Proceedings of the Seventh Annual U.S. Army Conference on Applied Statistics, 24-26 October 2001

Barry A. Bodt, Edward J. Wegman
EDITORS

Hosted by:
LOS ALAMOS NATIONAL LABORATORY

Cosponsored by:
U.S. ARMY RESEARCH LABORATORY
U.S. ARMY RESEARCH OFFICE
UNITED STATES MILITARY ACADEMY
TRADOC ANALYSIS CENTER—WHITE SANDS MISSILE RANGE
WALTER REED ARMY INSTITUTE OF RESEARCH
UNIFORMED SERVICES UNIVERSITY OF THE HEALTH SCIENCES
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Cooperating Institutions:
RAND
LOS ALAMOS NATIONAL LABORATORY
GEORGE MASON UNIVERSITY
OFFICE OF NAVAL RESEARCH
INSTITUTE FOR DEFENSE ANALYSIS

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067 July, 2003

Proceedings of the Seventh Annual U.S. Army Conference On Applied Statistics, 24-26 October 2001

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

Edward J. Wegman, EDITOR
Center for Computational Statistics, George Mason University

Hosted by:
Los Alamos National Laboratory

Cosponsored by:
U.S. Army Research Laboratory
U.S. Army Research Office
United States Military Academy
TRADOC Analysis Center—White Sands Missile Range
Walter Reed Army Institute of Research
Uniformed Services University of the Health Sciences
National Institute of Standards and Technology

TABLE OF CONTENTS
SEVENTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

Short Course

General Session I

Special Session I

Special Session II: Urban Warfare

Contributed Session I

Contributed Session II

Contributed Session III

Clinical Session I

Contributed Session IV

Wilks Award Banquet Address

General Session II

Contributed Session V

Contributed Session VI

Contributed Session VII

Contributed Session VIII

Special Session III

General Session III

Contributed Session IX

General Session IV

SEVENTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

ABSTRACT

The Seventh U.S. Army Conference on Applied Statistics was hosted by Los Alamos National Laboratory (LANL), 24-26 October 2001 at the Bishop's Lodge in Santa Fe, NM. The conference was co-sponsored 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 (TRAC-WSMR), the Walter Reed Army Institute of Research (WRAIR), the Uniformed Services University of the Health Sciences (USUHS), and the National Institute for Standards and Technology (NIST). Cooperating organizations included the Los Alamos National Laboratory, George Mason University (GMU), Office of Naval Research (ONR), and the Institute for Defense Analysis (IDA). 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. Approximately ninety individuals attended and forty-six papers were given, continuing the success of the sixth conference at Rice University. This document is a compilation of available papers offered at the conference.

FORWORD

The Seventh U.S. Army Conference on Applied Statistics was hosted by Los Alamos National Laboratory (LANL), 24-26 October 2001 at the Bishop's Lodge in Santa Fe, NM. The conference was co-sponsored 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), the Uniformed Services University of the Health Sciences (USUHS), and the National Institute for Standards and Technology (NIST). Cooperating organizations included the Los Alamos National Laboratory, George Mason University (GMU), Office of Naval Research (ONR), and the Institute for Defense Analysis (IDA). 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.

The seventh ACAS offered a variety of topics important to DoD. The conference was preceded by a two-day tutorial, "Applied Logistic Regression," taught by David Hosmer of the University of Massachusetts. Approximately twenty-five students took part. Sallie Keller-McNulty of LANL opened the conference. The keynote address, "On a New Approach to Robust Estimation," was delivered by David Scott of Rice University. Invited, general session presentations were given by W.J. Conover, Texas Tech University; William Meeker, Iowa State University; Leo Breiman, University of California; Juergen Symanzik, Utah State University; and Bin Yu, University of California. Three special sessions were featured. As a follow-up to the well-attended tutorial from the previous conference, Alyson Wilson, LANL, organized a session on "Case Studies in Elicitation and Quantification of Expertise and Expert Judgement." Addressing other current issues, Paul Deason, TRAC-WSMR and Eugene Dutoit of the U.S. Army Infantry School organized a special session on "Urban Warfare" and Edward Wegman of

GMU organized a special session on "Information Assurance." Thirty contributed papers rounded out the program. The U.S. Army Wilks award was not given at the seventh ACAS. The tradition of the banquet, however, continued, and guests were treated to an engaging presentation by Jas. Mercer-Smith of LANL.

The Executive Board for the conference recognizes the Statistics Group at the Los Alamos National Laboratory for hosting the conference with special thanks to Alyson Wilson and Rachael Vigil for attending to conference details, Edward Wegman, GMU, for assembling the proceedings, Edmund Baur, ARL, for maintaining the web site, David Webb, ARL, for overseeing conference communications, and Barry Bodt of ARL for chairing the conference.

Executive Board of the U.S. Army Conference on Applied Statistics

Barry A. Bodt, Chair *U.S. Army Research Laboratory*	J. Robert Burge *Walter Reed Army Institute of Research*	David F. Cruess *Uniformed Services University of the Health Sciences*
Paul J. Deason *U.S.A. Training and Doctrine Command*	Lee S. Dewald, Sr. *Virginia Military Institute*	Eugene F. Dutoit *Troy State University*
Arthur Fries *Institute for Defense Analyses*	LTC Andrew G. Glen *United States Military Academy*	Jock O. Grynovicki *U.S. Army Research Laboratory*
David Kim *United States Military Academy*	Robert L. Launer *U.S. Army Research Office*	Wendy L. Martinez *Office of Naval Research*
Carl T. Russell *Joint National Test Facility*	Douglas B. Tang *Uniformed Services University of the Health Sciences*	Jacqueline K. Telford *Johns Hopkins University Applied Physics Laboratory*
David W. Webb *U.S. Army Research Laboratory*	Edward J. Wegman *George Mason University*	Alyson Wilson *Los Alamos National Laboratory*

SEVENTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

SHORT COURSE

Applied Logistic Regression

David Hosmer
University of Massachusetts

Abstract: This two-day short course will present an introduction to using the logistic regression model. Topics to be covered will include model formulation, parameter estimation, estimation and interpretation of odds-ratios and probabilities, model building strategies, assessment of goodness-of-fit, and presentation and interpretation of results. The course will consider the logistic regression model for binary, multinomial, and ordinal scaled outcomes.

The course will be taught by Professor David W. Hosmer of the Biostatistics Department of the University of Massachusetts. Prof. Hosmer has over 10 years experience teaching similar short courses to statisticians, epidemiologists, physicians and other subject matter scientists.

The course will be based upon selected chapters and sections in Professor Hosmer's recent text, *Applied Logistic Regression.* Co-authored by Professor Stanley Lemeshow of Ohio State University, the second edition of this widely referenced text was published in 2000 by John Wiley & Sons. Topics to be covered from this edition, with sections and page numbers noted, appear below:

1 Introduction to the Logistic Regression Model

1.1 Introduction, 1
1.2 Fitting the Logistic Regression Model, 7
1.3 Testing for the Significance of the Coefficients, 11
1.4 Confidence Interval Estimation, 17
1.5 Other Methods of Estimation, 21

2 Multiple Logistic Regression

2.1 Introduction, 31
2.2 The Multiple Logistic Regression Model, 31
2.3 Fitting the Multiple Logistic Regression Model, 33
2.4 Testing for the Significance of the Model, 36
2.5 Confidence Interval Estimation, 40
2.6 Other Methods of Estimation, 43

SEVENTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

General Session I

On a New Approach to Robust Estimation
David W. Scott, Noah Harding Professor of Statistics, Rice University

In this talk, I describe an alternative approach to robust estimation. Robust estimation provides a powerful solution to practical problems in applied statistics. Simple tasks such as data cleaning may be prohibitively expensive with large datasets. These techniques may also handle the difficult situation, where a dataset contains large clusters of outliers.

In order to use a robust estimation algorithm (such as the M-estimator described by Hampel and Huber), the shape and scale of the influence function must be specified. Tukey's biweight function is a popular choice but there are many, many possibilities. The scale may be determined by a simple robust method (such as the interquartile range), or by iteratively reweighting the data.

In our approach, maximum likelihood is replaced by a data-based minimum-distance criterion. I show that the specification of the shape and scale of the influence function can be replaced by a single choice of a distribution function for the data. This idea is illustrated for several common choices of data, including Gaussian.

This framework works well in both density and regression problems. Groups of multivariate outliers may be readily identified. Experimental design with messy data is facilitated. Semiparametric models such as mixtures of normals also fall within this paradigm. Several case studies are presented and actual code given.

Predicting and Understanding Complex Data
Leo Breiman, University of California, Berkeley

Abstract Unavailable

Special Session I: Case Studies in Elicitation and Quantification of Expertise and Judgement

Mary Meyer, Jerry Morzinski, Laura McNamara, Gregory Wilson, Los Alamos National Laboratory
The goal of this session is to discuss and demonstrate, using a case study, the elicitation of expertise (knowledge structure) and expert judgment (data). General information on knowledge systems will be presented, including ideas on formalizing how technical experts and communities of practice think about their problems, descriptions of different kinds of expert judgment, and the importance of advisor-experts. Using a case study, three specific phases of elicitation will be addressed: problem definition, capturing a problem-solving process, and gathering quantitative expert judgment.

Special Session II: Urban Warfare

Nuggetizing the Elephant: Managing Urban Complexity During Military Operations

Russell W. Glenn, RAND

Whereas there is no lack of confidence among tacticians when debating doctrine for fighting on open terrain, even experts approach discussions of how to deal with combat and non-combat missions in villages, towns, and cities with far less assurance. Urban undertakings today pose far greater challenges than just a few decades ago. Seoul, Republic of Korea, exemplifies the problem. The city was virtually an entity unto itself in 1950, separated from neighboring urban areas by expanses of rice paddies and lightly occupied terrain. Today Seoul is awash in a much larger metropolitan area. Its population has increased over tenfold to some thirteen million. More vehicles pack the same downtown area; more offices, apartments, and commercial enterprises fill a unit of space. The inflation of spatial densities has complements in similar increases in events per unit time. More infrastructure, people, and activity in less space mean that situations change more rapidly. A greater number of events occurs in a given period. More decisions per unit of time are demanded of military leaders when their operations draw them to such urban conglomerations. The overall effect is one of time and space compression. Military commanders have less time to analyze situations and alternatives, to position logistical support properly, and to determine how to best maintain the initiative in such environments.

How is the military to handle such complexity? Where should a commander employ his limited resources in the urban vastness? This briefing proposes but a small step toward a conceptualization of the problem. It is an initial attempt at developing a construct for envisioning and analyzing urban challenges. The first of its two primary elements suggests a means of determining those elements of the terrain, population, and infrastructure key to mission accomplishment: urban critical points. The second focuses on the aforementioned densities in space and time. Employing density as an analytical tool facilitates both the original identification of critical points and the analysis of those points after identification.

Networked Organizational Structure and Function in a Complex Megacity

Matt Begert, National Law Enforcement Technology Center

This presentation intends to address a continuing operational experimentation with a networked, multifunctional organization or EMON Emerging Multifunctional Organizational Network). The focus of effort in the creation, experimentation and function of this support network has been to create a fusion cell for intelligence and operations as common operating ground for agencies and organizations that might not otherwise coordinate their various operations with other functioning organizations. This offers, for instance, some insight and understanding in coordinating operations with nongovernmental organizations (NGO's) in not-crime/not-war situations.

In operation, the broad general functions are force protection, infrastructure protection and continuing stability and sustainment operations. Although this experimentation is domestic, it can directly relate to any environment. An element of this intelligence synthesis Is early identification of emerging threats The complex urban terrain of Los Angeles County, the diversity of population, interest, motivation and activity is a setting unmatched by any other population center for observation,lessons learned and experimentation.

The presentation will cite specific examples of large scale activity, special event planning, the ability to monitor, collect intelligence and deal with the restraints and constraints of working out coordinated effort of a myriad of functional agencies at a tactical, operational and strategic level.

Real-time Pseudo-Randomly Generated Features for Combat Experimentation in Urban Sprawl

Greg Tackett, U.S. Army Aviation and Missile Research, Development and Engineering Center

The DoD has the need for simulation representations of urban sprawl to support virtual experiments of urban combat. This need is problematic due to the large terrain database development and subsequent terrain complexity that is required for representation of large urban and suburban areas. A further complexity is the need to simulate the interior structure of a large number of buildings for dismounted combat.

One approach to creating generic suburban terrain is to generate, in real-time, a feature set representing realistic suburban cultural entities in the immediate vicinity of player entities in a distributed simulation. A server could distribute these features as objects to client machines across HLA or DIS interfaces. These features would include an assortment of houses, fences, utility buildings, pavement, trees, and other vegetation, objects, and structures combined to form an if-you've-seen-one-you've-seen-them-all-type subdivision of mathematically infinite dimensions. This same approach can also instantiate the internal structure of buildings when player entities come within immediate range, to allow entry and interaction between the interested entities and the internal features. The entities themselves could be generated using a random-number-seed approach that ensures that instantiation of features will be totally repeatable but variable in combination, placement, orientation, and internal layout.

The Aviation and Missile Command has developed the Pseudo-Random Urban Feature Entity Server (PRUFES) to demonstrate this approach. PRUFES uses a model set and rule set which together generate the cultural entities comprising a generic suburbs known as "Protoville".

PRUFES was recently evaluated for practical use through the execution of a simple experiment involving a single soldier at a virtual workstation given a mission requiring interior entry into a number of Protoville houses to conduct searches, simulating a Homeland Defense mission to find a terrorist device.

This paper discusses the use of cultural entities for suburban representation, PRUFES design, experimental findings, and interoperability issues between cultural entities and legacy manned simulators and SAFOR.

International Trends in MOUT Research

Danny C. Champion, TRADOC Analysis Center - WSMR

An international conference was recently held on dismounted close combat. In 20 years it is estimated that 70-80% of the world will live in urban environments. I will summarize some of the findings from that meeting, focusing on the current trends in MOUT Research.

Contributed Session I

AN EXPERIMENTAL DESIGN TO COLLECT HUMAN SCIENCE DATA FOR MODELING AND ANALYSIS

Eugene Dutoit and William Guest, Dismounted Battlespace Battle Lab, Fort Benning, GA
Michael Statkus, Natick Soldier Center
Arthur Garrett, Army Materiel Systems Analysis Activity
Luci Salvi, Army Research Laboratory

ABSTRACT

There is little data available to the modeling and analysis community for describing soldier performance in close combat/MOUT environments. Therefore, the objectives of this experiment were to obtain dismounted soldier performance data (e.g. target engagement and weapons firing) while learning about the process of collecting human performance data in a virtual combat environment. Because this project is on-going, this paper will describe the experimental design procedure, the data collection, and the follow-on statistical analysis procedures without presenting the actual data gathered.

INTRODUCTION

It is widely recognized by the Department of Defense and by the Department of the Army that many future battles will almost certainly unfold as close combat or military operations in urban terrain (MOUT). This prediction for close combat/MOUT engagements has been evidenced by the Army's recent MOUT Advanced Concept Technology Demonstration and the subsequent investment in MOUT technologies and equipment over the past few years. Of particular concern to the modeling community are the existing data gaps that need to be filled to more accurately describe engagements closer than 25 meters. Given the emphasis on close combat/MOUT and the need for more complete underlying data, this experiment was the first step in a multi-year effort to address this lack of data in the broad areas of move, shoot, and communicate; human behavior representation; and enabling data such as metabolic work load and fatigue. Ultimately, this data collection effort will support the modeling and analysis community's ability to conduct technology assessments, equipment trade-offs, and Basis of Issue analyses for the dismounted warrior.

This experiment was a joint effort involving the Natick Soldier Center (sponsoring agency and project lead), the Army Materiel Systems Analysis Activity, the Army Research Laboratory, and the Simulation Division of the Battlelab at Fort Benning, Georgia.

GENERAL EXPERIMENTAL OVERVIEW

The virtual experiment was conducted with 10 soldiers stationed at Fort Benning during a 2-week period of time. There were four tests/experiments scheduled with an overall total number of 282 replications. These separate tests used individual soldiers and two-man fire teams immersed in a virtual MOUT environment with the humans pitted against the computer-generated forces. A special data collection and analysis tool was developed to extract the experimental information from the computers logger files and put these data into spreadsheets suitable for analysis by commercial statistical programs (primarily SPSS(1) with some Excel assistance). The experimental plan summary for these four tests is presented in the table below.

TABLE 1
VIRTUAL ENVIRONMENT TEST PLAN SUMMARY

Test Number	Test Objective	Primary Issues	Number of Runs	Test Type	Lighting Conditions	# of enemy	# of Non-combatants
1	To measure a soldier's ability to detect and shoot a single computer generated force within a room during daylight operations.	Target Detection Target Engagement Weapons Firing	81	Individual soldier	Daylight	1 computer generated	0
2	To measure a soldier's ability to detect and shoot a single computer generated force within a room during nighttime operations.	Target Detection Target Engagement Weapons Firing	81	Individual soldier	Nighttime	1 computer generated	0
3	To measure a team of two soldiers' ability to detect and shoot 2 computer forces within a room during daylight operations.	Target Detection Target Engagement Weapons Firing	60	2man buddy team	Daylight	2 computer generated	0
4	To measure a team of two soldiers' ability to detect 2 targets during daylight operations and correctly identify the targets as either friend or foe and then successfully shoot the correct target (foe).	Target Detection Target Identification Target Engagement	60	2 man buddy team	Daylight	1 computer generated	1

Because the experimental design methodology for these four tests was *essentially* the same, this paper will focus on test 1 as shown above.

Virtual Combat Environment

The name of the virtual environment is the Squad Synthetic Environment (SSE). This is a squad-level man-in-the-loop simulation especially designed for dismounted Infantry applications to include individual tasks, fire team and squad level missions and urban scenarios in a virtual environment. A total of 13 full immersion soldier simulators and 10 desktop simulators are networked with computer generated forces. The SSE has been used to support analysis for training exercises, advanced concepts and requirements and research, development and acquisition. The SSE allows the dismounted soldier to move through the virtual battlefield, enter buildings, climb stairs, and move into standing, kneeling and prone positions. It also provides the capabilities for command, control, and communications. The user has a choice of terrain features to include; the McKenna MOUT site (including the details of the inside of the buildings), Camp LeJuene, dynamic terrain (put holes in buildings), detailed furniture inside of rooms and day and thermal imaging. Output analysis can be conducted on the number of rounds fired, casualties, number of targets shot, hit probability and distance to target information. Data are provided in spreadsheet format as well as video recordings.

Questionnaires

Before any of the experiments were conducted, the 10 soldiers were asked to fill out a Demographics, Experience and Training Questionnaire. The data obtained from this questionnaire will be used to provide correlated insights about the information collected on each soldier's performance in the SSE. The soldiers were also asked to fill out an After Action Questionnaire after they participated in each replication within the SSE. These questionnaires were attempting to gather additional insights from the soldiers such as; "which aiming technique did you use?"; "How difficult was it to detect the enemy?" Finally, each of the ten soldiers was asked to fill out an Exit Questionnaire when they completed *all* tests and replications. The soldiers were asked to give their general impression of the test experience and identify any problems they had with the SSE simulator.

Anthropometric Data

In addition to the questionnaire data cited above, basic anthropometric data were collected on each soldier. The study team thought that some of these measurements might be useful when determining target profiles for use in combat modeling.

TEST 1 PLAN

As stated above, this paper will focus on the "experimental design process" for this test. The other tests were similar and there is no need to repeat the same process three other times.

Objective

To measure a live soldier's ability to detect and engage a single computer-generated enemy within a room during daylight conditions.

Description

The live soldier will enter a virtual building and room that really exists at the McKenna MOUT site. The live soldier will then attempt to detect and engage (shoot) a single computer generated enemy. The enemy will be stationary but will shoot at the live soldier upon detection. Each of the nine soldiers (there are nine soldiers in an Infantry squad) will engage each of the three enemy scenario/positions in the room exactly three times. Therefore, this test will require a total of 81 replications i.e.

9 soldiers x 3 scenarios x 3 replications per scenario = 81 total replications.

Rules of Engagement for Soldiers within the Test

The following rules were followed during the course of each test. 1. Each soldier had to calibrate his weapon prior to each trial. 2. Each soldier was taught and expected to follow the standard scanning (search) techniques appropriate for city combat and to use the correct tactics, techniques and procedures. 3. Soldiers were taught to engage the enemy targets as soon as possible. 4. In order to keep soldiers from anticipating the order of presented targets they were instructed not to discuss their prior experience with any other squad members.

Pre-Test Conditions to Keep Constant

For this test, these conditions were applied to the enemy targets. 1. The enemy was stationary. 2. The enemy was placed in a kneeling position. 3. The enemy was placed behind a piece of furniture. 4. Target shape for the enemy was irregular. The right side of the body was shown from behind the furniture. 5. The enemy was set to fire on the live soldiers after being fired upon. 6. The light level was set for daytime conditions. 7. The enemy targets appeared in one of three locations/scenarios within the room; back left corner, center, or front right corner.

Measurements Obtained for Each Trial

These were the measures obtained for each trial. 1. Total time (in seconds) to detect, acquire and engage (shoot) the target. Time began when the live soldier crossed a predetermined point in the virtual room containing the enemy target. Time ended with the last trigger pull. 2. Number of virtual rounds fired by the live soldier. 3. Number of rounds fired to get the first hit. 4. Number of casualties for each trial (both live soldiers and virtual enemy.

TEST 1 PROTOCOL DESIGN AND EXECUTION

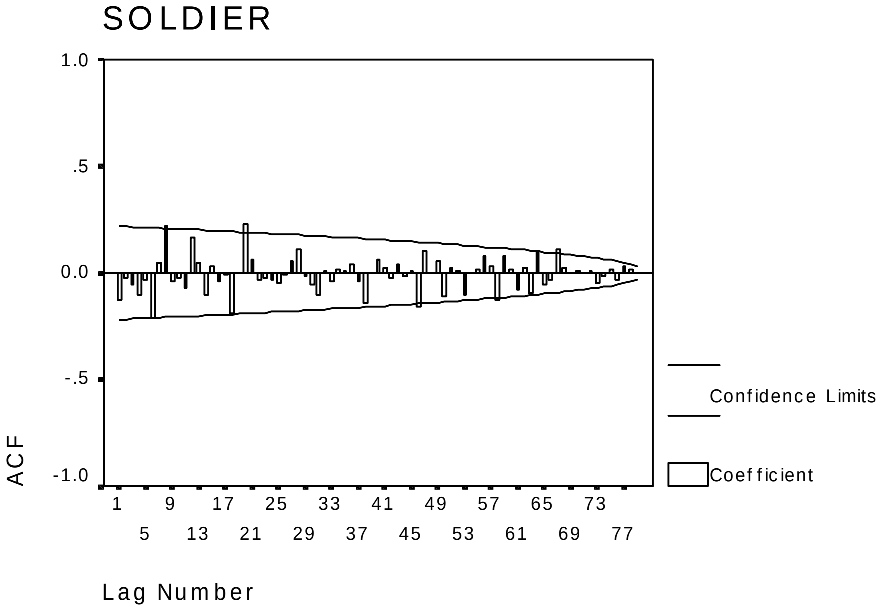
As stated above, this test was composed of three scenarios (the enemy target was located in either the left corner of the room, center of the room or the front right corner of the room). A nine-man squad of soldiers was available for the test. The number of combinations of 9 soldiers working in 3 scenarios is 27 (9 x 3). The study planners suggested that each of these 27 combinations be repeated 3 times giving a total of 81 trials for this test. A completely randomized protocol / design was constructed where each soldier was used 9 times in the test, seeing each of the 3 scenarios exactly 3 times. The randomization process was carried out using a large table of random numbers (2). For each trial (1-81) a soldier (1-9) was selected at random and then randomly assigned to a specific scenario (1-3). This process was carefully carried out until all soldiers were randomly assigned to each of the 3 scenarios exactly 3 times. The result of this randomization process is shown as Table 2 on the next page.

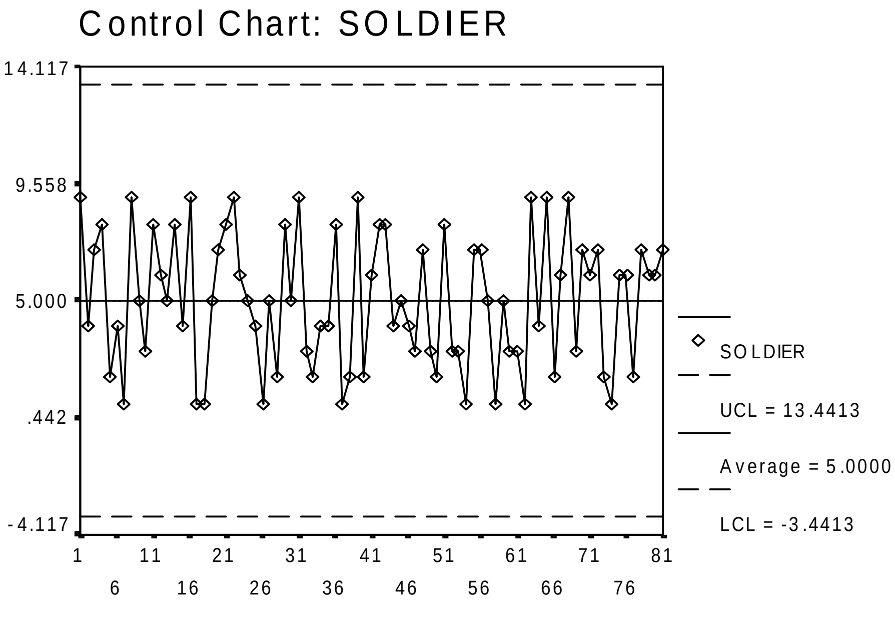
TABLE 2
RANDOM ASSIGNMENT OF SOLDIERS AND SCENARIOS TO EACH TRIAL

TRIAL	SOLDIER	SCENARIO	TRIAL	SOLDIER	SCENARIO	TRIAL	SOLDIER	SCENARIO
1	9	2	**28**	2	1	**55**	7	2
2	4	2	**29**	8	1	**56**	7	2
3	7	1	**30**	5	1	**57**	5	3
4	8	3	**31**	9	2	**58**	1	3
5	2	3	**32**	3	3	**59**	5	2
6	4	2	**33**	2	1	**60**	3	2
7	1	2	**34**	4	3	**61**	3	2
8	9	3	**35**	4	1	**62**	1	2
9	5	1	**36**	8	3	**63**	9	1
10	3	1	**37**	1	1	**64**	4	1
11	8	1	**38**	2	2	**65**	9	1
12	6	1	**39**	9	3	**66**	2	3
13	5	3	**40**	2	2	**67**	6	3
14	8	1	**41**	6	3	**68**	9	1
15	4	2	**42**	8	2	**69**	3	2
16	9	3	**43**	8	2	**70**	7	3
17	1	1	**44**	4	1	**71**	6	3
18	1	3	**45**	5	2	**72**	7	3
19	5	2	**46**	4	3	**73**	2	2
20	7	1	**47**	3	3	**74**	1	2
21	8	2	**48**	7	1	**75**	6	2
22	9	2	**49**	3	3	**76**	6	1
23	6	2	**50**	2	1	**77**	2	3
24	5	3	**51**	8	3	**78**	7	2
25	4	3	**52**	3	1	**79**	6	2
26	1	3	**53**	3	1	**80**	6	1
27	5	1	**54**	1	1	**81**	7	3

An independent verification for the randomization process was conducted. The methods used were described and recommended by the National Institute of Standards and Technology (3) by computing the autocorrealtion function for the sequences of soldiers and scenarios for the 81 trials. SPSS was used to compute these statistics. The autocorrelation functions were lagged from 1-81. There were no significant values for any of the autocorrelation functions (using a P value less than .05 as a criterion of statistical significance). In addition, a sequence quality control type plot was examined across trials (1-81) for the sequence of soldiers and scenarios to determine if there were perceptual clusters or gaps in the plotted series. None were visually apparent. For the sake of saving some space examples of these plots of the autocorrelation functions and sequence plots for the soldiers are shown in Figure 1 on the next page. The plots for the sequence of scenarios would be similar. Although the sequences of random number were drawn from a respected source, sometimes the sequence does not meet the autocorrelation criteria or perceptual quality control images. In those cases, the randomization process should be repeated and verified.

FIGURE 1
AUTOCORRELATION AND SEQUENCE PLOTS





Sigma level: 3

Brief Description of the Test

This test occurred during daylight hours. One live soldier entered the virtual room. One virtual enemy soldier was positioned in the room. He was stationary and positioned randomly in one of the three locations / scenarios described in the section titled **Pre-Test Conditions to Keep Constant**. The live soldiers were randomly assigned numbers from 1 through 9 in keeping with the protocol provided in Table 2. The trials 1 through 81 were completed in the sequencing of soldiers and scenarios as described in Table 2.

A Description of Summary Statistics and Procedures

The analysis procedure was conducted using SPSS. The following steps were taken for each of the tests described in Table 1.

1. Generate box plots and histograms for the data obtained for each of the three scenarios. These graphical procedures provide information to help characterize the data; i.e. is the variable approximately normally distributed? Are there outliers or extreme values? Are the variances approximately equal? Does some of the data need to be deleted because of blunders in recording. Are nonparametric methods preferred to parametric methods? The Explore subroutine and graphics routines of SPSS are useful for doing those exploratory data analysis (4) procedures that should precede inferential data analysis.
2. List and or tabulate the summary statistics for each variable. In this experiment the statistics tabulated for the variable called "engagement time (seconds)" were; arithmetic mean, median, variance, standard deviation, minimum value, maximum value and the range of the data values.
3. If parametric methods were determined to be appropriate (data approximately normal with equal variances), then the one way analysis of variance (ANOVA) was conducted comparing the performance between the three test scenarios. The Tukey post-hoc procedure was used to isolate pairwise differences between the scenarios if the overall ANOVA indicated statistically significant differences (.05 was set as the critical value).
4. If nonparametric methods were determined to be appropriate (data not normally distributed or variances not equal between groups) then the Kruskal-Wallace test was conducted. An appropriate nonparametric post-hoc procedure for statistically significant findings using the Kruskal-Wallace test is found in (5).

A Comment Concerning Engagement Times

An initial assessment of the engagement times indicated that they were larger than expected. These expectations were based on military experience and the results of similar experiments conducted at Camp Lejeune during the MOUT Advanced Technology Demonstration (ACTD). In order to re-evaluate these engagement times each of the engagements was played back and reviewed. It was apparent that a large proportion of the total engagement time was composed of “hallway maneuver” time in the squad synthetic environment. The engagement time required from this test was supposed to consist of the time the soldier was framed (standing) in the doorway of the room plus the time he was shooting his weapon. Based on the playbacks, the “hallway maneuver” time for each trial was subtracted from each total engagement time to provide an estimate of the desired engagement time. On the average, these adjusted engagement times differed by less than 1 second from the time obtained from the MOUT ACTD live experiments at Camp Lejeune.

CONCLUSION

There is little soldier performance data available for describing close combat in an urban environment that can be used by the modeling and analysis community. The purpose of this experiment was to learn about the process of collecting these data in a virtual combat environment. This paper described the experimental design *procedure*, data collection and the follow-on statistical analysis. It was the opinion of the participants that there was excellent coordination between all the members of the squad synthetic study team. The data collection effort was successful and the numerous lessons learned will improve future efforts. This simulator study will be validated with corresponding live experiments conducted at the McKenna MOUT site.

REFERENCES

(1). SPSS Base 8.0 User’s Guide, 1998.

(2). The RAND Corporation, (1955). A Million Random Digits with 100,000 Normal Deviates. The Free Press, New York, various pages.

(3). National Institute of Standards and Technology web site http://www.itl.nist.gov/div898/handbook/index2.htm .

(4). Tukey, John W. (1977). Exploratory Data Analysis. Addison-Wesley Publishing Company, Reading, various pages.

(5). Conover, W.J. (1980). Practical Nonparametric Statistics, 2 ed. John Wiley & Sons, New York, 230-231.

Analysis in the MOUT ACTD

W. M. Christenson, Institute for Defense Analyses

Seven years have elapsed since the Defense Science Board concluded that US National Security needs included the issues associated with probable commitment of Forces to areas of urban conflict. The first major effort that grew from their recommendations was the establishment of the Military Operations in Urban Terrain (MOUT) Advanced Concept Technology Demonstration (ACTD). Most members of the defense community are now aware that the MOUT ACTD has provided certain US Army and USMC units a suite of equipment to use in both training and possible real urban operational settings. The two-year period of this user evaluation will end in FY02. While the final MOUT ACTD report is nearing completion, there are lessons and observations emerging from the MOUT ACTD analyses and associated processes that can, and should, be shared with the community concerned with analysis of operational and technical experimentation, and with urban experiments in particular. This paper deals with those issues from the author's perspective as the lead analyst for the MOUT ACTD, backed by his experience as an analyst at IDA following a career as an army infantry officer. The issues include: the importance and difficulty of gathering data before, during and after urban experiments; visualization of and agreement on appropriate measures; the importance and difficulty of information superiority in urban operations; the difficulty in definition of operational needs; and, balancing those needs with technological capabilities. It is the author's view that while the focus on these issues is sharpened by increased awareness of urban operational difficulties, the solutions to problems presented by these issues is of growing general importance due to reduced force strengths, emergence of new threats and technologies, and smaller budgets. He challenges conference membership to participate in finding such solutions.

Human Factors Evaluation of the Digitized Battlefield (DCX Phase I)

JOCK O. GRYNOVICKI and KRAGG P. KYSOR
Human Research and Engineering Directorate, U.S. Army Research Laboratory

One of the U.S. Army Research Laboratory's (ARL's) Science and Technology Objective (STO) research projects is to develop standardized field-operational soldier performance metrics to quantify integrated soldier-information system performance on the digital battlefield. This research effort is intended to help the Army leadership assess the impact of digitization on individual soldier and staff performance. The paper describes efforts to define and measure Army Battle Command System (ABCS) information interface functionality and usability. The report explains how the evaluation methods and metrics were developed and improved to produce an evaluation package that can be used in other Advanced Warfighting Experiments (AWEs), Command Post Exercises (CPXs), and simulation exercises.

Key Words: Division Capstone Exercise, ABCS, performance metrics, soldier-system interface

1. Introduction

The U.S. Army Research Laboratory (ARL) supported the Battle Command Battle Laboratory and the TRADOC Analysis Center (TRAC) in studying Human Factor issues during the Division Capstone Exercise (DCX). Specifically, ARL's emphasis was on the Army Battle Command System (ABCS) software that was designed to enhance the 4th Infantry Division (ID) soldier and staff performance during the exercise by providing them a clear understanding of the current state of a battlefield situation with relation to the enemy and environment. In this study, we measured digital effects in terms of attitude change, behavior change, command staff task performance, and soldier-computer interface effectiveness.

To study and improve soldier-computer interface software design, a heuristic method of evaluation was used based on human-system interface research outlined by Molich and Nielsen (1990). The report describes how the evaluation methods and metrics were developed and improved to produce an evaluation package that can be transitioned for use in other Advanced Warfighting Experiments (AWEs), Command Post Exercises (CPXs), and simulation exercises.

1.2 Human Factors (HF) Issue Focus for the DCX

The focus of the HF Issue within the DCX was to develop an analytical understanding of how the commander and the battle staff use and interface with the ABCS. The HF Issues analysis was centered on the human dimension of digitized Battle Command by studying the ABCS human computer interface (HCI) 'usability' characteristics and the ability of the ABCS to provide the commander and his staff the required functionality for planning, information management, decision making, and control of the battle-space.

1. 3 Objective ABCS

The U.S. Army Battle Command System (see Figure 1) Capstone Requirements Document (CRD), Revision 3a (Draft, dated 23 November 1999), described the objective system as follows:

> *The ABCS will allow commanders to utilize dominant firepower systems more effectively to destroy enemy forces in an extended area of operations while protecting friendly forces. The firepower will be enhanced by providing the commander the ability to make quicker, more accurate decisions, and orchestrate combat power at critical times and places faster than an adversary. Additionally, the ABCS will enhance SA and enable friendly forces to share a common operational picture (COP) while communicating and targeting in real or near-real time. The ABCS will reduce the uncertainty of war situations, decrease decision-making time, and contribute to increased lethality, survivability, and operational tempo while reducing the potential for fratricide. The objective ABCS will use the Joint Common Database (JCDB) that will maintain the data elements required to build and provide the commander's COP of the battlefield.*

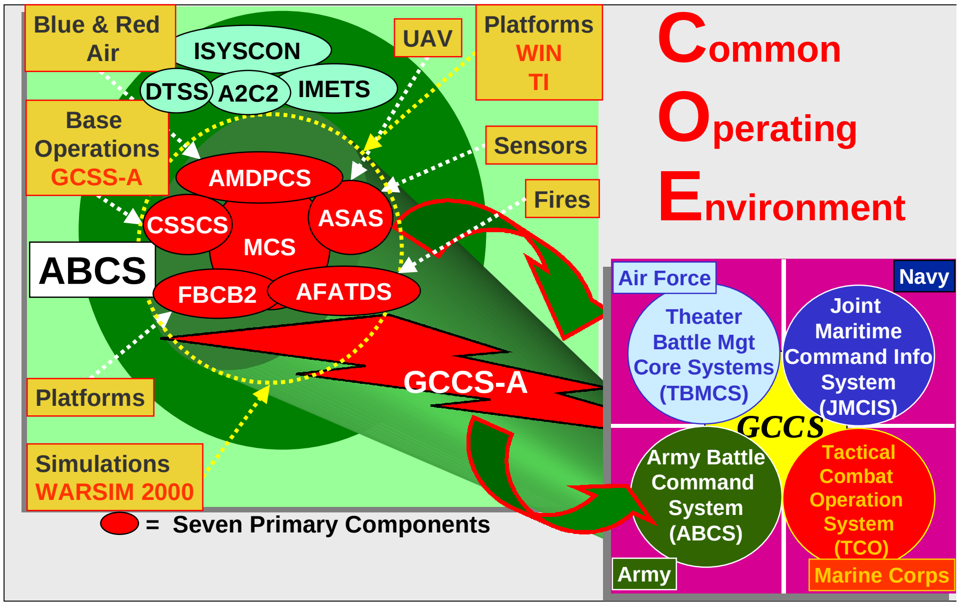


Figure 1. Objective ABCS

ABCS is an evolving "system of systems" that needs individual subsystem testing and evaluation. The entire family of systems will be assessed individually and collectively to ensure that the functional and usability requirements are met as well as the overarching commanders' decision-making and requirements.

