether or not the inputs given for the pairwise correlations constitute a consistent set must be answered. That particular question, though, is easy to answer, for it is well known that the correlation matrix must be nonnegative-definite in order for the pairwise correlations to be consistent, and there are relatively straightforward mathematical tests for this. The following questions, however, are much deeper:

(1) Given a valid correlation matrix (i.e., a consistent set of correlations) and a set of marginal distributions, does a multivariate distribution exist with the given marginals and correlation structure?

(2) Given a valid correlation matrix (i.e., a consistent set of correlations) and a set of marginal distributions, can we use the Monte Carlo process to generate random realizations of a multivariate distribution having those marginals and that correlation structure (assuming such a distribution exists)?

(3) Do the random realizations (assuming they can be generated) model any feasible multivariate distribution?

(4) If a number of multivariate distributions exist having those marginals and that correlation structure, can we tell which of them is being realized by the Monte Carlo process?

(5) Do the sums of the random realizations of the WBS-elements costs $X_1, X_2, \ldots, X_n$ model the total cost distribution?

FACTS WE WISH WEREN'T TRUE

Several disturbing mathematical facts preclude satisfactory answers to all five of the above questions. These facts range from involving routine details of the Monte Carlo random-number-generation process all the way to adapting to the complex nature of multivariate probability distributions.

A random-number-generating line of code in a computer program produces more or less independent, identically distributed (i.i.d.) observations from the continuous uniform distribution on the unit interval [0,1]. Standard transformation techniques, e.g., inversion of the distribution function when that is feasible and the Box-Müller procedure in the case of the normal distribution, applied to the i.i.d. uniform observations, yield simulated i.i.d. observations from the particular distribution of interest.

If the distribution of interest is the normal, it is relatively simple to transform the i.i.d. sequence into a sequence of correlated observations having any desired correlation structure. The most direct way to do this is to use Cholesky factorization, a method of finding a square root of the nonnegative-definite correlation matrix [Reference 9, pages 52-55]. (Of course, if the matrix is not nonnegative definite, no such square root exists, but then the matrix cannot represent a consistent set of correlations either.) We multiply the $n \times n$ square-root matrix by the $n \times 1$ vector of i.i.d. observations, and then the transformed observations will be correlated according to the given correlation matrix. See Reference 4 for specifics.

If the distribution of interest is not the normal, however, difficulties arise in applying the Cholesky and other matrix transformations to the vector of independent random numbers. First of all, except for the normal and perhaps a few other pathological distributions, the linear transformation effected by the square root of the correlation matrix does not preserve the form of the distribution. For example, if we generate independent random vectors that realize a collection of triangular distributions and then apply a matrix transformation to each of the vectors, the multivariate distribution thus realized will have the "correct" mean, variance, and correlation structure, but will not have triangular marginals.

Another problem involves the number of parameters specifiable for a multivariate distribution. In contrast with the multivariate normal distribution, which is uniquely determined by its mean vector, correlation matrix, and the requirement that its marginal distributions be normal, a general multivariate distribution is not thus uniquely determined, i.e., it is underdetermined by the correlation structure, or there

is no multivariate distribution having the given marginals, i.e., it is overdetermined by imposition of a correlation structure. The monograph by Johnson and Kotz [Reference 8, pages 260-268] provides several examples of different bivariate exponential distributions that may have the same correlation, as well as several that cannot have certain correlation values. Some of these examples are listed in Figures 4 and 5.

Figure 4
Correlated "Bivariate" Exponential Distributions

- **Gumbel's First Bivariate Exponential Distribution**

$$F(x,y) = 1 - e^{-x} - e^{-y} + e^{-(x+y+\theta xy)}$$

$$0 \le \theta \le 1$$

  - $\theta = 0$ means WBS-element costs are uncorrelated
  - Allowable range of correlation: $-0.40365 \le \rho \le 0$

- **Gumbel's Second Bivariate Exponential Distribution**

$$F(x,y) = \left(1 - e^{-x}\right)\left(1 - e^{-y}\right)\left(1 + \alpha e^{-x-y}\right)$$

$$-1 \le \alpha \le 1$$

  - Allowable range of correlation: $-0.25 \le \rho \le 0.25$

- **Gumbel's Third Bivariate Exponential Distribution**

$$F(x,y) = 1 - e^{-x} - e^{-y} + \exp\left[-\left(x^m + y^m\right)^{\frac{1}{m}}\right]$$

THE AEROSPACE CORPORATION

Figure 5
Other Bivariate Exponential Distributions

- **Freund's:** $-\frac{1}{3} \le \rho \le 1$

$$f(x,y) = \alpha(\alpha+\beta)\, e^{-(\alpha+\beta)y} \qquad \text{for } 0 \le x < y$$

$$= \beta(\alpha+\beta)\, e^{-(\alpha+\beta)x} \qquad \text{for } 0 \le y \le x$$

- **Marshall and Olkin's:** $\rho = \dfrac{\theta}{(\alpha+\beta+\theta)} > 0$

$$F(x,y) = \exp\{\alpha x + \beta y + \theta \min(x,y)\} \qquad \text{for } x \le 0,\ y < 0$$

- **Moran's:** $\rho = \omega^2 > 0$

$$f(x,y) = \sum_{j=o}^{\infty} \omega^{2j}\left[\prod_{k=1}^{2}\left\{\sum_{h=0}^{j}\left(-1^h\right)\binom{j}{h}(h!)^{-1}x_k^h\ e^{-x_k}\right\}\right]$$

- **Chi-Square:** $\rho = \frac{1}{2}$

$$f(x,y) = (2\pi)^{-3} e^{-\frac{x+y}{2}} \int_0^{\min(x,y)} t^{-1/2}(x-t)^{-1/2}(y-t)^{-1/2} e^{t/2}\, dt$$

THE AEROSPACE CORPORATION

Another illustrative example is provided by the version of the bivariate lognormal distribution that inherits its correlation characteristics from the underlying bivariate normal distribution. The fact that the

correlation ρ between two normal distributions must satisfy the inequality $-1 \le \rho \le 1$ forces restrictive bounds on the correlation between the two marginal lognormals. Details may be found in Reference 6, pages 364-365, and the results appear in Figure 6.

Figure 6
Correlated Lognormal Distributions

- If X_1 is $N(P_1,Q_1)$ and X_2 is $N(P_2,Q_2)$, then $Y_1 = e^{X_1}$ and $Y_2 = e^{X_2}$ have lognormal distributions
- If $Corr(X_1,X_2) = \rho$, then

$$\lambda = Corr(Y_1,Y_2) = \frac{e^{\rho Q_1 Q_2} - 1}{\sqrt{e^{Q_1^2} - 1}\ \sqrt{e^{Q_2^2} - 1}}$$

- The fact that $-1 \le \rho \le 1$ forces bounds to be imposed on λ, viz.

$$\frac{e^{-Q_1 Q_2} - 1}{\sqrt{e^{Q_1^2} - 1}\ \sqrt{e^{Q_2^2} - 1}} \le \lambda \le \frac{e^{Q_1 Q_2} - 1}{\sqrt{e^{Q_1^2} - 1}\ \sqrt{e^{Q_2^2} - 1}}$$

- In particular if $Q_1 = 1$ and $Q_2 = 1$,

$$-\frac{1}{e} \le \lambda \le 1$$

THE AEROSPACE CORPORATION

Conclusions that can be drawn from the above discussion can be summarized in the following two sentences: (1) A vector of random numbers generated from given marginal distributions and then transformed via a given correlation matrix may overdetermine a multivariate distribution and therefore may not in fact be a realization of any feasible multivariate distribution (and so the sums of the components of such random vectors may not represent a feasible total-cost distribution); and (2) Even if the cost distribution is a feasible one, we may not be able to determine which one of several possible underdetermined distributions it is.

P.M. Lurie and M.S. Goldberg [References 10 and 11] appear to have obtained the currently best available results in this area. Starting with Cholesky factorization, they proceeded to develop an approximate method of producing properly correlated random numbers that do a pretty good job in certain cases of modeling given characteristics of a multivariate distribution and provide approximately correct values of the mean, standard deviation, and percentiles of the total-cost distribution.

RANK CORRELATION TO THE RESCUE?

Up to now, we have been discussing our inability to generate a Monte Carlo sequence of random vectors with correlated components that simulate realizations of a multivariate distribution with arbitrary pre-specified marginals and correlation matrix. This inability apparently has led the commercial software industry to choose to satisfy the market demand for correlated random numbers by providing the capability to generate rank-correlated random numbers. The reason behind the industry preference for Spearman rank correlation over Pearson product-moment correlation is a simple one: it is known how to generated rank-correlated random numbers. A distribution-free technique was published in 1982 by R.L. Iman and W.J. Conover [Reference 7]. While it seems reasonable that the Iman-Conover technique has been adopted by the commercial vendors of Monte Carlo software, we cannot be sure that this so, because they respond to inquiries by stating that their methods are "proprietary." In any case, the vendors have not acknowledged the Iman-Conover work.

The bad news, however, is that rank-correlated random numbers do not satisfy the needs of cost analysts. While it is true that there are no problems of existence, feasibility, and non-uniqueness in the case of modeling rank correlation, this is due to the fact that rank correlations are not parameters of the multivariate distribution. Because the Iman-Conover technique is distribution-free, no particular multivariate distribution is being modeled – indeed the analyst need not even propose a particular multivariate distribution. He or she can be confident that one or more multivariate distributions are sure to exist with the given rank-correlation structure. As we have noted earlier, one may not necessarily exist with a given Pearson correlation structure.

The fact that no particular multivariate distribution is being modeled is not, by itself, bad for cost estimating. It would be sufficient for cost-estimating purposes to have the simulation produce reasonably correct values of the mean, standard deviation, and percentiles of the total-cost distribution. The real problem with rank correlation, though, is that it is not related in any known* way to the standard deviation (and therefore the percentiles) of the total-cost distribution. This means that modeling the distribution of total project cost using rank-correlated random numbers leads to a distribution that is not likely to contain the "correct" degree of uncertainty in the cost. In fact, a cost analyst who inputs a correlation matrix of Pearson correlations and then has these correlations interpreted by his or her software as if they were Spearman correlations will likely be surprised by the numerical value of the total-cost standard deviation that is reported by the software. If the software computes the standard deviation of the sums of the rank-correlated random numbers, which are supposed to be realizations of the total cost, it will not obtain the same numerical value as an analyst using Equation (2) to calculate the total-cost standard deviation.

P.R. Garvey, author of a recent book on cost-uncertainty analysis [Reference 6], has been seeking for some time a relationship between the pairwise rank correlations on the one hand and the standard deviation and range of the total-cost distribution on the other. However, he has been unable to find one that he considers satisfactory and, as a result, has been recommending to the cost-analysis community that rank correlation not be used by cost analysts. Nevertheless, his advice has not been received favorably, because rank-correlation-based software exists, is easy to use, and provides pages and pages of impressive-looking computer output – just what analysts need.

Lack of enthusiasm among cost analysts for Garvey's recommendations reveals that the commercial availability of software products that calculate random vectors consisting of rank-correlated random numbers has proved to be a "red herring" – a distraction that has served to take cost analysts' eyes off the hard statistical problems that have to be solved if cost uncertainty is to be "correctly" modeled in the future.

SUMMARY

Unfortunately for cost analysts, the probabilistic approach turned out to require the estimation of correlations between program-element costs as part of an effort to model total program costs. For its part, the concept of correlation among components of random vectors is more sophisticated mathematically than it appeared to cost analysts at first glance. Further complicating the situation is the impossibility of expressing the total-cost probability distribution in closed analytic form and the resultant need to model it by Monte Carlo simulation of the sum of pairwise correlated random variables. While generating the required sequences of correlated random numbers in support of this objective is a nontrivial problem, existence of multivariate distributions with given marginal and correlation structure is a more fundamental difficulty. Now that several commercial software products that purport to circumvent the problem by providing sequences of rank-correlated random numbers are easily available, interest in solving the modeling problem correctly has diminished among cost analysts.

---

* by me.

REFERENCES

1. Assistant Secretary of Defense (Program Analysis and Evaluation), *Cost Analysis Guidance and Procedures*, Washington DC: Department of Defense, DOD 5000.4-M, December 1992, vi+44 pages.

2. S.A. Book, "Do Not Sum 'Most Likely' Cost Estimates," *Proceedings of the 17th Annual Conference of International Society of Parametric Analysts (ISPA)*, San Diego, CA, 30 May-2 June 1995, pages RISK 16-41.

3. S.A. Book and P.L. Smith, "Reducing Subjective Guesswork and Maintaining Traceability When Reporting 'Risk' Associated with a Cost Estimate," The Aerospace Corporation, 2350 E. El Segundo Blvd., El Segundo, CA 90245, presented to American Statistical Association 1996 Joint Statistical Meetings, Chicago, IL, 4-8 August 1996.

4. S.A. Book and P.H. Young, "Monte Carlo Generation of Total-Cost Distributions When WBS-Element Costs are Correlated," The Aerospace Corporation, 2350 E. El Segundo Blvd., El Segundo, CA 90245, presented to the 24th Annual Department of Defense Cost Analysis Symposium, Leesburg, VA, 5-7 September 1990.

5. R. Fairbairn, "A Method for Simulating Partial Dependence in Obtaining Cost Probability Distributions," *Journal of Parametrics*, Vol. 10, No. 3 (October 1990), pages 17-44.

6. P.R. Garvey, *Probability Methods for Cost Uncertainty Analysis*, New York: Marcel Dekker, 2000, xv+401 pages.

7. R.L. Iman and W.J. Conover, "A Distribution-free Approach to Inducing Rank Correlation among Input Variables," *Communications in Statistics - Simulation, Computation*, Vol. 11, No. 3(1982), pages 311-334.

8. M.E. Johnson, *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions*, New York: John Wiley & Sons, Inc., 1987, 240 pages.

9. N.L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*, New York: John Wiley & Sons, Inc., 1972, xiv+333 pages.

10. P.M. Lurie and M.S. Goldberg, "Simulating Correlated Distributions with Bounded Domains," IDA Paper P-2732, Institute for Defense Analyses, 1801 N. Beauregard St., Alexandria, VA 22311-1772, September 1992.

11. P.M. Lurie and M.S. Goldberg, "A Method for Simulating Correlated Random Variables from Partially Specified Distributions," *Management Science*, Vol. 44, No. 2 (February 1998), pages 203-218.

STRATEGIES FOR DATA MINING

David L. Banks
Bureau of Transportation Statistics
Department of Transportation, Room 3430
400 Seventh St. SW, Washington DC 20590

ABSTRACT

Data mining is a hard, complex problem. This paper stitches together and reviews previous work on three separate fronts in data mining. The first issue concerns the preanalysis of one's data. The second issue concerns the choices and tradeoffs one must make in matching a particular nonparametric inferential technique to a given data set. The third issue concerns how one might estimate the local dimensionality of one's data.

INTRODUCTION

Data miners are like the blind Indian sages who were invited to inspect an elephant. Each found a true but distinct aspect of the elephant, and the real challenge was to integrate the different perspectives into a common understanding. In that spirit, this paper pulls together a decade of experience in working with statistical inference for superlarge data sets. Out of that, three particularly crucial problem areas emerge.

The first problem one encounters is the preanalysis of the data. This consists of all work that must be done in order to get the data into a form to which appropriate statistical software can be applied. This phase of the analysis can easily consume 80% of one's time and resources, but is one of the most understudied aspects of statistical inference. The discussion in this paper draws heavily upon the work of Banks and Parmigiani (1992).