(1) The Advanced Field Artillery Tactical Data System (AFATDS). *AFATDS provides automated decision support for the fire support (FS) functional subsystem, which includes both Joint and Combined fires (naval gunfire, close air support, etc.). AFATDS provides a fully integrated FS C2 System, giving the FS coordinator (FSCOORD) automated support for the planning, coordination, control, and execution of close support, counter-fire, interdiction, and air defense (AD) suppression fires.*

(2) All Source Analysis System (ASAS). *ASAS is the Intelligence and Electronic Warfare (IEW) component from battalions to echelons above corps (EAC). ASAS receives and rapidly processes large volumes of combat information and sensor reports from all sources to provide timely and accurate targeting information, intelligence products, and threat alerts. It consists of evolutionary modules that perform system operations management, system security, collection management, intelligence processing and reporting, high value/high payoff target processing and nominations, and communications processing and interfacing. The ASAS Remote Workstation provides automated support to the doctrinal functions of intelligence staff officers (G2/S2) from EAC through battalion, including Special Operations Forces (SOF).*

(3) Combat Service Support Control System (CSSCS). *CSSCS provides critical, timely, integrated, and accurate automated combat service support (CSS) information to include all classes of supply, field services, maintenance, medical, personnel, and movements to CSS, maneuver and theater commanders and logistic and special staffs. Critical resource data is drawn from both manual resources and the Standard Army Management Information Systems (STAMIS) at each echelon.*

(4) Forward Area Air Defense/Air and Missile Defense Workstation (FAADC2/AMDWS). *The FAADC2/AMDWS integrates Air Defense (AD) fire units, sensors, and C2 centers into a coherent system capable of defeating/denying the aerial threat (Unmanned Aerial Vehicles (UAVs), helicopters, fixed wing). The system provides the aerial dimension SA component of the COP. Initially, the Air and Missile Defense Workstation (AMDWS) will provide elements from EAC to battalions the capability to track the air and missile defense battle Force Operations (FO).*

(5) Maneuver Control System (MCS). *MCS is the primary battle command (BC) source, providing the COP, decision aids and overlay capabilities to support the tactical commander and the staff via interface with the force level information database built from the other Battlefield Automated Systems (BASs). MCS provides the functional common applications necessary to access and manipulate the JCDB.*

2. Method

2.1 Data Collection: Subject Matter Expert (SME) Observers and Data Analysts

ARL provided the following resources: (1) an issue proponent analyst manager, three HFE SME observers, and three HF analysts to serve throughout the simulation exercises (SIMEXs) and the AWE. (2) A sub-set of SMEs was assigned by the Operational Test Command (OTC) to support ARL in collecting HF related observations. ARL developed an HF Observer's Guide and provided the HF military SMEs with training just prior to the AWE start of the exercise (STARTEX). The SMEs conducted HF-focused observations throughout the AWE that were recorded on laptop computers for daily downloading to the OTC DCX database repository. (3) ARL executed analysis oversight of the HF observations including the resolution of anomalous observations. (4) ARL developed a two-part HF-focused questionnaire survey. The first part focused on general human factors aspects of command and control issues (e.g., setting up a tactical operations center (TOC), situational awareness (SA), COP, battle-tracking, timely commander decision making, and interoperability of the subsystems). The second part of the survey focused on soldier-computer interface usability aspects of the ABCS in general as well as the specific ABCS subsystems. The two-part survey was administered to the 4th ID ABCS users by the OTC following the DCX end of experiment (ENDEX). (5) Interview questions were developed by ARL and administered by OTC and TRAC to the 4th ID commanders and staff. In summary, the data sources consisted of SME observations, ABCS user 'HF Survey' responses, supplemented by commander and staff interview responses documented by OTC.

2.2 Materials

2.2.1 ARL's ABCS Issues Section of the Survey and Guide

The Universal Joint Task List (UJTL) was used to identify essential tasks that a combat commander is required to perform in exercising command and control. This list serves as an interoperability tool to help commanders construct their joint mission essential task list. It is a comprehensive hierarchical listing of the tasks that can be performed by a joint military force. UJTL is organized into four separate parts by the level of war: (1) Strategic level-National military tasks, (2) Strategic level-Theater tasks, (3) Operational level, and (4) Tactical level tasks. Each task in the UJTL is individually indexed to reflect its placement in the structure. Thus, the UJTL provided a Command Staff task baseline around which ARL developed its standardized soldier performance metrics research efforts

Utilizing the Department of Defense (DOD) UJTL for command and control (C2) as a foundation, ARL's HF C2 issues section of the survey or guide focused on the interrelationship between the division staff functions or processes required for effective command and control decision making as supported by ABCS software. ARL’s survey metrics methodology involved a cross-linking of FM 101-5 (Staff Organization & Operations, 1997) military decision-making processes (MDMP) with the ABCS software modules believed to support critical command and staff task execution. The U.S. Army’s field manual (FM101-5) states that a staff supports the 'Science of Control' in four primary ways: (1) gathers and provides information to the commander, (2) makes estimates of the set of actions required, (3) prepares plans and orders, and (4) measures organization behavior. To perform this type of support, the staff and commanders use various time-dependent decision-making and information management processes that require extensive staff coordination between and within echelons. Shortcomings in command, control, communications, and intelligence (C3I) automation functionality can lead to serious tactical failures such as inadequate battle plans, inadequate reporting, lack of coordination, and inadequate situation awareness that can result in fratricide.

2.2.2 HF SME Observer Guide

The guide provided information for the SME on which to focus personnel and digitized equipment factors to help answer each associated HF Sub-Issue. The HF issues and examples of Essential Elements of Analysis (EEAs) of the SME Observer’s guide are outlined in Table 1.

TABLE 1
Human Factors Issue Observer Guide for the ABCS Subject Matter Experts

Issues and EEAs	**Description**
(A) Issue HF 01. How adequate, efficient, and user-friendly are the ABCS information interfaces in enhancing soldier and staff performance.	This is part of the U.S. Army's attempt to assess the value of the ABCS for heavy force military operations, it is important to understand the effectiveness of the individual soldier-ABCS system interface.
(1) EEA HF 01.01 Did the soldier-computer information interfaces of the ABCS enhance soldier-operator and staff performance?	Consider the various aspects and features of the screen displays and presentation of information for ease of use in accomplishing and enabling the commander's mission tasks.
B. Issue HF 02. Does the *First Digitized Division Priority 1 computer architecture* support the task and cognitive processes needed to enhance commanders and staff's performance?	
(1) EEA HF 02.01 Did ABCS support adaptive commander or staff by permitting timely development and sharing of commander's intent, facilitate vertical and horizontal cohesion, that support commander or staff teams despite leader changes during the phases of the MDMP regardless of unexpected events?	Consider the ease or difficulty of using the ABCS information formats to obtain battle tracking data to support the commander in making timely and effective decisions by being responsive to unexpected changes during the planning, preparation, and execution phases of the MDMP.
C. Issue HF 03. How do ABCS system reliabilities affect the staff's performance?	
(1) EEA HF 03.01 Do the ABCS information presentation formats and computer interface enhance the staff's ability to quickly and accurately access distributed data sources at any time?	Consider the reliability of the ABCS in supporting the commander and his staff in the execution of C4ISR tasks and accessing distributed data sources during the course of a battlefield mission.
D. Issue HF 04. Is the ABCS architecture effective in the pulling, pushing, and assimilation of information and maintaining a COP and supporting battlefield visualization?	
(1) EEA HF 04.01 How well do the ABCS digitized data format designs and soldier-computer interface help the staff develop, maintain, distribute, and assimilate the COP?	Consider the ease of use of the ABCS media (e.g., VTC) and information presentation formats in supporting the commander and his staff in working collaboratively with other echelons in the development of a COP.

2.2.3 ARL’s HF ABCS General and Specific ABCS Subsystem User’s Survey.

In the ARL ABCS HF Survey's application, a heuristic methodology was used (a method of usability analysis in which users are presented with an interface design and then requested to comment on it). For the DCX, the 4th ID ABCS operators were asked to rate each usability characteristic (sub-issue item) on a scale from 1 to 5 to rate the ABCS software design as it attempts to support effective execution of critical TOC Staff tasks.

TABLE 2
ABCS Human Computer System Usability Characteristics

Tempo
Utility
Flexibility in Use
Prevent Fatigue
Use Army Doctrine
Provide Process Shortcuts
Consistency Between Modules
Minimize Demand on Human Memory
Provide Feedback
Good Error Recovery
Common Framework
Intuitiveness

TABLE 3
Tactical Operations Center (TOC) Staff Tasks

Setting up Local Area Network (LAN) Addresses
Using Communications Networks
Developing Situational Awareness
Determining the Commander's Critical Information Requirements
Determining Locations of Enemy & Friendly Units
Building Overlays & Templates
Creating, Editing, Updating Data Bases
Building Friendly & Enemy Order of Battle
Building & Modifying Synchronization Matrix
Preparing Unit Task Organizations
Computing Force Ratios
Determining Equipment & Personnel Resources
Coordinating Joint Services Defense Resources
Preparing Defense Assessments
Developing Courses of Action
Making Accurate & Timely Decisions
Preparing Briefings
Preparing Operation Orders & Reports
Sending & Receiving Information

The 'usability factor' has a direct impact on staff performance because shortcomings in system usability lead to underlying error patterns, attention deficits, and excessive workload which can be linked to inappropriate decisions and priorities, serious delays in operational tempo, and failures in effective staff coordination and communications. This ABCS HF Survey was guided by many human-computer system issues (see Table 2) that have been defined in the research literature (Nielsen & Molich, 1990; Molich & Nielsen, 1990; Nielsen & Levy, 1994; Smith & Mosier, 1986) as reflecting hardware and software design with good interface usability. The usability characteristics include: whether the computer system contains simple and natural dialogue, applications reflect military doctrine, 'speaks' the user language, minimizes user memory load, remains consistent between different modules and across applications, provides user feedback, provides clearly marked exits from modules, provides process shortcuts, and prevents errors. Examples of more complex staff tasks involving cognitive aspects of decision-making (see Cannon-Bowers & Salas, 1998) are presented in Table 3.

These metrics addressed critical functional dimensions of staff performance within the Military Decision Making Process that included: (1) Mission Analysis, (2) Course of Action (COA), (3) Information Assimilation, (4) Generation of Messages and Reports, (5) Workload Distribution, and (6) Development, Distribution and Maintenance of Situation Awareness.

3. Results and Discussion

3.1 Analysis of Data

The ABCS User’s Survey responses were obtained from using a five-point Likert-type scale to quantify the systems’ functionality and usability. Chi-Square analyses were performed to determine the significance of the percentage of responses in each of the five rating-scale cells. To obtain adequate statistic power in cases where there were very small sample sizes, the number of response category cells was collapsed to meet the power requirements of the Chi-Square statistical method. Descriptive data was obtained from SME observers and operator comments. Their documented narrative responses were analyzed using the HF issue observer guidelines (Table 1) and the ABCS usability characteristics listed in Table 2.

3.2 Issue HF 01.01 - How adequate, effective, and user-friendly are the ABCS information interfaces in enhancing soldier and staff performance?

3.2.1 AFATDS

In general, the AFATDS was effectively used to produce fire support products more quickly than using non-digital means. Operators thought it was easy to construct graphics using AFATDS tools. However, some SMEs noted that the lack of interoperability between the ABCS sub-systems for graphics and overlays caused the targeting officer to have to manually input the fire support overlays.

Set-up, initialization instructions, new user, master list, unit ID’s, status, communication, LAN modification were system start up tasks that were considered to be the most complicated processes of the AFATDS system. Operators used a “cheat card” with 21 steps, each step requiring up to 4 sub-steps to complete each major step.

Operators thought this process was too complex, requiring 30 minutes under the best conditions.

The SA picture generated by the AFATDS was not felt to be timely and was too cluttered to be used. The AFATDS screen display was hindered by the sheer volume of information being presented which made the COP difficult to understand. Unit icons were superimposed on one another, on top of obstacle graphics, and on top of general axes of advance, which made it difficult to pick one specific icon and retrieve information about it.

The AFATDS feedback regarding help and prompts was adequate for some functions such as troubleshooting, but incomplete for others such as graphics production. Operators reported that if they got an error message there was not a clear indication on how to fix the problem. On-screen instructions, prompts, and menus were generally good, but operators agreed that error messages should be adequately supported by information or methods that correct the error. Abbreviations, acronyms, codes, icons, and symbols were good. They directly replicated artillery symbology

The interface facility with avoiding input errors, and showing selected attributes was good. For example, if an operator entered an incomplete grid number, the system would not permit him to proceed. The system was good at preventing accidental keystrokes. System prompts requested the operator to verify execution of keystroke errors.

Operators suggested that AFATDS was a fine tool for fire support planning and mission processing. For example, SMEs reported that the 2nd Brigade Combat Team Fire Support Element (FSE) was able to maintain timely and accurate status of all firing units. as well as the Blue and Red SA provided by ASAS.

The concern for system reliability (e.g., lock ups, false error messages) required the operators to insert or duplicate tasks with manual or analog methods. This caused an increase in operator and staff workload that resulted in a decrease in their effectiveness.

Generally, required information was on the data displays, but various improvements were suggested. The interface could be improved if the operator didn’t have to flip back and forth between screens during "Calls-for-Fire" to verify certain target characteristics. The number of menus and screens needed to complete processes were generally adequate. However, in the case of sending free-text messages, the operator was required to perform eight steps. The operators report that this process was too complex and needed to be simplified in a way that is similar to using commercial e-mail systems.

3.2.2 ASAS

It was reported that technical problems prevented the ASAS link with the Joint Common Data Base (JCDB) which caused the S2 to use manual tracking methods to organize Spot Reports and perform critical actions.

ASAS operators had no problems with functions involving system set-up (e.g., initialization instructions, new user identification, master list, unit ID’s, status, communication, LAN modification) and TOC relocations. An adequate interface was provided for working with maps, but the software increased the operator's workload compared to the prior version’s Terrain Evaluation Module (TEM) versus the Joint Mapping Tool Kit (JMTK). These problems resulted in the majority of the ASAS operators (69%) rating the “maps drawing tools as unfriendly. Although the standard report formats were not difficult to use, the operators preferred creating free-text reports because it was easier and more familiar for them. The majority of the operators rated the use of the standard report format as being adequate.

The ASAS did not provide adequate feedback to allow the average operator to tell what effects his actions were having on the system. Although the software was generally good at helping the operator avoid data entry mistakes, there appeared to be a potential problem regarding visual cues and selected data entry attributes.

On-screen instructions, prompts, and menus were generally good. Operators found the message prompts to be useful. The task for creating a message distribution list was easy. Abbreviations, acronyms, codes, icons, and symbols were good except for their representation in the military symbols (MILSYM) manager module. There is a substantial display of icons in MILSYM, but they have no labels. Less experienced troops cannot identify them at first glance.

The amount of frustration and stress experienced appeared to increase when the ASAS or the overall digital system did not function properly or the system crashed. Operators felt that the software products and ASAS interface were “unstable” (e.g., system freezes and function failures). Consequently, they could not use their system to directly communicate with other ABCS systems.

The number of menus and screens needed to complete processes was adequate, but generally thought to be too numerous. For example, during a spoiling attack, the ASAS operator needed to perform a communication operation on the system to verify connectivity. To do this the user had to navigate through more than four menus and sub-menus before he could actually contact the recipient.

Operators (77%) reported that inputting information into the RWS databases was easy. They almost exclusively used the short form that appeared to be adequate for their purposes. They suggested that if they had the ability to enter battle damage assessments (BDA) it would improve their processing interface and more timely support the command regarding BDA. INTEL staff tracked the BDA manually.

3.2.3 CSSCS

In general, CSSCS operator tasks were considered fairly intuitive regarding automated processes. Some CSSCS operators reported the software was flexible and allowed them to modify their processes. These operators liked the ability to change echelon reporting levels so one could look at specific assets of interest. However, other operators reported that the software did not give users options to modify the processes or sequences of support task requirements. As a result, the operator had to use MS EXCEL instead of the CSSCS for maintenance reporting.

CSSCS reported digitization increased their speed regarding receiving OPORDs compared to non-digital methods. The CSSCS software interface was considered to be soldier friendly. Operators reported consistent interface controls, presentation, familiar words and menus. CSSCS "drag and drop" procedures were very soldier-friendly. The CSSCS main menu bar and pull-down menus were easy to use.

CSSCS fatigue levels were reduced by using digitized versus non-digitized methods. Operators saw a fatigue reduction with specific digital functions such as logistics statistics (LOGSTAT) reporting and Unit Task Organization (UTO) processes. Operators reported that the UTO automated update process for planning and execution was a useful tool. CSSCS colored displays were easy to use. Operators found the “Gumball” formatted display screen showing logistical resources to be very helpful.

Standard report formats were not difficult to use, especially using the Rapid Data Entry option. However, other operators used free-text messaging because they thought they were getting inaccurate and outdated data. Consequently, they often used voice means to get current data. It was easy to use the Equipment-Force Echelon Status Report. Likewise, the Equipment-Item Status report, the Battle Loss Unit Summary Report and the Personnel Daily Summary Force Echelon Report were easy to use. The process to obtain a Class III Bulk Force Echelon Report was easy. The Baseline Resource Item List (BRIL) and Critical Tracked Items List (CTIL) forms were easy to use.

The CSSCS provided some feedback to assist operator functions, but improvements need to be made. Prompts were sometimes vague by identifying an error but not providing information to correct the error. CSSCS was reported as being "fair" to "good" in providing error prevention or recovery capability. On-screen instructions, prompts, and menus were generally good. Some operators considered the prompts incomplete and needed to be more useful by providing the necessary information to resolve a fault, failure, or error. Operators particularly liked the icons, codes, and acronyms. Being able to click on icons and symbols for identification was a great help to the user.

Overall, operators rated CSSCS messaging (e.g., receiving, preparing, & sending) less than adequate. They said the process should be as easy as commercial e-mail. The process of addressing and sending required too many steps. The CSSCS Message address screen was easy to use, but it did not contain all the addresses required for message distribution. The number of menus and screens needed to complete processes was generally thought to be too numerous.

The system was good at preventing accidental keystrokes, with the exception of the “power” key. Accidentally striking the “Power” key locked up the system. In order to unlock a subsystem, the entire system had to be rebooted.

The CSSCS placed a "moderate" to "high" demand on human memory. The multiple number of menus to perform certain tasks (e.g., messaging) was excessive. When some CSSCS operators processed volumes of information they made "cheat sheets" to remember the required operations.

3.2.4 FAADC2 / AMDWS

AMDWS allowed the operator to monitor current air operations while assisting the commander to plan for future events. The commander had complete SA during the Air Force close air support (CAS) mission conducted during enemy advancement into a sector. The air defense (AD) battle captain at the Division Tactical Analysis Center (DTAC) command center had a live feed and was able to share information with the staff and the Forward Air Controller. This system gave the staff SA of the deep fight when used in conjunction with the Joint Surveillance Target Attack Radar System (JSTARS) picture. Status updates by the system FAAD engagement operations (EO) were generally good. Air tracks were timely as long as radars were functioning. The air defense artillery

(ADA) cell provided the air SA picture to the DTAC and DMAIN. The SA picture was clear and concise, aided the staff in identifying enemy air activities, and accelerated the MDMP.

The users of the system stated that the start up and sharing of information provided by the system would be improved if the internet protocol (IP) addresses used to identify subordinate units were more user friendly. If the user did not know the intended recipients' IP address he was not able to directly send e-mail to the individuals who needed the information. Instead, he had to place the information on the TAC web and hope the appropriate user looked and found the information in a timely manner. Operators could not modify the communications table or change the node configuration. These modifications were performed by the contractor.

The graphic user interface allowed for operator flexibility. Shortcuts capabilities were helpful, but FAADC2 had better shortcut capabilities than AMDWS. The graphics and drawing tools for developing products were considered to be adequate.

AMDWS was reported to be adequate regarding error prevention and helping operator recovery. The interface for avoiding input errors, and showing selected attributes was adequate. Certain tools (e.g., grid locations in line-of-sight analysis) could use more "Help" analyses because the error involved was not obvious to the operator.

The demand on human memory needed to complete tasks was not excessive for the operator but could be further reduced by eliminating some of the windows needed to complete a task. On-screen instructions, prompts, and menus were generally good. Abbreviations, acronyms, codes, icons, and symbols were also good. There were no problems in the use of the mouse to click and double-click on functions. The system was good at preventing accidental keystrokes.

The common message processor (CMP) messaging system met specified requirements but the operators preferred to use free-text messages indicating a need to improve user friendliness. The system prioritized users' incoming messages and permitted the user to prioritize his messages, but operators did not use these capabilities. The process of creating a message distribution list was considered adequate.

3.2.5 MCS

The software was fairly consistent in function. Operators reported that their general fatigue level was less with MCS than with non-digital means, but mental fatigue may be greater. The MCS appeared to require moderate demands on human memory. Memory demands were high for tasks requiring multiple commands, menus, and screens, especially during peak information periods.

The graphics and drawing tools for developing products were reported by SMEs to be adequate for performing many critical tasks. In general, operators preferred using the automated overlay tools rather than producing overlays manually. The benefits of the automated methods were decreased task time and workload.

Many of the operators (23%) felt the graphics and drawing tools to be difficult to operate. There were some difficulties with naming conventions, overlay construction, and problems with drawing boundaries. Operator suggestions to improve the interface were to allow the immediate transfer of the commander's sketch to an "overlay" to save time and ensure "actual" commander representation.

MCS users reported that the interface for the development of the COP was good. They believed the COP to be the best application of the MCS.

The MCS filter function interface was operator friendly. Operators reported that filter functions supported force level control and SA throughout the Division. If the display was not too cluttered then it was easy to identify military units by clicking on the appropriate icon representation for the unit. Operators (46%) reported that the icon identification interface was friendly

The interface facility with avoiding input errors, and showing selected attributes was accurate, but in order for the operator to determine whether he had selected the correct attributes in the system, he had to physically stand up in the shelter, pull the system keyboard out to its furthest position, and look on the top of the keyboard above the number keys to determine whether or not the correct attributes had been selected. Suggestions to improve the interface were to display the attributes on the monitor, or make the keyboard more accessible. The system was good at preventing accidental keystrokes. System prompts requested the operator to verify execution of keystroke errors.

SMEs reported that the system was adequate for receiving and preparing messages, but not for sending messages. Surveyed operators (58%) found that the message handling capabilities of the MCS were unfriendly or only adequate. The system tended to lock-up with multiple addressing, so operators often had to send messages one at a time which added to their workload. In addition, acknowledgement of their message was by voice that also increased the messaging time and workload.

Operators reported that MCS was not very good at providing error prevention and recovery. Operators (64%) felt that the interface for unlocking a subsystem was only adequate or unfriendly. They reported that the system crashed too easily and would like the system to have more processing power.

Some operators felt that the use of the MCS increased the speed of performing certain critical tasks compared to non-digital means. The plans and orders were transmitted electronically by the system to the organizational structure. However, other operators reported that the system did not increase their task performance compared to non-digital means. They cited problems with system reliability, long initialization time (45 min.), and that reports using system information may be outdated.

3.2.6 ABCS Data Filters

Filters reduce screen clutter and thereby minimize workload associated with readability and understandability of information. When displaying targets on the COP, the screen becomes extremely cluttered and masks key graphical information from the commander. TTPs must be established for data display filtering at each echelon of command to ensure that only relevant material is displayed. Filtering capabilities need to be developed for detecting specific obstacles such as mine fields. During the DCX, a lack of enemy minefield information resulted in fratricide (simulated).

3.2.7 Does the ease of use of the ABCS information designs and interfaces help distribute products faster and distribute them to the proper places?

ABCS allows units to distribute information products (overlays, reports, and messages) among units that have the core ABCS systems, but limits in bandwidth prevented efficient passage of graphics and other large products. The limited bandwidth caused overlays to be sent in pieces that leaves many opportunities for failure with both the sender and receiver. However, a digitized unit has far greater internal messaging capability than a non-digitized unit. The majority of survey respondents indicated that ABCS had a positive impact on their ability to receive critical graphs and messages in a timely fashion. Generally, digitization enabled the command to disseminate products to more places and faster.

3.3 Issue HF 02. Does the *First Digitized Division Priority 1 computer architecture* support the task and cognitive processes (e.g., information assimilation, situational awareness, and decision making) needed to enhance commanders and staff's performance?

3.3.1 General

While ABCS does not currently have all the requirements specified for the objective system, the ABCS has the capacity to support the commander and staff's task and cognitive processes better than the manual systems involving acetate covered maps and "yellow canary" message books. However, standing operating procedures (SOPs) linking the current features of the system to the team and individual cognitive processes that comprise the command and staff operations could improve the system's performance. The sources of situational awareness necessary to conduct these processes are supported by the following three ABCS capabilities:

(1) ABCS can display friendly unit locations in near real time in each command post and operations center throughout the Tactical Internets (TI), and at a level that suits the requirements of the decision maker at each location. ABCS makes unit location data available to the system operators. The human factors challenge is for battle staff members to know what unit information they need and how to display it on the appropriate screens.

(2) With the exception of file size constraints in sending graphics files from ABCS to FBCB2, ABCS can create and disseminate battlefield geometries among the command posts and operations centers. The system permits operators to create lines, shapes and symbols relatively easily and to disseminate them quickly through several procedures. Thus, the operational graphics received at each location are exact copies of the original, a feature that is not possible when overlays have to be reproduced manually, one at a time. Also, the graphics are disseminated electronically, a faster process than courier distribution. The drawing tools are less facile than the human hand, consequently, the operational control measures are more difficult to tailor to the exact flow of terrain features, e.g., stream beds, ridgelines, country roads. The drawing tools are also more time-consuming than hand graphics.

(3) Currently, ABCS's capability to support the intelligence analysis and fusion processes is superior to earlier manual processes. ASAS's All Source Correlated Database (ASCDB) has an impressive capability to receive, store, and display (based on analysts' queries) a wide range of combat information provided by sensor systems. The principal shortcoming is that overlays of enemy unit locations developed in ASAS cannot be

combined in one composite display with the current friendly unit situation and the battlefield geometries. Instead, the enemy unit overlay must be displayed on a separate screen. From a human factors perspective, this is not satisfactory. However Tactics, Techniques, and Procedures (TTPs) documents could be developed to integrate the enemy unit locations into the friendly unit display.

3.3.2 Critical Events, Unexpected Changes, and Uncertainty. The timeliness and detail in the friendly unit situation display on the COP permitted the staff to monitor critical events in the current operations order better than manual systems. However, the system had no special features to monitor changes in the enemy situation. Initial information on emerging enemy activities was generally disseminated verbally or by message before the analysts were able to post appropriate enemy unit icons to the COP. The capability of the system to create and disseminate text or graphic files rapidly and accurately greatly enhances the units' ability to react to unexpected changes. ABCS has no special features to facilitate the commander and staff managing uncertainty.

3.3.3 Support to Decision Making. ABCS provided indirect support to decision making during the planning process. The maps and graphics tools and the office products allow the operational planning teams (OPT) to visualize the operation and to quickly create planning products throughout the flow of the planning process. The synchronization matrix allows the OPT to capture the decisions made during the course of action wargame, and easily translate the decisions into the task and purpose statements in the execution paragraph of the operations order. During the execution phase, the fact that unit locations are automatically updated in the system, by the Force XXI Battle Command Brigade and Below (FBCB2) computer, freed the commander and staff to concentrate on synchronizing the subordinate units' shaping and decisive actions, and anticipate subsequent actions necessary to maintain operational tempo. In contrast, in non-ABCS operations centers, the staff must expend considerable effort updating each subordinate unit's current location while concurrently attempting to concentrate on the tactical elements of the operation.

3.4 Issue HF 03. How do ABCS system hardware and/or software reliabilities affect the staff's performance?

ABCS Reliable Throughout DCX I. The networks worked well throughout the DCX. At almost no point did staff members lack connectivity for purposes of sending or receiving data over the networks. The fact that the networks were highly reliable is a success for ABCS. Although network reliability is a technical issue, testimony to the system reliability was the ability of the units to quickly re-establish their internal and external networks after displacing their TOC or TAC.

3.5 Issue HF 04. Is the ABCS architecture effective in the pulling, pushing, and assimilation of information and maintaining a Common Operational Picture (COP) and supporting battlefield visualization?

3.5.1 Key to COP is the Joint Common Database (JCDB). The data comprising the COP is extracted from the JCDB. The JCDB receives and distributes three elements of information that are the core of the COP: (1) friendly unit locations, (2) current battlefield geometries, and (3) enemy unit locations. ABCS performed very well on friendly unit locations and the current battlefield geometries, but due to technical shortcomings, the battle staffs had to devise alternate solutions to display a reasonable view of the current enemy situation.

3.5.2 "Blue" Situational Awareness. The "friendly unit locations" function is the only one of the three that is executed almost entirely by the system. The key is the FBCB2 computer. The reporting signal moves from the lower TI to the upper TI where "Embedded Battle Command" translation software on each ABCS system, converts the FBCB2 data to a format that can be read by the ABCS systems. The data goes to the JCDB, and automatically populates the ABCS Common Operational Picture at that location

3.5.3 Operational Graphics. MCS-Light is effective in preparing and distributing graphics and overlays. Once created, the basic overlay is saved via the maps and overlays function on MCS-Light, and when saved initially, is actually saved in an MS Access database identified in the MSC-Light Microsoft (MS) Explorer window as a "JCDB" folder. Depending upon the filter settings, appropriate graphics are created on the COP based on data stored in the JCDB.

3.5.4 "Red" Situational Awareness. The major shortfall in the COP was the difficulty displaying the current enemy situation. The JCDB was not able to populate the COP with enemy unit locations on a timely basis. Operators believe that the enemy unit locations are transferred from the All Source Correlated Database (ASCDB) in ASAS to the JCDB. The system is designed so that the Blue unit locations are refreshed on the COP display first, and the Red unit locations, second. The Blue refresh rate is apparently set for more frequent intervals and will override a Red refresh cycle in progress. Thus, the Red unit locations were continuously overridden by the Blue refresh rate. The alternate solution was to display the ASCDB version of the current enemy locations in a separate window on the COP display. Consequently, the Blue and Red displays were adjacent but not superimposed. Accuracy (defined as where the enemy unit is now) is more a TTP problem than system design. HF areas of improvement include: (1) ensuring that Spot Reports are integrated into the digitized system, (2) data filters are used appropriately, and (3) intelligence analysts are trained to a high degree of proficiency.

References

Cannon-Bowers, J.A. & Salas, E. (Eds.). (1998). *Making Decisions Under Stress: Implications for Individual and Team Training.* Washington, DC: American Psychological Association.

Leedom, D. & Simon, R. (1995). Improving Team Coordination: A Case for Behavior-Based Training, *Military Psychology*, **7**(2), 109-123.

Molich, R. & Nielsen, J. (March 1990). Improving a human-computer dialogue: What designers know about traditional interface design. *Communication of the ACM,* **33(3)**.

Nielsen, J. & Molich, R. (1990). Heuristic Evaluation of User Interfaces. *Computer Human Interface (CHI) 90 Proceedings,* 249-255.

Prairie Warrior 96 Experimenting Agencies. (1996). *Prairie Warrior 96 Advanced Warfighting Experiment Final Report.* U.S. Army TRADOC Analysis Center (in press).

Smith, S. & Mosier, J. (August 1986). *Guidelines for Designing User Interface Software. Report MTR-10090.* The Mitre Corp., Bedford, MA.

Contributed Session II

Advantages of NDE Data Over Destructive Testing Data

C. Shane Reese and Mike Hamada, Los Alamos National Laboratory

We present a framework for exploring the statistical advantages of non-destructive evaluation (NDE) data over destructive testing (D-test) data. This framework allows for quantitative comparison and relative merit of the two different testing scenarios. We evaluate both testing scenarios under both discrete and continuous cases. Included in the results are suggestions for an equivalent number of D-tests for a given number of NDE tests.

Statistical Artifacts in the Ratio of Discrete Quantities

Roger G. Johnston, Los Alamos National Laboratory

The ratio is a familiar statistic, but it is often misused. One frequently overlooked problem occurs when ratioing two discrete (digital) variables. Fine structure appears in the histogram of the ratio that can be very subtle, or can sometimes even dominate the histogram. It disappears when the numerator and/or denominator become continuous. This statistical artifact is not a binning error, nor is it removed by taking more data. It is important to be aware of the artifact in order to avoid misinterpretation of ratio data. Examples of the statistical artifact appear in the areas of flow cytometry, data acquisition, digital-to-analog conversion, computer modeling, light scattering, and baseball. There are a number of ways to avoid or minimize the problems that the artifact can cause.

Statistical Assessment of Aging Materials

Joanne Wendelberger, Los Alamos National Laboratory

As materials age, their physical properties may change over time. Statistical methods are developed to assess material aging and degradation. Aging data may include a variety of different types of data. With the development of new and improved chemical instruments, aging data often include measurements in the form of curves or spectra. Methods for utilizing various types of data in assessing aging phenomena are explored. Examples encountered in the examination of materials aging data will be used to illustrate the proposed methods.

Contributed Session III

Simulating Survival Probabilities of a Lander Mission on Mars

Karen Kafadar, University of Colorado-Denver

Lockheed-Martin Company (LMCO) must design lander vehicles to survive their landings on the rock-strewn surface of the planet Mars. The failure of the Mars'98 lander, though for reasons other than the planet's surface, increased LMCO's attention to this problem. Given the basic structure of the lander and the characteristics of rocks that are likely to cover the planet's surface, what is the probability that the lander will survive its mission? Mathematics students at CU-Denver during the Spring 2000 semester developed an algorithm to simulate this probability. We describe the components of this algorithm and propose a possible Monte Carlo Swindle to increase its efficiency.

This project involved many themes that typically arise in scientific modeling problems: exploratory data analysis on the features of rocks measured from previous Viking Lander missions, models of rocks using contaminated normal distributions for the “killer rocks”, and some theoretical analysis needed to develop a Monte Carlo swindle for this problem.

LA-UR-02-0014

Approved for public release; distribution is unlimited.

Title: FUNCTIONAL SENSITIVITY ANALYSIS FOR COMPUTER MODEL OUTPUT

Author(s): Katherine Campbell

Submitted to: Proceedings of the Seventh Army Conference on Statistics,
Santa Fe, NM,
October 22-26, 2001

Functional Sensitivity Analysis for Computer Model Output

Katherine Campbell
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM 87545

Abstract

The outputs of computational models are often time series or functions of other continuous variables (space, angle, etc.) For the purposes of model sensitivity and uncertainty analysis, it makes little sense to treat individual points on these curves as scalars. Of much greater interest is the effect of model input choices and uncertainties on the overall shapes of such curves. We explore a range of methods for characterizing a set of functions generated by a series of model runs for the purposes of exploring the relationships between these function and the model inputs.

Introduction

The outputs of computational models are often time series or functions of other continuous variables (space, angle, etc.) In this paper, we propose that sensitivity analysis of such outputs be carried out by means of expansion of the functional outputs in an appropriate functional coordinate system, i.e., in terms of an appropriate set of basis functions, followed by sensitivity analysis using any standard method of the coefficients of the expansion. The only new problem, therefore, is choosing an appropriate coordinate system in which to apply the selected sensitivity analysis methods. We consider both pre-defined basis sets and data-adaptive basis sets, with their associated advantages and disadvantages. We devote only passing mention to some related, but important problems, such as increasing the interpretability of the results by appropriate preprocessing of the functional outputs (in particular, curve registration), and by enforcing some degree of smoothness when data-adaptive bases are used.

Figure 1 shows a simple made-up example. This is a set of four-parameter curves, that is, the model in this case is just a function with four parameters, *a*, *b*, *c* and *d*:

$$f(\theta) = 10 + a \exp\left(-\frac{(\theta - b)^2}{\mathrm{K}_1\, a^2 + c^2} \right) + (b + d) \exp\left(\mathrm{K}_2\, a\, \theta\right). \qquad (1)$$

We interpret these functions as output from a problem with azimuthal symmetry, say a shock wave problem or an implosion problem. The independent variable θ is a polar angle ranging from -90º to 90º. The model was run 81 times, using a complete 3^4 factorial design for the four input parameters. The 81 output curves are color coded in Figure 1 according to the value of the parameter *a*. Eq. (1) (which of course in a real example would be unknown) shows that this parameter controls the height of the central peak (and also, but less strongly, its width as well as the scale of the right-hand tail.)