The second problem is how to decide which of the many new wave nonparametric methods best applies to one's data set. Recent years have seen the invention of many new techniques (MARS, CART, ACE, PPR, Loess, neural nets, and so forth), and the analyst must have some informed basis for choosing among these methods. This portion of the paper summarizes work described in Banks, Olszewski, and Maxion (1999).

The third problem concerns the estimation of the local dimensionality of one's data set. For reasons that are discussed later, local dimensionality drives the entire analysis. If local dimensionality is too large, then essentially no statistical method can produce good inference except under highly artificial circumstances. But if local dimensionality is low, then with intelligence, sufficient data, and a little luck, there is a good chance that the statistician can discover what hidden structure exists. The discussion of this relies chiefly on a paper by Banks and Olszewski (1997).

The next three sections elucidate these issues. The intention is to provide a survey and a point of entry to the literature. For that reason, much of the discussion is heuristic and aimed at a general audience.

PREANALYSIS

The stoop labor of statistics lies in the preparation of one's data for analysis. This task can absorb enormous amounts of one's time, yet for large data sets it is nearly impossible ever to be sure that one has weeded out all the outliers, adequately handled the high-leverage points, and caught and corrected all transpositions, missing values, and other problematic data. This problem grows rapidly more acute as the size of the data set increases. With a few dozen observations, much can be checked by eye. With a few million multivariate observations, it quickly becomes necessary to devise automatic procedures and adaptive rules for reviewing and correcting data.

This kind of data cleaning problem for large data sets arises in many contexts. Examples include intrusion detection for computer networks, analysis of interactions as shown from gene chips, and inference on complex manufacturing processes. The first and last examples have time series structure, which is an additional complication. Our discussion is motivated in terms of manufacturing, because the original data discussed in Banks and Parmigiani (1992) were obtained from PPG Industries and Alcoa.. However, the basic principles and approach

apply broadly. The core idea is that one must spend time to develop an explicit and nearly automatic protocol for data cleaning.

The following data cleaning protocol is intended as a guideline for the preparation and initial analysis of data that arise from complex continuous or batch-continuous manufacturing processes. In any specific application, one expects to tailor the protocol to the problem. With this caveat, the key steps are:

1. Data Input. Reformat the data; all numerical input is read into a common floating point representation. All non-numerical input is represented consistently. The size of the representations is determined by the largest possible (signed) value.
2. Time Scale. Create or identify the time scale. All other observations are then keyed to the value of the time variable. The data record thus becomes a set of vectors indexed by time, or a multivariate time series. Each vector contains the measurements on all control or response variables at a given time point. If variables are not measured simultaneously, unmeasured values are assigned a code to indicate that they are missing as part of normal process operation.
3. Missing Values. All data values known to be missing are assigned special codes. These indicate whether the data are missing due to plant shutdown, failure of a monitoring device, unknown causes, or routine operation of the process inspection schedule.
4. Sample Size Check. Verify the numbers of observations for each variable. The input is the expected length of each time series and the expected length of each data record, accounting for known missing values. The output is a list of variables for which the expected and actual sizes do not match.
5. Impossible Values. List and adjust all observations with impossible values. The determination of allowable values requires guidance from a process engineer or other domain expert knowledgeable about the data. With this list, the possible actions are to verify, approximate, code as missing data, or correct for a missing sign, zero or data-entry transposition.
6. Synchronicity. Reindex the database so that values are synchronous. This ensures that measurements keyed to a specific time actually pertain to common process characteristics. With unsynchronized data, control measurements recorded at a given time do not affect quality measurements until the end of the production cycle. To properly assess the interplay between process control and output quality, the analysis must first link corresponding values into common records, and this usually requires expert judgment about process lags.
7. Missing Value Chart. Produce a chronological guide to missing values. Often it is useful to subtract out known interruptions, which requires information from the process engineer regarding plant or subsystem shutdowns, or analogous information in non-manufacturing environments.
8. Unequal Frequencies and Data Imputation. Use data imputation procedures, smoothing, or expert knowledge to fill short gaps in the data sequence. In most cases, the missing data arise because not all variables are recorded at the same frequency. In other cases gaps are caused by sensor failures or similar extraneous circumstances. Sometimes it happens that the short gaps are informative, especially when the values are probably near the boundary of the historical operating region; these require expert judgment.
9. Extreme Value Chart. Generate a chart showing the extremes in the variation of process variables. This extreme value chart helps to flag outliers in step 10, identify data flaws that have not been previously detected, and point up patterns of variation that require examination. The chart may be viewed as a global Shewhart chart, and thus also has value for day-to-day process management.
10. Outlier Detection. The extreme value chart can discover univariate outliers. These should be checked and confirmed, or else replaced by an estimate. Additionally, one should search for multivariate and possibly time series outliers.
11. Descriptive Statistics. Test the time series for nonstationarity, and examine model consistency between plant shutdowns or other natural process interruptions. Do Q-Q plots to assess departures from normality, and perhaps make normalizing transformations of the data.
12. Exploratory Analysis. This begins to phase into the conventional analysis. We find that some exploration is needed to plan subsequent work, and the specific methods used are sensitive to the context of the study. Rather than describing such generality, see Banks and Parmigiani (1992) for examples of tools used in previous applications.

As a final step, we urge that one rethink all that's been done. A process engineer should review the adjustments and ratify the cleaning process. This sounds trivial, but our experience indicates that it is inevitably useful.

Many of these twelve points require some additional discussion. For example, the first step sounds trivial. However, consistent formatting makes all subsequent analyses easier to implement; in particular, one can move rapidly between different packages and special purpose software. Also, important flaws can be detected in this stage; for example, in a PPG data file, a particular variable field did not allocate enough space for large numbers, and so adjacent values were either read as a single number (free format input), or else negative numbers became positive (column input). It took two days to pinpoint the error, and thus it is time effective to write a simple program that automatically searches for such flaws. Similarly, missing values often arise when changes in the data collection methods are not properly reflected in the data format. Finally, if one inspects the records visually, a regular format magnifies the ability of the human eye to discover errors.

Steps 1, 2, 3, 5, 6, 8, and 10 all create a new database from the preceding database. These new databases refine the original by reformatting, indicating the kind of missing values that occur, relagging the database time index to enable meaningful comparisons, and imputing discovered deficiencies. The result is a pseudo-database that approximates the data set one would have liked to have been given for study in the first place.

In Step 3, the most common cause of missing data is differential measurement frequency. For example, in an Alcoa data set some variables were measured every half-hour, whereas others were measured each hour. Thus the time index changed in half-hour increments, and some variables showed missing values in every other observation. This kind of gap is viewed as missing as part of the normal operation of the process. Similarly, planned interruptions should be coded to distinguish them from failures in the data capture effort.

Step 7 produces the missing value chart, a graphic output that aids process management and informs the data cleaning process. It displays patterns of omission by category of omission. Specifically, the chart:

- ignores data that is absent as part of the normal operation of the process.
- shows the number of observations missing at a given time period.
- shows the number of consecutive time periods with missing values.
- flags time periods in which certain subsets of missing values occur, such as those pertaining to specific subsystems in the production process.
- flags time periods according to known causes of missing data, such as plant shutdowns or sensor failure.

In our experience, process experts can read this chart easily, recognizing data gaps of which they were previously aware, and discovering new problems that require attention.

In specific applications, other kinds of missing data may occur. For example, data may be missing because the values have fallen outside the range of sensitivity of the measuring instrument. Figure 1 in Banks and Parmigiani (1992) provides and example of a missing value chart, and the surrounding discussion shows how that chart's information was used to discover hidden aspects of the PPG data set.

In realistically large applications, the time span is too long and the numbers of variables too great to allow missing values to be represented in a single chart. Instead, one should use a color monitor; white pixels mark normal values, colored pixels code the categories of missing data. This compact display accommodates about 750 variables and 1000 time points on a standard workstation. For greater time spans one can scroll the screen, divide the time span into intervals or collapse intervals without missing data.

Step 8 adjusts the database by imputing the missing values that arise from normal process operation, and, perhaps, for the sprinkling of isolated unpatterned missing values that occur for unknown reasons. In our work with Alcoa and PPG, most adjustment is made to account for differential sampling rates. There are sophisticated strategies for interpolating or imputing such missing values. If one has knowledge of the underlying time series model, or if the process history permits tight record matching, then one could use either a form of forecasting/backcasting or hot deck imputation, respectively. However, our experience indicates that situations with this degree of structure are rare, so we prefer to estimate values by linear interpolation of the data.

The advantages of linear interpolation, as opposed to the more sophisticated smoothing routines described by Friedman, Grosse and Stuetzle (1983), are both computational and theoretical. First, linear interpolation is easy to program and quick to implement in almost any software system that might be used. Second, linear interpolation makes a strictly local adjustment of the data; distant values have no influence on the correction. This seems a

desirable feature when cleaning data without a well-defined probability model. Even when there is information that would enable slight improvement over simple linear interpolation, it may not be cost-effective to pursue such refinement; i.e., if one's ultimate inference from the mountain of available data is sensitive to the rule for imputing missing values, then the entire solution is surely unstable. Spending effort on optimal imputation takes resources for other phases of the analysis, making this a case of the best being the enemy of the good.

Whatever imputation method one uses, the output from Step 8 is a new, artificial database that spans data gaps in order to:

- remove very short sequences of missing data whose absence is deemed random and uninformative.
- obviate difficulties in subsequent analysis caused by data missing as part of the usual operation of the process.

The first adjustment is appropriate when clerical error or sensor failure causes missing data, but the process is under control and nearby values are likely to be similar to those that are missing. The second is useful when not all observations are recorded at the same sampling rate; here one can interpolate the least frequent (data inflation) or average the most frequent (data reduction). If one variable is measured daily and another is measured hourly, we recommend interpolating the daily values to create hourly proxies, since the alternative, compressing the hourly data to daily averages, destroys information.

A caution is needed. When large amounts of missing data are estimated, the variance of the values in the artificial database is too small. Thus the outlier tests in Step 10 will be oversensitive, and probably require some tuning.

Step 9 produces an extreme value chart; in analogy with the missing value chart, it looks for patterns over time in the least and largest values of the control and quality variables. The vertical axis identifies the control and quality variables. In many applications it is useful to group these according to process sequence or process subsystems. The horizontal axis is the time index of the data. For a given variable at a given time, we plot blank space if the datum is within $\pm$ 2.5 σ of the mean for that control or quality variable. Otherwise, we plot a stroke if the datum exceeds the upper 2.5 σ bound, and a dot if it falls below the lower 2.5 σ bound.

In our experience, a $\pm$ 2.5 σ rule captures data oddities without overloading the visual display; for other applications, the customary $\pm$ 3 σ zone might serve best. For data sets with very large numbers of variables, it is more compact to code the direction of extreme deviations with color pixels, and display very large numbers of variables simultaneously upon a color monitor.

In general, the extreme value chart shows both isolated data points and runs of data (often affecting downstream variables) that fall outside the zone of normal operating values. The former look like potentially influential data flaws, and should be checked to ensure that no coding glitch has occurred; even if no identifiable cause can be found, it may improve the analysis to replace these outliers by interpolated values. In contrast, runs of extreme values carry a great deal of information about process capability beyond the usual operating region. These should be preserved in the database, and closely studied in the subsequent analysis phase.

The extreme value chart enables managers to identify which variables are prone to fall out of control, and which quality variables are sensitive to such departures. These charts can also discover faulty control loops or differences in performance between operating shifts. For some applications, it merits note that the extreme value chart responds to planned changes in control set points, so one should calculate control means only over intervals between process innovations.

Step 9 has produced a chart that guides Step 10. Here one identifies outliers, and remedies these in the way most appropriate to the problem. One should verify their values, and then either replace them by less extreme estimates (i.e., treat them as missing values), or else prepare a robust analysis. To increase the complexity, there is also concern about multivariate outliers and outliers with respect to the time series structure.

Multivariate outliers occur when two or more variables take values that are jointly improbable. Hawkins (1982) describes methods of multivariate outlier detection. We feel that simple Mahalanobis distance is broadly sensitive and often adequate. The Mahalanobis distance for the total process and the process subsystems may be

included as rows in the extreme value chart. If such outliers are found in those rows, they should be checked by process experts, and either retained or replaced by interpolation.

Univariate and multivariate time series outlier detection is difficult, but potentially informative. Univariate time series outliers can often be recognized as runs of outside values on the extreme value chart, and multivariate outliers can be detected as patterns of such runs. No useful theory is available for the multivariate case. From an exploratory standpoint, inspection of the extreme value chart may be the most practical way to identify multivariate outlier behavior. In the PPG application described in Banks and Parmigiani (1992), these showed up visually as the process engineers chased the drifting system.

Step 11 is conventional methodology; too much detail sidetracks the purpose of this survey. In brief, Step 11 moves the preanalysis beyond data cleaning and lays the foundation for addressing substantive questions. One has finally constructed an artificial database that resembles what classroom lectures assume one receives in the first place, and now it becomes appropriate to begin applying textbook tools. For example, if one contemplates an analysis based upon normal theory methodology, such as fitting a linear model or extracting the principal components, then it is proper to undertake Q-Q plots and look for normalizing transformations.

Step 12 moves the preanalysis firmly towards conventional analysis. Although methods depend on the application, we have found it useful to build two types of charts, both aimed at detecting data structure in the region of optimal operation. These tools provide useful guidance in planning an analysis when there is little physical understanding of the process model. Details on these charts, and our experience in implementing them, are described in Banks and Parmigiani, (1992).

As a final point, we emphasize that the 12-step protocol is a guideline. In a formal sense, the preanalysis is part of the total analysis---if one has a good model, then it should drive one's decisions during the preanalysis phase. Since well-defined models are rare for complex processes, it usually happens that the chief concern is to be sure that the preanalysis does not entail any alterations of the data that will hinder subsequent examination of central research questions.

CHOOSING A NONPARAMETRIC METHOD

Classical statistical inference works well when the dimensionality of the data is low. However, regression analysis in high dimensions quickly becomes extremely unreliable; this phenomenon is called the "Curse of Dimensionality" (COD). There are three nearly equivalent formulations of the COD, each offering a usefully different perspective on the problem:

- The complexity of the possible regression structure increases super-exponentially with dimension.
- In high-dimensions, nearly all data sets are sparse.
- In high dimensions, nearly all data sets show multicollinearity (and its nonparametric generalization, concurvity).

Detailed discussion of this topic and its consequences for regression may be found in Hastie and Tibshirani (1990) and in Scott and Wand (1991).

Historically, multivariate statistical analysis sidestepped the COD by imposing strong model assumptions that restricted the potential complexity of the fitted models, thereby allowing sample information to have non-local influence. But now there is growing demand for data analytical techniques that make weaker model assumptions and use larger data sets. This has led to the rapid development of a number of new methods; however, the comparative performance of these methods is poorly understood.