In analyzing this "model output" we are typically less interested in what affects the values at, say, 45 degrees, than in questions such as: what shifts the curves up and down? left or right? What makes the central peak wider or narrower? the right-hand tail higher or lower? We could, of course, pick some appropriate functionals for answering these questions. The last, for example, we might address by examining the sensitivity of the

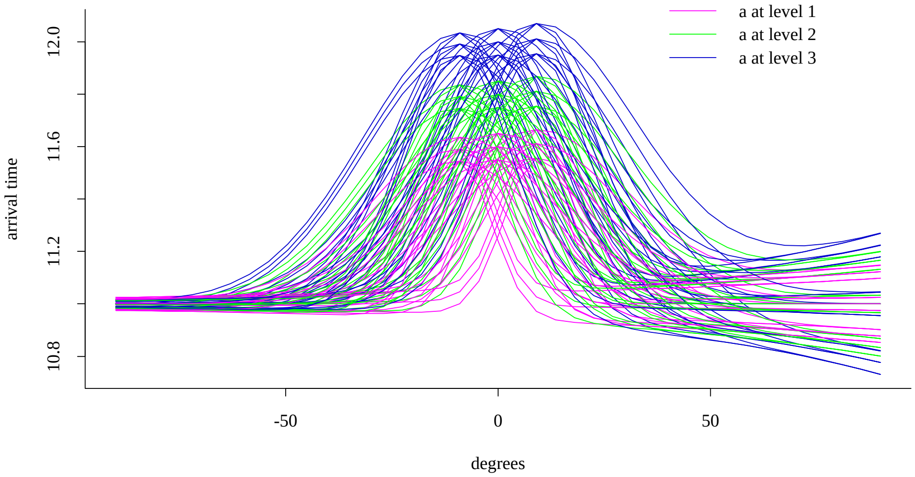


Figure 1. 81 runs of the four-parameter (4-P) model

values at 90º to the four input parameters. In order to address questions such as peak width we could devise some surrogate measurement that could be computed on each curve and then study its sensitivity to the input parameters. However, such choices are highly problem specific.

The curves in Figure 2 are outputs from 102 runs of an accelerator beam transport model, from a study that varied 18 input parameters. Twelve of these parameters describe the shape of the input beam in six dimensions (x and y position and momentum, phase and energy), while six are perturbations of the parameters of the transport line elements (x and y position and angular orientation of two quadrupole magnets). The output profiles are

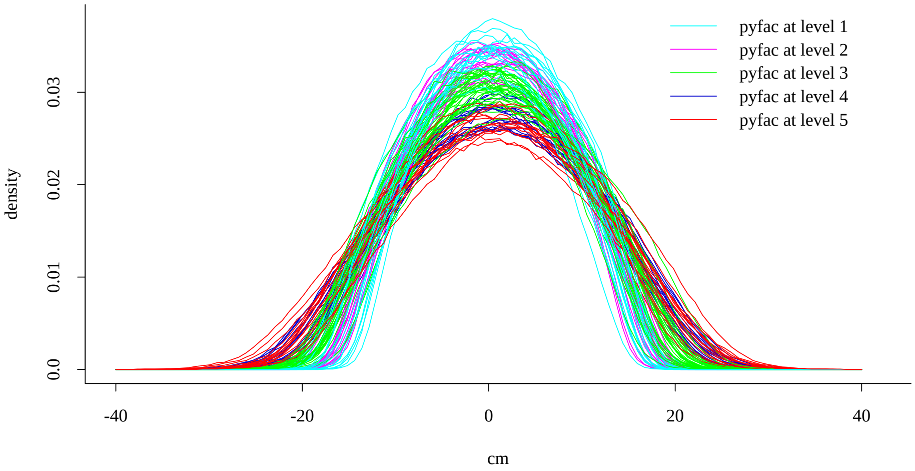


Figure 2. 102 runs of the beam transport model

compared with measurements made by a wire scanner inserted into the beam, moving from one side to the other of the beam. That is, one scan provides T=41 values along a one-dimensional projection of the beam intensity at a given point along the length of the transport line, a function of one variable which we again call θ. All the curves are normalized to have the same area underneath them.

With the benefit of hindsight, the curves have been color-coded according to the level of the input parameter which determines the width of y momentum distribution in the input beam. The principal effect of this parameter is to make the basically Gaussian shape of the intensity distribution of the output beam at the wire scanner either wider or narrower.

Transforming functional data

It might seem natural to regard functions provided on a grid of T points as T dependent variables for the purposes of sensitivity analysis. However, this is unsatisfactory for many reasons:

- The T variables are highly correlated with one another, so this natural coordinate system is inefficient for statistical methods like discriminant analysis, sensitivity analysis, or almost anything other multivariate statistical method. Results are redundant from one value of θ to another.
- The results obtained in this way are often not particularly interpretable for the underlying physical or modeling problem.
- Even though the data are the output of a computer model, the different runs may not have generated outputs at the same times or points θ. Alternatively, identical model output times may not be physically comparable because, as a function of the input parameters, the modeled process may be evolving faster in one run than another. So we may need to register the output curves (rescale time) in some physically more interpretable manner before proceeding with analysis.
- Finally, smoothness of the true curves may be a physical expectation that is not preserved by multivariate procedures in the original coordinate systems.

All of these problems can be addressed by transforming the functional output in one way or another. For sensitivity analysis, the most useful approach is expanding the output functions in terms of some basis functions (after rescaling time, if necessary) and then applying the statistical method of interest—in our case, a sensitivity analysis method—to the coefficients of that expansion. Different types of bases can be considered. There are familiar, predefined bases such as Legendre polynomials or other orthogonal polynomials, trigonometric functions, Haar functions, or wavelet bases. Adaptive basis functions include principal components, partial least squares components, and bases extracted adaptively from overcomplete dictionaries. If the columns of $\Phi_{T\times K}$ (K≤T) are a proposed set of basis functions, then the original functional output from N model runs, an N ×T matrix Y, can be rewritten as

$$Y - \overline{Y} = H\Phi^{T}\text{, i.e., } y_i(t) - \overline{y}(t) = \sum_{k=1}^{K} h_{ik}\varphi_k(t) \text{ for } 1 \le i \le N, \qquad (2)$$

where the mean function $\overline{y}(t)$ is computed as the mean of the $y_i(t)$ for each t.

Most standard basis systems are orthonormal. For example, the Legendre polynomials are orthonormal with respect to Lebesgue measure on [-1,1]. But the Legendre polynomials in sin(t), which are used in the examples below, are not orthonormal with respect to ordinary Lebesgue measure $d\theta$, but only with respect to a weighted measure $\cos\theta d\theta$. Adaptive bases functions may be orthonormal by construction, or not. Orthonormality of the basis functions is a nice property, since then the total variance is naturally partitioned among the variances of the coefficients:

$$\sum_{i=1}^{N}\|y_i\|^2 = \sum_{i=1}^{N}\left(\sum_{t=1}^{T} y_i(t)^2\right) \sim \sum_{i=1}^{N}\left(\sum_{k=1}^{K} h_{ik}^2\right) = \sum_{i=1}^{N}\|h_i\|^2 . \tag{3}$$

(Usually the basis functions are ordered so that the first few capture most of the total variance.) However, even when the basis functions are not orthonormal the total variance captured by the expansion in terms of the first k (k≤K) basis functions can be computed, and orthonormality may be less important than some other features when it comes to sensitivity analysis.

Legendre analysis for 4-P example

Since the first example is being interpreted as a set of functions of angles from -π/2 to π/2, the Legendre expansion in sin(t) is a natural choice among standard expansions. Figure 3 shows how the coefficients $\{h_{ik}\}$ of the expansions of the 81 functional outputs depend on the parameters, for k=1, 2, ..., 6. The Legendre polynomials are alternately symmetric and anti-symmetric around zero, as shown in the top row of Figure 3. The first k polynomials define a k-dimensional subspace of the 41-dimensional space in which the output functions are vectors. The percentages at the top show how much of the total variance in the original family of functions lies in this subspace for k up to 6. Note for future reference that the six-dimensional subspace defined by the first six polynomials still includes less than 90% of the total.

In the second row, the Legendre polynomials are interpreted as perturbations of the overall mean of the 81 output functions. The mean function is shown in blue, the mean plus a multiple of the Legendre polynomial in green, and the mean minus the same multiple in magenta.

The remaining rows contain box plots showing dependencies of the coefficients on the four parameters. Of course, we are not proposing sensitivity analysis by inspection as a serious method for sensitivity analysis, but SA methodology is not the main goal of this paper. The figures are intended to suggest what more formal sensitivity analysis would indicate. For the beam transport example, the displayed subset of five of the 18 input parameters was selected based on the results the partial correlation coefficient (cf. McKay, 1997, for example) on the PLS components.

For the 4-P example, variability in the coefficients of the Legendre polynomials of even order is controlled largely by *a*, although *c* and *d* influence the constant, zero-order term. The odd orders are controlled mostly by *b* with some influence of *d* on the first-order term.

Legendre polynomials and other standard expansions are well understood by many modelers, and this is an advantage not to be abandoned lightly. The other main advantage of using a consistent, non-adaptive basis system arises when a series of problems is being considered. The differences among corresponding analyses are then localized to the coefficients, instead of being partitioned out between the coefficients and the basis functions themselves.

The disadvantages arise in the case where the selected basis functions are not particularly well suited to the problem at hand. The Legendre basis, for example, is not a particularly

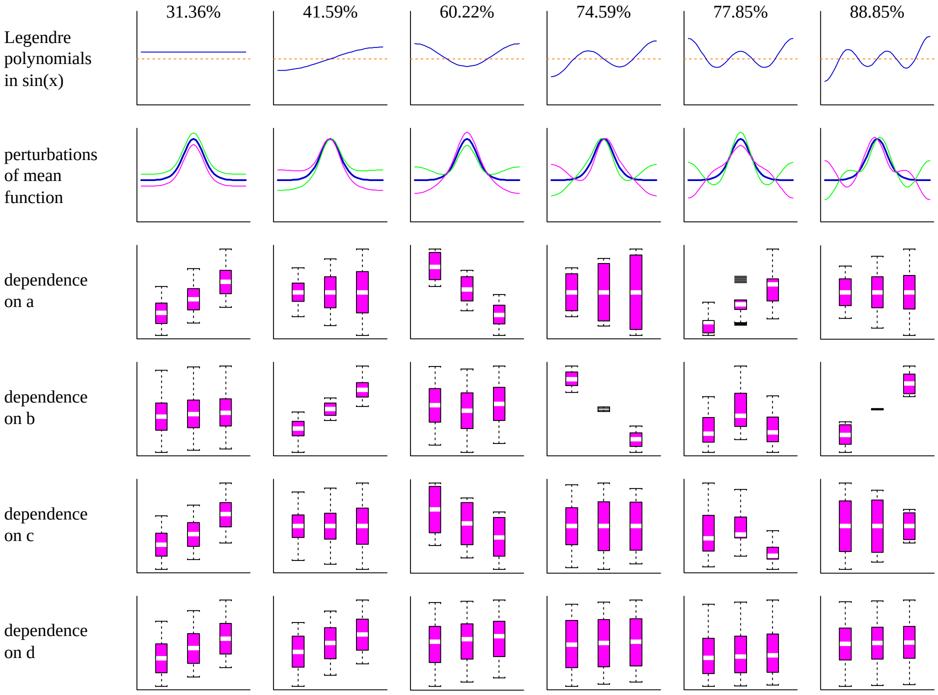


Figure 3. Dependence of the coefficients of the Legendre expansions for the four-parameter model on the parameters

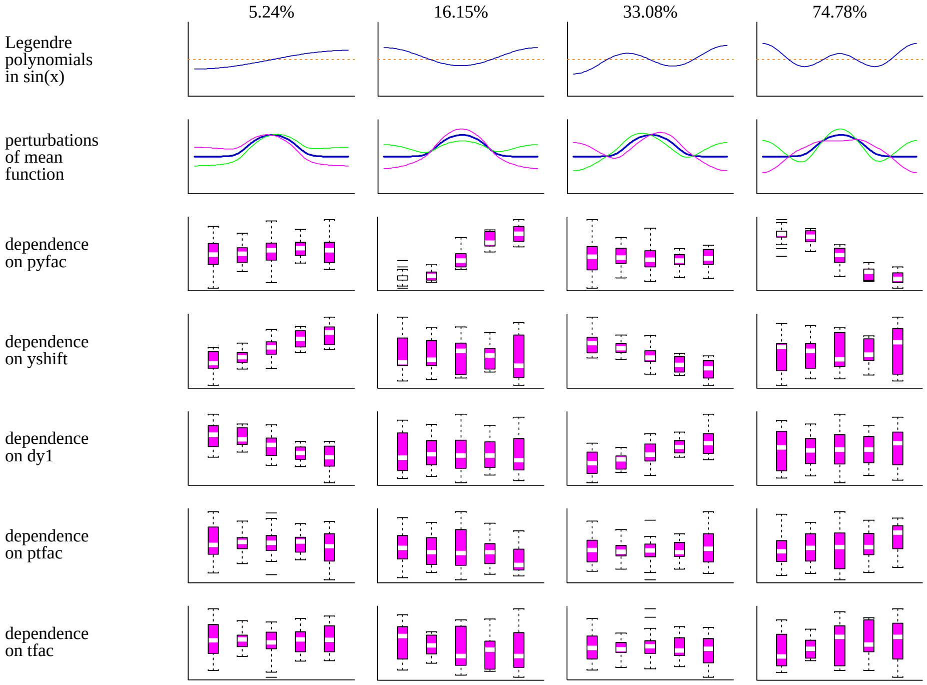


Figure 4. Dependence of the coefficients of the Legendre expansions for the beam transport model on the parameters

good choice for a problem in which one of the main effects is neither symmetric nor antisymmetric, as for the 4-P example. The dispersion in the right-hand tail by comparison with the tight left-hand tail is not well captured by any single polynomial but spread out over several of them. The other disadvantage is that a relatively simple effect may be spread over several terms. For example, in this problem the effect of *b*, responsible for the left-right shift of the main peak, is spread out over all polynomials of odd order.

Figure 4 is a similar plot for the second through fifth Legendre polynomials for the beam transport example. (Because the uniform up and down shift accounts for less than one percent of the total variance, the polynomial of order zero is omitted from the plot.) As for the 4-P example, seven terms are needed to capture 90% of the total variance. Most of the action is in the even order, width-controlling terms, and the most important variable for these is *pyfac* (spread in y-momentum in the input beam), while *yfac* (spread in y-position in the input beam, not shown in Figure 4) is a distant second. The second important effect is left-right shift, which is controlled by *yshift* (the y-position shift of the input beam) and by *dy1* (the misalignment in the y-position of the first of the two quadrupole magnets.) Being carried along in Figure 4 for comparison with later methods are a couple of other scaling factors for the spread of input energy (*ptfa*c) and phase shift in the input beam (*tfac*).

Adaptive bases computed by principal components analysis

The principal components of Y, considered as N observations in a T-dimensional space, are themselves T-vectors. They form an orthonormal basis for the T-dimensional space (or for a subspace of T-dimensional space, if N<T) that is specifically adapted to maximize the variance of the projection of the data onto the first basis vector, then onto the subspace spanned by the first and second basis vectors, etc. Thus expansions in the PC basis for sensitivity analysis should at least achieve some compression, avoiding one of the more serious problems with the Legendre polynomial, namely the allocation of a fairly simple effect (e.g., width changes or left-right shifts) to several components.

For the family of curves in Figure 1, the first principal component is basically an up-down shift, but unlike the first Legendre function this shift is not constant across all angles. (Refer to Figure 5.) The subspace spanned by this one function accounts for about 46% of the total variance in the family of curves, compared with about 31% for the Legendre polynomial of order zero. Like the zero-order Legendre coefficient, the coefficient of the first principal component depends on all four parameters. The second principal component for this example is a left-right shift accounting for another 34% of the total variance and controlled primarily by the *b* parameter. A similar amount of the total variance was spread across the Legendre polynomials of odd orders. The third principal component is devoted explicitly to the right-hand tail and accounts for 11% of the total variance. It is clearly controlled by the *d* parameter, something that could not be extracted from the Legendre analysis.

These first three terms capture over 90% of the total variance, compared to seven terms required by the Legendre analysis. The fourth component, which accounts for another 5% of the total variance, is a symmetric kurtosis or tail-fattening component depending most strongly on *a* and *c*.

For the beam profile example, just two terms account for over 90% of the variance (Figure 6). The first principal component is basically a widthing term, the second a left-right shift. From the point of view of the beam modelers, the third and fourth terms are also very interesting, both affecting the shapes of the tails. However, the box plots don't

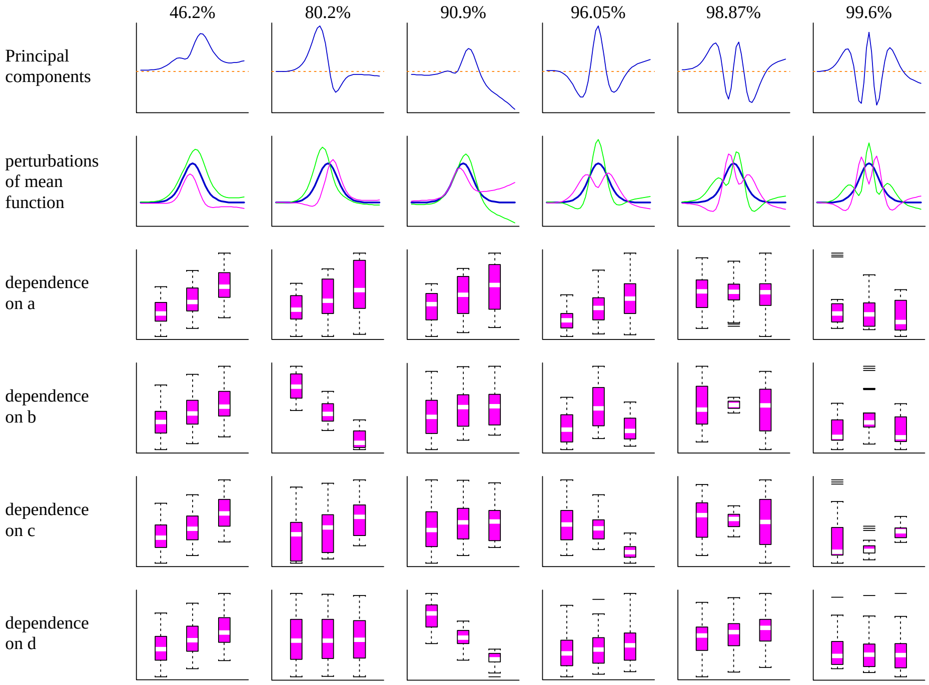


Figure 5. Dependence of the coefficients of PCA expansions for the four-parameter model on the parameters

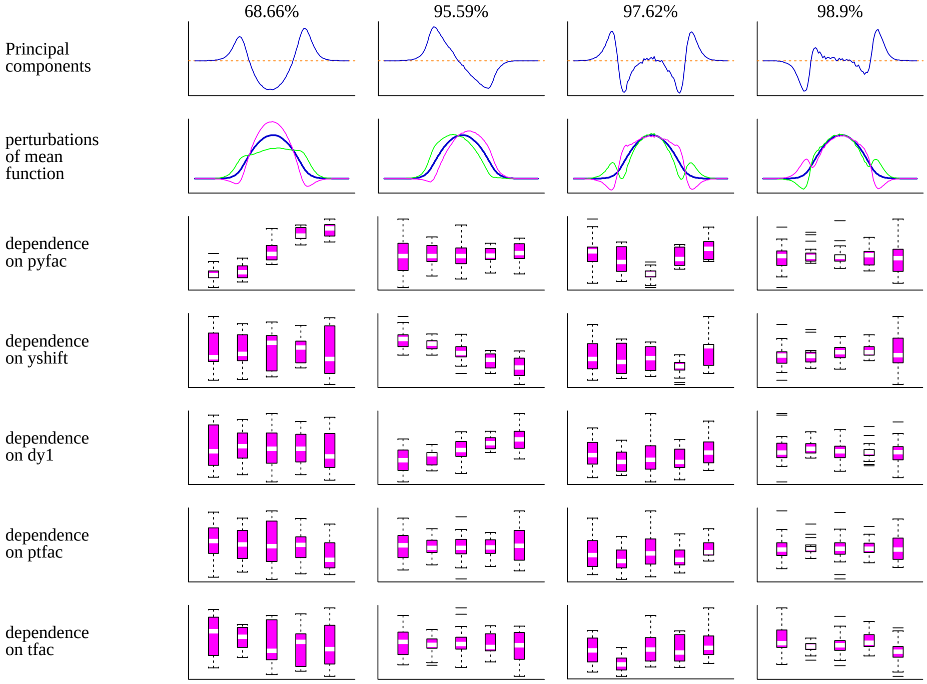


Figure 6. Dependence of the coefficients of PCA expansions for the beam transport model on the parameters

show exactly where these come from; the top variables for these components are the same as those for components 1 and 2, respectively.

Adaptive bases computed by partial least squares

Partial least squares (PLS) regression was invented to handle near-collinearity among the independent variables, which is not usually a problem in analyzing computer experiments, assuming a reasonable experimental design. Thus, PLS is really a technique for decomposing the design matrix. (For a review or PLS regression, see Frank and Friedman, 1993.) However, PLS simultaneously provides a transformation of the dependent variables in such a way that the first PLS component of the dependent variables is has the maximum variance that can be predicted by a linear combination of the independent variables. The second PLS component is computed using the residuals from the prediction of the first, and has the maximum variance that can be predicted by a second, orthogonal component of the independent variables, etc. So one might think of PLS as "peeking" at the explanatory variables while doing something that is similar to a PC analysis of the dependent variable. Note that while the PLS components of the independent variables are orthogonal, the PLS components of the dependent variables are not, in general.

While there is no *a priori* guarantee that PLS results will be interesting for functional sensitivity analysis, in the event they often seem to be fairly revealing. In particular, in the two examples they pull out some dependencies that were overshadowed by more important terms in both Legendre and principal component analyses.

For the four-parameter example, the PLS components (Figure 7) are somewhat more readily interpretable than the PCA components (Figure 5). The first component is an up-down shift of the middle of the curve, depending as before on all four parameters. (The first PLS component should be the same as the first PCA component if the independent variables are standardized, which is to be recommended; it is only with the extraction of the second component that the algorithms diverge.) The second PLS component is a left-right shift, almost entirely a function of *b*, compared to the second PCA component which had more substantial contributions from *a* and *c* as well. The third PLS component is pure right-hand tail, dependent on *d*. The fourth is primarily a widthing term, although it also includes a small left-right shift component, and depends on *a* and *c*. As there are only four input parameters, the PLS algorithm can provide only four component vectors, but this four-dimensional subspace captures almost 96% of the total variability in this family of curves, which is almost as much as the first four PCA components. By comparison, the first four Legendre components captured only about 75% of the total variance.

For the beam transport example (Figure 8), the first two PLS components are almost identical to the first two PCA components (Figure 6)—a widthing term and a left-right shift. The third component, no longer restricted to being orthogonal to the first two, turns out to be very similar to the first except for an arbitrary sign reversal, i.e., it is another widthing term. However, the explanatory direction for this component *is* required to be orthogonal to those already selected and turns out to have a strong dependence on *ptfac*, which controls the heterogeneity of the energy of the beam. Specifically, larger values of *ptfac* (increased heterogeneity or spread in the energy distribution) narrow the observed peak (the coefficients of this component are negative for larger values, corresponding to the direction of the red perturbation in the second line of the plot.) The fourth component is almost a pure tail-fattening component, and has a strong dependence on *tfac*, which controls the heterogeneity of the phase of the beam. Larger values fatten the tails. These

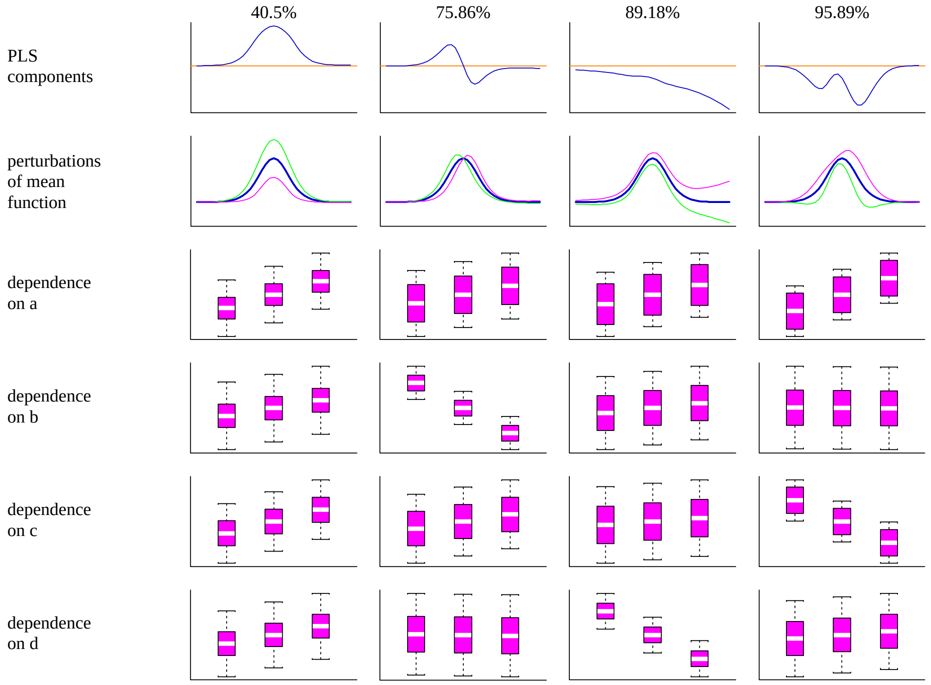


Figure 7. Dependence of the coefficients of PLS expansions for the four-parameter model on the parameters

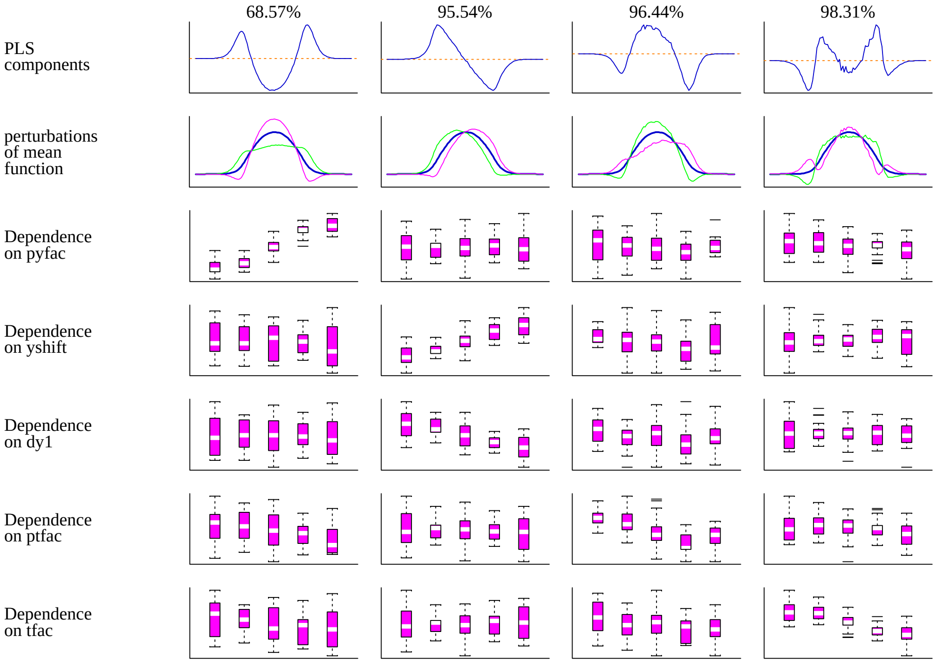


Figure 8. Dependence of the coefficients of PLS expansions for the beam transport model on the parameters

dependencies were not readily observable using the other functional transformations (although more sophisticated methods for sensitivity analysis than the simple box plots used here for illustration might have discovered them!)

The advantages and disadvantages of adaptive bases are pretty much the inverse of those for standard bases. The main advantage is good compression of the information; it is usually necessary to do sensitivity analysis on only the first few coefficients. The basis functions are also frequently more interpretable in physical terms. In a series of related problems, it may be interesting to study how the shapes of the component functions evolve (as well as their coefficients.) Of course, the down side to this is that shapes and coefficients are evolving simultaneously, which may lead to interpretation problems. In some cases it may make sense to pool all of the output functions for the series to extract a common set of principal components or PLS components, so that the evolution of their coefficients through the series can be studied in the same way as the evolution of the coefficients of a fixed basis set, such Legendre polynomials, could be examined.

Other considerations

Penalty methods can be used to enforce a degree of smoothness on adaptive basis functions. For example, some meaningless high-frequency information appears in the higher-order PCA and PLS components for the beam transport model output (Figures 6 and 8.) Orthonormality is lost when this is done, but the results are probably more interpretable, and curve comparison across problems, or between model output and noisy data, becomes easier. Ramsay and Silverman (1997) discuss the enforcement of smoothness in PCA (Chapter 7), and the technique is readily extended to PLS.

Curve registration may be needed or advisable when the parameters affect the time- or space-scale or when the functions not sampled at identical times in different runs. We may be interested in studying both the sensitivity of the scaling to the input parameters, independently of the variability in the functional outputs after adjusting for these scaling effects. Again, Ramsay and Silverman (1997) address this problem in detail, proposing a series of methods from parametric location and/or scale change, through feature or landmark registration methods, to the estimation of general monotonic transformation.

Summary

The purpose of this paper has been to suggest that sensitivity analysis for functional computer model outputs, correctly performed, is not significantly more difficult than for scalar outputs. The basic method is the expansion of the functional outputs in an appropriate functional coordinate system, i.e., in terms of an appropriate set of basis functions, followed by sensitivity analysis of the coefficients of the expansion using any standard method. The main art, then, is in choosing the appropriate coordinate system. We have considered both standard, pre-defined basis sets and data-adaptive basis sets. The examples tend to favor the latter because of the compression and interpretability of the results, but the former may have value, depending on the problem or set of problems and the customer.

References

Frank, I. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics* 35, pp. 109-135.

M. D. McKay, “Nonparametric Variance-based Methods of Assessing Uncertainty Importance,” *Reliability Engineering and System Safety*, 57, 267-279, 1997.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer, New York

Introduction to Data Mining with Military Applications
Cadet First Class Christopher Jeffreys and Major James Wisnowski, USAF Academy, CO

Data mining is the step in the knowledge discovery in databases (KDD) process that uses algorithms and other methodologies to discover potentially useful relationships in large databases. KDD and data mining have received considerable attention across a wide variety of communities in recent years. Although some of its popularity may be more marketing hype than actual useful applications of the techniques, we show the methods do have many potential uses in Department of Defense programs. One tenant of the KDD process is that more often than not, these methods show unexpected insights and answer hypotheses not originally explored. This presentation gives a brief overview of the history of the data-mining field, describes its relationship to classical statistics, discusses the common data mining objectives and tasks, and illustrates with military examples. We explore several of the most commonly used methods including neural networks, decision trees, association rules, cluster analysis, and data visualization. We discuss decisions trees in depth using relevant military cases and demonstrate how to perform the analysis with a common PC-based statistical software package.

Clinical Session I

Danny C. Champion, TRADOC Analysis Center – WSMR
Louis A. Fatale, Topographic Engineering Center (TEC)

In 1998, we conducted a study titled "The Effects of Vegetation on Line-of-Sight (LOS) for Dismounted Infantry Operations" which examined methods for depicting LOS in vegetated areas for combat models and simulations (M&S). We did this by using empirical information to create parameters for exponential decay curves. These parameters were then recorded in look-up tables as a function of geographic area (called biomes). This limited an entire biome to a single exponential decay parameter (or all LOS within a biome was the same).

In this follow-on work, we are attempting to provide a more robust method for determining LOS using remotely collected data (i.e. world soil type, annual rainfall, and undergrowth prediction obtained from overhead satellite imagery). This should allow for more accurate prediction of LOS in vegetated areas. This is especially important for denied areas. In this study, we: 1) defined and collected additional data from field collection surveys and variables from stereo high-resolution spaceborne imagery; 2) conducted correlation analyses to define the important variables; and 3) develop a regression equation using the defined variables to improve LOS prediction.

The problems are occurring in the correlation and regression analysis. We were limited to 56 collections points (scattered throughout North America) because of cost. However, we have over 50 independent variables selected because they intuitively are correlated with plant growth (i.e. rainfall, canopy closure, tree height, undergrowth depiction, soil information). For the most part, the correlations are remarkable poor. Have I overlooked something? If the decay parameters are used in an exponential decay, should I consider non-linear dependant variables? (This provided little improvement). What is the best way to partition the outliers?

On the Implementation of a Replication Paradigm for Calculating Confidence Intervals on Latin Hypercube Statistics

Vicente J. Romero, Sandia National Laboratories

Latin Hypercube Monte Carlo sampling has a well established reputation for generally being more efficient than Simple Random sampling for propagating uncertain (random variable) model inputs into random variable outputs, as indicated by smaller sampling variance on calculated statistics of response, such as mean, variance, and cumulative probabilities.

However, the efficiency savings of LHS cannot be broadly realized until a method for calculating LHS confidence intervals (CI) on calculated statistics is definitively established. Though for a given number of samples, LHS CI are at least as small as (i.e. are bounded by) CI calculated from the classical formula for Simple Random sampling, this is not helpful when it is desired to meet some level of confidence with the least number of samples necessary. This is especially dismaying when CI requirements must be met on important calculations and each sample requires the run of an expensive computer model. Thus, in order to broadly exploit the efficiency advantages of LHS, a dependable method of quantifying sampling variance must be developed for various statistics under fairly general conditions, so that sampling can be terminated as soon as sampling accuracy goals are met.

A seemingly general "Replicated LHS" paradigm that divides and groups LHS samples to generate CI on various calculated statistics has been proposed at Sandia. However, the methodology has only been used to attach CI to mean calculated cumulative distribution function values based on three replicates. Moreover, the validity of the manner in which the paradigm was implemented has never been empirically verified. Empirical verification is necessary because no theoretical basis or empirical rules of thumb have been established or suggested regarding the optimal or required number of replicates and number of Latin Hypercube samples per replicate for calculating valid LHS CI in this manner -for any of the types of commonly calculated statistics such as mean, variance, and cumulative probability.

The goal of the presentation in the clinical session will be to: 1) summarize the proposed LHS CI methodology; 2) propose some prospective rules of thumb for suitable number of replicates and number of samples per replicate that should be used under various conditions; 3) propose a methodology for implementing the Replicated LHS CI paradigm in general and for evaluating the validity of the implementation through hypothesis testing; 4) elicit critical review and advice from the statistician panelists on the proposed rules of thumb and implementation and evaluation methodologies; and 5) reach a consensus on the best implementation and evaluation approaches toward the purpose of establishing an empirically verified general LHS CI methodology.

Contributed Session IV

Bootstrapping a Stochastic Process: Time-Indexed Risk Profile Analysis of an Index Fund[1]

James R. Thompson, Noah Harding Professor of Statistics
Edward E. Williams, Henry Symonds Professor of Management
Rice University

Abstract. It is demonstrated how resampling can be used to obtain a risk profile analysis of an index portfolio. A variant of the partially privatized Social Security System concept is examined using a nonparametric analysis.

Introduction The use of resampling techniques for market forecasting has not proved fruitful. Indeed, the forecasting of the future value of a portfolio by any technique has defied the experts. In this paper, we take a new tack. Instead of forecasting the value of a portfolio at a future time, we forecast the entire stochastic process characterizing the risk profile of the portfolio.

At any given time, there will be this or that investment fund which is performing well above the average of the United States stock market. Some of these, such as Warren Buffett's Berkshire-Hathaway have outperformed the overall market for many years. But that is unusual.

The creator of the Vanguard S&P 500 fund, John C. Bogle [1] has long noted that investment funds tend not to outperform the weighted average of the overall market. Their stock selections, if better than the market, tend not to hold up over time. By simple chance, some funds at any given time will appear to perform wondrously well. Over time, these outperforming funds fade like flowers in winter.

The large management fees required by fund managers put them at a disadvantage relative to the performance of funds based simply on the broad market index funds. Index funds require management fees in the 0.2% per year range.

Now, if it were possible to find some funds which consistently (absent management fees) underperformed the market, that would be as useful as discovering a good fund, for then we could find what the investment policy of the bad fund was and bet against it. Unfortunately, we do not seem to have found any such funds.

Discussion

If there is no magic forecasting device to give good predictions of stock prices, then the investor still has two weapons at his disposal. One weapon is that of diversification among a number of securities. If we have ten stocks, each with the same growth rate and each with the same volatility, dividing our investment among the ten stocks rather than putting all our investment in any one of them is almost a "free lunch." Of course, the lunch is not entirely free. Such diversification should save us from losing everything in an Enron but it might kill our hopes of becoming a Microsoft millionaire (as many of Microsoft's secretaries, who had retirement plans,

[1] This research was supported in part by the Army Research Office (Durham) under (ARO-DAAD19-99-1-0150) at Rice University.

not unlike those of Enron's secretaries, became). Diversification of this sort has been used for a long time (in the nineteenth century many farmers planted corn as well as wheat in the event that hail storms zapped the more profitable wheat).

But in a bear market, the overwhelming majority of stocks decline in value. We have treated this elsewhere [2], [3], and [4] by adding Poisson bear jumps to the Gaussian walk part of a model of stock performance. Just as an extended drought will zap both corn and wheat, a bear market will hurt stocks generally. (An old politically incorrect adage of Wallstreet is "When the paddy wagon comes, good girls are arrested as well as the bad.") What other variable can we use for "diversification"? The answer is **time**.

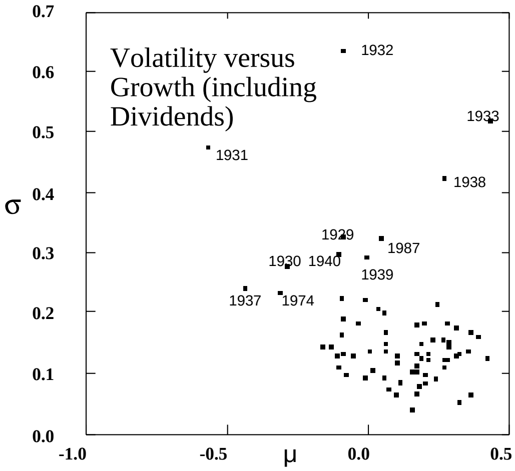


Figure 1. 75 Years of Ibbotson Index Growth and Volatility.