We focus upon comparisons among ten commonly used methods:

- Multiple Linear Regression (MLR).
- Stepwise Linear Regression (SLR).
- Additive Model (AM). Described in Hastie and Tibshirani (1990), this fits sums of nonlinear univariate functions of the explanatory variables.
- LOESS. This fits a local regression to the data, as described by Cleveland (1979),

- Alternating Conditional Expectations (ACE). The method was invented by Breiman and Friedman (1985), and is similar to AM except that it allows transformations of the dependent variable to maximize the correlation between the left- and right-hand sides.
- Additivity and Variance Stabilization (AVAS). A modification of ACE developed by Tibshirani (1988) that incorporates a variance stabilizing transformation.
- Projection Pursuit Regression (PPR). This is like AM, except that the nonlinear univariate functions depend upon linear combinations of the explanatory variable (Friedman and Stuetzle, 1981).
- Recursive Partitioning Regression (RPR). This is similar to the CART methodology pioneered by Breiman, Friedman, Olshen and Stone (1984); it finds rectangular regions in which local averages fit well.
- Multivariate Adaptive Regression Splines (MARS). Developed by Friedman (1992), this is combines features of RPR and PPR.
- Neural Network (NN). We use a procedure called CASCOR, which automatically chooses the number of hidden nodes (Fahlman and Lebiere, 1990).

These methods were chosen because people use them, because they have relatively sophisticated strategies, and because code is easily available to implement them.

Currently, our understanding of comparative regression performance consists of a scattering of theoretical and simulation results. The key sources are:

- Donoho and Johnstone (1989), who develop asymptotic results which imply that projection based regression methods (PPR, MARS) perform significantly better for radial functions, whereas kernel based regression (LOESS) is superior for harmonic functions.
- Friedman (1991), who reports simulation studies of MARS alone, and related work described in the discussion.
- Tibshirani (1988) and Hastie and Tibshirani (1990), who report simulation results for AVAS.
- Breiman (1991), who describes benchmark results for high-dimensional regression for five functions, and pertinent discussions in the following section.
- Barron (1993), who shows that in a somewhat narrow sense, the mean integrated squared error of neural net estimates for a certain class of functions has order that is subexponential in the dimension; similar results were subsequently obtained by Zhao and Atkeson (1992) for PPR.
- Ripley (1993) describes simulation studies of neural network procedures, usually in contrast with statistical methods.

These short, often asymptotic, explorations do not provide sufficient understanding for a practitioner to make an informed choice among regression techniques

To address this gap, Banks, Maxion, and Olszewski (1997) describe a designed experiment that compares the ten regression techniques. The basis for the comparison is the mean integrated squared error (MISE) of the different techniques, assessed across a range of functions and conditions. The results of the simulation experiment are too lengthy to present here, but are available on the web at www.cs.cmu.edu/~bobski/. The simulation was performed on two advanced workstations (a DECstation 300 and an HP Apollo 715/75) that ran standard code for regression methodology over a period of nearly 19 months.

The simulation experiment had six factors:

- Regression Method. The ten levels of this factor are MLR, SLR, MARS, AM, PPR, ACE, AVAS, RPR, LOESS, and NN, as previously described.
- Function. The five kinds of functions that were examined were hyperflats, multivariate normals with zero correlation, multivariate normals with all correlations .8, two-component mixtures of multivariate normals with zero correlation, and a function proportional to the product of the explanatory variables. These situations correspond to the kinds of structure one encounters in practice.
- Dimensionality. The three levels of this factor take the dimensionality of the explanatory variable space to be 2, 6, and 12. In higher dimensions, methods perform so poorly that it differences between them are of little practical interest.
- Sample Size. The three levels of this factor take the sample size to be $n=2^p k$, where p is the dimensionality and $k = 4, 10, 25$. This scales across dimensionality, so that the different values of k correspond to small, medium, and large samples, respectively.

- Noise. This determines the variance in the additive Gaussian error associated with each observation. The standard deviations of the error variance are $\sigma = 0\ 02, 0.1, 0.5$.
- Model Sparseness. This determines the proportion of explanatory variables that are functionally related to the response variable. The different levels consist of all variables, half of the variables, and none of the variables. The latter case is especially important in practice, since it shows how much spurious structure the methods will find in pure noise.

Note that not all combinations of this design are realizable. Specifically, when Model Sparseness is set so that none of the variables pertain to the response variable, then the level of the Function is irrelevant.

The simulation experiment proceeded according to the following steps:

1. Generate a uniform random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ inside the unit hypercube in $\Re^p$.
2. Generate a sample of random errors $\varepsilon_1, \ldots, \varepsilon_n$, all iid $N(0, \sigma^2)$.
3. Calculate $Y_i=f(\mathbf{X}_i)+\varepsilon_i$, where $f:\Re^p \rightarrow \Re$ is the target function determined by appropriate combinations of levels of Function, Dimensionality, and Model Sparseness.
4. Apply one of the regression techniques in Factor 1 to obtain f^*, an estimate of f.
5. Estimate the integrated squared error of f^* over the unit cube. Call this m.
6. Repeat steps 1-5 20 times. The average of the 20 resulting m values is an estimate of the MISE; we also calculate the standard error of this estimate for use in subsequent comparisons.

From each combination of factor levels, we thus obtain an estimate of the MISE and its standard error. Regression methods whose MISE values are significantly lower than competing methods are superior. We wish to understand which methods are best for which functions, dimensions, sample sizes, and noise levels.

The experiment produced a many findings, but those of most importance to practitioners are:

- For the linear relationship, MLR and SLR do well, as expected. More interestingly, GAM also performs well.
- For the normal model, LOESS generally does well.
- For the correlated normal model, ACE, PPR, and LOESS do well, though there is great variation across situations.
- For the mixture of normals model, MARS performs well; when all variables are used, LOESS also does well.
- For the product model, the results are very mixed. A weak conclusion is that when MISE is low, MARS is often among the best, but when MISE is high, LOESS is among the best.
- In low dimensions, MARS does consistently well when the explanatory variables are not related to the response variable.
- In high dimensions with large samples, most of the methods do well when the explanatory variables are not related to the response variable. When samples are smaller, the more computationally sophisticated models do slightly worse, discovering spurious structure.
- RPR, and presumably the commercial version of CART, generally have higher levels of MISE than competing methods. But RPR is often the winner in very high dimensional problems.

More detailed comparisons are given in the technical report at www.cs.cmu.edu/~bobski/.

In applications, there is a more direct and practical way to select a regression procedure. Hold out a portion of the data. Fit a model to the remaining data using each regression technique. Use that model to predict the values of the holdout sample. Then use the regression technique that achieves the smallest MISE on the holdout data. (If one's data set is not large enough to support holding a portion of the data out, then one probably ought not be doing high dimensional regression.)

LOCAL DIMENSIONALITY

None of the methods compared in the preceding section actually avoid the Curse of Dimensionality (except NN and PPR, in an impractical and narrow sense). Instead, they fit a flexible model that assumes some specific form for the hidden structure in the data set. If the model happens to approximate reality, then the method performs well. The methods are not truly nonparametric; they all make model assumptions of varying degrees of weakness.

These weak assumptions typically involve the kind of locally low dimensional structure of the data. For example MARS, PPR, and neural nets (and wavelets) search hard to select a small set of variables that are influential in a particular region (MARS and wavelets) or a particular direction (PPR and neural nets). MARS and

wavelets do not assume that the same variables are active everywhere, which is a great boon, but it is unclear that they are particularly successful in handling regions where the local functional dependence involves more than about four variables. Similarly, PPR and neural nets are most effective when the regression surface in a given direction is dominated by the behavior of a small number of explanatory variables. In an alternate path to a comparable end, GAM, ACE, AVAS all assume a simple additive structure that precludes the interactions that make high-dimensions so difficult.

Our perspective is that many problems are intrinsically too hard for statisticians to solve with the data available. It then becomes valuable to find a screening tool to identify data sets having simple structure that makes them potentially amenable to a modern computer-intensive analysis, as opposed to data sets that are so complex that no method has realistic hopes of discovering useful information. To this end, we describe a strategy for screening data sets that applies to either regression or structure discovery, and show the results of a designed experiment that implements our strategy for the structure discovery form of the problem.

The simplest and first case to consider is that of structure discovery. The data consist of vectors $\mathbf{X}_1,\ldots,\mathbf{X}_n$ in $\Re^p$, and one wants to discover whether there is some hidden pattern. We call this the ``crumpled napkin problem" because of the following illustration. Suppose during dinner with Persi Diaconis (a mathematician who is also a magician), he draws a large number of dots on his napkin, crumples it loosely, and causes the napkin to become invisible, so that only the dots may be seen. From cursory observation, it appears that the dots are a featureless blob in space; but if one looks very closely, and if there are enough dots, one could in principle discover that the points lie on a hidden two-dimensional manifold that is the surface of the napkin. The clue that enables that insight is the fact that in sufficiently small regions of the volume containing the dots, the dots tend to lie on nearly flat two-dimensional surfaces. Thus the local dimensionality is in fact two except near the folds in the napkin, despite the apparent three-dimensionality.

Following this direction, our analytical strategy is to pass a small hypersphere in $\Re^p$ over the volume of the data; at random intervals, we stop the sphere's passage, record the points that lie inside the sphere, and perform a principal components analysis on them. The number of eigenvalues that contribute appreciably to the trace of the covariance matrix indicates the approximate local dimensionality of the data. (In our work, we estimate the local dimensionality as the number of eigenvalues needed to sum to 80% of the trace, but this is more a choice of convenience than principle.) Finally, we average the local estimates of dimension; if the result is much smaller than p, we have good reason to hope that there is simple structure hidden in the data, and thus the data set might repay further study. But if the result is not much smaller than p, this suggests that no currently available analysis can succeed, and our time would be better spent on other projects.

The previous strategy is still a bit too simple. Suppose that one's data fill out the locus between two p-dimensional spheres of different radii, each centered at $\mathbf{0}$. Then the moving sphere will typically find either no points in its ambit (when it lies in empty region of the smaller sphere), or it will estimate a local dimensionality of p, when it is located inside the outer sphere but outside the inner sphere. Thus the proposed strategy does not notice structure caused by regions of data sparsity. To repair this deficiency, we count the number of observations within the sphere at each random stop. If the number falls below a threshold (we use $2p$), then the region is declared sparse and no further work is done at that site; otherwise we proceed with the principal components analysis. At the end of the passage, we record the proportion of stops that were sparse, and the average local dimensionality at the stops that were not. If the first is large or second number is small, this implies the presence of interesting structure.

We want to extend the structure discovery approach to regression analysis. The way to do this is to replace the principal components analysis at each stop of the small, moving hypersphere with a principal components stepwise regression. The number of components that are selected for inclusion in the model represents the local dimensionality of the functional relationship between the explanatory variables and the response. When this number is low, there is hope that MARS, ACE, GAM, PPR, or one of the other new wave methods might be tuned to discover the kind of relationship that exists.

Of course, this requires a bit more judgment. One must determine the levels of significance in the repeated tests for the stepwise regression's variable selection. Also, one would probably want to set a minimum value for the coefficient of determination, below which one does not feel that the strength of the local relationship is sufficiently

strong to warrant further study. In this spirit, one could calculate the local average coefficient of determination, which, if sufficiently small, would persuade one that an analysis is bootless despite a low average local dimensionality.

This experiment calibrates the method we described previously for the situation in which data on a q-dimensional manifold is presented in $\Re^p$, for $p \geq q$. The particular structure of the manifold that we consider is the q-boundary of the unit p-dimensional hypercube. For example, uniform data on the 1-boundary of a 3-dimensional cube would have points that lie (apart from noise) on the 12 edges of the cube. In contrast, uniform data on the 2-boundary of the cube would have points that lie (apart from noise) on the 6 faces of the cube. And clearly, when $q = p = 3$, there is no structure present that anyone would be interested to discover. We wish to discover how well average local dimensionality and sparsity do in flagging the cases in which $q < p$.

The experiment consists of 20 replications at each combination of the following factor levels:

- Dimension. We take all values of (p, q) such that $7 \geq p \geq q \geq 1$.
- Noise. All sample observations (uniform on the q-surface are corrupted by independent p-variate Gaussian noise with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}$ where σ takes two levels: .1 and .5.
- Sample Size. We consider two levels; the general formula is $n = 2^q k$, where $k = 10, 15$.

This design is not a stringent test of the strategy, but rather helps us to establish the general sensitivity of our tuning choices in developing the algorithm. These choices include the sphere size and the threshold value for the sum of the eigenvalues.

The first result of the experiment is shown in Table 1. The entries show the average of the average local dimensionality over the 20 replications of the treatment combinations. The standard error on these estimates is small, on the order of 0.01 or less.

Table 1: Estimated Average Local Dimension of q-Surfaces on p-Cubes

q=7							5.03
q=6						4.25	4.23
q=5					3.49	3.55	3.67
q=4				2.75	2.90	3.05	3.18
q=3			2.04	2.24	2.37	2.50	2.58
q=2		1.43	1.58	1.71	1.80	1.83	1.87
q=1	*0.80*	*0.88*	*0.92*	*0.96*	*0.95*	*0.95*	*0.98*
	p=1	*p=2*	*p=3*	*p=4*	*p=5*	*p=6*	*p=7*

The key result in Table 1 is that the estimated local dimensionality is quite stable across different values of p. Of course, the estimated dimensionality is lower than the true value of q, which is inevitable since the threshold rule sums only those eigenvalues needed to explain 80% of the variability in the data. But since the estimate is so stable, bias-correction methods, empirical or theoretical, should improve that performance. And the exact estimate of q is not essential---our purpose is simply to decide whether the data warrant further study.

CONCLUSIONS

This review has outlined three issues that arise in data mining. The first problem, involving preanalysis, is perhaps the most difficult to resolve. It is very hard to get around the need to invest substantial amounts of time in data preparation when mining for structure. The second issue is becoming better understood. There is a growing body of experience, theory, and simulation that can provide guidance on selection of regression methods. The third issue is somewhat novel, but grows out of the fact that truly complex data is beyond our grasp. We need to develop better screening tools to identify which data sets are worth our time.

Besides these three issues, there are other problems that arise, or should arise, in the minds of data miners. One concerns the problem of data compression. This has ramifications for information theory, storage, and

automatic pattern recognition. A second problem is to find typical values in a data set, thereby producing a quick index to the range of phenomena in the data set.

ACKNOWLEDGEMENTS

Portions of this work were performed with the support of Alcoa, PPG, NSF Grant #IRI-9224544, and DARPA Grant #F30602-96-1-0349..