Investors over longer periods of time, have the advantage of the fact that in roughly 70% of the years, the index of large cap U.S. stocks rises rather than falls. And there is the further encouraging news that in over 40% of the years, the index rises by over 20%. In 30% of the years, the market rises by over 25%. And in 25% of the years, the index has risen by over 30%. Over the roughly 75 year period such records have been kept, the United States has lived through the Great Depression, the Second World War, the Cold War, Korea, Vietnam, assorted massive sociological changes, shifts toward and away from free markets, and assorted epidemics. These can all be viewed as the political/economic/sociological analogs of major "droughts." It is true that we have yet to experience Martian invasion, attacks by genetically engineered viruses or suitcase nuclear devices, or the costs of mounting the Sixth Crusade. We hope such events do not occur, but events of comparable angst have occurred to other countries of the West over the past 75 years. Poland was occupied by Russia and Germany in September of 1939, and the Russian occupation only ended (sort of) in June of 1989. It is hard to imagine a market hedge (other than taking oneself and ones money out of Poland and moving to, say, the United States) which would have saved an investor in the Warsaw Stock Exchange. And it is hard today to imagine a safe harbor for oneself or ones property in the event that the United

States falls. Past performance is not an infallible guide for predicting a risk profile and we do not claim it to be. But it is surely a guide which all should at least consider.

Table 1. Ibbotson Large Stock Index
μ (Including Dividends) and σ

Year	μ	σ	Year	μ	σ
1926	0.10993	0.11798	1964	0.15255	0.03985
1927	0.31838	0.13038	1965	0.11734	0.08560
1928	0.36193	0.16555	1966	-0.10603	0.11051
1929	-0.08796	0.32487	1967	0.21495	0.11965
1930	-0.28635	0.27484	1968	0.10490	0.12837
1931	-0.56810	0.47468	1969	-0.08883	0.13045
1932	-0.08545	0.63357	1970	0.03932	0.20436
1933	0.43172	0.51663	1971	0.35837	0.13481
1934	-0.01450	0.22225	1972	0.17379	0.06531
1935	0.38981	0.15963	1973	-0.15853	0.14316
1936	0.29207	0.14341	1974	-0.30748	0.23396
1937	-0.43124	0.23919	1975	0.31627	0.17478
1938	0.27094	0.42202	1976	0.21382	0.13136
1939	-0.00411	0.29154	1977	-0.07451	0.09550
1940	-0.10292	0.29437	1978	0.06354	0.16708
1941	-0.12319	0.14373	1979	0.16924	0.13337
1942	0.18515	0.14602	1980	0.28081	0.18309
1943	0.23032	0.15564	1981	-0.05035	0.12923
1944	0.18024	0.07712	1982	0.19400	0.18413
1945	0.31071	0.12841	1983	0.20302	0.09729
1946	-0.08414	0.18971	1984	0.06081	0.13610
1947	0.05553	0.09503	1985	0.27884	0.11859
1948	0.05354	0.19931	1986	0.16949	0.17930
1949	0.17219	0.10062	1987	0.05098	0.32354
1950	0.27543	0.10740	1988	0.15538	0.09988
1951	0.21527	0.11992	1989	0.27376	0.12009
1952	0.16865	0.11214	1990	-0.03221	0.18407
1953	-0.00995	0.09333	1991	0.26659	0.15397
1954	0.42278	0.12566	1992	0.07390	0.07315
1955	0.27429	0.12017	1993	0.09522	0.06076
1956	0.06354	0.14693	1994	0.01301	0.10559
1957	-0.11406	0.12720	1995	0.31794	0.05080
1958	0.36019	0.06137	1996	0.20758	0.08288
1959	0.11297	0.08002	1997	0.28788	0.15119
1960	0.00469	0.13557	1998	0.25138	0.21275
1961	0.23815	0.08793	1999	0.19095	0.12391
1962	-0.09135	0.20038	2000	-0.09552	0.16284
1963	0.20539	0.09662			

In the plot of μ versus σ we note some interesting years outside the apparent cluster. The two years with both high volatility and high growth are 1933 (the "Happy Days Are Here Again" optimism which characterized the start of the Roosevelt era) and 1938 (the year after the bottom of the Great Depression). Nine of the eleven outliers are depression years. Both 1974 and 1987 had significant bear epochs. With these eleven years removed, the correlation between μ and σ is $-.142$. With all 75 years left in the data base, the correlation is $+.184$.

Let us consider a portfolio with initial value of \$100,000. We pick five of the 75 index annual growths at random (with replacement), say, μ_1,μ_2,μ_3,μ_4,μ_5. Then a simulated portfolio value after the five years is given by

$$V = \$100,000 \exp(\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5)$$

In Figure 2, we show the picture obtained by sorting a thousand such simulations according to percentiles.

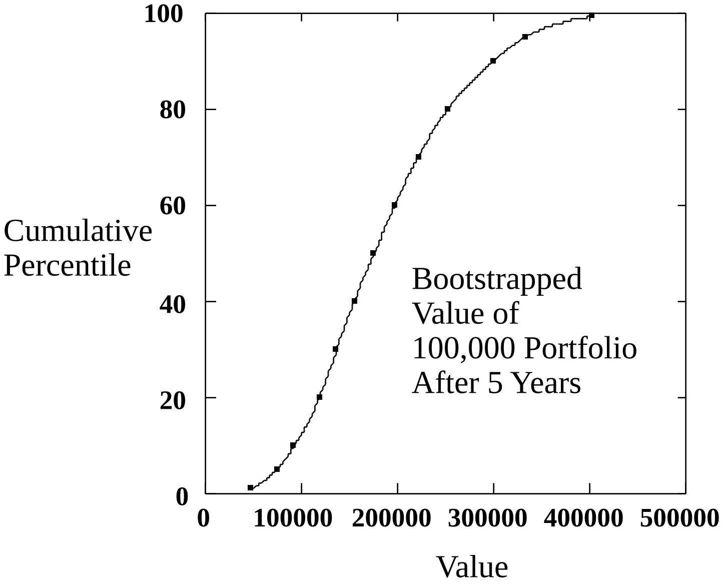


Figure 2. Distribution Function of Portfolio after 5 Years Using Resampling.

The mean value of a \$100,000 portfolio after five years is \$192,676. The median value is \$175,530 (growth rate of .1125). However, the lower ten percentile is \$92,747 (growth rate of $-.015$).

Next, we consider the same scenario except looking 20 years into the future. The results are quite optimistic. The median value is \$873,100, an annual increase of 10.8%. Even the lower ten percentile value of \$285,590 represents a growth rate of 5.2%.

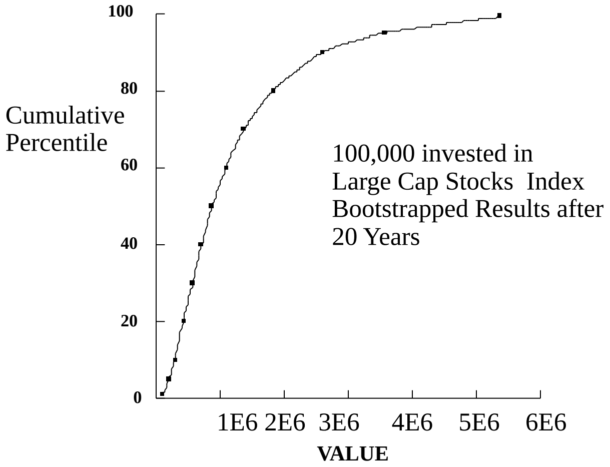


Figure 3. Distribution Function of Portfolio after 20 Years Using Resampling.

A Partially Privatized Social Security Plan President George W. Bush has suggested a partial privatization of Social Security. Under the present plan, workers and their employers together "contribute" roughly 15% of a worker's salary to the FICA fund. The worker's portion of the contribution is not tax exempt. After roughly 45 years of employment, the worker may start to draw a Social Security stipend until the time of his/her death. Part of the stipend may be subject to income tax, even though no tax exemption was given the worker while paying his FICA tax.

President Bush has suggested that a worker might elect to use a portion of his FICA setasides to invest in the stock market. Typically, it is assumed that some restrictions leaning toward fiscal conservatism will be applied. So, we will use, as an example, a contribution of $2,000 per year for each worker over a period of 45 years. We will invest the money in something like the Ibbotson Index. So, let us note in Figure 2 what the bootstrapped results look like when we take 5,000 resampled concatenations of the index, assuming that $2,000/year will be added to the fund. The results are quite promising. The mean value of such a fund is $4.84 million dollars. The median value is $2.724 million. Even the lower ten percentile is $695 thousand. The lower five percentile is $464 thousand. The lower one percentile is $225 thousand. Of course, there is the problem of inflation. Even so, we realize that the "contributions" would be indexed on inflation. And, naturally, we could index the Ibbotson index as well.

Another objection could be that the kind of massive infusion of monies into the stock market as might be caused by a partial privatization could inflate stock values in the shorter term and might lead to a collapse in the longer term. That is, persons already with substantial holdings in the stock market would receive an immediate benefit as the new funds from partial privatization poured into the market causing a rise in stock prices. A partial privatization similar to that considered here of social security has already taken place in Sweden (instituted by the socialist government there). No apparent inflation in world wide markets has been noted. However, the number of new Swedish investors in the market is quite small compared with that which would be experienced in the United States.

It is unlikely that potential problems of market inflation will be the actual reason for not giving workers the option to put some of their FICA assets into bonds or securities. The actual reason will be the loss of control of money by the political elites. Who implicitly owns assets is not nearly as important as who controls them. The attempt to take over the health care system in the United States by the federal government during the Clinton Administration was unsuccessful, in part, because it was learned by the citizenry that over 10% of the American economy would have been transferred from the private sector to that of the state. History has shown an extreme reluctance of politicians to give up control of assets once they have achieved such control. And the Social Security program in the United States is nearly 70 years old. On the other hand, it is possible that, like the Swedish socialists, the American bureaucracy will recognize that they have little choice but to privatize an increasingly expensive and inefficient program. In any event, even if the partially privatized FICA plan is never introduced, the study in this section could be of use to a person thinking of making regular investments into a tax deferred index fund.

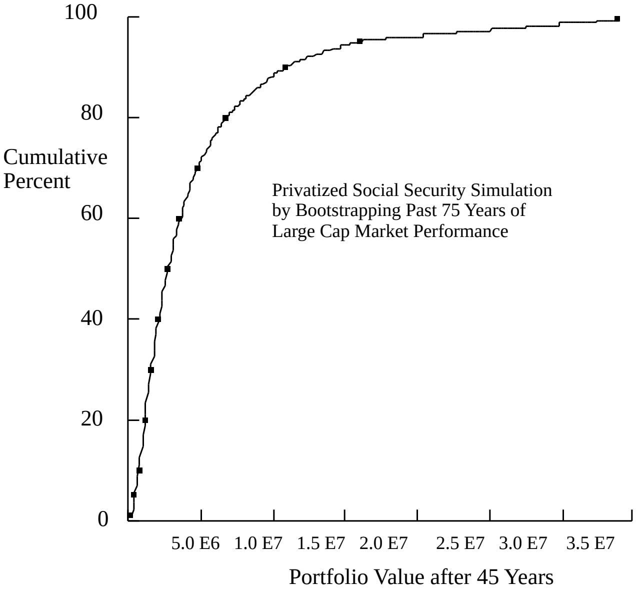


Figure 4. Resampling Based Distribution Function of Privatized Social Security Index Fund After 45 Years of Work.

Index Funds As High Interest Paying Money Market Accounts

Figure 3 and Figure 4 give some support for investing in an index fund broadly composed of large cap corporations. Note that we used an approach which has few modeling assumptions in both these cases. We have assumed that the future increases in such an index will be similar to those in the past.

Some may object to taking single year rates from the Ibbotson history. What happens when we have long patches of decline? Might not inclusion of these in an appropriate fashion introduce more pessimism into both long term investments in index funds and the hypothetical privatized Social Security plan. In Figure 5 we show the cumulative percentiles one obtains when examining a $100,000 investment for five contiguous years starting in 1926 and going through 1996 (we are limited by 1996, since we are looking at 1996 and the four following years).

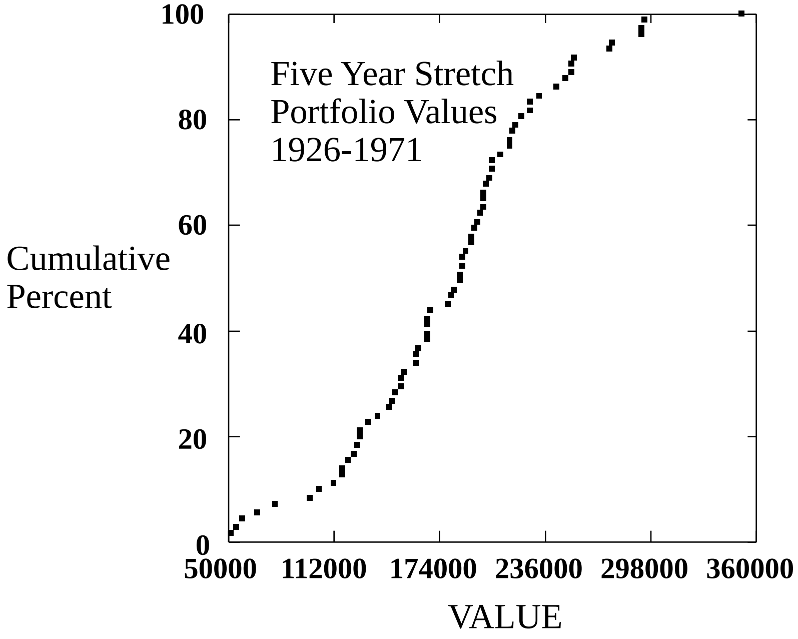


Figure 5. Resampling Based Distribution Function of $100,000 invested for Five Contiguous Years.

Now the lower ten percentile represents almost no gain at all. On the other hand, the median and mean both correspond to an annual gain (again, including dividends) of around 12%. Still, there is the troubling lower ten percentile.

Next, we carry out 10,000 resamplings of two five year stretches from the Ibbotson Index with an initial $100,000 investment in the index. This time we show a histogram of the results in Figure 6. The lower five percentile is slightly better than break even. The lower ten percentile now pays over 2.7%. The lower twenty percentile pays 4.7%. The median pays 11.3%. The mean pays 11.7%. Perhaps we can say that at ten years we really have reached the point where we can talk meaningfully about “the long term.” An investor in the index fund for ten years would appear likely to be pleased with his/her end results and has little chance of awful results. In other words, risk would appear to have been reduced to bearable levels. We note the shape of the histogram (which has been based on no distributional assumptions) has the characteristic shape of the log normal density function.

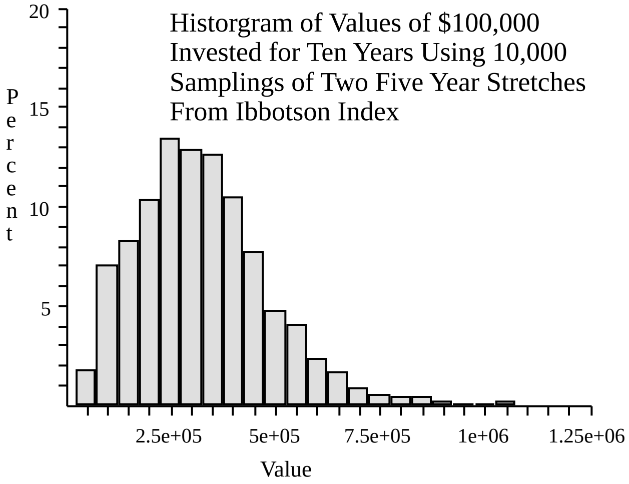


Figure 6. Resampling Based Distribution Histogram of $100,000 invested for 2 Five Contiguous Year Stretches.

Next, let us look at an index fund starting with $100,000 randomly selecting four five year stretches from the Ibbotson index of large cap stocks. We show these results in Figure 6.

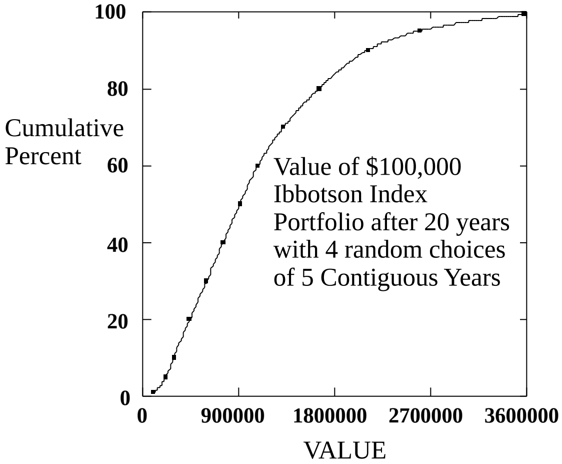


Figure 7. Resampling Based Distribution Function of Initial $100,000 Invested in Ibbotson Index for 20 Years.

The riskiness is reduced still further. The lower ten percentile of performance is a growth of 5.3%. The median growth shows over 11% annual growth and the mean over 12%.

Finally, we return to the notion of bootstrapping a 45 year investment from an annual $2,000 per year invested in the index fund. Here we pick randomly 9 five year stretches. For each year, we add $2,000 and we suppose the annual increase in the fund is the average of the five year stretch in which the year lies. We note here that the lower ten percentile is $1.32 million, the mean $4.4653 million and the median $2.5380 million.

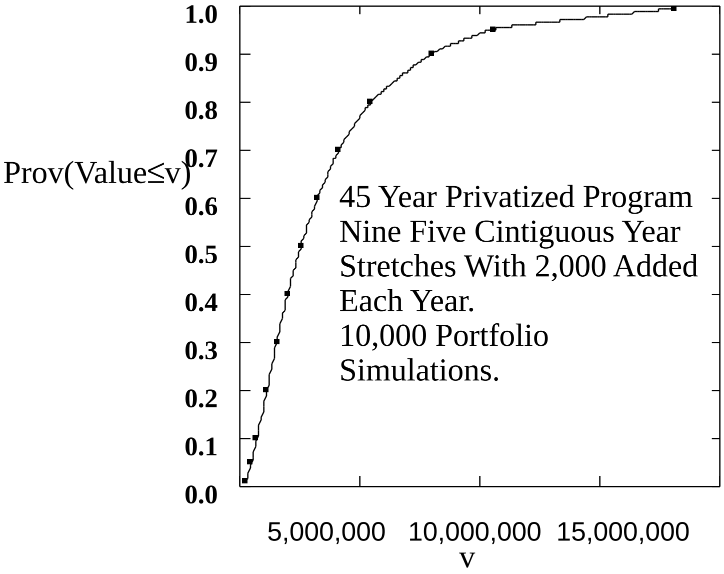


Figure 8 . Resampling Based Distribution Function of Privatized Social Security Index Fund After 45 Years of Work by Selecting Randomly 9 Five Contiguous Year Stretches from Ibbotson Index.

In any event, it would appear that if history of growth is the best guide to the future, large cap index funds appear to be very attractive. Our results show that, over the long haul, they appear to act like money market accounts paying high rates of interest (over 10%) with relatively small chance that the investor will be disappointed.

References

[1] Bogle, J.C. (1999). *Common Sense and Mutual Funds: New Imperatives for the Intelligent Investor.* New York: John Wiley & Sons.

[2] R.G. Ibbotson Associates, 2001. *Stocks, Bonds, Bills, and Inflation 2001 Yearbook: Market Results for 1926-2000.* Chicago: R.G. Ibbotson Associates, pp. 200-

[3] Thompson, J.R. (1999). *Simulation: A Modeler's Approach.* New York: John Wiley& Sons, 115−142.

[4] Thompson, J.R. and Williams E.E. (1998).“A Post Keynesian analysis of the Black-Scholes option pricing model,” in *The Journal of Post Keynesian Economics*, Winter, pp. 251-267.

Analyzing Vulnerability Results for Tags and Tamper-Indicating Seals

Roger G. Johnston and Anthony R.E. Garcia

Vulnerability Assessment Team
Los Alamos National Laboratory
MS J565, Los Alamos, NM 87545

Keywords: tamper detection, vulnerability assessments,
security, tampering, design of experiments

Introduction

Tamper-indicating devices, often called "seals", are meant to detect unauthorized access, entry, or tampering (NFESC, 1997; NFESC, 2000; Johnston, 1997c). Seals are widely used for a variety of applications including access control, cargo security, pilferage detection, banking, courier services, document and records integrity, customs, law and drug enforcement, hazardous materials accountability, nuclear safeguards & nonproliferation, counterespionage, counterterrorism, and consumer protection (Johnston, 2001d; Tyska, 1999). The U.S. Army frequently uses seals to detect pilferage and tampering with weapons during storage and shipment, and also to secure ammunition, medical supplies, soldier's personal property, courier packages, and classified documents.

Unlike locks, seals are not designed to resist, complicate, or delay unauthorized access. Instead, they record that it took place. Also unlike locks, seals must be inspected, either manually or electronically, to do their job. Seals differ from intrusion detectors ("burglar alarms") in that unauthorized access or entry is not reported immediately. This has both advantages and disadvantages (Johnston, 2001c).

Seals take a variety of forms. They can be frangible foils or films, plastic wraps, pressure-sensitive adhesive tapes, crimped cables or other (theoretically) irreversible mechanical assemblies; security containers or enclosures that give evidence of being opened; devices or materials that display irreversible damage or changes when manipulated; and electronic or electrooptic devices and systems that continuously monitor for changes, such as a break in an electrical cable or fiber-optic bundle. Perhaps the most familiar everyday example of seals is the tamper-evident packaging found on over-the-counter pharmaceuticals.

A tag is a device, or an applied or intrinsic feature, used to identify an object or container. A familiar example of a tag is the license plate on a car. When used for security purposes, a tag should be difficult or expensive to counterfeit, as well as difficult to lift. To "lift" a tag means to remove it from one object and attach it to another without damaging the tag and without being detected.

Tags and seals are related in that an effective security tag must be able to detect tampering. An effective seal, in turn, must have a unique (tag-like) characteristic or "fingerprint", such as a serial number. This is necessary so that it is not trivial to remove the seal from an object or container and replace it with a duplicate.

The Vulnerability Assessment Team at Los Alamos National Laboratory (LANL) is involved in trying to improve tamper detection for a variety of different applications, commercial as well as government. We conduct vulnerability assessments on tags, seals, and tamper detection programs, develop new tags and seals, and work on ways to improve existing tamper detection methods (VAT, 2002).

A "vulnerability assessment" (Jones, 1996; Johnston, 1997b) involves discovering and demonstrating ways to defeat security devices, systems, or programs. It may also involve suggesting countermeasures. To "defeat" a seal means to open it, then reseal (using either the original seal or a counterfeit) <u>without being detected</u>. Similarly, to "defeat" a tag means to counterfeit or lift it, <u>without being detected</u>. "Attacking" a tag or seal involves undertaking a sequence of actions designed to defeat it.. A successful attack is also called a "defeat".

There are two aspects of vulnerability assessments on tags, seals, and tamper detection programs that are particularly tricky: (1) designing effective vulnerability assessment experiments and (2) analyzing and reporting the results. This paper examines some of the complications, problems, and issues associated with these matters.

Complications, Problems, and Issues

Quite a number of different complex factors affect the design of experiments for testing tags, seals, and tamper detection. These factors and others also greatly complicate the statistical analysis and reporting of results.

One of the major problems in performing vulnerability assessments is that the field of tamper detection is not well developed. Although tags and seals have been used for at least 7,000 years (Johnston, et al., 2001b), they are poorly understood (Johnston, 2001d). There is little in the way of formal theory, few meaningful standards for performance or testing, and considerable confusion among end users about how to use tags and seals effectively (Johnston, 2001a). We have observed considerable misunderstanding among security professionals about tamper detection concepts, strategies, and terminology. There is frequently, for example, confusion about the difference between locks, seals, and tags. Expectations for tamper detection are often vague or unrealistic.

In our experience, there is also often a lot of confusion and wishful thinking associated with vulnerability assessments. Many security managers believe that a vulnerability assessment should ideally find no vulnerabilities. Our view is that multiple vulnerabilities are always present in ANY security device, system, or program. Discovering some of these vulnerabilities provides the opportunity to mitigate or eliminate them. Thus, the discovery of vulnerabilities should be viewed as good news, not bad news. Indeed, a vulnerability assessment that finds no problems has zero value.

Security managers (or their supervisors or auditors) often feel that if countermeasures and recommendations arising from a vulnerability assessment are implemented, this is an admission that they have been negligent or incompetent in the past. Vulnerability assessors must make allowances for such a mindset, and attempt to counter it. This can be quite a challenge.

A related problem that often plagues vulnerability assessments is the prevalence of absolutist ideas about security. To many people, a security device, system, or program is either secure, or it has vulnerabilities and is insecure. In reality, security is a continuum. Nothing is either fully secure nor completely insecure. A binary view of security is both unrealistic and dangerous (Johnston, 2001a).

Unfortunately, we in the Vulnerability Assessment Team are all too familiar with another problem with vulnerability assessments: "Shoot the Messenger Syndrome". It is all too common when security problems are discovered for the vulnerability assessors (often called "black hatters"!) to be viewed as the problem, rather than the vulnerabilities themselves. The Nobel prize-winning physicist Richard Feynman has written amusingly about this phenomenon (Hutchings, et al., 1985). Sometimes security programs are evaluated be personnel who may damage their own careers if they report significant or numerous security problems. Effective vulnerability assessments rarely occur in such an environment.

In many cases, security managers and supervisors, or manufacturers and vendors of security products do not want vulnerability assessments done because they highlight problems. This attitude is not conducive to effective security. Because security professionals (and security programs) are often judged by the lack of problems, however, this situation is difficult to avoid.

Often, vulnerability assessments are conducted by personnel who have a serious conflict of interest. It is not unusual to find that a security product or system has been analyzed and tested by proponents, manufacturers, or vendors of that product or system. Not surprisingly, few vulnerabilities are typically found in such analyses. Sometimes, vulnerability assessors are chosen who are clearly unqualified or unimaginative. They also tend to uncover few, if any, vulnerabilities.

A problem that we have frequently encountered in conducting vulnerability assessments is that the use protocol for a given tag or seal may be poorly defined, inconsistent, or not formalized. The "use protocol" is exactly how a tag or seal is used. This includes the entire lifecycle of the tag or seal including procurement, shipping, storage, checkout, installation, inspection, removal, disposal, data handling, analysis, interpretation, postmortem forensics (if any), and training of security personnel. A tag or seal is no better than its use protocol (Johnston, 1977a). Discovering and demonstrating defeats is difficult if it is not clear what use protocol must be defeated.

The human factors associated with defeating tags and seals are particularly tricky to handle. While defeating a lock, safe, or vault (for example) is mostly about beating hardware, defeating a tag or seal is primarily about fooling people. That psychological factor can be difficult to model, predict, or analyze. It can also be quite difficult to model and predict human error, yet human error is responsible for most security failures.

One frustrating problem with vulnerability assessments is that tags and seals are often exposed to a vulnerability assessment only after the design is finalized and the product is in production. By then, it is usually too late to make any changes in the design that might mitigate or eliminate vulnerabilities. Ideally, vulnerability assessments should be conducted on a tag or seal iteratively, throughout the various design phases of the product. This has the additional advantage of make the vulnerability assessors part of the design team, rather than the "enemy"—thus increasing the chances that their warnings and recommendations will be heeded.

Another very common problem with conducing vulnerability assessments on tags and seals is the paucity of samples made available to the assessment team. We have often been asked to find the vulnerabilities in a seal when we have been given exactly one sample—which must be returned to its owner undamaged. The most effective vulnerability assessments require dozens, if not hundreds of seal samples. Some, though not all, of the test seals usually need to be destroyed during the assessment process. It is usually not reasonable, in our view, to assume that an adversary won't have access to large numbers of seals for testing and practice—especially since most seal manufacturers give away samples for free.

Designing and executing a vulnerability assessment on a tag or seal is often constrained by time and funding. Adversaries who might try to attack tags or seals may not be so constrained. A time- or budget-limited vulnerability assessment, moreover, requires some kind of prioritization of the hundreds of possible attacks. Time and money will usually not be available to study them all. Not all possible attacks will be relevant or ultimately prove to be successful, and some of the attacks that do ultimate work may end up consuming more time and money to develop than they are worth. There are usually no obvious guidelines for how to prioritize attacks, though experience seems to be helpful.

A related complication is that we don't automatically know when the best attacks have been found. The best attacks may forever go undiscovered, or be discovered only at a later date by a different set of vulnerability assessors (or adversaries). Vulnerability assessments thus have no clear-cut end point, and so it is never clear when the "experiment" is over.

It can be difficult in analyzing vulnerabilities to know what adversaries might attempt. The exact identity of all possible adversaries, and the resources/capabilities available to them are usually unknown and can only be speculated about.

There are often "recursion" problems with vulnerability assessments. By this we mean problems associated with iteration and with hitting a moving target. Often, for example, after being shown a serious seal vulnerability, a distressed security manager will ask if there isn't a simple countermeasure. Usually there is. Once reassured that the vulnerability can be easily dealt with, the security manager will relax—yet never actually implement the countermeasure! Another example of recursion occurs when the recommended countermeasure, whether a change in the seal use protocol or to the seal design itself, introduces new problems and vulnerabilities. It is very difficult to fully foresee in the original vulnerability assessment, all the vulnerabilities associated with a theoretical tamper detection regime that may come into existence as a result of implementing some or all of the initial recommendations. Ideally, a new vulnerability assessment should be conducted after recommended changes are actually made, but this is rarely done.

"Compliance mode" is another problem that threatens security and can complicate vulnerability assessments and the implementation of recommendations that arise form them.. Sometime security managers or other security personnel become (or are forced to become) so focused on satisfying auditors, regulations, and formal requirements that they lose sight of real-world security issues (Johnston, 2001a). Compliance mode is difficult to avoid in large organizations and bureaucracies, in old established operations, and for security programs that do not encourage security personnel to be flexible, creative, or proactive.

Vulnerability assessments, like any kind of security analysis, suffer from ambiguities associated with the choice of metrics and with the difficulty of conducting a cost/benefit analysis. In the security field, success is defined as nothing happening. That is a vary bizarre metric, and it makes tradeoffs between costs and benefits difficult to rigorously analyze.

A very common problem in security is that security managers and planners often have a very different idea of what is happening in a security program than what is really going on (Johnston, 1997c). Vulnerability assessors should ideally try to analyze the true security program, not the mental image of the program that exists at high levels in the organization. Doing this, however, can create a lot of resistance from high-level security managers and planners.

Many real-world attacks on security devices or systems rely on false alarming, fault analysis, or "watch and pounce" methods (defined below). Yet because these are anomalous, rare events, they can be quite difficult for vulnerability assessors to observe, model, predict, or replicate. It can also be difficult to sufficiently control related parameters.. The classic example of a *false alarm* attack is for burglars to shake the windows outside a bank building one night to set off the alarm. When they first do this, the police arrive in a hurry to check out the alarm. The next night, and for each of the next 3 nights, the burglars generate the same false alarm. After 5 nights in a row, because of all the false alarms, the police either are slow to arrive, fail to arrive entirely, or the alarm system has been turned off because it has become a nuisance. That is when it is safe for the burglars to actually enter the bank in order to steal money. *Fault analysis* is a method used by an adversary to learn about a security device or system (especially a complex one) and its vulnerabilities by studying how the device or system behaves when exposed to unusual conditions or probes. *Watch and pounce* attacks involve the adversary passively waiting and observing until security personnel make a mistake, then leaping into action to exploit that mistake. In general, it can be very difficult to control experimental parameters for rare events so that meaningful results can be obtained.

It is generally quite difficult to obtain realism when testing or demonstrating vulnerabilities with tags, seals, or security programs. This is particularly true inside high security facilities. One major reason is that for critical applications, such as guarding nuclear weapons, highly realistic tests may be too risky or difficult to arrange inside the facility. Indeed, one of the best times for adversaries to attack a facility is when security personnel think a drill is underway. Another factor that limits realism for tests inside a high security facility—but a factor we welcome—is the need to avoid putting vulnerability testers at risk of injury or death. Real adversaries may not feel so constrained. It is also usually quite difficult to arrange realistic tests inside a large secure facility without alerting security personnel or security committees that such tests are underway. Advance notice can distort the experiment.

There are problems with achieving realism even for tests or demonstrations on seals outside the facility in which they are used. For one thing, security managers often send their best seal inspectors to participate in experiments, rather than their average or mediocre inspectors. In the tests and demonstrations, the inspectors are unavoidably on high alert and often don't use the same seal use protocols they routinely employ. Obviously this can skew results.

A particularly interesting challenge in trying to maintain realism has to do with the artificial paranoia we typically see in experiments for Type 2 vs. Type 1 errors (Johnston, et al, 2001b). In a Type 2 (false accept) error, the seal inspector fails to detect that a seal has been attacked. In a Type 1 (false reject) error, the inspector incorrectly believes a seal has been attacked when it really has not. Usually there is some kind of inherent tradeoff between Type 1 and Type 2 errors, such as shown schematically in figure 1.

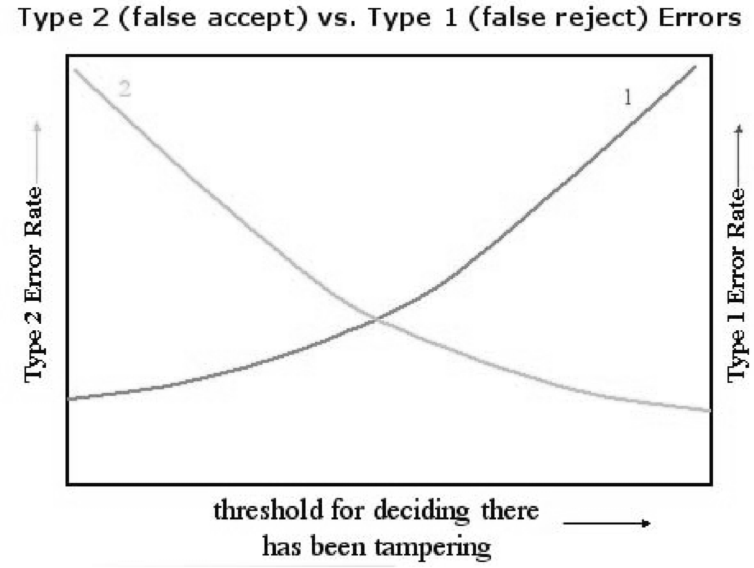


Figure 1 - Low rates of Type 2 errors typically come at the cost of relatively high rates for Type 1 errors, and vice versa

For most tamper detection applications, Type 1 errors are approximately as undesirable as Type 2 errors. Seal inspectors usually recognize that accusing a seal of having been attacked, when it has not, can have serious repercussions. (There are, however, applications where Type 1 errors are much less serious than Type 2 errors. If, for example, the items being monitored for tampering are relatively inexpensive and can be readily discarded if suspicious, or if extensive postmortem forensics are available to reliably check for tampering, high rates of Type 1 errors may be entirely acceptable.)

Another common problem is that sponsors of vulnerability assessments are rarely willing to commit the time and money necessary to conduct thorough and rigorous blind or double blind tests of whether seal inspectors can detect the attacks. Sponsors are usually content with hearing a description of the attacks, seeing them demonstrated, or examining attacked seals—followed by a discussion of possible countermeasures (Johnston, et al., 2002).

A final complication in designing and executing vulnerability assessments has to do with the fact that tags and seals are rarely used in isolation. They are often employed in conjunction with a number of other security devices, personnel, and various nested security layers. Seals, for example, may be used inside a facility that is surrounded by security fences and protected by a guard force and intrusion detectors. The interactions and coupled vulnerabilities among these various security layers can be extraordinarily difficult to analyze.

Reporting Results

One of the major complications in reporting the results of vulnerability assessments is that a defeat of a tag, seal, or tamper detection program is a matter of degree and of probability (Johnston, et al., 2002). A crude attack will not necessarily be detected with 100% probability, nor will a subtle attack always be missed.

We have attempted to deal at least partially with this problem by developing what we call the "Los Alamos Defeat Categorization Scheme" (Johnston, et al., 2002; Johnston, 1997b). Under this scheme, a defeat is classified as being of type 1, 2, 3, or 4 depending on whether it fools the seal inspector when he/she: follows the nominal, usual, or recommended inspection procedures (type 1 defeat), does an unusually careful visual inspection of the exterior of the seal (type 2 defeat), opens the seal and does a careful visual inspection of the seal interior and exterior (type 3 defeat), or uses state-of-the-art forensics techniques to analyze the seal and look for signs of tampering (type 4 defeat). A type 4 defeat is problematic in that it is not possible to *prove* there is no technology capable of detecting the attack. We have nevertheless categorized about 15% of the 289 seal defeats we have demonstrated on 198 different seal designs as type 4 because we are unable to identify any method or technology that could be used to spot the attack (Johnston, et al., 2002).

We have attempted to identify the attributes of an effective vulnerability assessment of a tag or seal, and how to most effectively present the findings and recommendations (Johnston, 1997b). Some of the information that we consider essential in reporting on vulnerability assessments include the following:

- Who did the vulnerability assessment and what is their background and qualifications?

- Do the vulnerability assessors have any potential conflict of interest?

- How many attacks were devised, partially demonstrated, fully demonstrated, and practiced to perfection?

- What was the time and cost to devise, develop, and practice each attack?

- What is the time and cost to execute each attack?

- How much off-site preparation time is needed to execute each attack?

- What inside information, if any, was used for each attack?

- Is each attack high-tech or low-tech in terms of methods, tools, and attack personnel?

- What are the size, weight, and nature of the attack tools and materials for each attack?

- What are the countermeasures and recommendations that arise from the study?

- Vulnerability assessors should try to provide samples of attacked seals, as well as in-person or video demonstrations of the attacks.

- Vulnerability assessors should also provide a sanitized (unclassified) summary of the vulnerability assessment that is devoid of sensitive details. This permits others to judge the thoroughness of the assessment without giving away vulnerability information that might assist adversaries.

Conclusion

Conducting vulnerability assessments of tags, seals, and tamper detection programs is a complex and challenging process. Issues of how to design vulnerability experiments, analyze the results, reach rigorous conclusions, and present findings in a statistically meaningful way are largely unresolved. Addressing these issues is important because of the continuing need for effective tamper detection.

References

Hutchings, Edward (Editor), Leighton, Ralph, Feynman, Richard Phillips, and Hibbs, Albert, 1985, *'Surely You are Joking, Mr. Feynman': Adventures of a Curious Character* (Batam, New York, 1985), pp. 119-137.

Johnston, Roger G. and Garcia, Anthony R.E., 1977a, "Vulnerability Assessment of Security Seals," Journal of Security Administration, Vol. 20, No. 1, June 1997, pp. 15-27, http://lib-www.lanl.gov/la-pubs/00418796.pdf.

Johnston, Roger G., 1997b, "Effective Vulnerability Assessment of Tamper-Indicating Seals," Journal of Testing and Evaluation, Vol. 25, July 1997, pp. 451-455, http://lib-www.lanl.gov/la-pubs/00418792.pdf.

Johnston, Roger G., 1997c, "The Real Deal on Seals," Security Management, Vol. 41, September 1997, pp. 93-100, http://lib-www.lanl.gov/la-pubs/00418795.pdf.

Johnston, Roger G., 2001a, "Tamper Detection for Safeguards and Treaty Monitoring: Fantasies, Realities, and Potentials," The Nonproliferation Review, Vol. 8, No. 1, Spring 2001, pp. 102-115, http://lib-www.lanl.gov/la-pubs/00367047.pdf.

Johnston, Roger G., Martinez, Debbie D., and Garcia, Anthony R.E., 2001b, "Were Ancient Seals Secure?," Antiquity, Vol. 75, No. 288, June 2001, pp. 302-303, http://lib-www.lanl.gov/la-pubs/00818331.pdf.

Johnston, Roger G., 2001c, "The 'Town Crier' Approach to Monitoring," Report LAUR-01-3726 (Los Alamos, NM: Los Alamos National Laboratory, July 2001).

Johnston, Roger G., 2001d, "Tamper-Indicating Seals for Nuclear Disarmament and Hazardous Waste Management," Science & Global Security, Vol. 9, No. 3, 2001, pp. 105-107, http://lib-www.lanl.gov/la-pubs/00818333.pdf.

Johnston, Roger G., Garcia, Anthony R.E., and Pacheco, Adam N., 2002, "Efficacy of Tamper-Indicating Devices," The Journal of Homeland Security (in press).

Jones, James L., 1996, "Improving Tag/Seal Technologies: the vulnerability assessment component," Report 95/00599, (Idaho Falls, ID: Idaho National Engineering and Environmental Laboratory, December, 1996).

Naval Facilities Engineering Services Center (NFESC), 1997, "Antipilferage Seal User's Guide" (Port Hueneme, CA: October, 1997), http://locks.nfesc.navy.mil/Security_seals/guides/seal_ug/rev_sealguide.pdf.

Naval Facilities Engineering Services Center (NFESC), 2000, "DoD Training Course on Effective Seal Use" (Port Hueneme, CA: Spring, 2000), http://locks.nfesc.navy.mil/Security_seals/security_seals/sp2086.pdf.

Tyska, Lou (Editor), 1999, Guidelines for Cargo Security & Loss Control (Annapolis, MD: National Cargo Security Council, 1999), pp. 29-38.

Vulnerability Assessment Team (VAT), 2002, Los Alamos National Laboratory, http://pearl1.lanl.gov/seals.

LIARS, DAMNED LIARS AND STATISTICIANS

Arthur Fries
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
afries@ida.org

ABSTRACT

Nowhere is the need for insightful and accurate statistics more profound than in the arena of national policy making, yet statisticians routinely are excluded from the table during public policy debates. Why? Drawn from personal encounters, this paper presents examples of recent questionable "statistical analyses" – originally intended to support national-level decision-making within the U.S. Department of Defense, but instead having the effect of further perpetuating the widespread impression that the statistical community should not be entrusted with such a critical role.

INTRODUCTION

The all too familiar refrain goes: "There are three kinds of lies: lies, damned lies, and statistics." But it is not the "statistics" per se that ever are truly at fault, it is the individuals that produce, describe, publish and disseminate them. Nowhere is the need for insightful and accurate statistics more profound than in the arena of national policy making, yet statisticians routinely are excluded from the table during public policy debates.[1] Why?

Drawn from personal encounters, this paper documents examples of recent questionable "statistical analyses" – originally intended to support national-level decision-making, but instead having the effect of further perpetuating the widespread impression that the statistical community should not be entrusted with such a critical role. The examples motivate specific "lessons learned" and raise various "ethical questions" that readily generalize. Although most statisticians are likely to consider the associated principles of conduct to be self-evident, their application in practice obviously has not been uniformly successful. As a discipline and as a community of individuals fully capable of effectively supporting government officials and policy-makers, we can and must do better.

All of the cited examples are based on actual occurrences, but, to simplify the exposition and promote ready comprehension, the situational contexts and other details have been altered. The central observations, concerns and statistical issues, however, have been maintained. To repeat, while a specific example is introduced here as, say, a “Marine Corps Helicopter Effectiveness Study,” the underlying authentic circumstances may in fact involve, say, Army jeeps. This approach also serves to shield and protect the “guilty”, i.e., avoiding direct identification of the specific parties involved.

Five examples are presented and discussed in turn below, all dealing with high-level national defense or security issues. The examples are followed by a brief discourse of conclusions and lessons learned.

The setting for each example is similar. In response to a high-profile Department of Defense (DoD) study, an independent review was convened at government expense. The review panel comprised one or more assigned statisticians, *in toto* or as an integral subgroup of a more expansive collective of subject matter experts. The independent review panel was provided complete and unrestricted entrée to all aspects of the DoD study in question – including study data, detailed supporting data and analyses, briefings from study authors, access to study authors for follow-on questions, etc. Upon completion of their independent review, the review panel issued an assessment report to its sponsoring agency.

EXAMPLES

MARINE CORPS HELICOPTER EFFECTIVENESS STUDY

Based on observed test results on a series of developmental helicopters, the Marine Corps had determined that a new helicopter was suitable to proceed into full-rate production, despite some acknowledged unresolved performance issues. A critical component of their conclusion was a set of helicopter flight effectiveness curves that traced out aerodynamic performance across flight envelope regions.

An independent review panel of statisticians held several working sessions, interviewed the study authors, and scrutinized data sets. Its official assessment did not support the conclusions of the Marine Corps study. One cited reason was that that the flight effectiveness curves presented in the original study lacked confidence intervals, and that the observed features and reported performance could be merely manifestations of randomness.

When the review panel's assessment report was formally released, the original study authors objected to this particular cited shortcoming in their work. While it was indeed true that no confidence intervals appeared in their study *per se*, the authors had in fact constructed such curves – utilizing bootstrap procedures based on replicated observations, i.e., completely model independent. These definitively established that the reported curves were not attributable to randomness. Moreover, the study authors explicitly had briefed these exact results to the review panel and, at the review panel's request, provided paper copies of the entire briefing.

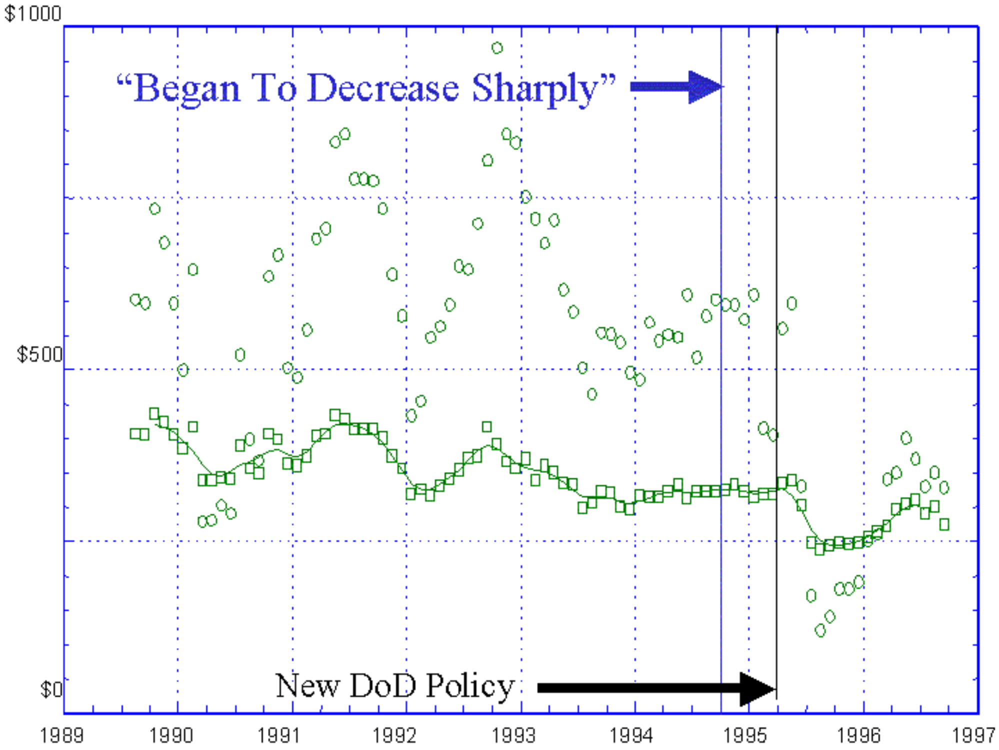
Despite the protest of the original study authors, the review panel did not rescind or even edit its assessment report. Neither did they apprise their government sponsors of the study authors' objections. Instead, the lead statistician reported back that *the review panel had been tasked to assess the original study report, and that what they had written in their assessment was technically correct and not an "error of fact"!*

DOD BUSINESS PRACTICES

In early 1995 the DoD instituted new business practices that significantly reduced procurement costs for certain classes of items, bringing expenditures down to a level commensurate with public sector commercial transactions. This conclusion was supported by a DoD study that offered a version of Figure 1 as substantiation. Here the "oval" and "rectangular"-shaped data points respectively depict per item costs for DoD purchases and for commercial non-DoD purchases. The new DoD business practices were officially implemented March 1995 (denoted by a vertical black line). Prior to that point, DoD costs generally greatly exceeded comparable public sector costs.

The group of independent reviewers declined to attribute the achieved cost reduction to the implementation of the new DoD procurement practices instituted formally in the first quarter of 1995. Instead, they argued that the costs depicted in Figure 1 actually "began to decrease sharply" in September of 1994 (denoted by a vertical blue line), a full half-year before the new DoD policy. Accordingly, they contended, something else other than the new DoD policy could just as well be the root cause for the decline in DoD procurement costs.

Now it is difficult for me, and every other individual that has been shown Figure 1, to imagine how any reasonable analysis could conclude that DoD costs started their rapid fall in the early fall of 1994. What had happened is that the review team had relied on a slightly different depiction of Figure 1 – identical in all respects except that the DoD cost data had been supplemented with a smoothing curve (similar in nature to the one displayed in Figure 1 for the public sector transaction costs). The smoothing curve was derived from a simple running average of nine neighboring points – the current point, the four preceding ones, and the four following ones. As such, the curve "anticipates" declines in future data and begins to decrease before the data actually do. This property had been explicitly noted in briefings to the independent reviewers, but it was not specifically cited in the published DoD study (although the particulars of the smoothing curve calculations were given).



DoD = "Oval"   Public Sector = "Rectangle"

Figure 1. Dollar Transaction Costs – DoD & Public Sector

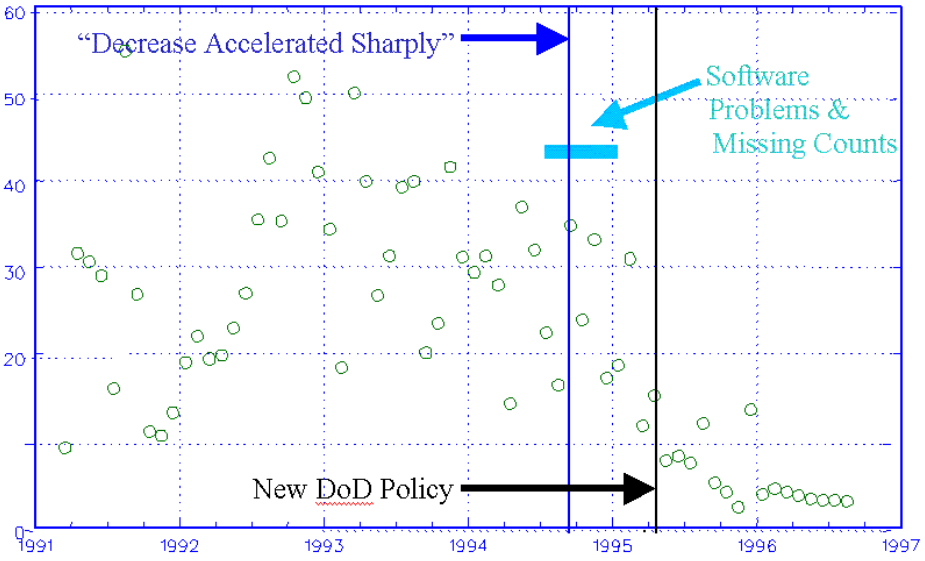


Figure 2. Counts of Monthly DoD Purchases

One aspect of the new DoD procurement policy was the shift away from many small purchases to fewer larger transactions reflecting volume discounts. The original DoD study pointed to the relatively low purchase counts post-March 1995 in Figure 2, and argued that, in tandem with Figure 1, this logically supported the contention that the new policy was effective. The independent assessment team, however, reached an entirely different conclusion. They argued instead that the purchase counts actually began to decrease well beforehand and, further, that the "decrease accelerated sharply" in September 1994 (depicted by the blue vertical line). While this observation correctly characterized Figure 2 as presented in the original DoD study, it ignored additional information made available to the review panel. The counts depicted in Figure 2 had been extracted from an automatic purchase tracking system. Although unknown to the DoD study team at the time of their original report, they subsequently learned that the computerized tracking system had experienced software problems during the latter half of 1994 (depicted by the light blue shading in Figure 2) resulting in an undercounting of the number of recorded transactions. This fact, as well as another set of independently derived confirmatory data, had been briefed explicitly to the assessment panel.

Despite the DoD study team's "reminders," the independent reviewers and their lead statistician declined to revise the assessment report. Neither did they apprise their government sponsors of any objections. To the consternation of the DoD researchers, attempts to appeal to the prestigious national committee of statisticians that furnished the assessment's team lead statistician also proved to be counterproductive. The committee's chair downplayed the significance of the divergent conclusions, referring to them as *merely "differences in data interpretation."*

AIR FORCE COST-BENEFIT COMBAT STUDY

The Air Force conducted a cost-benefit combat study in which it assessed the relative effectiveness of recent air campaigns in terms of successful mission sorties per thousands dollars of support and execution costs. An independent review panel scrutinized the Air Force's study and issued its own report. Its main conclusion was that the Air Force results are suspect because the Air Force may have selectively excluded data from its analyses:

> The Air Force study states that multiple air operations, with no kills, were not included because *"Further examination has shown that these operations were not, in themselves, major operations."* What does it mean to be a 'major' operation? Are actions deemed to be 'major' *ex ante*, or are they deemed *ex post* if they are observed to be sufficiently successful?

Here the reviewers extracted a quote from the Air Force study (in italics) to suggest that the Air Force analysis may have been biased. They did not independently review the descriptions of specific operations to determine whether any may have been improperly excluded from the calculations. Nor did they ask the Air Force analysts directly as to what the rules had been for inclusion in the computations.

One standard that the Air Force analysts employed was to not count logistical operations towards target kill effectiveness measures, since the mission did not include engaging targets. On the other hand, any associated tactical operation would be penalized the dollar cost of its supporting logistical operations. For example, Operation Desert Shield entailed the massive deployment of U.S. forces to Saudi Arabia as a prelude to the Gulf War against Iraq. The Air Force study did not score it as a "0-kill operation" and did not incorporate it into "benefit" analyses quantifying average mission accomplishment. The costs of Operation Desert Shield, however, were combined with those of the ensuing Operation Desert Storm, the actual combat portion of the Gulf War, in determining the "cost" of the latter operation. The original Air Force study indirectly alluded to this procedure [*emphasis added*]:

> Further examination has shown that these operations were not, in themselves, major operations. *Rather they established the international cooperation, increased deployed assets, and provided training that was later employed effectively in the follow-on major operations.*

This complete Air Force study citation provided explicit logical rationale for the exclusion of specific operations. The independent assessment, however, chose not to report these. When confronted with these facts, the lead statistician from the review team stated that their *assessment report was "accurate and fair."*

NAVY CIGARETTE SMOKING CESSATION STUDY

A group of Navy analysts had been assigned the task of assessing the degree of success of various programs aimed at assisting enlisting men curtail or cease altogether their cigarette smoking. Unfortunately, personal data on individual smoking frequencies had not been chronicled systematically. The analysts augmented the available data by examining indirect indicators of smoking, *i.e*., surrogate variables, including official records of sick days and hospital visits. One Navy initiative that was assessed was whether increasing the cost of cigarettes charged to sailors aboard ship lead to less smoking. In their final study, the analysts noted that data limitations precluded them from computing direct estimates of the price elasticity of demand (*i.e*., percent reduction in total consumption per percent increase in cigarette cost). However, they were able to estimate price elasticities for the cigarette usage indicators. Table 1 presents their study findings. The estimated elasticities are negative, synonymous with increases in cigarette prices being associated with fewer sick days and fewer hospital visits.

The Navy study focused on the "Best Estimates *e*" − describing how they were calculated and providing graphical depictions (with confidence intervals). It also drew a clear distinction between the price elasticities of the indirect indicators (that had been estimated) and the price elasticity of the demand for cigarettes (that had not been estimated):

> Assuming equal errors in the two variable, 10,000 bootstrapped replications of the data yield a best estimate of $e = -0.63$, a 95th percentile confidence interval of $-0.86 < e < -0.50$, and a standard least squares regression coefficient of R = -0.71." …
> The table summarizes the estimates of price elasticity as determined from the two indirect measures of cigarette use. While it is apparent that the 'elasticities' of the indirect usage indicators developed above are each related to the price elasticity of demand for cigarettes, those relationships are imprecisely understood. [*emphasis added*]

These careful characterizations were not faithfully duplicated in the subsequent descriptions thereof offered by a selected group of independent statistician reviewers. Their rendition appears in Table 2. Note that the Navy "Best Estimates *e*" are nowhere reported. Further, the presented values are incorrectly asserted to be estimates for the demand for cigarettes, exactly what the Navy study had insisted they were not estimating. The independent assessment continued to criticize the Navy study's "estimates of demands," essentially for the very reasons that had led the Navy analysts to abandon estimation of demand.

When informed of these discrepancies, the statisticians argued: "There had been some uncertainty on their part as to which set of estimates to report (*i.e*., *e* or R)," "They were accustomed to regression-based estimates of elasticity", and "These are not 'errors of fact'." The Navy analysts countered that if the statisticians were uncertain, all they had to do is request clarification. They further noted that the estimates *e* also had been derived from regression analyses. And they continued to press for a corrected assessment report.

Rather than admit any errors, the statisticians offered an "addendum for clarification." They did not withdraw or correct the copies of their report that had already been published and distributed (*e.g*., to the government sponsor that paid for their services). Instead, they published new report copies and inserted, with limited amplifying explanation, a solitary page after the References section to "reproduce the original Navy study table." This "reproduction" appears here in Table 3. Note that the sign of the "Best Estimate" for Hospital Visits erroneously is missing a negative sign, and that the estimate is mistakenly labeled as "*c*" vice "*e*."

Summary of Price Elasticity Estimates for Indicators of Cigarette Usage

Usage Indicator	**Best Estimate**	**95% Confidence Interval**	**Regression Coefficient**
Hospital Visits	-0.63	$-0.86 < e < -0.50$	-0.71
Sick Days	-0.60	$-0.98 < e < -0.29$	-0.51

Table 1. Navy Study Table [*emphasis added*]

Estimated Price Elasticities of Demand in the Navy Study

Data Set	**Estimated Price Elasticity of Demand**
Hospital Visits	-0.71
Sick Days	-0.51

SOURCE: Navy Study

Table 2. Independent Assessment Table [*emphasis added*]

Summary of Price Elasticity Estimates for Indicators of Cigarette Usage

Usage Indicator	**Best Estimate**	**95% Confidence Interval**	**Regression Coefficient**
Hospital Visits	0.63	$-0.86 < c < -0.50$	-0.71
Sick Days	-0.60	$-0.98 < e < -0.29$	-0.51

Table 3. Addendum to Independent Assessment

MULTI-SERVICE MEDICAL STUDY

Two military services conducted studies of how best to treat a particular health malady prevalent among combat veterans. Service A deduced "Treatment Regimen 1 was much more cost effective than Treatment Regimen 2," while Service B determined that "Treatment Regimens 1 and 2 were approximately equally cost effective." The inexact conclusion espoused by Service B's study was accompanied by an explicit acknowledgement that all available pertinent data were imprecise.

An independent review team of subject matter experts was convened. They judged neither of the service's studies to be plausible or persuasive. One statistician on the review panel subsequently authored an open-literature publication with main points illustrated by supporting examples. In recalling the multi-service medical study, the statistician reported "Service A concluded Treatment Regimen 1 was much more effective than Treatment Regimen 2, while Service B concluded exactly the opposite." [*emphasis added*]

The Study B analysts objected to this "incorrect characterization" of their overall conclusion, noting that they had never stated that Treatment Regimen 2 was much more effective than Treatment Regimen 1. The statistician, however, countered that his text was correct because in mathematical logic the opposite of "X" is "not X." In other words, the opposite of "Treatment Regimen 1 was much more effective than Treatment Regimen 2" is "Treatment Regimen 1 was not much more effective than Treatment Regimen 2." Apparently the statistician deemed this latter statement to be sufficiently consistent with Service B's originally reported conclusion, and was unconcerned that his particular words (*i.e.*, "Service B concluded exactly the opposite") were prone to misinterpretation.

DISCUSSION

Defense and national security debates are of critical importance. Relevant data may be limited and sparse, or, at the other extreme, voluminously intractable. Interpretational difficulties and analytical complexities may abound. The realm of statistics and our community of statisticians have much to offer – to properly frame and characterize the essential content of the available data, to focus attention on practically insightful issues, and to propose additional potentially illuminating data collection initiatives.

Despite our potential to contribute substantively, statisticians generally are excluded from defense and national security discussions – even when the primary issue is the analysis of a particular set of data or a specific statistical hypothesis is in question. Moreover, high-level decision makers rarely are inclined to suggest that statisticians should be consulted. Indeed, the exact opposite reaction is more typical. I have been at defense-oriented meetings and conferences in which the mere mention of "statisticians" (e.g., to announce an upcoming defense-related workshop sponsored by a statistical organization) spontaneously prompted condescending laughter and derisive commentary, including, for instance, "Do you *really* think that statisticians will be helpful?" What have we done to deserve such a reputation?

Judging by the examples cited above, the answer is in my opinion "plenty." It is always true that members of special panels and committees formed to address sensitive policy issues are charged with the grave responsibility of conducting impartial, objective, competent and comprehensive assessments. I argue, however, that individual statisticians granted the rare opportunity to participate in such endeavors must be held to an even higher accounting, for they represent not just themselves but also our entire discipline. If they fail to uphold exemplary standards – whether by direct act, complicity, or omission – they discredit us all.

Clearly statisticians supporting national-level panels and committees must ensure that all statistical analyses appearing in their reports are sound, whether they themselves are the architects of those analyses or not. But they should not restrict themselves to such a narrow focus. They must be cognizant of and comfortable with the basis of and derivation of overall conclusions. They must be sensitive to potential fairness and objectivity issues – including incomplete or unbalanced panel/committee membership, actual conflicts of interest or the possible perception thereof, over reliance on pre-conceived or initial impressions, and the exclusion of inputs from and comprehensive scrutiny by independent subject matter experts. When reviewing the work of other analysts, those analysts should be provided an opportunity to comment on emerging products – especially with respect to the clarity and accuracy of the portrayals of the analysts' work.

Practicing statistical consultants learn early on that in order to be effective they must work hand-in-hand with their clients, closely and repeatedly interacting to develop a proper appreciation for the details of the problems at hand. It is my personal observation that "academic" statisticians entrusted with national policy studies all too often pursue the exact opposite tact, minimizing interactions with the very individuals whose reports they are to review. The first preference for some seems to be to just scrutinize an actual report, without any personal interchanges at all with the authors. Others occasionally may schedule *pro forma* briefings, but generally few extended two-way dialogues, especially at any detailed level, ensue.

Observers of this process might infer from such behavior that the statisticians either already have made up their mind or that they do not consider the matter at hand to be of any particular importance. Neither impression paints a flattering picture. When the statisticians' final product is itself problematic, as in the specific examples cited above, what are policy makers to conclude? Why should they even consider giving statisticians another chance?

ACKNOWLEDGEMENTS & DISCLAIMER

Portions of the authors' research were conducted at the Institute for Defense Analyses (IDA) – utilizing Professional Development funding, under various task orders supporting the Director of Operational Test and Evaluation (DOT&E) within the Office of the Secretary of Defense (OSD), and under a task order supporting the Executive Agent for the DoD Counterdrug Technology Developmental Program, under the sponsorship of the Office of the Undersecretary of Defense/Policy – Counternarcotics (OUSD(P)) in OSD.

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

REFERENCES

1. Ellenberg, J. H. (2000), "Statistician's Significance," *Journal of the American Statistical Association*, 95, 1-8.

Wilks Award Banquet Address: ***Overview of Stockpile Stewardship***

Jas. Mercer-Smith, Technical Staff Member and former Deputy Associate Laboratory Director for Nuclear Weapons, Los Alamos National Laboratory

Abstract unavailable

General Session II

DEVELOPMENT OF SOME LOCALLY MOST POWERFUL RANK TESTS FOR CORRELATION

W.J. Conover

Texas Tech University
Lubbock, Texas 79409

ABSTRACT

The Hájek-Sidák (1967, p.71) theorem for locally most powerful rank tests (LMPRT) is extended in this paper to the bivariate case. This enables the locally most powerful rank test for correlation to be developed for continuous random variables under some fairly mild restrictions. Four examples are given to illustrate the ease and practicality of the procedure. The first two examples deal with the bivariate exponential models of Mardia (1970) and Gumbel (1970). The third example uses the bivariate normal distribution, and the fourth example derives the LMPRT for a general correlation model of Morgenstern (1956).

1. INTRODUCTION

There are two primary difficulties in developing tests with good power properties for testing the null hypothesis of independence between two variables X and Y, based on a bivariate random sample (X_i, Y_i), i = 1, 2, ..., n, against the alternative hypothesis of correlation. One difficulty is in finding a suitable model for the bivariate distribution, and the other is in developing a powerful test for correlation once the model is selected. Some of the results in this paper were previously published in a paper by Shiratata (1974) and by Conover (2001).

The bivariate normal distribution is a convenient model to use for many reasons. The parameter rho is the linear correlation coefficient, so correlation is convenient to address in this model.

The most powerful test for correlation is well known, and the locally most powerful rank test (LMPRT) uses Fisher-Yates expected normal scores.

But some types of data do not fit the bivariate normal distribution very well. Therefore other classes of bivariate distributions have been developed in an attempt to find something other than the bivariate normal distribution to fit the data, while retaining some of the nice analytical properties found in the bivariate normal distribution. In this paper the general bivariate density function $h(x,y;\Theta)$ is considered, with some fairly general restrictions.

One family of bivariate distributions was proposed by Morgenstern (1956). Let $F(x)$ and $G(y)$ be the marginal distribution functions. A bivariate distribution function with those marginals is given by

$$H(x,y;\Theta) = F(x)G(y)[1 + \Theta\{1 - F(x)\}\{1 - G(y)\}] \qquad (1.1)$$

where Θ is the parameter that governs the degree of dependence between the random variables. Farlie (1961) found Spearman's rho to be the optimal correlation coefficient for Morgenstern's model (1.1), and studied the efficiency of less than optimal coefficients. Our derivation of the same result is much simpler and with fewer restrictions on the model.

Morgenstern's model (1.1) was generalized by Farlie (1960) to

$$H(x,y;\Theta) = F(x)G(y)[1 + \Theta A(x)B(y)] \qquad (1.2)$$

where $A(x)$ and $B(y)$ are bounded functions such that $A(\infty) = B(\infty) = 0$. The model (1.2), and thus (1.1) also, is a special case of the general model studied in this paper.

Konijn (1956) studied correlation tests for the hypothesis $\Theta_2 = \Theta_3 = 0$ in the model

$$X = \Theta_1 W + \Theta_2 Z \qquad Y = \Theta_3 W + \Theta_4 Z \qquad (1.3)$$

where W and Z are independent random variables. Correlation tests for a similar class of alternatives

$$X = (1 - \Theta)U + \Theta Z \qquad Y = (1 - \Theta)V + \Theta Z \qquad (1.4)$$

where U, V and Z are independent random variables, and $\Theta \to 0$, were investigated by Bhuchongkul (1964). Hájek and Sidák (1967, p.75) discuss the nearly identical model

$$X = U + \Theta Z \qquad Y = V + \Theta Z \qquad (1.5)$$

These models are more restrictive in their application than the more general model considered in this paper.

In this paper the general alternative distribution $h(x,y;\Theta)$ is investigated. A theorem is presented that enables the locally most powerful rank test to be derived under some fairly general conditions. Four examples are given to illustrate the usefulness of this result. While the development stops short of finding the efficiencies of the obtained tests, in some cases the tests are well known, and their efficiencies have already been studied.

2. THE LOCALLY MOST POWERFUL RANK TEST FOR CORRELATION

Let (X,Y) have the joint density function $h(x,y;\Theta)$ under H_a and the density $h(x,y;\Theta_0) = f(x)g(y)$ under H_0, the independence hypothesis, where f(x) and g(y) are the marginal density functions of X and Y respectively. In order to derive the locally most powerful rank test for H_0 against H_a, a bivariate rank version of the Neyman-Pearson lemma will be developed, followed by a useful theorem that will enable us to derive such a test.

Neyman-Pearson lemma for bivariate rank tests

Let $X_1, \ldots, X_n$ be i.i.d., with density f(x) and ranks $\mathbf{R} = (R_1, \ldots, R_n)$. Similarly, let $Y_1, \ldots, Y_n$ be i.i.d. with density g(y) and ranks $\mathbf{Q} = (Q_1, \ldots, Q_n)$. The most powerful size α rank test for

$$H_0\text{: the joint density of the X's and Y's is } \prod_{i=1}^{n} f(x_i)g(y_i) \qquad (2.1)$$

against some simple alternative H_a is given by the critical region defined by the index function

$$\begin{aligned} \Phi(\mathbf{r},\mathbf{q}) &= 1 \text{ if } P(\mathbf{R} = \mathbf{r},\ \mathbf{Q} = \mathbf{q} \mid H_a) > k \\ &= a \text{ if } P(\mathbf{R} = \mathbf{r},\ \mathbf{Q} = \mathbf{q} \mid H_a) = k \\ &= 0 \text{ if } P(\mathbf{R} = \mathbf{r},\ \mathbf{Q} = \mathbf{q} \mid H_a) < k \end{aligned} \qquad (2.2)$$

where k and a are chosen so $E\{\Phi(\mathbf{R},\mathbf{Q})\} = \alpha$ under H_0.

Proof: The proof follows from the fact that $P(\mathbf{R} = \mathbf{r},\ \mathbf{Q} = \mathbf{q})$ is equal for all points $(\mathbf{r},\mathbf{q})$ under H_0, where $\mathbf{r}$ and $\mathbf{q}$ represent permutations of the ranks 1, ..., n. Then the critical region with the most power is the region that consists of those points with the greatest probability when H_a is true. Randomization with probability a for points with boundary probabilities k is used only to achieve a significance level exactly equal to α.

Now we are ready to develop a theorem for finding a locally most powerful rank correlation test. Consider independent copies (X_m,Y_m), $m = 1, \ldots, n$ of (X,Y), with ranks R_m for X_m, and Q_m for Y_m as before. Let the density of (X,Y) be $h(x,y;\Theta)$ under H_a, and $h(x,y;\Theta_0) = f(x)g(y)$ under H_0: $\Theta = \Theta_0$, where $f(x)$ and $g(y)$ are the marginal densities. Also define the scores

$$a(i,j;h) = E\left[\frac{\delta \ \{h(X_n^{(i)},Y_n^{(j)};\Theta)\}/\delta\Theta \ |_{\Theta=\Theta0}}{h(X_n^{(i)},Y_n^{(j)};\Theta_0)} \ \Big| H_0 \right] \qquad (2.3)$$

where $X_n^{(i)}$ and $Y_n^{(j)}$ are the i th and j th order statistics in a random sample of size n from $f(x)$ and $g(y)$ respectively.

Theorem for locally most powerful rank correlation tests

Let J be some open interval around Θ_0. If

1. $h(x,y;\Theta)/h(x,y;\Theta_0)$ exists for $\Theta \epsilon J$,
2. $\delta \ \{h(x,y;\Theta)\}/\delta\Theta \ |_{\Theta=\Theta0}$ exists for $\Theta \epsilon J$,
3. $h(x,y;\Theta_0) = \lim_{\Theta\to\Theta0} h(x,y;\Theta)$ exists for $\Theta \epsilon J$, and
4. $\lim_{\theta\to\theta0} \iint \mid \delta \ \{h(x,y;\Theta)\}/\delta\Theta \mid dx\ dy$

$$= \iint \left| \delta \ \{h(x,y;\Theta)\}/\delta\Theta \ |_{\Theta=\Theta0} \right| dx\ dy < \infty,$$

then the locally (δ -> 0) most powerful rank test of H_0: $\Theta = \Theta_0$ against H_a: $\Theta = \Theta_0 + \delta$ is given by the test with the critical region

$$\sum_{m=1}^{n} a(R_m,\ Q_m;\ h) > k \tag{2.4}$$

where k is chosen so the test will have an appropriate size α.

Outline of the Proof: The proof resembles the proof on p. 71 of Hájek and Sidák (1967), except the integral is a 2n-fold integral over the region defined by both **R** and **Q**, and the Neyman-Pearson lemma for bivariate rank tests is invoked where appropriate.

Comment 1. This theorem and its preceding lemma are easily generalized to the p-variate setting, p > 2.

Comment 2. A slight variation of this theorem and proof allows us to consider the regression alternative, where the density of (X_m, Y_m) is $h(x,y{:}c_m\Theta)$, and results in the critical region

$$\sum_{m=1}^{n} c_m\, a(R_m,\ Q_m;\ h) > k \tag{2.5}$$

which is more in the spirit of the theorem on p.71 of Hájek and Sidák (1967).

Comment 3. This theorem is more general than the bivariate version given on p.75 of Hájek and Sidák (1967), which derived the LMPRT only for the model given by (1.5).

Comment 4. If X and Y are not continuous random variables, then the LMPRT is derived in the same manner described above, but with the joint density $h(x,y;\Theta)$ replaced by the bivariate (or multivariate) Radon-Nikodym derivative of $H(x,y;\Theta)$ with respect to $H(x,y;\Theta_0)$, in the same manner as in Section 6 of Conover (1973) in the univariate case.

Comment 5. The statistic defined by (2.4) is asymptotically normal under some general conditions on the scores. Asymptotic results are discussed in Section 4.

Implementation of the previous theorem to find the scores $a(i,j;h)$ associated with the locally most powerful rank test for correlation involves the following steps.

1. Find the partial derivative of $h(x,y;\Theta)$ with respect to Θ and set Θ equal to Θ_0.
2. Divide the result in Step 1 by $h(x,y;\Theta_0) = f(x)g(y)$.
3. Substitute $X_n^{(i)}$ for x and $Y_n^{(j)}$ for y in the quotient in Step 2, where $X_n^{(i)}$ and $Y_n^{(j)}$ are the i th and j th order statistics in random samples of size n from f(x) and g(y) respectively.
4. Find the expected value of the random variable in Step 3 under H_0. That is, integrate the product of
   - (a) the result of Step 2,
   - (b) the density function of the ith order statistic from f(x),
   - (c) and the density function of the jth order statistic from g(y),

over the entire range of values of X and Y.

3. FOUR EXAMPLES OF LOCALLY MOST POWERFUL RANK TESTS

These four examples show the ease with which the theorem of Section 2 can be applied to obtain locally most powerful rank tests for correlation. In all four examples the resulting test statistic is known, and the literature citations can be consulted to find tables for small sample sizes, and asymptotic approximations for large sample sizes.

The first two examples involve bivariate distributions, where both marginal distributions are exponential. The model in the first example allows only nonnegative correlations, and may be used when the alternative hypothesis is one of positive correlation. The model in the second example allows only nonpositive correlations, and may be used when the alternative hypothesis is one of negative correlation. In both examples, the locally most powerful rank test uses the top-down correlation coefficient of Iman and Conover (1987). The third example involves the bivariate normal distribution, and the fourth example looks at a very general bivariate distribution.

Example 1. Mardia (1970) presents a bivariate exponential distribution

$$h_1(x,y;\Theta) = \frac{1}{1-\Theta} e^{-\frac{x+y}{1-\Theta}} \sum_{r=0}^{\infty} \left[\frac{\Theta xy}{(1-\Theta)^2}\right]^r / r!r! \;;\; x>0,\ y>0,\ 0<\Theta<1 \qquad (3.1)$$

which has exponential marginal densities exp(-x) and exp(-y), and which degenerates to the product of those marginal densities exp{-x-y} for $\Theta = 0$, representing the case of independence. The correlation coefficient between X and Y is Θ.

This model first appeared in Mardia (1962) as a special case of a bivariate gamma distribution that appeared in Kibble (1941). It has been attributed to various authors, such as to Downton (1970) by Hawkes (1972) and others, and to Nagao and Kadoya (1971) by Cordova and Rodreguez-Iturbe (1985), Johnson and Kotz (1972), and others. It is widely used as a model for the bivariate exponential distribution. A parametric test of the null hypothesis of independence is apparently unknown. The locally most powerful rank test is derived in the following.