REFERENCES

1. Banks, D.L. and Parmigiani, G. "Preanalysis of Superlarge Datasets." Journal of Quality Technology, vol. 24, p. 115-129, 1992.
2. Banks, D.L., Maxion, R. and Olszewski, R.J. "Comparing Methods for Nonparametric Regression." Submitted to the Communications in Statistics: Simulation, 2001.
3. Banks, D.L. and Olszewski, R. J. "Estimating Local Dimensionality." In Proceedings of the Statistical Computing Section of the American Statistical Association, p. 782-788, American Statistical Association, Alexandria, VA, 1997.
4. Barron, A.R. "Universal Approximation Bounds for Superpositions of a Sigmoidal Function." IEEE Transactions on Information Theory, vol. 39, 930-945, 1993.
5. Breiman, L. "The Π Method for Estimating Multivariate Functions from Noisy Data," Technometics, vol. 33, 124-144, 1991.
6. Breiman, L. and Friedman, J. "Estimating Optimal Transformations for Multiple Regression and Correlation." Journal of the American Statistical Association, vol. 80, 580-619, 1985.
7. Breiman, L., Fiedman, J. Olshen, R.A., and Stone, C. Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.
8. Cleveland, W. "Robust Locally Weighted Regression and Smoothing Scatterplots." Journal of the American Statistical Association, vol. 74, 829-836, 1979.
9. Donoho, D. and Johnstone, I. "Projection Based Approximation and a Duality with Kernel Methods." Annals of Statistics, vol. 17, 58-106.
10. Fahlman, S.E. and Lebiere, C. "The Cascade-Correlation Learning Architecture." In Advances in Neural Information Processing Systems 2, ed. by D. Touretzky, p. 525-533, Morgan Kaufman, San Mateo, CA 1990.
11. Friedman, J. "Multivariate Adaptive Regression Splines." Annals of Statistics, vol. 19, 1-66, 1991.
12. Friedman, J., Grosse, E. and Stuetzle, W. "Multidimensional Additive Spline Approximation." SIAM Journal of Scientific and Statistical Computing, vol. 4, 291-301, 1983.
13. Friedman, J. and Stuetzle, W. "Projection Pursuit Regression." Journal of the American Statistical Association, vol. 76, 817-823, 1981.
14. Hastie, T.J. and Tibshirani, R.J. Generalized Additive Models. Chapman and Hall, New York, NY, 1990.
15. Hawkins, D.M. Identification of Outliers. Wiley, New York, 1982.
16. Ripley, B.D. "Statistical Aspects of Neural Networks." In Networks and Chaos—Statistical and Probabilistic Aspects, ed. by O.E. Barndorff –Nielsen, J.L. Jensen, and W.S. Kendall, p. 40-123, Chapman and Hall, London, 1993.
17. Scott, D.W. and Wand, M.P. "Feasibility of Multivariate Density Estimates." Biometika, vol. 78, 197-206, 1991.
18. Tibshirani, R.J. "Estimating Optimal Transformations for Regression Via Additivity and Variance Stabilization." Journal of the American Statistical Association, 83, 394-405, 1988.
19. Zhao, Y. and Atkeson, C.G. "Some Approximation Properties of Projection Pursuit Networks." In Advances in Neural Information Processing Systems 4, ed. by J. Moody, S.J. Hanson, and R.P. Lippmann, p. 936-943, Morgan Kaufmann, San Mateo, CA, 1992.

ATTENDANCE LIST
U.S. ARMY CONFERENCE ON APPLIED STATISTICS
19–21 OCTOBER 1999

COL David Arney, Ph.D.
Dept of Math Sci.
United States Military Academy
West Point NY 10996

MAJ Keith Arthur
AMSAM-RD-AA-R
Bldg 401 Lee Blvd
Ft. Eustis VA 23604-5577

Dr. David Banks
Bureau of Transportation Statistics
400 Seventh Street SW
Washington DC 20590

LTC Philip Beaver
Dept of Math Sci.
United States Military Academy
West Point NY 10996

Dr. Barnie Bissenger
281 W. Main St
Middletown PA 17057

Dr. Barry Bodt
US Army Research Laboratory
ATTN:AMSRL-CI-CD
Aberdeen Proving Ground MD 21005-5067

Dr. Stephen Book
The Aerospace Corp.
M4/921
P.O. Box 92957
Los Angeles CA 90009-2957

Mr. Robert Burge
7134 Smooth Path
Columbia MD 21045

Mr. Tony Colby
Sterling Software
Beeches Technical Campus
RT 26N, Rome NY 13440

Mrs. Susan Conover
5213 85th St
Lubbock TX 79424

Prof. W. J. Conover
College of Bus. Admin
Texas Tech U
Lubbock Texas 79409

Prof. David Cruess
6123 Camel Back Lane
Columbia MD 21045

Dr. Paul Deason
US Army TRAC-WSMR
ATTN ATRC-WAD
WSMR NM 88002-5502

MAJ Curt Doescher
4817A Faith Court
Fort Mead MD 20755

Dr. Gene Dutoit
3544 Statler Dr
Columbus GA 31907

Prof Doug Frank
Dept of Math
Indiana U of Penn
233 Stright Hall
Indiana PA 15705

Dr. Jock Grynovicki
US Army Research Laboratory
ATTN: AMSRL-HR-SD
Aberdeen Proving Ground MD 21005-5425

Prof. Bernie Harris
Dept of Statistics
U of Wisconsin-Madison
Madison WI 53706

Prof. Paul Hshieh
36 Case St
Gaithersburg MD 20878

Mr. Richard Kass
US Joint Forces Command
Suffolk VA 23434

Ms. Keri Kenney
US Army Yuma Proving Ground
ATTN: STEYP-MT
Yuma AZ 85365-9110

BG Fletcher M Lamkin Jr
United States Military Academy
West Point NY 10996

Dr. Robert Launer
US Army Research Office
P.O. Box 12211
Research Triangle Park NC 27709-2211

Ms. Valerie Lee
17 Penny Lane
Perryville MD 21903

Prof Wei-Yin Loh
Dept of Statistics
U of Wisconsin-Madison
Madison WI 53706

Mr. Richard Maruyama
12413 Hatchery CT
Woodbridge VA 22192

Prof. Charles McCulloch
439 Warran Hall
Cornell University
Ithaca NY 14853-7801

Mr. Don Neal
34 Blueberry Hill LN
Sudbury MA 01776

Prof. Ingram Olkin
Dept of Statistics
Stanford University
Stanford CA 94305-4065

COL David Olwell, Ph.D.
Navel Postgraduate School
Monterey CA 93943-5221

Mr. Mike Prather
USA Evaluation Center
4120 Susquehanna Ave
Aberdeen Proving Ground MD 21005-3013

Dr. Daryl Pregibon
21 Sherman Avenue
Summit NJ 07901

Dr. Carl Russell
JNTF/SE 730 Irwin Ave
Schriever AFB CO 80912-7300

Prof. Frank Samaniego
University of California-Davis
Division of Statistics
Davis CA 95616

Prof. Jayaram Sethuraman
Dept of Statistics
Florida State University
Tallahassee FL 32306-4330

Prof. Nozer Singpurwalla
Dept of Operations Research
George Washington University
707, 22nd St N.W.
Washington DC 20052

Mr. Samir Soneji
Harmony Hall 206
Columbia U
544 W. 110th Street
New York NY 100215

Dr. Doug Tang
Uniformed Services University
of the Health Sciences
4301 Jones Bridge Road
Bethesda MD 20814

Ms. Deloris Testerman
US Army Yuma Proving Ground
ATTN: STEYP-MT
Yuma AZ 85365-9110

Prof. Jim Thompson
Dept. of Statistics
Rice University
Houston Texas 77251-1892

Mr. Tom Walker
US Army Evaluation Center
4120 Susquehanna Ave
Aberdeen Proving Ground MD 21005-3013

Dr. R. John Weaver
16007 Prairie Ronde
Schoolcraft MI 49087

Mr. David Webb
Director, USARL
AMSRL-WM-BC (WEBB)
Aberdeen Proving Ground MD 21005-5066

Prof. Ed Wegman
Ctr. for Comp. Stat.
George Mason U.
MS4A7
Fairfax VA 22030

Ms. Connie Whitener
US Army Yuma Proving Ground
ATTN: STEYP-MT
Yuma AZ 85365-9110

Dr. Alyson Wilson
PO Box 1663 MS F600
Los Alamos NL
Los Alamos NM 87545-0600

INTENTIONALLY LEFT BLANK.