It is easy to show that

$$a(i,j;h_1) = E\{(X_n^{(i)}-1)(Y_n^{(j)}-1) \mid H_0\} = (s_n(i)-1)(s_n(j)-1) \qquad (3.2)$$

where $s_n(i)$ and $s_n(j)$ are the expected values of order statistics from the exponential distribution. That is, step 1 in the previous section involves finding the derivative of $h_1(x,y;\Theta)$ with respect to Θ, and setting $\Theta = 0$. This gives

$$\frac{\delta}{\delta\Theta} h_1(x,y;\Theta)\Big|_{\Theta=0} = e^{-x-y}(x-1)(y-1) \qquad (3.3)$$

The second step is to divide by $f(x)g(y) = e^{-x-y}$, which gives

$$(x-1)(y-1)$$

In the third step the ith and jth order statistics from the exponential distributions f(x) and g(y) replace x and y respectively. Thus the expected values, in step 4, give the LMPRT scores in (3.2).

The scores in (3.2) are given by the formula

$$s_n(i) = \sum_{j=0}^{i-1} \frac{1}{n-j} \tag{3.4}$$

and are sometimes called Savage scores because they were introduced by Savage (1956). Their use in a rank correlation coefficient

$$r_T = \frac{\Sigma\ s_n(R_m)s_n(Q_m) - (\Sigma s_n(i))^2/n}{\Sigma\ [s_n(i)]^2 - (\Sigma s_n(i))^2/n} = \frac{\Sigma\ s_n(R_m)s_n(Q_m) - n}{n - s_n(n)} \tag{3.5}$$

was studied by Iman and Conover (1987), and called the top-down correlation coefficient r_T because of its tendency to emphasize the tail values. Exact tables for r_T for $n \le 14$ are given by Iman (1987). Therefore the locally most powerful rank test of H_0: $\Theta = 0$ against H_a: $\Theta > 0$ in the bivariate exponential distribution given by (3.1) rejects H_0 if and only if $r_T > k$ for a suitably chosen value of k.

Example 2. Gumbel (1960) introduced another bivariate exponential distribution

$$h_2(x,y;\Theta) = \{(1+\Theta x)(1+\Theta y)-\Theta\}\ e^{-x-y-\Theta xy} \quad x>0,\ y>0,\ 0\le\Theta\le 1 \tag{3.6}$$

with non-positive correlation coefficient

$$-1 + \int_0^\infty \frac{e^{-y}}{1 + \Theta y}\ dy \tag{3.7}$$

Note that the correlation coefficient is zero when $\Theta = 0$, and it decreases monotonically as Θ increases. Therefore the LMPRT for correlation is also the LMPRT for Θ. This distribution degenerates to $\exp\{-x-y\}$ under H_0: $\Theta = 0$. This widely known model was studied further by Gumbel (1961) and has been used more recently by Wei (1981) and Barnett (1983). As with the previous model, a parametric test of the null hypothesis of independence is apparently unknown. The locally most powerful rank test is derived in the following.

The optimal scores are again found to be functions of the

Savage scores. Specifically the scores are

$$a(i,j;h_2) = E\{-(X_n^{(i)}-1)(Y_n^{(j)}-1) \mid H_0\} = -(s_n(i)-1)(s_n(j)-1) \quad (3.8)$$

which leads to the locally most powerful rank test that rejects H_0 when $r_T < k$ for some suitably chosen negative number k. Note that the negative value for k is due to the model, which allows only negative correlation in the restricted parameter range for Θ.

Example 3. The all-important bivariate normal distribution has density

$$h_3(x,y;\Theta) = (2\pi(1-\Theta^2))^{-1/2}\exp\{-(x^2+y^2-2\Theta xy)/2\} \quad (3.9)$$

and correlation coefficient Θ. The scores for the locally most powerful rank test are given by

$$a(i,j;h_3) = E(Z_n^{(i)})E(Z_n^{(j)}) \quad (3.10)$$

where $Z_n^{(i)}$ and $Z_n^{(j)}$ are order statistics from the standard normal distribution. These scores are used in the well-known normal scores statistic first given by Fisher and Yates (1957). This derivation of the locally most powerful rank test for the bivariate normal distribution is much simpler than the previous ones, and uses a more general model than the rather restrictive models (1.3), (1.4) and (1.5).

Example 4. The class of bivariate distributions introduced by Morgenstern (1956) has the bivariate distribution function

$$H(x,y;\Theta) = F(x)G(y)[1 + \Theta\{1 - F(x)\}\{1 - G(y)\}] \quad (3.11)$$

for any marginal distribution functions F(x) and G(y). This model has been extended by Plackett (1965) and often appears in discussions of bivariate distributions (see for example Mardia, 1970, or Johnson and Kotz, 1972). Due to the unspecified nature of F(x) and G(y) no parametric test is possible. However, rank tests are possible. In fact the locally most powerful rank test is easily derived, as shown in the following.

When H(x,y) is continuous then the density function is

$$h_4(x,y;\Theta) = f(x)g(y)[1 + \Theta\{1 - 2F(x)\}\{1 - 2G(y)\}] \tag{3.12}$$

which reduces to the independence case f(x)g(y) when $\Theta = 0$. This example shows the full power of the method introduced in this paper for finding the locally most powerful rank test for independence. The scores $a(i,j;h_4)$ in this case reduce to

$$\begin{aligned} a(i,j;h_4) &= E\{(2F(X_n^{(i)}) - 1)(2G(Y_n^{(j)}) - 1)\} \\ &= (2E\{U_n^{(i)}\} - 1)(2E\{U_n^{(j)}\} - 1) \end{aligned} \tag{3.13}$$

where $U_n^{(i)}$ and $U_n^{(j)}$ represent order statistics from the uniform distribution on (0,1). These are the scores used in the Spearman rank correlation coefficient, so Spearman's rho is the locally most powerful rank test for correlation for the entire class of Morgenstern distributions, assuming only that the bivariate distributions are continuous. This result was first obtained by Farlie (1961), but this method of proof is much simpler.

Note that $h_4(x,y;\Theta)$ is a density function with marginal densities f(x) and g(y) for all density functions f and g. In particular if f and g are exponential density functions, h_4 is another form of a bivariate exponential distribution. In this case the correlation coefficient is $\Theta/4$ (Gumbel, 1960) and it varies only within the narrow domain [-.25, .25]. Since the correlation coefficient is a monotonic function of Θ, the LMPRT for correlation in this bivariate exponential model uses Spearman's rho, instead of the top-down correlation coefficient of the previous two bivariate exponential models.

4. CONCLUDING REMARKS

Asymptotic normality for the special cases of the test statistic given in the previous section is already known. In general, asymptotic normality results from the following theorem.

Theorem showing asymptotic normality

1. Let $a(i,j;h) = E\{\phi(U_1,V_1)\,|\,R_1=i,\ Q_1=j\}$, where X_m and Y_m are independently distributed according to $F(x)$ and $G(y)$ respectively, $1 \le m \le n$, and where $U_m = F^{-1}(X_m)$ and $V_m = G^{-1}(Y_m)$.
2. Assume $0 < \iint [\phi(u,v) - \phi]^2 dudv < \infty$, where $\phi = \iint \phi(u,v)dudv$ and where integration is over the unit square.
3. Assume H_0 is true. Let $S = \Sigma\, a(R_i,Q_j;h)$, where R_i and Q_j are the ranks of X_i (hence U_i) and Y_j (hence V_j) respectively.

Then S is asymptotically (as n gets large) normal with mean

$$\mu = \Sigma_i\Sigma_j a(i,j;h)/n$$

and variance given by either

$$\sigma^2 = (n - 1) \iint [\phi(u,v) - \phi]^2 dudv$$

or

$$\sigma^2 = (n-1)^{-1}\Sigma_i\Sigma_j[a(i,j;h) - a(\cdot,j;h) - a(i,\cdot;h) + a(\cdot,\cdot;h)]^2$$ where the dot notation refers to averages over the missing arguments.

Proof. Introduce $T = \Sigma\, \phi(U_i,V_i)$ and let $\mathbf{U}^{(\cdot)}$ and $\mathbf{V}^{(\cdot)}$ be the vectors of order statistics for U and V respectively. Then

$$E\{(S - T)^2|\ \mathbf{U}^{(\cdot)} = \mathbf{u}^{(\cdot)} \text{ and } \mathbf{V}^{(\cdot)} = \mathbf{v}^{(\cdot)}\}$$
$$= E\{\Sigma[a(R_i,Q_i;h) - \phi(u^{(Ri)},v^{(Qi)})]\}^2$$

Let $b(i,j) = a(i,j;h) - \phi(u^{(i)},v^{(j)})$. Then by Theorem a on p.57 of Hájek and Sidák (1967) the above expression is equal to

$$(n-1)^{-1}\Sigma_i\Sigma_j[b(i,j) - b(\cdot,j) - b(i,\cdot) + b(\cdot,\cdot)]^2$$
$$\le (n-1)^{-1}\Sigma_i\Sigma_j[b(i,j) - b(\cdot,\cdot)]^2 = n^2(n-1)^{-1}\mathrm{Var}\{b(R_1,Q_1)\}$$
$$\le n^2(n-1)^{-1}E\{a(R_1,Q_1;h) - \phi(u^{(R1)},v^{(Q1)})\}^2.$$

Therefore, unconditionally,

$$E\{(S - T)^2\} \le n^2(n-1)^{-1}E\{a(R_1,Q_1;h) - \phi(U_1,V_1)\}^2$$

and
$$E\{\frac{(S - T)^2}{\sigma^2}\} = \frac{n^2}{(n-1)^2}\ \frac{E\{a(R_1,Q_1;h) - \phi(U_1,V_1)\}^2}{\iint[\phi(u,v) - \phi]^2 dudv}$$

Because $E\{a(R_1,Q_1;h) - \phi(U_1,V_1)\}^2$ converges to zero for square integrable functions (see Theorem a on page 157 of Hájek and Sidák, 1967) and because $\sigma^2 > 0$, S and T are asymptotically identically distributed. However, T is asymptotically normal by the central limit theorem, which proves the theorem. The alternative form for σ^2 is found on p.57 of Hájek and Sidák (1967).

Hájek and Sidák (1967, p.221) were unable to derive the asymptotic distribution of correlation statistics under the alternative hypothesis suggested by the model (1.5). We, also, were unable to achieve those results under our more general model. This prevents computing asymptotic relative efficiencies for our model. However, under the model discussed by Farlie (1961), the efficiency of Spearman's rho when Fisher-Yates scores are optimal, or vice-versa, is $(3/\pi)^2 = .912$. Similarly it can be shown that the efficiency of Spearman's rho when the top-down correlation coefficient is optimal, or vice-versa, is $(3/4)^2 = .5625$.

BIBLIOGRAPHY

Barnett, V. (1983), "Reduced Distance Measures and Transformations in Processing Multivariate Outliers," ***Australian Journal of Statistics, 25(1),*** 64-75.

Bhuchongkul, S. (1964), "A Class of Nonparametric Tests for Independence in Bivariate Populations," ***The Annals of Mathematical Statistics, 35,*** 138-149.

Conover, W.J. (1973), "Rank Tests for One Sample, Two Samples, and k Samples Without the Assumption of a Continuous Distribution Function," ***The Annals of Statistics, 1(6),*** 1105-1125

Conover, W.J. (2001), "Some Locally Most Powerful Tests for Correlation," ***Journal of Modern Statistical Methods, 1.***

Cordova, J.R. and Rodriguez-Iturbe, I. (1985), "On the Probabilistic Structure of Storm Surface Runoff," ***Water Resources Research, 21,*** 755-763.

Downton, F. (1970), "Bivariate Exponential Distributions in Reliability Theory," ***Royal Statistical Society Journal, 32,*** 408-417.

Farlie, D.J.G. (1960), "The Performance of Some Correlation Coefficients for a General Bivariate Distribution," ***Biometrika, 47,*** 307-323.

Farlie, D.J.G. (1961), "The Asymptotic Efficiency of Daniels' Generalized Correlation Coeffients," ***Royal Statistical Society Journal, 23,*** 128-142.

Fisher, R.A., and Yates, F. (1957), ***Statistical Tables for Biological, Agricultural and Medical Research, (3rd Ed.),*** Darien, Conn: Hafner.

Hájek, J., and Sidák, Z. (1967), ***Theory of Rank Tests,*** New York: Academic Press.

Hawkes, A.G. (1972), "A Bivariate Exponential Distribution with Applications to Reliability," ***Royal Statistical Society Journal, 34,*** 129-133.

Iman, R.L. (1987), "Tables of the Exact Quantiles of the Top-Down Correlation Coefficient for n = 3(1)14," ***Communications in Statistics, Theory and Methods, 16(5),*** 1513-1540.

Iman, R.L., and Conover, W.J. (1987), "A Measure of Top-Down Correlation," ***Technometrics, 29(3),*** 351-357.

Johnson, N.L., and Kotz, S. (1972), ***Distributions in Statistics: Continuous Multivariate Distributions,*** New York: John Wiley.

Kibble, W.F. (1941), "A Two-Variate Gamma Type Distribution," ***Sankhya, 5,*** 137-150.

Konijn, H.S. (1956), "On the Power of Certain Tests for Independence in Bivariate Populations," ***The Annals of Mathematical Statistics, 27,*** 300-323.

Mardia, K.V. (1962), "Multivariate Pareto Distributions," ***The Annals of Mathematical Statistics, 33,*** 1008-1015.

Mardia, K.V. (1970), ***Families of Bivariate Distributions,*** Darien, Conn: Hafner.

Morgenstern, D. (1956), "Einfache Beispiele Zweidimensionaler Verteilungen," ***Mitteilungsblatt für Mathematische Statistik, 8,*** 234-235.

Nagao, M., and Kadoya, M. (1971), "Two-Variate Exponential Distribution and its Numerical Table for Engineering Applications," ***Bull. Disaster Prev. Res. Inst., Kyoto Univ., 20,*** 183-215.

Plackett, R.L. (1965), "A Class of Bivariate Distributions," ***Journal of the American Statistical Association, 60,*** 516-522.

Savage, I.R. (1956), "Contributions to the Theory of Rank Order Statistics - the Two-sample Case," ***The Annals of Mathematical Statistics, 27,*** 590-615.

Shiratata, S. (1974), "Locally Most Powerful Rank Tests for Independence." ***Bulletin of Mathematical Statistics, 16,*** 11-21.

Wei, L.J. (1981), "Estimation of Location Difference for Fragmentary Samples," ***Biometrika, 68(2),*** 471-476.

The exponential distribution is widely used for waiting times, and for times to failure.

The bivariate exponential distribution may be an appropriate model for some of the following cases:

1. X = the time interval between the arrival of a customer, and the arrival of the previous customer
   Y = the time it takes for the newly-arrived customer to get served

2. X = the time in service of an item (e.g., light bulb) until it fails and requires replacement
   Y = the time it takes to replace the item

3. X = the length of service of the first bulb in a two-bulb overhead projector
   Y = the length of service of the spare bulb in a two-bulb overhead projector

4. X = the time a telephone is available (not in use) until it rings (assuming no call waiting)
   Y = the length of the telephone call until the phone is again available

If the correlation (r) between X and Y is positive, then Mardia's (1970) model may be appropriate. See Example 1. In that case the top-down correlation is the locally most powerful rank test.

If the correlation (r) between X and Y is negative, then Gumbel's (1960) model may be appropriate. See Example 2. In that case the top-down correlation is again the locally most powerful rank test.

The Asymptotic Relative Efficiency of Spearman's rho, when the top-down correlation is the locally most powerful rank test, is 0.5625.

Exact quantiles for the top-down correlation coefficient (from Iman and Conover, 1985)

n	0.001	0.005	0.01	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999
4					-0.7536	-0.7101	-0.5362	-0.3188	-0.2319	-0.0580	0.0870	0.4203	0.7246	0.8696	0.9420				
5			-0.7919	-0.7612	-0.6999	-0.5976	-0.4852	-0.3062	-0.1989	-0.0915	0.0772	0.2536	0.5731	0.7520	0.9054	0.9591	0.9770		
6		-0.7953	-0.7429	-0.6936	-0.6372	-0.5488	-0.4201	-0.3004	-0.1736	-0.0700	0.0473	0.2063	0.4894	0.6757	0.8095	0.8866	0.9430	0.9687	
7	-0.7751	-0.7335	-0.7099	-0.6503	-0.5879	-0.5000	-0.3711	-0.2693	-0.1555	-0.0480	0.0541	0.1836	0.4013	0.6216	0.7382	0.8361	0.9062	0.9354	0.9771
8	-0.7441	-0.6968	-0.6642	-0.6059	-0.5470	-0.4629	-0.3419	-0.2429	-0.1436	-0.0448	0.0556	0.1703	0.3480	0.5751	0.6917	0.7792	0.8651	0.9037	0.9596
9	-0.7138	-0.6616	-0.6296	-0.5716	-0.5124	-0.4329	-0.3187	-0.2243	-0.1337	-0.0417	0.0546	0.1620	0.3152	0.5301	0.6538	0.7415	0.8262	0.8712	0.9359
10	-0.6874	-0.6326	-0.5995	-0.5425	-0.4847	-0.4078	-0.2988	-0.2094	-0.1245	-0.0383	0.0523	0.1543	0.2909	0.4921	0.6201	0.7066	0.7925	0.8404	0.9129
11	-0.6630	-0.6070	-0.5739	-0.5174	-0.4611	-0.3868	-0.2823	-0.1971	-0.1168	-0.0356	0.0501	0.1471	0.2732	0.4609	0.5887	0.6767	0.7624	0.8122	0.8899
12	-0.6412	-0.5845	-0.5514	-0.4957	-0.4407	-0.3687	-0.2683	-0.1867	-0.1102	-0.0332	0.0483	0.1406	0.2588	0.4347	0.5604	0.6497	0.7354	0.7861	0.8679
13	-0.6216	-0.5645	-0.5316	-0.4767	-0.4230	-0.3530	-0.2561	-0.1777	-0.1046	-0.0311	0.0467	0.1348	0.2467	0.4124	0.5353	0.6250	0.7111	0.7621	0.8470
14	-0.6039	-0.5465	-0.5139	-0.4598	-0.4073	-0.3392	-0.2455	-0.1699	-0.0996	-0.0292	0.0453	0.1298	0.2362	0.3932	0.5129	0.6023	0.6890	0.7400	0.8273

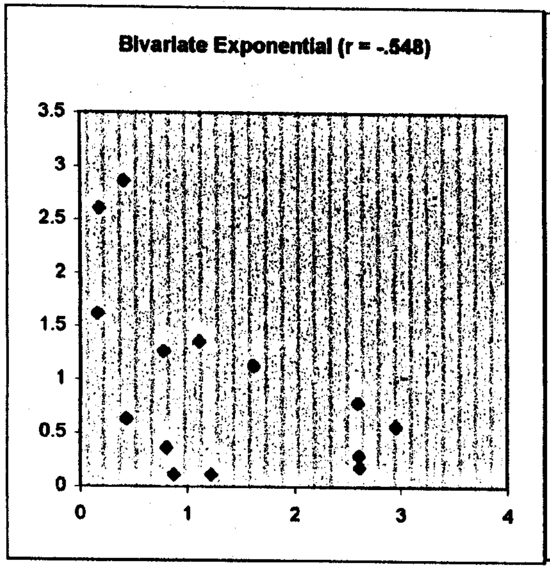
For an approximation for n > 14 use either the standard normal quantile divided by SQRT(n-1), or more exact approximate tables in Iman and Conover (1987).

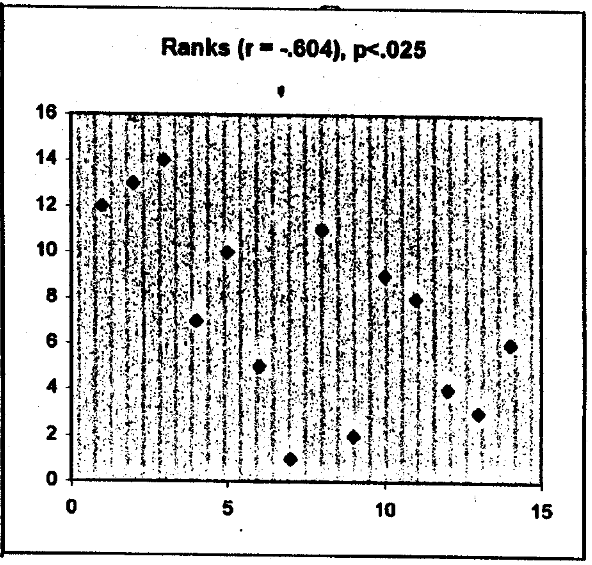
AN EXAMPLE OF TOP-DOWN CORRELATION WITH n = 14.

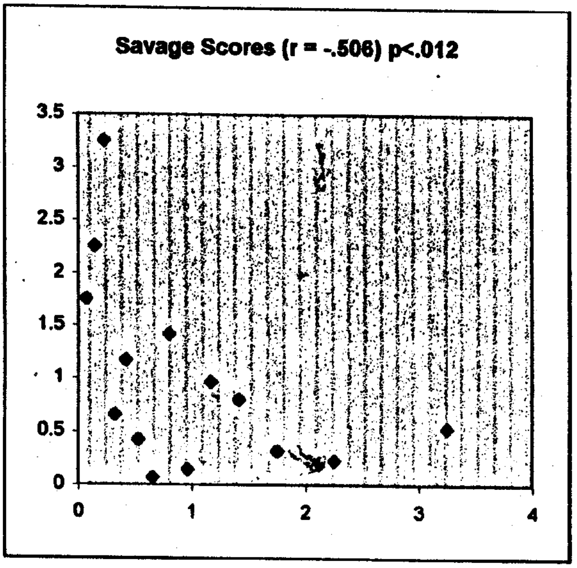
Uniform on (0,1)		=-LN(1-A3)		=RANK(A3,A$3:A$16,1)		Savage Scores	
0.355602	0.466018	0.439438	0.627392	4	7	0.322594	0.658705
0.926145	0.256691	2.605654	0.296644	12	4	1.751562	0.322594
0.925596	0.544969	2.598244	0.78739	11	8	1.418229	0.801562
0.802149	0.675954	1.620239	1.126871	10	9	1.168229	0.968229
0.706076	0.111026	1.224435	0.117688	9	2	0.968229	0.148352
0.926939	0.168096	2.616456	0.184038	13	3	2.251562	0.231685
0.584918	0.110508	0.879279	0.117104	7	1	0.658705	0.071429
0.948088	0.431745	2.958206	0.565186	14	6	3.251562	0.533705
0.163274	0.926084	0.178259	2.604828	2	13	0.148352	2.251562
0.151585	0.80166	0.164386	1.617774	1	12	0.071429	1.751562
0.553606	0.306223	0.806553	0.365604	6	5	0.533705	0.422594
0.541856	0.718528	0.780572	1.267722	5	10	0.422594	1.168229
0.671194	0.742393	1.112286	1.356321	8	11	0.801562	1.418229
0.337565	0.942595	0.411833	2.857618	3	14	0.231685	3.251562
Pearson's Correlation	-0.57516	Pearson's Correlation	-0.54818	Spearman's Correlation	-0.6044	Top-down Correlation	-0.50597

HOW TO FIND SAVAGE SCORES

I	=1/(n+1-I)	=SUM(N$3:N3)
1	0.07143	0.071429
2	0.07692	0.148352
3	0.08333	0.231685
4	0.09091	0.322594
5	0.1	0.422594
6	0.11111	0.533705
7	0.125	0.658705
8	0.14286	0.801562
9	0.16667	0.968229
10	0.2	1.168229
11	0.25	1.418229
12	0.33333	1.751562
13	0.5	2.251562
14	1	3.251562



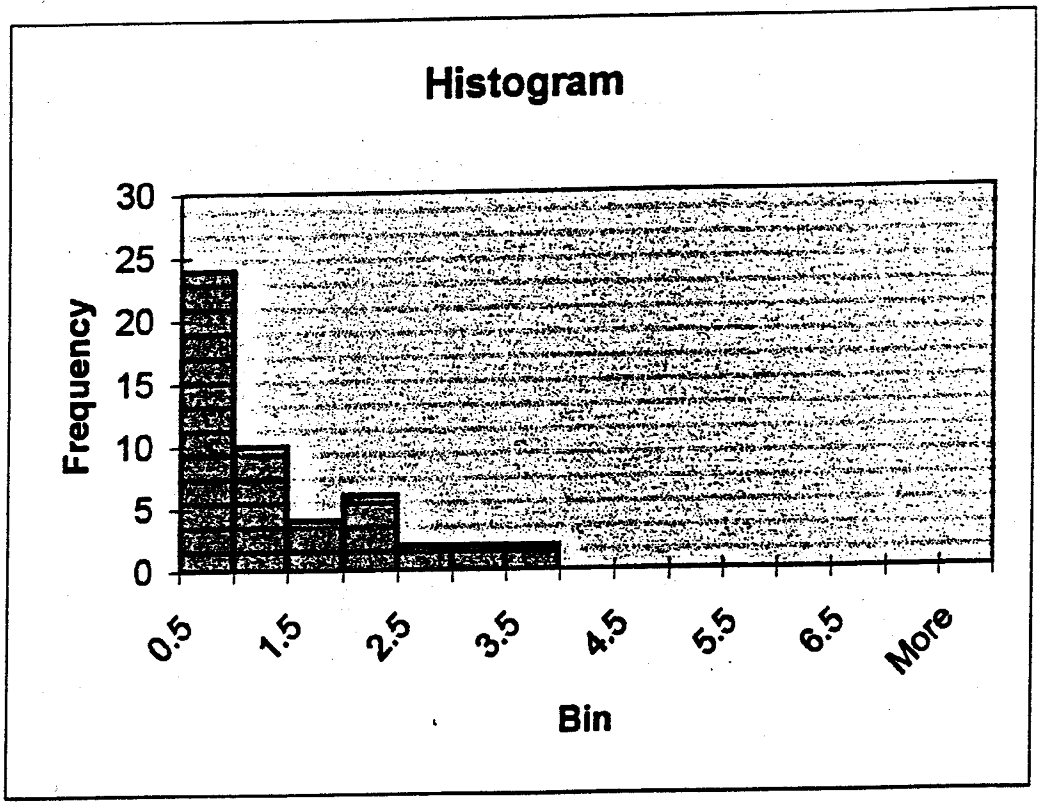


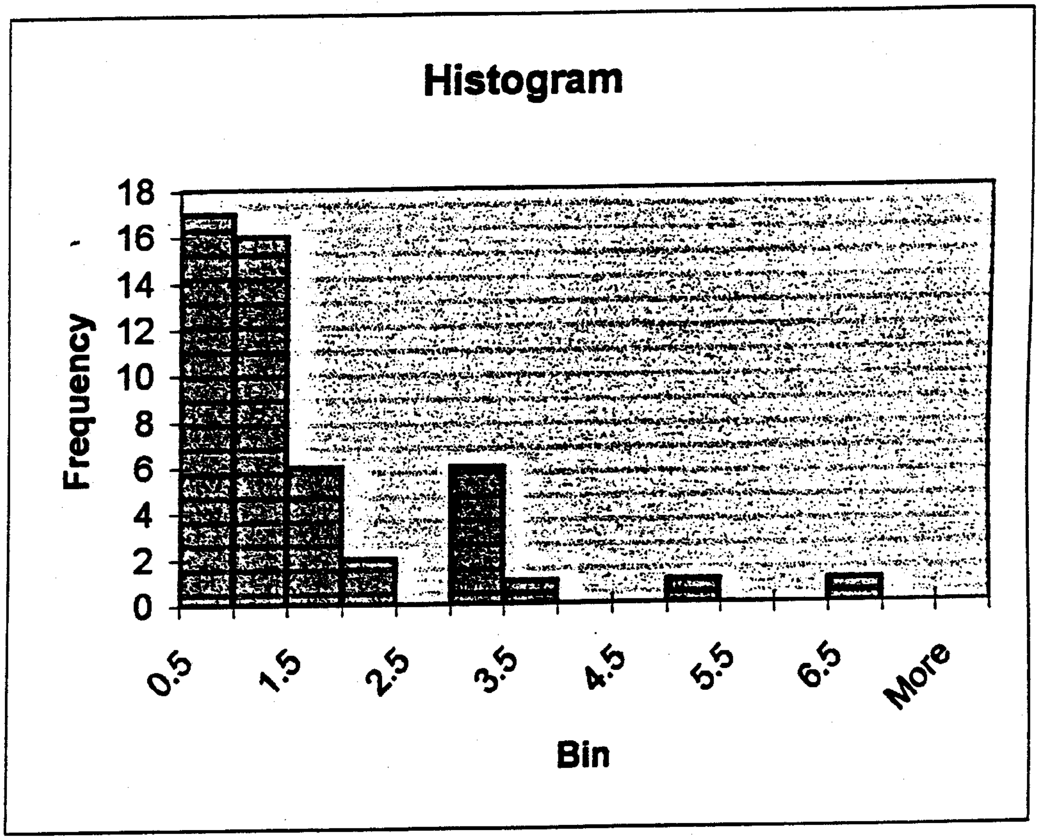


Data Set 1: Positive Correlation

Uniform Distribution		Exponential Distribution		Ranks		scores	scores
0.382	0.355602	0.481267	0.439438	23	16	0.607749	0.380995
0.951689	0.926145	3.0301	2.605654	49	44	3.499205	2.049205
0.756157	0.925596	1.411231	2.598244	38	43	1.395995	1.906348
0.67156	0.802149	1.113401	1.620239	36	40	1.247643	1.570237
0.850429	0.706076	1.899983	1.224435	43	38	1.906348	1.395995
0.65685	0.926939	1.069587	2.616456	35	45	1.180976	2.215872
0.601886	0.584918	0.921017	0.879279	34	29	1.118476	0.853847
0.545518	0.948088	0.788598	2.958206	30	47	0.901466	2.665872
0.259865	0.163274	0.300923	0.178259	11	8	0.245662	0.172463
0.014222	0.151585	0.014324	0.164386	1	7	0.02	0.149207
0.265145	0.553606	0.308082	0.806553	14	27	0.324646	0.764914
0.943815	0.541856	2.879113	0.780572	48	25	2.999205	0.683247
0.350017	0.671194	0.430809	1.112286	19	34	0.47196	1.118476
0.414594	0.337565	0.53545	0.411833	26	15	0.723247	0.352424
0.202582	0.489242	0.226376	0.67186	9	22	0.196272	0.572034
0.125889	0.602466	0.134548	0.922475	6	31	0.126479	0.951466
0.539445	0.225105	0.775323	0.255027	29	12	0.853847	0.271303
0.332774	0.229682	0.404626	0.260952	17	13	0.410407	0.297619
0.930357	0.927061	2.66437	2.618129	47	46	2.665872	2.415872
0.894345	0.998169	2.247575	6.302833	46	50	2.415872	4.499205
0.218421	0.430342	0.246439	0.562718	10	20	0.220662	0.504218
0.526933	0.920164	0.748517	2.527775	28	42	0.808392	1.781348
0.87994	0.150853	2.119765	0.163523	45	6	2.215872	0.126479
0.856441	0.598682	1.941009	0.913	44	30	2.049205	0.901466
0.167089	0.685781	0.182828	1.157666	8	36	0.172463	1.247643
0.362743	0.813227	0.450582	1.67786	21	41	0.537552	1.670237
0.378277	0.689261	0.475261	1.1688	22	37	0.572034	1.319072
0.81106	0.052797	1.666325	0.054242	41	3	1.670237	0.061241
0.0983	0.992523	0.103474	4.895919	4	49	0.082518	3.499205
0.262673	0.215155	0.304724	0.24227	13	11	0.297619	0.245662
0.753624	0.436689	1.400897	0.573924	37	21	1.319072	0.537552
0.260292	0.190527	0.3015	0.211372	12	9	0.271303	0.196272
0.325571	0.387371	0.39389	0.489996	16	17	0.380995	0.410407
0.280221	0.549089	0.328811	0.796485	15	26	0.352424	0.723247
0.56563	0.399518	0.833859	0.510022	32	18	1.004097	0.44071
0.046297	0.050508	0.047403	0.051828	3	2	0.061241	0.040408
0.556139	0.001495	0.812243	0.001497	31	1	0.951466	0.02
0.798975	0.194739	1.604324	0.216588	39	10	1.479328	0.220662
0.960997	0.06711	3.244126	0.069468	50	4	4.499205	0.082518
0.112796	0.532334	0.119681	0.760002	5	24	0.104257	0.644786
0.808893	0.744102	1.654922	1.362978	40	39	1.570237	1.479328
0.357006	0.953429	0.441619	3.066772	20	48	0.504218	2.999205
0.154027	0.408094	0.167268	0.524407	7	19	0.149207	0.47196
0.39198	0.67391	0.497547	1.120581	24	35	0.644786	1.180976
0.457289	0.610858	0.611179	0.943812	27	32	0.764914	1.004097
0.596088	0.1377	0.906557	0.148151	33	5	1.059653	0.104257
0.031892	0.491134	0.032411	0.675571	2	23	0.040408	0.607749
0.816156	0.553758	1.69367	0.806895	42	28	1.781348	0.808392
0.397412	0.617573	0.506522	0.961216	25	33	0.683247	1.059653
0.335673	0.286447	0.408981	0.337498	18	14	0.44071	0.324646

	Column 1	*Column 2*		*Column 1*	*Column 2*		*Column 1*	*Column 2*		*Column 1*
Column 1	1		Column 1	1		Column 1	1		Column 1	1
Column 2	0.166783	1	Column 2	0.207295	1	Column 2	0.18213685	1	Column 2	0.175186


Histogram
Frequency
30
25
20
15
10
5
0
0.5
1.5
2.5
3.5
4.5
5.5
6.5
More
Bin


Histogram
Frequency
18
16
14
12
10
8
6
4
2
0
0.5
1.5
2.5
3.5
4.5
5.5
6.5
More
Bin

Data Set 2: Positive Correlation

Uniform		Bivariate Exponential		Ranks		Savage Scores	
0.466018	0.303903	0.627392	0.362267	24	11	0.644786	0.245662
0.256691	0.679647	0.296644	1.138332	16	31	0.380995	0.951466
0.544969	0.674978	0.78739	1.123862	28	30	0.808392	0.901466
0.675954	0.948515	1.126871	2.96647	32	48	1.004097	2.999205
0.111026	0.386334	0.117688	0.488304	9	15	0.196272	0.352424
0.168096	0.005158	0.184038	0.005171	10	1	0.220662	0.02
0.110508	0.558916	0.117104	0.81852	8	27	0.172463	0.764914
0.431745	0.967467	0.565186	3.425509	21	49	0.537552	3.499205
0.926084	0.381664	2.604828	0.480724	44	14	2.049205	0.324646
0.80166	0.695242	1.617774	1.188238	40	32	1.570237	1.004097
0.306223	0.227607	0.365604	0.258262	19	7	0.47196	0.149207
0.718528	0.76693	1.267722	1.456417	37	36	1.319072	1.247643
0.742393	0.795495	1.356321	1.587165	38	38	1.395995	1.395995
0.942595	0.15302	2.857618	0.166078	45	4	2.215872	0.082518
0.511582	0.245125	0.716583	0.281203	25	9	0.683247	0.196272
0.980438	0.806055	3.934148	1.64018	47	40	2.665872	1.570237
0.103946	0.796197	0.109755	1.590603	7	39	0.149207	1.479328
0.693106	0.884823	1.181252	2.161286	34	45	1.118476	2.215872
0.254158	0.758171	0.293242	1.419526	15	35	0.352424	1.180976
0.554918	0.832606	0.809497	1.787405	29	41	0.853847	1.670237
0.294473	0.460952	0.34881	0.61795	18	18	0.44071	0.44071
0.531144	0.027345	0.75746	0.027725	26	2	0.723247	0.040408
0.04944	0.054598	0.050704	0.056145	3	3	0.061241	0.061241
0.043733	0.222968	0.044718	0.252274	2	6	0.040408	0.126479
0.989959	0.64687	4.601119	1.04092	49	29	3.499205	0.853847
0.993774	0.50029	5.079057	0.693727	50	21	4.499205	0.537552
0.86578	0.747276	2.008272	1.375458	43	34	1.906348	1.118476
0.308939	0.946837	0.369527	2.934388	20	47	0.504218	2.665872
0.535844	0.486953	0.767535	0.667388	27	20	0.764914	0.504218
0.985809	0.78341	4.25514	1.52975	48	37	2.999205	1.319072
0.213904	0.582293	0.240677	0.872975	13	28	0.297619	0.808392
0.094119	0.157109	0.098847	0.170918	5	5	0.104257	0.104257
0.849574	0.51149	1.894286	0.716396	41	22	1.670237	0.572034
0.713584	0.318979	1.250309	0.384163	36	12	1.247643	0.271303
0.862239	0.87289	1.982238	2.062706	42	44	1.781348	2.049205
0.270363	0.40492	0.315209	0.519059	17	17	0.410407	0.410407
0.947844	0.39375	2.953514	0.500463	46	16	2.415872	0.380995
0.243294	0.539262	0.27878	0.774926	14	24	0.324646	0.644786
0.176305	0.557878	0.193955	0.81617	11	26	0.245662	0.723247
0.70745	0.479141	1.229118	0.652275	35	19	1.180976	0.47196
0.098849	0.321421	0.104083	0.387754	6	13	0.126479	0.297619
0.684805	0.518601	1.154563	0.731059	33	23	1.059653	0.607749
0.083865	0.88525	0.087591	2.165003	4	46	0.082518	2.415872
0.769921	0.871456	1.469332	2.051484	39	43	1.479328	1.906348
0.599841	0.992553	0.915894	4.900009	30	50	0.901466	4.499205
9.16E-05	0.250984	9.16E-05	0.288995	1	10	0.02	0.220662
0.664602	0.720847	1.092436	1.275996	31	33	0.951466	1.059653
0.438154	0.866176	0.576528	2.011232	23	42	0.607749	1.781348
0.179174	0.236579	0.197444	0.269946	12	8	0.271303	0.172463
0.436964	0.552263	0.574412	0.803549	22	25	0.572034	0.683247

	Column 1	*Column 2*		*Column 1*	*Column 2*		*Column 1*	*Column 2*		*Column 1*
Column 1	1		Column 1	1		Column 1	1		Column 1	1
Column 2	0.30382	1	Column 2	0.058951	1	Column 2	0.281633	1	Column 2	0.049993

Data Set 3: Negative Correlation

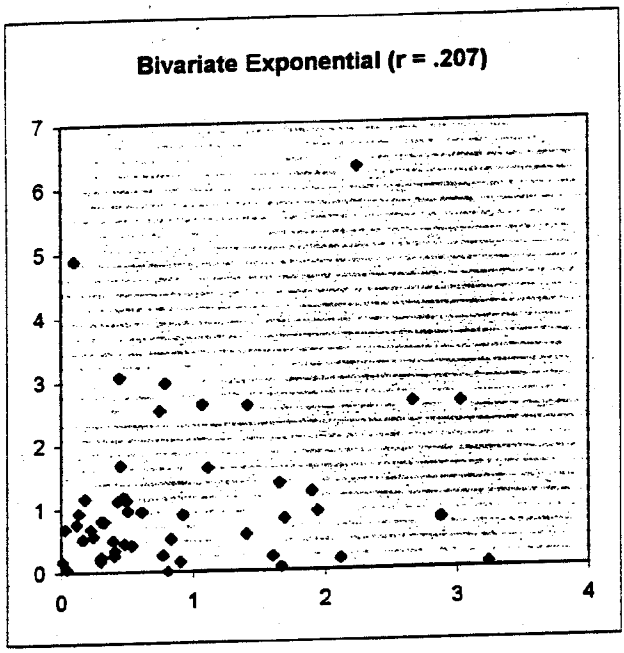
Uniform Distribution		Exponential Distribution		Ranks		scores	
0.382	0.466018	0.481267	0.627392474	23	24	0.607749	0.644786
0.951689	0.256691	3.0301	0.296643686	49	16	3.499205	0.380995
0.756157	0.544969	1.411231	0.787389783	38	28	1.395995	0.808392
0.67156	0.675954	1.113401	1.126871237	36	32	1.247643	1.004097
0.850429	0.111026	1.899983	0.11768767	43	9	1.906348	0.196272
0.65685	0.168096	1.069587	0.18403817	35	10	1.180976	0.220662
0.601886	0.110508	0.921017	0.117104229	34	8	1.118476	0.172463
0.545518	0.431745	0.788598	0.56518564	30	21	0.901466	0.537552
0.259865	0.926084	0.300923	2.604828266	11	44	0.245662	2.049205
0.014222	0.80166	0.014324	1.617773592	1	40	0.02	1.570237
0.265145	0.306223	0.308082	0.365604298	14	19	0.324646	0.47196
0.943815	0.718528	2.879113	1.267721547	48	37	2.999205	1.319072
0.350017	0.742393	0.430809	1.356321126	19	38	0.47196	1.395995
0.414594	0.942595	0.53545	2.857618361	26	45	0.723247	2.215872
0.202582	0.511582	0.226376	0.71658322	9	25	0.196272	0.683247
0.125889	0.980438	0.134548	3.934147733	6	47	0.126479	2.665872
0.539445	0.103946	0.775323	0.109754648	29	7	0.853847	0.149207
0.332774	0.693106	0.404626	1.18125244	17	34	0.410407	1.118476
0.930357	0.254158	2.66437	0.293241694	47	15	2.665872	0.352424
0.894345	0.554918	2.247575	0.809496874	46	29	2.415872	0.853847
0.218421	0.294473	0.246439	0.348810377	10	18	0.220662	0.44071
0.526933	0.531144	0.748517	0.75745989	28	26	0.808392	0.723247
0.87994	0.04944	2.119765	0.050703979	45	3	2.215872	0.061241
0.856441	0.043733	1.941009	0.044718141	44	2	2.049205	0.040408
0.167089	0.989959	0.182828	4.60111944	8	49	0.172463	3.499205
0.362743	0.993774	0.450582	5.079057197	21	50	0.537552	4.499205
0.378277	0.86578	0.475261	2.008272019	22	43	0.572034	1.906348
0.81106	0.308939	1.666325	0.369526995	41	20	1.670237	0.504218
0.0983	0.535844	0.103474	0.767534553	4	27	0.082518	0.764914
0.262673	0.985809	0.304724	4.255139785	13	48	0.297619	2.999205
0.753624	0.213904	1.400897	0.240676653	37	13	1.319072	0.297619
0.260292	0.094119	0.3015	0.09884742	12	5	0.271303	0.104257
0.325571	0.849574	0.39389	1.894285784	16	41	0.380995	1.670237
0.280221	0.713584	0.328811	1.250309241	15	36	0.352424	1.247643
0.56563	0.862239	0.833859	1.982238233	32	42	1.004097	1.781348
0.046297	0.270363	0.047403	0.31520878	3	17	0.061241	0.410407
0.556139	0.947844	0.812243	2.953513507	31	46	0.951466	2.415872
0.798975	0.243294	1.604324	0.278779891	39	14	1.479328	0.324646
0.960997	0.176305	3.244126	0.193955484	50	11	4.499205	0.245662
0.112796	0.70745	0.119681	1.229118211	5	35	0.104257	1.180976
0.808893	0.098849	1.654922	0.104082946	40	6	1.570237	0.126479
0.357006	0.684805	0.441619	1.154563258	20	33	0.504218	1.059653
0.154027	0.083865	0.167268	0.087591397	7	4	0.149207	0.082518
0.39198	0.769921	0.497547	1.469332364	24	39	0.644786	1.479328
0.457289	0.599841	0.611179	0.91589407	27	30	0.764914	0.901466
0.596088	9.16E-05	0.906557	9.15597E-05	33	1	1.059653	0.02
0.031892	0.664602	0.032411	1.092436143	2	31	0.040408	0.951466
0.816156	0.438154	1.69367	0.576527916	42	23	1.781348	0.607749
0.397412	0.179174	0.506522	0.197444335	25	12	0.683247	0.271303
0.335673	0.436964	0.408981	0.574411743	18	22	0.44071	0.572034

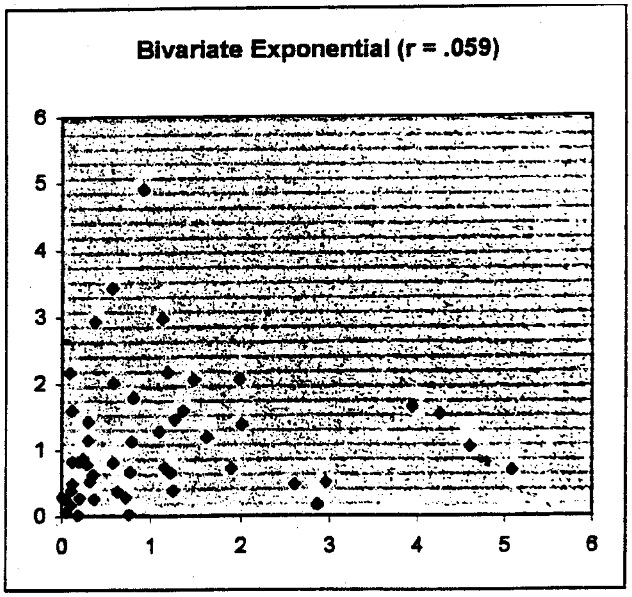
	Column 1	*Column 2*		*Column 1*	*Column 2*		*Column 1*	*Column 2*		*Column 1*
Column 1	1		Column 1	1		Column 1	1		Column 1	1
Column 2	-0.439328	1	Column 2	-0.36762093	1	Column 2	-0.40389	1	Column 2	-0.341062

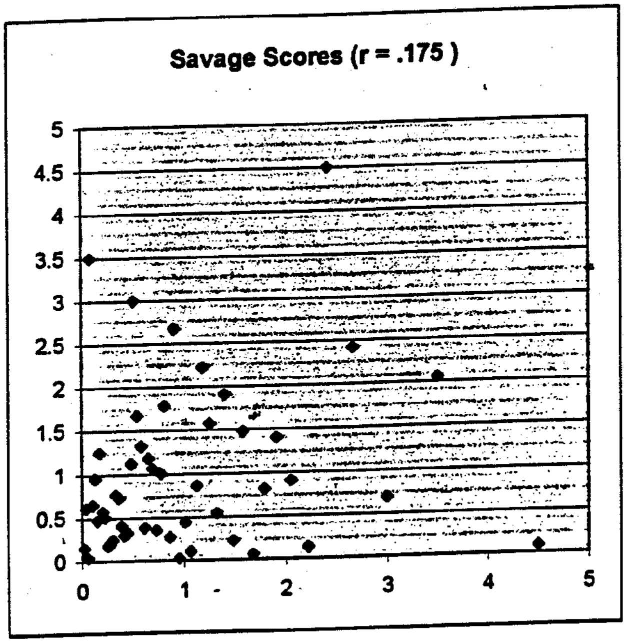
Data Set 4: Negative Correlation

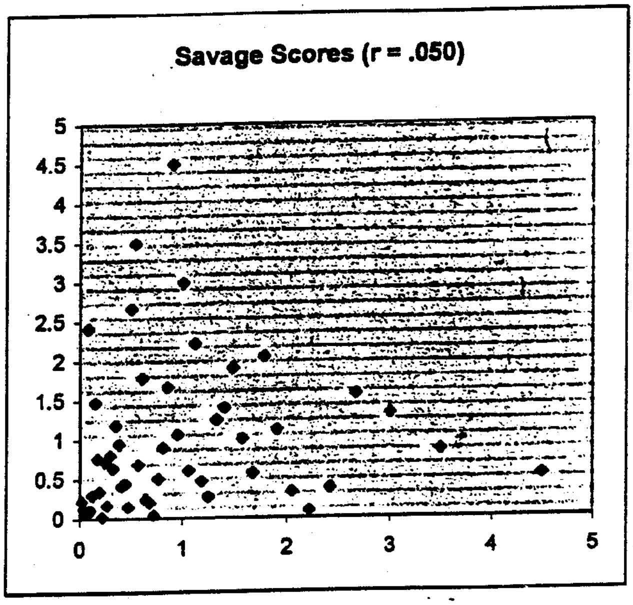
Uniform		Bivariate Exponential		Ranks		Savage Scores	
0.910306	0.303903	2.411353	0.362267	47	11	2.665872	0.245662
0.100314	0.679647	0.10571	1.138332	6	31	0.126479	0.951466
0.903867	0.674978	2.342019	1.123862	46	30	2.415872	0.901466
0.789026	0.948515	1.556018	2.96647	44	48	2.049205	2.999205
0.40144	0.386334	0.513229	0.488304	24	15	0.644786	0.352424
0.451765	0.005158	0.601052	0.005171	27	1	0.764914	0.02
0.497024	0.558916	0.687214	0.81852	29	27	0.853847	0.764914
0.250771	0.967467	0.28871	3.425509	13	49	0.297619	3.499205
0.808313	0.381664	1.651893	0.480724	45	14	2.215872	0.324646
0.284555	0.695242	0.33485	1.188238	15	32	0.352424	1.004097
0.958678	0.227607	3.186359	0.258262	50	7	4.499205	0.149207
0.223121	0.76693	0.25247	1.456417	12	36	0.271303	1.247643
0.576434	0.795495	0.859045	1.587165	30	38	0.901466	1.395995
0.695029	0.15302	1.187537	0.166078	41	4	1.670237	0.082518
0.405713	0.245125	0.520393	0.281203	26	9	0.723247	0.196272
0.270211	0.806055	0.315	1.64018	14	40	0.324646	1.570237
0.62215	0.796197	0.973259	1.590603	34	39	1.118476	1.479328
0.382153	0.884823	0.481514	2.161286	21	45	0.537552	2.215872
0.947508	0.758171	2.947098	1.419526	49	35	3.499205	1.180976
0.033418	0.832606	0.033989	1.787405	1	41	0.02	1.670237
0.656911	0.460952	1.069765	0.61795	38	18	1.395995	0.44071
0.62804	0.027345	0.98897	0.027725	36	2	1.247643	0.040408
0.072329	0.054598	0.075078	0.056145	5	3	0.104257	0.061241
0.774468	0.222968	1.489294	0.252274	43	6	1.906348	0.126479
0.636311	0.64687	1.011456	1.04092	37	29	1.319072	0.853847
0.403394	0.50029	0.516498	0.693727	25	21	0.683247	0.537552
0.159276	0.747276	0.173492	1.375458	11	34	0.245662	1.118476
0.043214	0.946837	0.044176	2.934388	3	47	0.061241	2.665872
0.334574	0.486953	0.407328	0.667388	17	20	0.410407	0.504218
0.59801	0.78341	0.911329	1.52975	32	37	1.004097	1.319072
0.114872	0.582293	0.122023	0.872975	8	28	0.172463	0.808392
0.470717	0.157109	0.636233	0.170918	28	5	0.808392	0.104257
0.043519	0.51149	0.044495	0.716396	4	22	0.082518	0.572034
0.376202	0.318979	0.471928	0.384163	20	12	0.504218	0.271303
0.134861	0.87289	0.144865	2.062706	9	44	0.196272	2.049205
0.687185	0.40492	1.162144	0.519059	40	17	1.570237	0.410407
0.396954	0.39375	0.505762	0.500463	23	16	0.607749	0.380995
0.657308	0.539262	1.070922	0.774926	39	24	1.479328	0.644786
0.590869	0.557878	0.89372	0.81617	31	26	0.951466	0.723247
0.364879	0.479141	0.45394	0.652275	18	19	0.44071	0.47196
0.319742	0.321421	0.385284	0.387754	16	13	0.380995	0.297619
0.721122	0.518601	1.27698	0.731059	42	23	1.781348	0.607749
0.042085	0.88525	0.042996	2.165003	2	46	0.040408	2.415872
0.386273	0.871456	0.488205	2.051484	22	43	0.572034	1.906348
0.102145	0.992553	0.107747	4.900009	7	50	0.149207	4.499205
0.938353	0.250984	2.786324	0.288995	48	10	2.999205	0.220662
0.372143	0.720847	0.465442	1.275996	19	33	0.47196	1.059653
0.136784	0.866176	0.14709	2.011232	10	42	0.220662	1.781348
0.599048	0.236579	0.913913	0.269946	33	8	1.059653	0.172463
0.626179	0.552263	0.983978	0.803549	35	25	1.180976	0.683247

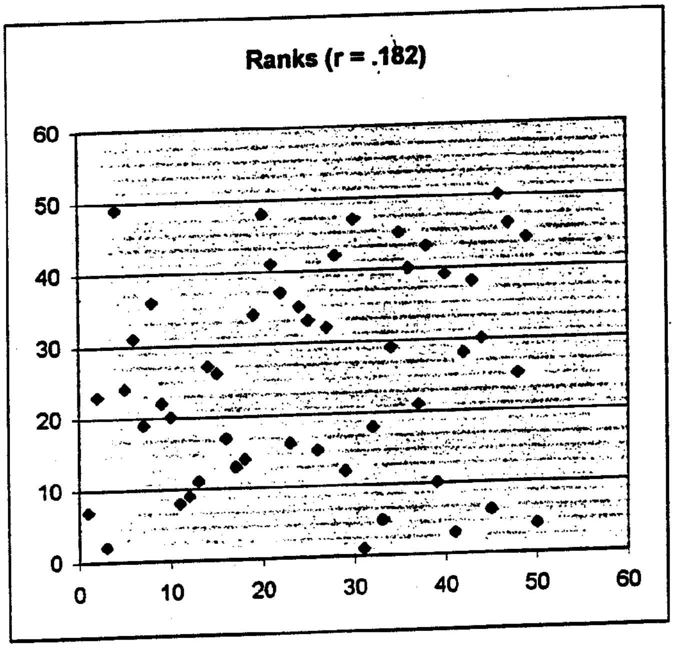
	Column 1	*Column 2*		*Column 1*	*Column 2*		*Column 1*	*Column 2*		*Column 1*
Column 1	1		Column 1	1		Column 1	1		Column 1	1
Column 2	-0.36194	1	Column 2	-0.29665	1	Column 2	-0.4206	1	Column 2	-0.31448

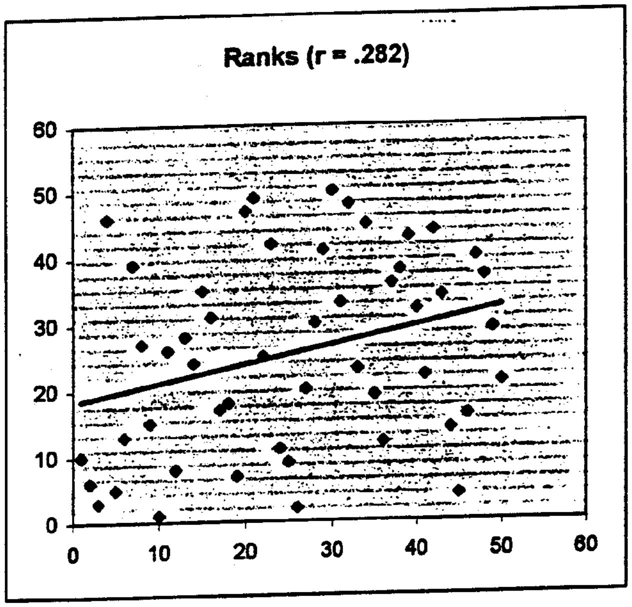

Bivariate Exponential (r = .207)

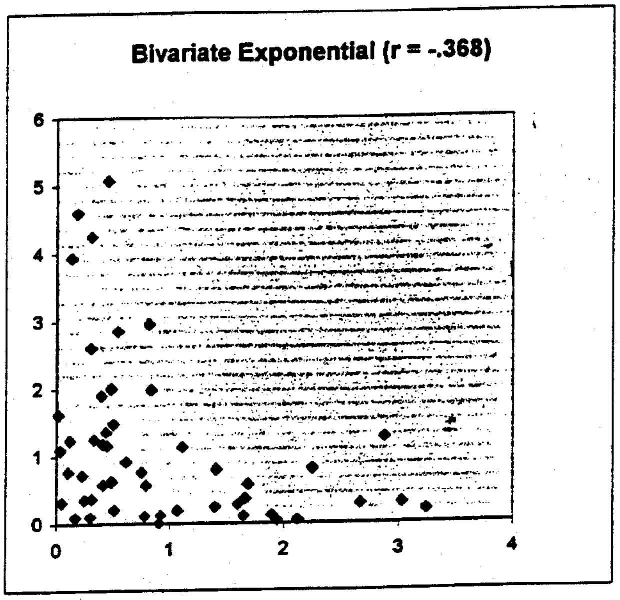

Bivariate Exponential (r = .059)


Savage Scores (r = .175)

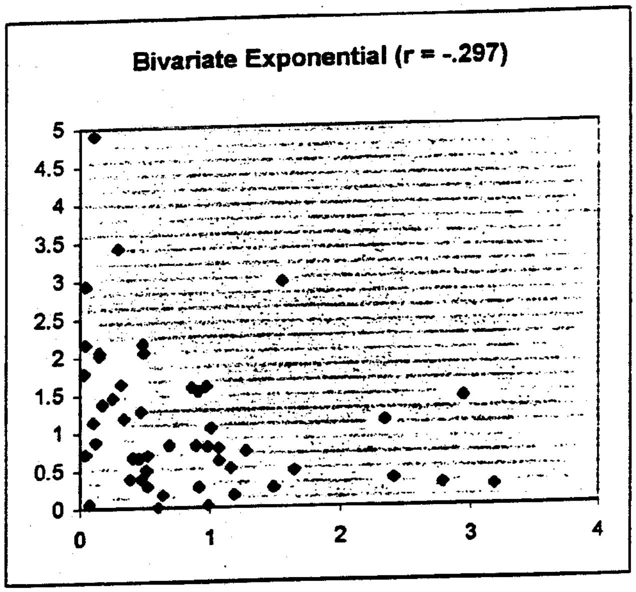

Savage Scores (r = .050)


Ranks (r = .182)

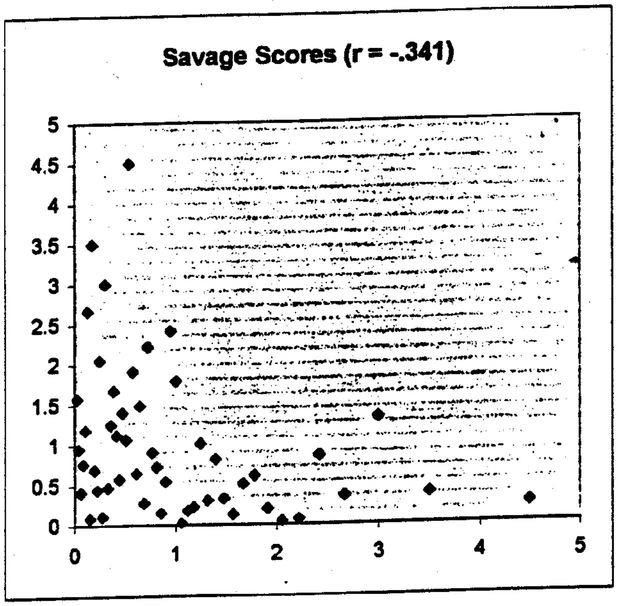

Ranks (r = .282)



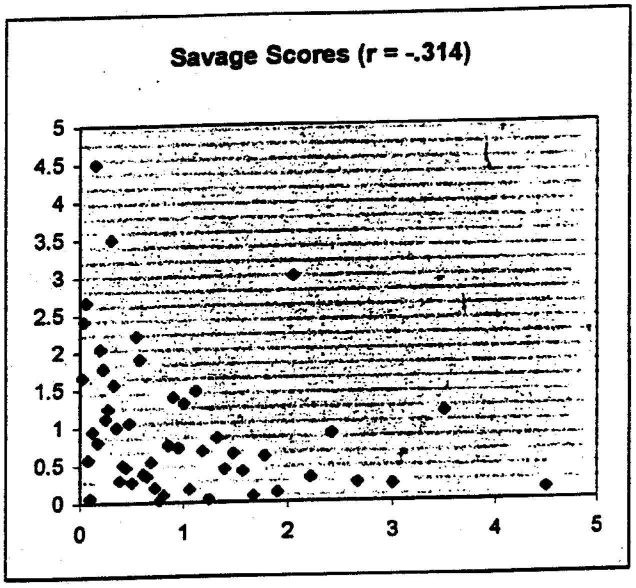
Bivariate Exponential (r = -.368)



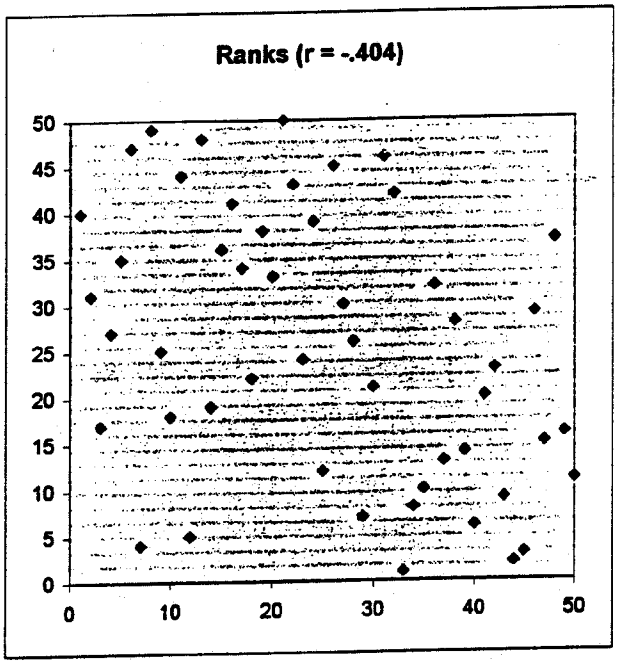
Bivariate Exponential (r = -.297)



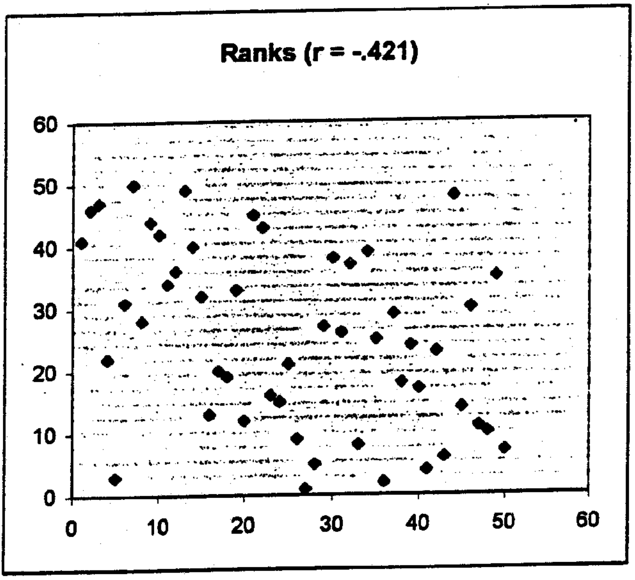
Savage Scores (r = -.341)



Savage Scores (r = -.314)



Ranks (r = -.404)



Ranks (r = -.421)

Boosting with the L2 Loss: Regression and Classification

Bin Yu, University of California, Berkeley

In this talk, we investigate a variant of boosting: L2 Boost, which is constructed from a functional gradient descent algorithm with the L_2loss function. Based on an explicit stage wise refitting expression of L2 Boost, the case of (symmetric) linear weak learners is studied in detail in both regression and two-class classification. In particular, with the boosting iteration m working as the smoothing or regularization parameter, a new exponential bias-variance trade off is found with the variance (complexity) term bounded as m tends to infinity. When the weak learner is a smoothing spline, an optimal rate of convergence result holds for both regression and two-class classification. And this boosted smoothing spline adapts to higher order, unknown smoothness. Moreover, a simple expansion of the 0-1 loss function is derived to reveal the importance of the decision boundary, bias reduction, and impossibility of an additive bias-variance decomposition in classification. Finally, simulation and real data set results are obtained to demonstrate the attractiveness of L2 Boost, particularly with a novel component-wise cubic spline as an effective weak learner.

Contributed Session V

Subsampling of Biased Statistics with Application to Nonparametric Density and Intensity Estimation

Andreas FUTSCHIK
Univ. of Vienna and UC Berkeley
Universitätsstr. 5/9, A-1010 Vienna, Austria

Abstract

We consider the application of bootstrap and subsampling to cases where biased statistics are of interest. A situation where such statistics frequently occur, is in nonparametric estimation problems. We focus on the case of nonparametric density/intensity estimation. An application is the identification times of maximum arrival, for instance of ships in a harbor.

1 Introduction

A frequent statistical goal is to approximate the distribution of some functional $R_n = R_n(\mathbf{X}_n, \theta_n(P))$ of the data $\mathbf{X}_n = (X_1, \ldots, X_n)$ and some parameter sequence $\theta_n(P)$. This is useful when constructing confidence and prediction intervals, doing hypothesis tests, among others. For instance, a confidence interval for the parameter $\theta_n(P) = \theta(P) = \mathbf{E}_P X_1$, for i.i.d. real valued random variables $X_1, \ldots, X_n$ can be obtained by approximating the distribution of $R_n(\mathbf{X}_n, \theta(P)) := n^{1/2} \frac{\bar{X}_n - \theta(P)}{S_n}$, where $\bar{X}_n$ is the sample mean and $S_n = [\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2]^{1/2}$ the sample standard deviation. Obviously $P(-\gamma \leq R_n \leq \gamma) = 1 - \alpha$ implies that $[\bar{X}_n - \gamma \frac{S_n}{\sqrt{n}}, \bar{X}_n + \gamma \frac{S_n}{\sqrt{n}}]$ is an $1 - \alpha$ confidence set. Alternatively, confidence intervals could be based on the α trimmed mean

$$\bar{X}_{n,\alpha} = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor}^{n - \lfloor \alpha n \rfloor} X_{[i]}$$

instead of $\bar{X}_n$. An approximation of the distribution of

$$R_n(\mathbf{X}_n, \theta(P)) := n^{1/2} \left(\bar{X}_{n,\alpha} - \theta(P) \right)$$

leads to confidence interval based on the trimmed mean.

As elaborated for instance in Beran and Ducharme (1991), the bootstrap approximation of the distribution of $R_n(\mathbf{X}_n, \theta_n(P))$ is given by the distribution of

$$R^*_{n,n} = R_n(\mathbf{X}^*_n, \theta_n(P_n)),$$

where P_n is an estimate of P and $\mathbf{X}_n^*$ is a sample of size n from P_n. If P_n is chosen to be the empirical distribution function, the resulting approximation is called nonparametric bootstrap. Alternative estimates of P, like a smoothed version of the empirical cdf or a parametric estimate of P lead to different types of bootstrap. Furthermore, if $\theta_n(P_n)$ is not well defined in a problem, it is often replaced by a suitable estimate of $\theta_n(P)$.

Bootstrap is said to "work", if the asymptotic distribution of $R_n(\mathbf{X}_n, \theta(P))$ is reproduced correctly, as $n \to \infty$. This is a minimal condition for the applicability of bootstrap, stating that the distribution estimated by bootstrap approaches the correct distribution as the sample size increases.

Unfortunately there seems to be no easy to check, generally applicable rule to find out whether bootstrap works in a particular situation. However, several examples have been found where bootstrap fails, and classes of functionals and distributions have been identified where the bootstrap is known to work. (See for instance Bickel et al. (1997).)

An example where the bootstrap fails, is for the family of distributions $\mathcal{P} = \{P : \mathbf{E}_P X < \infty\}$ when the goal is to estimate $|\theta(P)|$ with $\theta(P) = \mathbf{E}X_1$, and P is such that $\mathbf{E}_P(X) = 0$. Consider the functionals $R_n(\mathbf{X}_n, \theta(P)) = \sqrt{n}(|\bar{X}_n| - |\theta(P)|)$. Then, with $Z \sim N(0, \sigma^2[X_1])$, $R_n(\mathbf{X}_n, \theta(P))$ converges in distribution against $|Z|$. The bootstrap approximation $R_n^*(\mathbf{X}_n, \theta(P_n))$, on the other hand, converges weakly against $|Z^* + Z| - |Z|$. (See Beran & Srivastava (1985), Dümbgen (1993).)

Subsampling (also known as m-bootstrap) provides an alternative to bootstrap that is known to work under much weaker conditions than bootstrap does. When subsampling is done, the distribution of $R_n(\mathbf{X}_n, \theta_n(P))$ is approximated by that of

$$R_{m,n}^* := R_m(\mathbf{X}_m^*, \theta_m(P_n)),$$

where $\mathbf{X}_m^*$ is a sample of size m from P_n.

Subsampling can be done with or without replacement. Politis and Romano (1994) show that for a sequence of functionals of type

$$R_n(\mathbf{X}_n, \theta(P_n)) = \tau_n(\theta(P_n) - \theta(P)),$$

weak convergence of R_n against some nondegenerate limiting distribution $\mathcal{L}(P)$ is sufficient for subsampling without replacement to work, provided that $\tau_n \to \infty$, $m \to \infty$ and $m/n \to 0$. Often subsampling also works when done with replacement. A detailed discussion concerning consistency and advantages of bootstrap and subsampling with and without replacement can be found in Bickel et al. (1997).

2 Bootstrap and Subsampling in the context of nonparametric estimation problems

Here we focus on resampling applied to density and/or intensity estimation. Intensity functions occur in the context of inhomogeneous Poisson processes, applications include arrival and departure times of ships etc. Nonparametric estimates have been proposed by several authors, including Diggle (1985), Diggle and Marron (1988), and Leadbetter and Wold (1983). See Figure 1 below for an example involving an estimated intensity function.

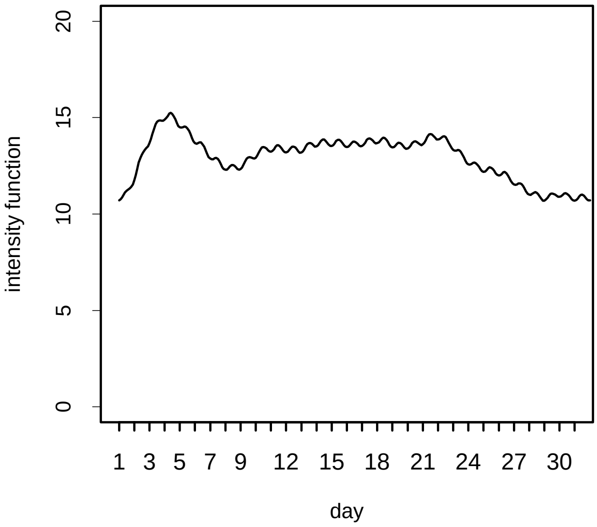


Figure 1: Estimated intensity function for ships arriving at Keelung harbor (Taiwan) in January.

Estimating intensity functions is related to the problem of estimating probability densities. To see this, we will state the problems more formally. When estimating intensities, we assume to have a random number N of observations $X_1, \ldots, X_N$ from an inhomogeneous Poisson process $X(t)$ with intensity function λ on some time interval. To avoid trivialities, assume that λ is positive at least on some interval. Then the intensity density λ^* is related to the intensity function via

$$\lambda^*(x) = \frac{\lambda(x)}{\int_0^t \lambda(s)\, ds}, \quad 0 \leq t \leq 1.$$

Since conditionally on $N = n$, the jump points X_j have the same joint distribution as the order statistics of a sample of n independent observations $X_1, \ldots, X_n$ from density λ^*, the problems of intensity and density estimation are related. Indeed, intensity functions can be estimated in two steps. First estimate λ^* by using a kernel density estimate and then estimate a normalizing constant (the cumulative intensity). An asymptotic analysis of the behavior of intensity estimates is possible in two ways. It might either be assumed that the intensity function is periodic and that the number of observation periods increases. Or, more technically, one might assume that there is a sequence of unknown intensities $\lambda_l = l\mu$ for some density μ and let $l \to \infty$. Both approaches lead to similar results, but we will focus on the first approach involving periodic intensities.

In nonparametric estimation problems, bias is typically non-negligible. Consider for instance a distribution P with twice differentiable density f whose second derivative is bounded. The classical kernel density estimate on $\mathbb{R}^d$ for a sample of $X_1, X_2, \ldots, X_n$ of size n is given by

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Assume that the bandwidth h is chosen of optimal order $h = cn^{-1/(d+4)}$. It is then well known that $\hat{\tau}_n[\hat{f}_n(x) - f(x)]$, with

$$\hat{\tau}_n = \left(\frac{\hat{f}_n(x) \int K^2(u)\,du}{nh^d}\right)^{-1/2}$$

has an asymptotic normal $N(\gamma_f, 1)$ distribution, with $\gamma_f = c^{(d+4)/2} \frac{B_x}{V_x^{1/2}}$,

$$V_x = f(x) \int K^2(u)\,du$$

and

$$B_x = \frac{1}{2} \sum_{|\alpha|=2} D^\alpha f(x) \int u^\alpha K(u)\,du.$$

(See for instance Rosenblatt (1991).) Thus the bias does not vanish asymptotically.

As pointed out by Hall (1992), the bootstrap fails in this context, unless the asymptotic bias is zero. To see why, consider the functional

$$R_n = R_n(\mathbf{X}_n, \theta_n(P)) = \hat{\tau}_n[\hat{f}_n(x) - f(x)]$$

whose distribution is of interest when constructing confidence sets. As shown above, R_n converges in distribution to a normal $N(\gamma_f, 1)$ distribution. The bootstrap version is

$$R_n^* = \hat{\tau}_n[\hat{f}_n^*(x) - \hat{f}_n(x)],$$

with $\hat{f}_n^*(x) = \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{x - X_i^*}{h_n}\right)$ calculated from a resample $X_1^*, X_2^*, \ldots, X_n^*$ with replacement of size n. Since

$$R_n^* = \hat{\tau}_n[\Big(\hat{f}_n^*(x) - \mathbf{E}^* \hat{f}_n^*(x)\Big) + \Big(\mathbf{E}^* \hat{f}_n^*(x) - \hat{f}_n(x)\Big)] = \hat{\tau}_n[\hat{f}_n^*(x) - \mathbf{E}^* \hat{f}_n^*(x)],$$

we see that $R_n^* \to N(0,1)$ in distribution.

Subsampling on the other hand gets the bias right, irrespectively whether it is done with or without replacement. To see why, notice that the subsampling version of R_n is given by

$$R^*_{m,n} = \hat{\tau}_n[\hat{f}^*_{m,h_m}(x) - \hat{f}_n(x)],$$

where $\hat{f}^*_{m,h_m}(x) = \frac{1}{mh_m^d}\sum_{i=1}^m K\left(\frac{x-X_i^*}{h_m}\right)$ is based on a resample $X_1^*, X_2^*, \ldots, X_m^*$ of size m. Now

$$\begin{aligned} R^*_{m,n} &= \hat{\tau}^{-1}_{m,h_m}[\hat{f}^*_{m,h_m}(x) - \mathbf{E}^*\hat{f}^*_{m,h_m}(x)] + \hat{\tau}^{-1}_{m,h_m}[\mathbf{E}^*\hat{f}^*_{m,h_m}(x) - \hat{f}_n(x)] \\ &=: U_{m,n} + \tilde{U}_{m,n}. \end{aligned}$$

Since $U_{m,n} \to N(0,1)$ in distribution, the bias is approximated correctly, if $\tilde{U}_{m,n} \to \gamma_f$ in probability. But this follows since

$$\begin{aligned} \tilde{U}_{m,n} &= \hat{\tau}_m\left[(\mathbf{E}f_m(x) - f(x)) + (\mathbf{E}^*f^*_m(x) - \mathbf{E}f_m(x)) + [f(x) - f_n(x)]\right] \\ &= \frac{B_x}{V_x^{-1/2}}(mh_m^{d+4})^{1/2} + o_P(1). \end{aligned}$$

I.e. subsampling works, if the rate-optimal bandwidth is adapted together with the sample size.