NO. OF COPIES	ORGANIZATION
2	DEFENSE TECHNICAL INFORMATION CENTER DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	HQDA DAMO FDT 400 ARMY PENTAGON WASHINGTON DC 20310-0460
1	OSD OUSD(A&T)/ODDR&E(R) DR R J TREW 3800 DEFENSE PENTAGON WASHINGTON DC 20301-3800
1	COMMANDING GENERAL US ARMY MATERIEL CMD AMCRDA TF 5001 EISENHOWER AVE ALEXANDRIA VA 22333-0001
1	INST FOR ADVNCD TCHNLGY THE UNIV OF TEXAS AT AUSTIN 3925 W BRAKER LN STE 400 AUSTIN TX 78759-5316
1	DARPA SPECIAL PROJECTS OFFICE J CARLINI 3701 N FAIRFAX DR ARLINGTON VA 22203-1714
1	US MILITARY ACADEMY MATH SCI CTR EXCELLENCE MADN MATH MAJ HUBER THAYER HALL WEST POINT NY 10996-1786
1	DIRECTOR US ARMY RESEARCH LAB AMSRL D DR D SMITH 2800 POWDER MILL RD ADELPHI MD 20783-1197

NO. OF COPIES	ORGANIZATION
1	DIRECTOR US ARMY RESEARCH LAB AMSRL CI AI R 2800 POWDER MILL RD ADELPHI MD 20783-1197
3	DIRECTOR US ARMY RESEARCH LAB AMSRL CI LL 2800 POWDER MILL RD ADELPHI MD 20783-1197
3	DIRECTOR US ARMY RESEARCH LAB AMSRL CI IS T 2800 POWDER MILL RD ADELPHI MD 20783-1197
	ABERDEEN PROVING GROUND
2	DIR USARL AMSRL CI LP (BLDG 305)

NO. OF COPIES	ORGANIZATION
1	USMA DEPT OF MATH SCI COL D ARNEY WEST POINT NY 10996
1	FT EUSTIS AMSAM RD AA R MAJ K ARTHUR BLDG 401 LEE BLVD FT EUSTIS VA 23604-5577
1	BUREAU OF TRANSPORTATION STATISTICS D BANKS 400 SEVENTH ST SW WASHINGTON DC 20590
1	USMA DEPT OF MATH SCI LTC P BEAVER WEST POINT NY 10996
1	B BISSENGER 281 W MAIN ST MIDDLETOWN PA 17057
1	THE AEROSPACE CORP S BOOK PO BOX 92957 LOS ANGELES CA 90009-2957
5	R BURGE 7134 SMOOTH PATH COLUMBIA MD 21045
1	STERLING SOFTWARE BEECHES TECHNICAL CAMPUS T COLBY RT 26N ROME NY 13440
1	S CONOVER 5213 85TH ST LUBBOCK TX 79424
1	TEXAS TECH U COLLEGE OF BUS ADMIN W J CONOVER LUBBOCK TX 79409

NO. OF COPIES	ORGANIZATION
1	D CRUESS 6123 CAMEL BACK LANE COLUMBIA MD 21045
5	US ARMY TRAC WSMR ATRC WAD P DEASON WSMR NM 88002-5502
1	MAJ C DOESCHER 4817A FAITH COURT FORT MEADE MD 20755
1	G DUTOIT 3544 STATLER DR COLUMBUS GA 31907
1	INDIANA U OF PENN DEPT OF MATH D FRANK 233 STRIGHT HALL INDIANA PA 15705
1	U OF WISCONSIN-MADISON DEPT OF STATISTICS B HARRIS MADISON WI 53706
1	P HSHIEH 36 CASE ST GAITHERSBURG MD 20878
1	US JOINT FORCES COMMAND R KASS SUFFOLK VA 23434
1	USA YUMA PROVING GROUND STEYP MT K KENNEY YUMA AZ 85365-9110
5	US ARMY RESEARCH OFFICE R LAUNER PO BOX 12211 RESEARCH TRIANGLE PARK NC 27709-2211

NO. OF COPIES	ORGANIZATION
1	V LEE 17 PENNY LANE PERRYVILLE MD 21903
1	U OF WISCONSIN-MADISON DEPT OF STATISTICS W LOH MADISON WI 53706
1	CORNELL UNIVERSITY C MCCULLOCH 439 WARRAN HALL ITHACA NY 14853-7801
1	D NEAL 34 BLUEBERRY HILL LN SUDBURY MA 01776
1	STANFORD UNIVERSITY DEPT OF STATISTICS I OLKIN STANFORD CA 94305-4065
1	D PREGIBON 21 SHERMAN AVENUE SUMMIT NJ 07901
5	JNTF SE C RUSSELL 730 IRWIN AVE SCHRIEVER AFB CO 80912-7300
1	U OF CA DAVIS DIVISION OF STATISTICS F SAMANIEGO DAVIS CA 95616
1	FLORIDA STATE U DEPT OF STATISTICS J SETHURAMAN TALLAHASSEE FL 32306-4330
1	GEORGE WASHINGTON U DEPT OF OR N SINGPURWALLA 707 22ND ST NW WASHINGTON DC 20052

NO. OF COPIES	ORGANIZATION
1	COLUMBIA U S SONEJI HARMONY HALL 206 544 W 110TH ST NEW YORK NY 10025
5	USUHS D TANG 4301 JONES BRIDGE RD BETHESDA MD 20814
1	USA YUMA PROVING GROUND STEYP MT D TESTERMAN YUMA AZ 85365-9110
1	RICE UNIVERSITY DEPT OF STATISTICS J THOMPSON HOUSTON TX 77251-1892
1	R WEAVER 16007 PRAIRIE RONDE SCHOOLCRAFT MI 49087
1	GEORGE MASON U CTR FOR COMP STAT E WEGMAN MS4A7 FAIRFAX VA 22030
1	USA YUMA PROVING GROUND STEYP MT C WHITENER YUMA AZ 85365-9110
3	LOS ALAMOS NL A WILSON PO BOX 1663 MS F600 LOS ALAMOS NM 87545-0600

NO. OF
COPIES ORGANIZATION

ABERDEEN PROVING GROUND

2 USAEC
M PRATHER
T WALKER

16 DIR USARL
AMSRL CI
B BODT (5 CPS)
A BRODEEN
AMSRL WM
D WEBB (5 CPS)
AMSRL HR
J GRYNOVICKI (5 CPS)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project(0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2001	Final, 19-21 October 1999

4. TITLE AND SUBTITLE
Proceedings of the Fifth Annual U.S. Army Conference on Applied Statistics, 19–21 October 1999

5. FUNDING NUMBERS
AH80

6. AUTHOR(S)
Barry A. Bodt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: AMSRL-CI-CT
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-SR-110

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
United States Military Academy
West Point, NY 10996

10.SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The fifth U.S. Army Conference on Applied Statistics was hosted by the United States Military Academy, West Point during 19-21 October 1999. The conference was cosponsored by the U.S. Army Research Laboratory (ARL), the U. S. Army Research Office (ARO), the United States Military Academy (USMA), the Training and Doctrine Command (TRADOC) Analysis Center-White Sands Missile Range, the Walter Reed Army Institute of Research (WRAIR), and the National Institute for Standards and Technology (NIST). The U.S. Army Conference on Applied Statistics is a forum for technical papers on new developments in statistical science and on the application of existing techniques to Army problems. This document is a compilation of available papers offered at the conference.

14. SUBJECT TERMS
applied statistics, experimental design, statistical inference

15. NUMBER OF PAGES
110

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

INTENTIONALLY LEFT BLANK.

USER EVALUATION SHEET/CHANGE OF ADDRESS

This Laboratory undertakes a continuing effort to improve the quality of the reports it publishes. Your comments/answers to the items/questions below will aid us in our efforts.

1. ARL Report Number/Author ARL-SR-110 (Bodt) Date of Report July 2001

2. Date Report Received _____

3. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which the report will be used.) _____

4. Specifically, how is the report being used? (Information source, design data, procedure, source of ideas, etc.) _____

5. Has the information in this report led to any quantitative savings as far as man-hours or dollars saved, operating costs avoided, or efficiencies achieved, etc? If so, please elaborate. _____

6. General Comments. What do you think should be changed to improve future reports? (Indicate changes to organization, technical content, format, etc.) _____

CURRENT ADDRESS

_____
Organization

_____
Name E-mail Name

_____
Street or P.O. Box No.

_____
City, State, Zip Code

7. If indicating a Change of Address or Address Correction, please provide the Current or Correct address above and the Old or Incorrect address below.

OLD ADDRESS

_____
Organization

_____
Name

_____
Street or P.O. Box No.

_____
City, State, Zip Code

(Remove this sheet, fold as indicated, tape closed, and mail.)

(DO NOT STAPLE)