Notice that whereas the classical nonparametric bootstrap does not work in this situation, smoothed bootstrap provides another alternative that works under suitable assumptions.

In the context of intensity estimation bootstrap faces the same bias problems as described above. See Cowling, Hall and Phillips (1996) for details. Their resampling method 3 is equivalent to the method described here. Subsampling is able to approximate the distribution of kernel based intensity density estimates correctly. If our goal is to estimate the intensity function itself and assume that it has period 1 and observations are collected up to time t, correct results are obtained by multiplying the kernel density estimate by N/t, where N denotes the total number of observations.

References

[1] Beran R., and Ducharme, G.R. Asymptotic theory for bootstrap methods in statistics. Les Publications CRM, 1991.

[2] Beran, R., Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. Annals of Statistics, 13, 95–115.

[3] Bickel, P. J., Götze, F. and Van Zwet, W. R. (1997) Resampling fewer than n observations: gains, losses and remedies for losses. Statistica Sinica, 7, 1–31.

[4] Cowling, A., Hall, P., Phillips, M.J. (1996). Bootstrap confidence regions for the intensity of a Poisson point process. JASA, 91, 1516-1524.

[5] Diggle, P. (1985). A kernel method for smoothing point process data. Applied Statistics, 34, 138–147.

[6] Diggle, P. and Marron, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. JASA, 83, 793–800.

[7] Leadbetter, M. R., Wold, D. (1983) On estimation of point process intensities. In: Contributions to Statistics: Essays in Honor of Norman L. Johnson, 299–312, ed. P.K. Sen. Amsterdam: North Holland.

[8] Hall, P. (1992) Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann. Statist. 20 675-694.

[9] Politis, D.N., Romano, J.P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. Ann. Statist. 22, 2031-2050.

[10] Rosenblatt, M. Stochastic curve estimation. NSF-CBMS Regional conference series, Hayward, 1991.

The Biggest, The Oldest, and Other Such Extremes

Bruce J. West, US Army Research Office

We investigate the relation between the underlying dynamics of a randomly evolving system and the extrema statistics of such systems. Failure modes, as an exemplar of an extreme property, are considered in independent processes, Fokker-Planck processes and Lévy stable processes. Using the Kolmogorov-Sinai entropy we find a relation between dynamical chaos and the ubiquitous inverse power-law distribution. Applications of these ideas to biomedical data sets are used as examples throughout the discussion.

Reducing the Error in Estimating Production Costs of Multiple-Unit Procurements

Scott M Vickers, MCR Federal Inc.

Standard cost-estimating practice involves application of a cost-improvement factor, or “learning” rate, to account for management, engineering, and production improvements that save money as successive units of a multiple-unit procurement are produced. A combination of circumstances makes it difficult to determine *a priori* what the "correct" learning rate will be for any particular procurement. The most common method of forecasting production costs is to estimate from a set of data a "theoretical” first-unit cost (“T1”), which is then used as the independent variable in an exponentially decreasing “learning curve,” whose dependent variable is the average per-unit cost of all units produced. In this report, we provide computational evidence that the variance of T1 is larger than that of any other independent variable on which an estimate of total production cost may reasonably be based. Naturally this variance is passed through to the total-production estimate, so to minimize the error of estimating total-production cost, a different independent variable must be used. Using data from missile-production programs, we show that “T250” (the average unit cost of the first 250 production units) has smaller variance than T1 based on the same data and has additional properties that further reduce the error of estimating the total cost of multiple units. The mathematical method we use is a version of Iteratively Reweighted Least Squares that was first applied to learning curves by Dr. M.S. Goldberg of the Center for Naval Analyses.

Contributed Session VI

Cybernetic Ballistic Missile Defense Systems

Robert L. Fry, Johns Hopkins University Applied Physics Laboratory

The complexity of ballistic missile defense systems suggest that new system engineering paradigms are needed to formally address issues relating to their performance estimation, architectural trades, designs, and perhaps most importantly, their operation and adaptation. A new system engineering approach is proposed based on the formalization of cybernetic concepts advocated in part by Norbert Wiener. The described cybernetic approach is being developed and evaluated for the purpose of formalizing the engineering of ballistic missile defense systems with the goal of designing an adaptive predator that can capture and kill its ballistic threat prey. A cybernetic approach provides a means of flowing down top-level requirements to the functional elements of a system. In addition, cybernetics provides a basis for the design of intelligent algorithms capable of learning and adaptation within these systems. One consequence of a cybernetic approach is that each ballistic missile defense system possess a probabilistic real-valued wave function that, through the use of system sensor assets, can be collapsed in a controlled and coordinated manner among system elements as required for the use and direction of system energy assets of the system toward the negation of the threat.

A central thesis of this paper is that cybernetic systems are characterized through the notion of “objective subjectivity” whereby information and control represent dual and subjective properties of a system. Information and its acquisition characterize the operation of a system attempting to ascertain a “state-of-nature” posited to exist external to the system. Similarly, action and control describe the intent of a system to place a nature in a posited and purposeful state thereby having it acquire a particular “state-of-nature.” In both cases, the notion of “state” and “distinguishability” are subjective properties of the system as are probability and entropy. However, the logical rules afforded cybernetic systems that include the manipulation of probabilities and entropies can be rigorously derived and are universal in that they apply to all cybernetic systems. In this sense, cybernetic systems can objectively be designed and implemented. Simple examples serve to demonstrate the basic concepts as applied to a ballistic missile defense system and the closure of its firecontrol loop.

Sensitivity Analysis Using Design of Experiments in Ballistic Missile Defense

Jacqueline K. Telford
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, Maryland 20723-6099

A sensitivity analysis discovers the functional relationship between a response variable (Y) and many possible explanatory variables (X_1, X_2, … , X_k). The first step is a screening experiment to determine which explanatory variables need to be included in the function. Since only two levels are used, any non-linearity of the effects cannot be detected. The second step is to fit a response surface model by a second-degree polynomial of the important X's found in the screening experiment. Three types of designs were used in this study: central composite design, three-level fractional factorial design, and D-optimal design. These three types of designs were evaluated using the same number of additional experiments to determine the efficiency in estimating the response surface. These design of experiment techniques have been successfully applied to several Ballistic Missile Defense sensitivity studies to maximize the amount of information in a minimum number of computer simulation runs.

Keywords: Sensitivity analysis, Design of experiments, Simulations

1.0 Introduction. The basic situation is that of needing to evaluate some process with input variables called factors and with measured output variables called responses. This process could be a complex computer simulation model or a manufacturing process with raw materials and temperature and pressure settings as the inputs and a product being produced. If the input variables to the process are varied, the outputs will vary. The question is which input variables (factors) are causing the majority of the variability in the output (responses), in other words, which factors are the "drivers." It is desirable to answer the question of where the variability is coming from (also known as "sensitivities") with a minimum expenditure of resources.

Experimental design is an effective tool for maximizing the amount of information gained from a study while minimizing the amount of data to be collected, which, in this case, is minimizing the number of computer runs. Factorial experimental designs investigate the effects of many different factors in a single study, instead of conducting many separate studies, each varying one factor at a time. Factorial designs allow estimation of the sensitivity to each factor and also to combinations of two or more factors at a time.

This paper describes a process for identifying Ballistic Missile Defense (BMD) Family of Systems (FoS) needs using experimental design techniques and shows some of the findings from a first implementation of the methodology. The sensitivity analysis proceeds in two steps: a screening experiment to determine the main drivers and a response surface experiment to determine the shape of the effects (linear or curved). The following sections describe the methodology, data, modeling assumptions, and some results from the study.

2.0 Screening Design Methodology. Many factors are screened in a sensitivity analysis to determine which are the main drivers of system performance. However, as the number of factors increases, the total number of combinations increases geometrically. For this reason, studies employing experimental design should use a method such as the Fractional Factorial Method which produces high confidence in sensitivity results using very small fractions, that is, a small subset of the total number of combinations, in this case as small as 1 in 200 hundred billion.

2.1 *General Concepts of Experimental Design.* In experimental design, certain terminology is used. The controllable input variables to the experiment (in this case, the simulation program) are the factors. The performance measures output from the experiment are called responses. The polynomial equation that is frequently used to model the response variable (Y) as a function of the input factors (X's) is :

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} \beta_{ij} X_i X_j + \sum_{\substack{i=1 \\ i \neq j \neq k}}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_{ijk} X_i X_j X_k + \cdots \qquad (1)$$

where β_0 represents the overall mean response

β_i represents the main effects for each factor (i = 1, 2, ... , p)

β_{ij} represents the two-way interaction between the ith and jth factors

β_{ijk} represents the three-way interaction between the ith, jth, and kth factors.

Usually, two values of the X's (called levels) are used in the experiment for each factor, denoted by "high" and "low" and coded +1 and –1. The use of only two levels implies that the effects are monotonic on the response variable, but not necessarily linear. At least three levels of the factors would be required to detect curvature. Interaction is present when the effect of a factor on the response variable depends on the level of another factor. Graphically, this can be seen as two non-parallel lines when plotting the response means from the four combinations of high and low levels of the two factors. The β_{ij}'s account for the two-way interactions and it is desirable to be able to estimate these effects. Two-way interaction can be thought of as the correction to (or lack of) perfect additivity of the factor effects.

Experimental designs can be categorized by their resolution level. The higher the resolution level (as in optics, being able to separately resolve objects), the more terms in the regression analysis equation that can be estimated. Confounding is the opposite of resolution. Confounding occurs when only the summation of several effects can be estimated, not the effects separately. Confounded results require that additional experiments be conducted to untangle the results, to clearly identify which factor combinations are drivers and which are not (Daniel, 1962). Resolution levels are usually denoted by Roman numerals (III, IV, and V are the most commonly used). The effects in the regression analysis equation are not confounded if the sum of their "ways" is less than the resolution level of the design. In order to have all of the two-way interactions unconfounded from each other, the resolution level needs to be at least V. However, in the resolution level V, the main effects ("one-ways") are confounded with some four-way interactions and the two-way interactions are confounded with some three-way interactions. Therefore, it is usual to assume that most of the three-way and higher order interactions are negligible. The three-way and higher order interactions are not estimated separately, but their combinations are used to estimate the precision of the estimates and to compute the confidence intervals on the effects.

Table 1. Resolution Levels and Confounding Patterns

Resolution Level	Confounding Patterns
II	Main effects confounded with themselves
III	Main effects not confounded with themselves but confounded with two-way interactions
IV	Main effects not confounded with two-way interactions, but two-ways confounded with themselves
V	Main effects and two-way interactions not confounded except with higher order interactions

Factorial designs collect data at the vertices of a cube in p-dimensions (p is the number of factors being studied). If data is collected from all the vertices, the design is a full factorial, requiring 2^p runs and no confounding occurs. Fractional factorial designs collect data from a specific subset of all possible vetrices and require 2^{p-q} runs. Fractional factorial designs can determine which factors and their combinations have significant effects on the response variable. Fractional factorial designs yield sets of unconfounded effects (main, two-way, three-way, . . .), depending on the resolution level of the design. The minimum number of runs needed for Resolution IV and V designs for different numbers of factors are shown in Tables 2 and 3. Law and Kelton (2000) specifically "recommend that only designs of resolution V or higher be considered," page 639. However, as the number of factors increase, it may not be feasible to perform the Resolution V design. Since the significant two-way interactions are most likely combinations of the significant main effects, a Resolution IV design can be used, especially if the factors have monotonic effects on the response variable. A second, smaller Resolution V design using only the significant main effects (as determined from the Resolution IV experiment) can be performed to determine if there are any significant two-way interactions. Fractional factorial designs have been suggested for use in computer simulations in Jacoby and Harrison (1962), Hunter and Naylor (1970), Kleijnen (1975), and Biles (1979).

Table 2. Resolution IV Designs: All Main Effects Free of Two-way Interactions

Number of Factors	Minimum Number of Runs
1	2
2	4
3 - 4	8
5 - 8	16
9 - 16	32
17 - 32	64
33 - 64	128
65 - 128	256
129 - 256	512

Table 3. Resolution V Designs: All Main Effects and Two-way Interactions Unconfounded

Number of Factors	Minimum Number of Runs
1	2
2	4
3	8
4 - 5	16
6 - 7	32
8	64
9 - 11	128
12 - 17	256
18 - 22	512
23 - 31	1,024
32 - 40	2,048
41 - 54	4,096

If a factorial design is used in a screening experiment of many factors, there is no need to replicate the same combinations of factors. Factorial designs, including fractional factorials, have essentially built-in replication. More design points are preferable to replicating the same points. An experimental design is a matrix of +1's and –1's with one column for each factor, and one row for each set of factor combinations, called a design point, labeled “Run” in Table 4.

Table 4. Notational Experimental Design Matrix

Run	Factor 1	Factor 2	Factor 3	.	.	.	Factor 47
1	–1	–1	–1				–1
2	+1	–1	–1				+1
3	–1	+1	–1				+1
.							.
.							.
.							.
xx	+1	+1	+1	.	.	.	–1

The statistical textbooks such as those by Box, Hunter, and Hunter (1978) and Montgomery (2000) describe the concepts of this section in more detail. Law and Kelton (2000) also discuss factorial designs; however, Law and Kelton’s analysis requires replicating the experimental design, is not efficient, and is therefore not recommended. The analysis of factorial data found in statistics texts and in statistical software packages is preferred. The traditional statistical analysis of factorial designs has the feature that there is equal precision on the main effects and on the interactions being estimated. The confidence intervals on the main effects and on the interactions are of equal length and are as small as possible. Confidence intervals on the main effects will be shown in Figure 2.

2.2 *Survey of Other Approaches to Experimental Design.* There are many approaches to and types of experimental designs. The typical engineering experiment involves varying one factor at a time, while holding constant the other factors such as to the standard operating conditions. The shortcomings of this type of design are that there is no information at other combinations of operation conditions or on possible two-way interactions. Another type of design is the Random Balance design, but this design randomly confounds the effects that one is trying to study. The Random Balance design might be only a resolution II design. A Plackett-Burman design is used when only a very limited number of runs can be performed (for example in Grier et al., 1999), but confounds the two-way interactions with the main effects and is only valid if there are no two-way interactions. A Blocked design reduces variability by testing for effects on groups of similar experimental units, but is not normally needed in a simulation study. A Latin hypercube design employs blocking in multiple directions and, similarly to the Plackett-Burman design, assumes no multi-way interactions. Plackett-Burman and Latin hypercube are resolution III designs

2.3 *Application to Simulation Model.* Each row of the design matrix specifies the particular combination of high or low values of each of the factors to be run. In this case, there are 47 factors, so there is a high value (denoted by +1) or low value (denoted by –1) for each of the factors specified for each simulation run, which corresponds to one row of the design matrix. For the very large experiments illustrated in this paper, the experimental designs were generated and the data from the simulation runs were analyzed using the SAS® software, version 7. The exact factor combinations specified by the experimental design must be run to achieve the desired resolution level.

The steps to perform the analysis are described here. First, each of the input factors specified as +1 or –1 in the design matrix are converted to the actual engineering values and then the simulation is run for each design point. The response variables or Measures of Effectiveness (MOEs) are associated with each of the design points and the sensitivity results for each MOE are calculated using regression analysis (Equation 1). The sensitivity results are the change in MOE caused by changing each main effect (factor) from the "low" to "high" value, along with the 95% confidence limit. The two-way interactions and their confidence limits are also estimated if they can be separated, depending on the resolution level of the design. It should be noted that in executing this process, a number of scripts have been written to run the Extended Air Defense Simulation (EADSIM) automatically rather than via the Graphical User Interface (GUI). Figure 1 contains a pictorial representation of the analysis process using the EADSIM program. The UNIX script shown is an example with a single processor machine operation. There are other scripts required to employ multiple processor machines.

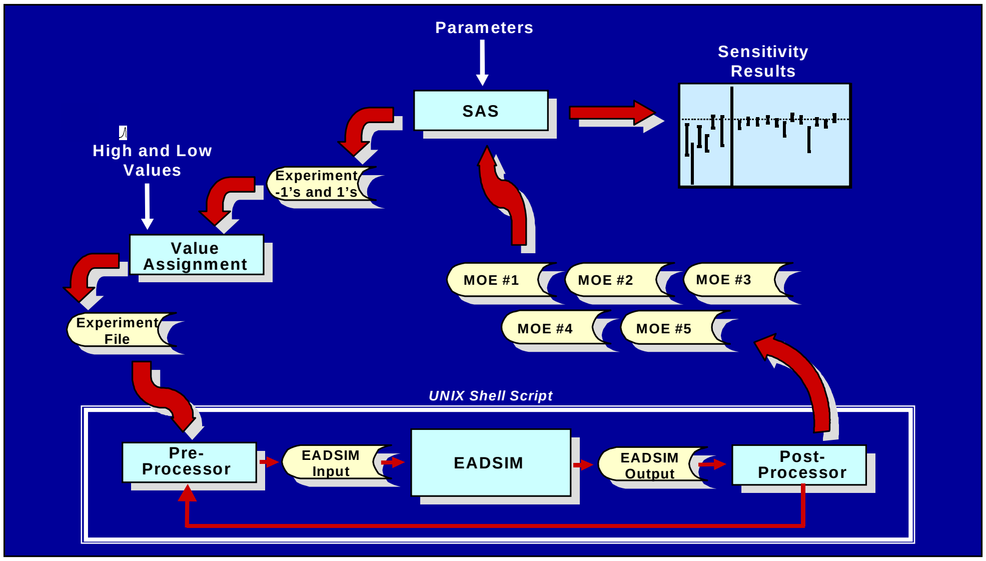


Figure 1. Process for Running the Sensitivity Analysis

We used a Fractional Factorial experimental design and EADSIM to screen 47 factors for their relative importance in far-term (i.e., 2010 timeframe) Northeast Asia (NEA) and Southwest Asia (SWA) scenarios over the first 10 days of the war. A three-tiered defense system was employed for both scenarios, including an Airborne Laser (ABL), a Ground-Based (GB) Upper Tier, and a Lower Tier comprised of both Ground-Based and Sea-Based (SB) systems.

The primary MOE for the study was FoS Protection Effectiveness and the secondary MOEs were inventory usage for each of the defensive weapon systems. We defined FoS Protection Effectiveness as the number of threats negated divided by the total number of incoming threats over the course of a scenario.

The sensitivities to the 47 factors were calculated from the MOEs at the completion of 10 days of warfare, but, as a check, the sensitivities at times less than 10 days were computed and the results were found to be comparable. This is important because it means the sensitivity results are roughly the same no matter what time is selected, and indicates the robustness of the method to variations in the scenario length.

Table 5 shows the 47 factors that were screened in the study. We selected these factors by doing a functional decomposition of the engagement process for each defensive weapon system (e.g., a radar must detect, track, discriminate, and assess the success of intercept attempts) and then by accuracy, reliability, and timeline factors associated with each of those functions.

Table 5. Factors to be Screened

Threat RCS	GB Lower Tier 2 Reaction Time
Satellite Cueing System Prob of Detection	GB Lower Tier 2 Pk
Satellite Cueing System Network Delay	GB Lower Tier 2 Vbo
Satellite Cueing System Accuracy	SB Lower Tier Time to Acquire Track
Satellite Cueing System Time to Form Track	SB Lower Tier Time to Discriminate
GB Upper Tier Time to Acquire Track	SB Lower Tier Time to Commit
GB Upper Tier Time to Discriminate	SB Lower Tier Time to Kill Assessment
GB Upper Tier Time to Commit	SB Lower Tier Prob of Correct Discrimination
GB Upper Tier Time to Kill Assessment	SB Lower Tier Prob of Kill Assessment
GB Upper Tier Prob of Correct Discrimination	SB Lower Tier Launch Reliability
GB Upper Tier Prob of Kill Assessment	SB Lower Tier Reaction Time
GB Upper Tier Launch Reliability	SB Lower Tier Pk
GB Upper Tier Reaction Time	SB Lower Tier Vbo
GB Upper Tier Pk	Network Delay
GB Upper Tier Vbo	Lower Tier Minimum Intercept Altitude
GB Lower Tier Time to Acquire Track	Upper Tier Minimum Intercept Altitude
GB Lower Tier Time to Discriminate	ABL Reaction Time
GB Lower Tier Time to Commit	ABL Beam Spread
GB Lower Tier Prob of Correct Discrimination	ABL Atmospheric Attenuation (j-95%)
GB Lower Tier 1 Launch Reliability	GB Upper Tier Downtime
GB Lower Tier 1 Reaction Time	GB Lower Tier Downtime
GB Lower Tier 1 Pk	SB Lower Tier Downtime
GB Lower Tier 1 Vbo	ABL Downtime
GB Lower Tier 2 Launch Reliability	

As in any factorial experimental design study, we selected “high” and “low” values for all factors to be screened and, in a full factorial design, we would have run EADSIM for all possible combinations of the high and low values. We selected high and low values in this study to cover a large range of operating conditions for each factor. Our goal was to assess FoS sensitivities resulting from large variations in the 47 factors. For example, we varied the Probability of Kill (Pk) for all weapon over a range of 30% between the low and high values. If no sensitivity for Pk is indicated in the screening analysis, we can say with reasonable confidence that Pk is not a driver of FoS. Follow-on response-surface analysis is warranted for those factors flagged as being drivers in the screening analysis to identify possible “knees in the curve” in FoS performance in response to smaller changes in those factors.

We conducted the NEA and SWA screening experiments to find the main factor (i.e., linear effects) and two-way interactions for the 47 factors. We assumed all three-way and higher interactions were insignificant. The number of required experiments (i.e., EADSIM runs) was driven by the number of factors, the precision needed to resolve technology drivers from underlying randomness in the problem, and the need for “unconfounded” results. A few of the factors in Table 5 were selected for reasons unrelated to technology issues. Weapon system downtimes are the best example of this. Firing units in this study experienced downtimes that were varied over a 20% range of the total scenario time. Future sensitivity studies could vary firing unit downtimes from 0% to 100%, effectively turning entire weapon systems on and off, to explore force structure and architectural issues.

We initially conducted 512 EADSIM runs to screen the sensitivities of the 47 factors in the NEA scenario. This is a Resolution IV design and resolves all the 47 main factors but leaves confounded most of the 1,081 possible two-way interactions. After analyzing results from the initial 512 runs, 17 additional, separate experimental designs were needed (for a total of 352 additional EADSIM runs) to resolve the confounding in the two-way interactions for FoS Protection Effectiveness.

We learned from the NEA screening study that more runs are warranted in the initial experiment to reduce or eliminate the number of additional experiments needed to untangle the results. The time saved by not having to untangle results is well worth the additional computer runtime. Thus, for the SWA screening study, we conducted 4,096 EADSIM runs to find the main factors and two-way interactions for the 47 factors, all unconfounded. This was a Resolution V design. An added benefit of conducting more experiments is that smaller error estimates are obtained (approximately one third less), meaning that the relative importance of the performance drivers can be identified with higher certainty.

Running EADSIM 4,096 times for the SWA analysis, each with a 25- to 40-minute runtime, was a formidable challenge. To make this feasible, we conducted the study as part of a cooperative effort with analysts at the Ballistic Missile Defense Simulation Support Center (BMD SSC) located at the Joint National Test Facility (JNTF) in Colorado Springs, Colorado. A majority of the 4,096 EADSIM runs for the SWA analysis were run on multi-processor computers operated by the BMD SSC.

Figure 2 illustrates the main factor sensitivities to the 47 factors for both NEA and SWA. The colored dots in Figure 2 represent the sensitivity to each factor and the error bars around the colored dots are 95% confidence bounds for the results. The y-axis is the difference in the average Protection Effectiveness for a factor between the "high" and "low" values. Factors are flagged as being FoS performance drivers if the 95% confidence bounds do not include zero as a probable result. Factors shown in Red in Figure 2 were found to be performance drivers in both the NEA and SWA scenarios. Factors shown in Blue were found to be drivers in NEA only, and factors shown in Green were found to be drivers in SWA only. Factors that were not found to be drivers in either scenario are shown in Grey.

The sensitivities in Figure 2 are ranked according to the relative importance of the factors in the NEA scenario. The Red factors all appear at or near the top of the figure, indicating that the same factors that are most important in the NEA scenario tend to be also the most important factors in the SWA scenario. The differences in the sensitivities between the two scenarios result from geometric and laydown differences inherent to those theaters.

If the initial experiment was designed to screen a large number of factors in a Resolution IV design, many two-way and three-way interactions are confounded, that is, only linear combinations of the two-way interactions can be estimated from the differences of means. When the data is reanalyzed for a smaller number of factors (for example, five factors), often all of the two-way and sometimes all of the three-way interactions among the five factors are unconfounded. However, the same combination of differences of means may estimate more than one of two-way and three-way interactions among the five factors. In this case, some of the interactions are confounded. Only by running an additional experiment to examine the specific five factors and their two-way and three-way interactions can be effects be separately estimated (resolved).

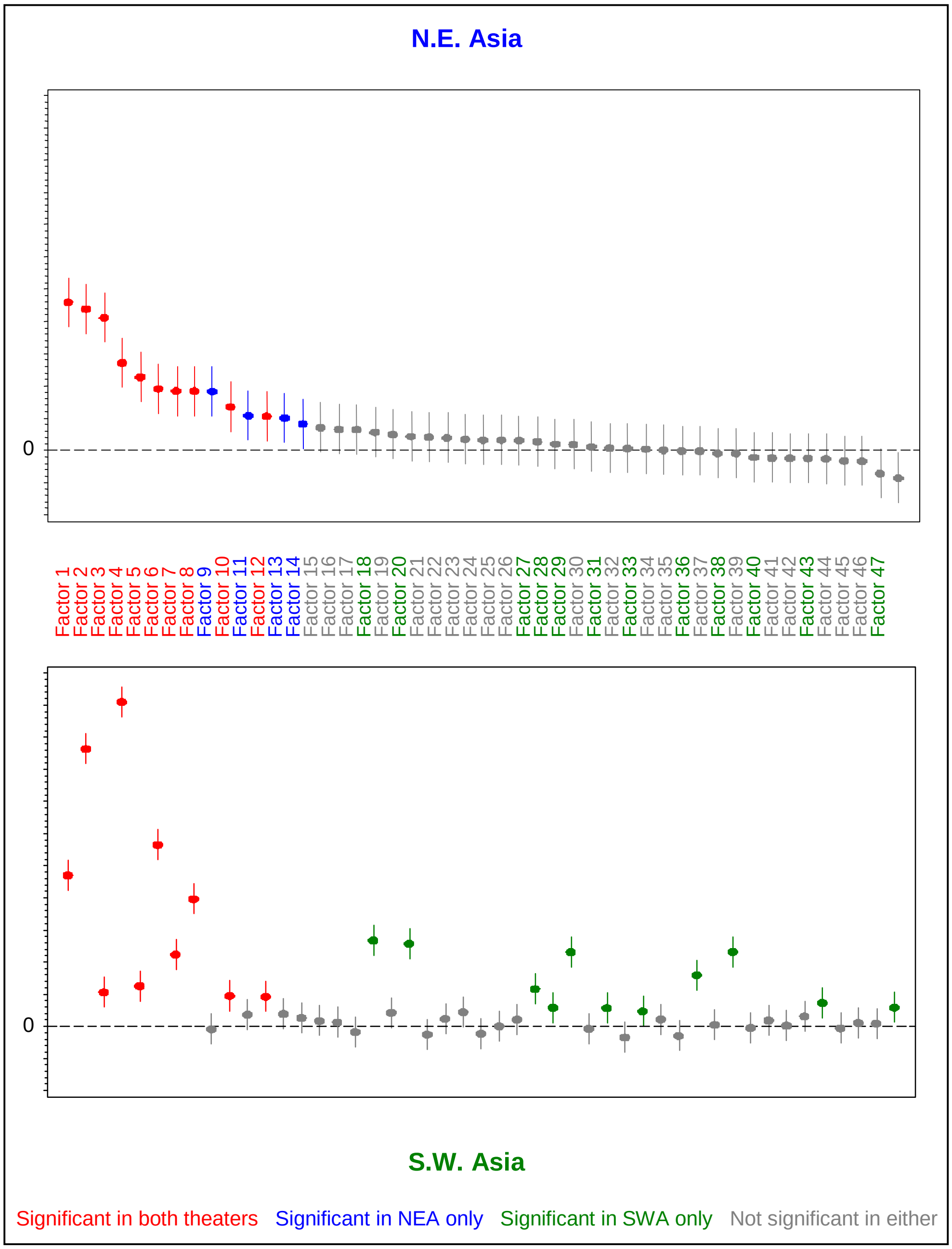


Figure 2. Protection Effectiveness: 47 Main Effects and 95% Confidence Limits

An example of a significant interaction effect can be seen in Figure 3, as the two lines in the interaction graph are not parallel. The increase in Protection Effectiveness from improving Factor 6 (denoted as F6 in the graph) is large if Factor 9 is at the low level, but essentially zero if Factor 9 is at its high level. (Factor 6 and Factor 9 are not the sixth and ninth values listed in Table 5.)

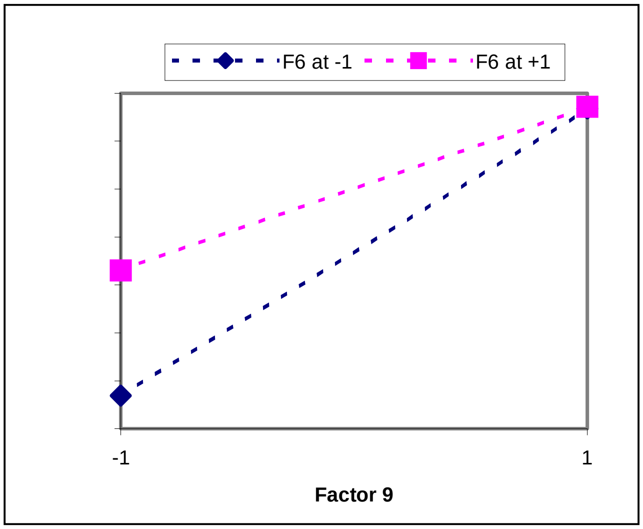


Figure 3. Protection Effectiveness: Two-way Interaction Between Factors 6 and 9 from the Screening Experiment

3.0 Response Design Methodology. Once the screening experiment has been performed and the important factors have been determined, the next step is to perform a response surface experiment. The polynomial equation that is frequently used to model the response surface is a quadratic model with cross-product terms is:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} \beta_{ij} X_i X_j + \sum_{i=1}^{p} \beta_{ii} X_i^2 \qquad (2)$$

where β_0 represents the overall mean response

β_i represents the main effects for each factor (i = 1, 2, ... , p)

β_{ij} represents the two-way interaction between the ith and jth factors

β_{ii} represents the quadratic effect for the ith factor.

In order to fit the fully second-degree polynomial in Equation (2), more than two levels for the X's variables are needed, usually three, that is, a "medium" as well as a "high" and "low" and these are coded +1, 0, and –1. The use of three levels can model possible curvature. A total of 3^k computer simulations are needed to take observations at all the possible combinations of the three levels of the k variables. If 2^k computer simulations is large, then 3^k computer simulations is much larger. This is the value of conducting the initial screening study to reduce k to a smaller number. Even so, 3^k computer simulations may still be prohibitively large.

There are three types of experimental designs that are commonly used for response surfaces: the central composite design, the three-level fractional factorial design, and the "optimal" designs. For the central composite design, the design points are the augmentation of the two-level fractional factorial with points on the faces of the hypercube (or further out if a rotatable design is desired) and at the center of the design space. For the three-level fractional factorial design, the design points are a subset of all the possible 3^p points in the design space. For the “optimal” design, the design points are selected by a statistical criterion such as minimizing the uncertainty on the estimated effects, the determinant of X′X, where X is the design matrix, which are called D-optimal designs. An “optimal” design is useful if too many points are required by a fractional factorial design or there is an irregular design space. Response surface methods are discussed in more detail in Box and Draper (1987).

The minimum number of runs needed for Resolution V designs for different numbers of factors are shown in Table 6. From the screening design, there are 11 main effects that were statistically significant and have at least a 1% effect on Protection Effectiveness. For 11 factors, Table 6 shows that a minimum number of 243 runs are needed. To provide a fair comparison among the three types of response surface designs, 243 new runs were made for each of the three types of designs. The central composite design is 10 replicates for each of the 22 faces of the hypercube plus 23 replicates of the center of the cube. The “optimal” design also contained 243 points.

Table 6. Three-level Resolution V Designs: All Main Effects and Two-way Interactions Unconfounded

Number of Factors	Minimum Number of Runs
1	3
2	9
3	27
4 – 5	$81 = 3^4$
6 – 11	$243 = 3^5$
12 - 14	$729 = 3^6$
15 – 21	$2{,}187 = 3^7$
22 – 32	$6{,}561 = 3^8$

The comparison of the three designs are shown in Tables 7 and 8. The standard deviations of the effects are generally minimized by the three-level fractional factorial design, with quadratic effect term having nearly twice as large standard errors for the central composite and “optimal” designs as compared with the fractional factorial design (Table 7). The number of statistically significant effects with effects estimated to be larger than 1% is generally largest for the three-level fractional factorial design, with fewer quadratic effects found by the central composite and “optimal” designs (Table 8). Therefore, the three-level fractional factorial seems to be the best design for estimating quadratic effects, which is the reason for the response surface experiment.

Table 7. Comparison of Three-level Resolution V Designs: Statistical Measures

Statistical Measure	3^{11-6} Fractional Factorial	Central Composite: Only Star and Center Points	D-Optimal
Det[(X′X)]	50	8 (no cross-products)	59
Standard Error:			
Main Effects	.0016	.0030	.0015
Two-way Interactions	.0019	--	.0016
Quadratic Effects	.0027	.0041	.0051

Table 8. Comparison of Three-level Resolution V Designs: Number of Significant Effects

Number of Significant Effects	3^{11-6} Fractional Factorial	Central Composite: Only Star and Center Points	D-Optimal
Main Effects	11	11	8
Two-way Interactions	7	5	8
Quadratic Effects	6	4	2

Examples of a significant quadratic main effect and a significant two-way interaction for the three-level fractional factorial response surface experiment are shown in Figure 4. Factor 6 and Factor 9 are not the sixth and ninth factors listed as in Table 5. Factor 6 is denoted as F6 in Figure 4.

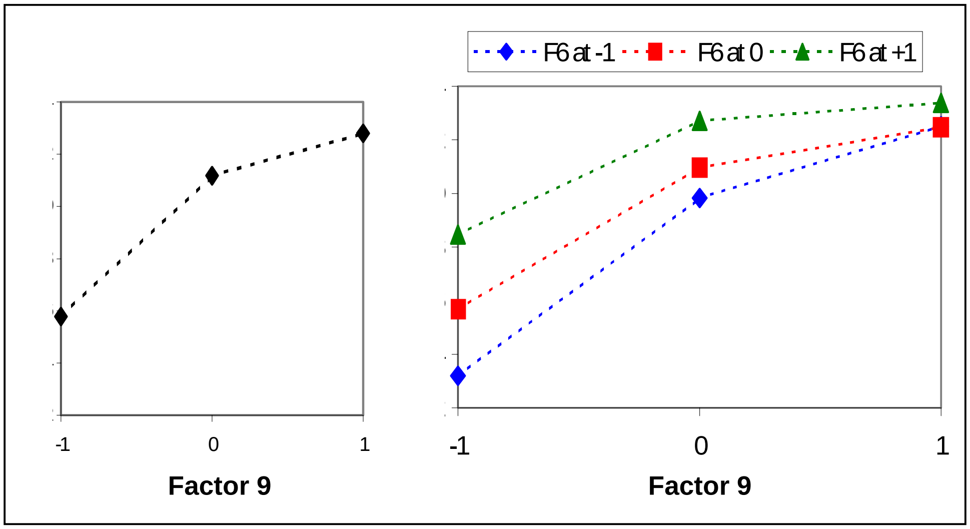


Figure 4. Protection Effectiveness: Quadratic Main Effect and Two-way Interaction Between Factors 6 and 9 from the Response Surface Experiment

The fitted model for Protection Effectiveness with quadratic and cross-product terms using the 3^{11-6} fractional factorial response surface experiment is as follows. The size of the effects are actually twice as large as the coefficients on the X terms since X has a range of 2 (from –1 to +1).

$$\begin{aligned}\text{P.E.} = {} & .938 + .035X_9 + .026X_{11} + .017X_5 + .016X_2 + .015X_6 + .014X_1 + .012X_7 + .011X_4 \\ & + .007X_3 + .006X_8 - .011X_6X_9 - .007X_8X_9 - .007X_2X_5 - .006X_5X_7 - .005X_3X_9 \\ & - .005X_5X_6 - .005X_1X_5 - .019X_9^2 - .011X_5^2 - .009X_{11}^2 - .008X_4^2 - .006X_3^2 - .006X_2^2\end{aligned}$$

Not only were there two theaters examined (NEA and SWA) but also at four force levels. All of the preceding analysis was conducted at a Force Level 4, which is comparable to the Desert Storm level of logistics support prior to the operation. Force Level 1 is a rapid response with no prior warning, and Force Levels 2 and 3 are intermediate between Force Levels 1 and 4. The response surfaces for the four force levels are shown in Figure 5. The individual graphs are the response surfaces for Factors 9 and 11, the two largest main effects for Force Level 4. There is very noticeable curvature for Factor 9, especially at the lowest two force levels. As the force level increases, Protection Effectiveness increases. The different color bands are 5% increments in Protection Effectiveness, with Red being between 65% and 70% and Orange being between 90% and 95%. Therefore, the response surfaces flatten out and raise up as the force level increases, and correspondingly Protection Effectiveness improves and is less sensitive to changes in the factors. As the force level increases, there are more assets of the same type, so the reliance on the performance of any individual asset diminishes.

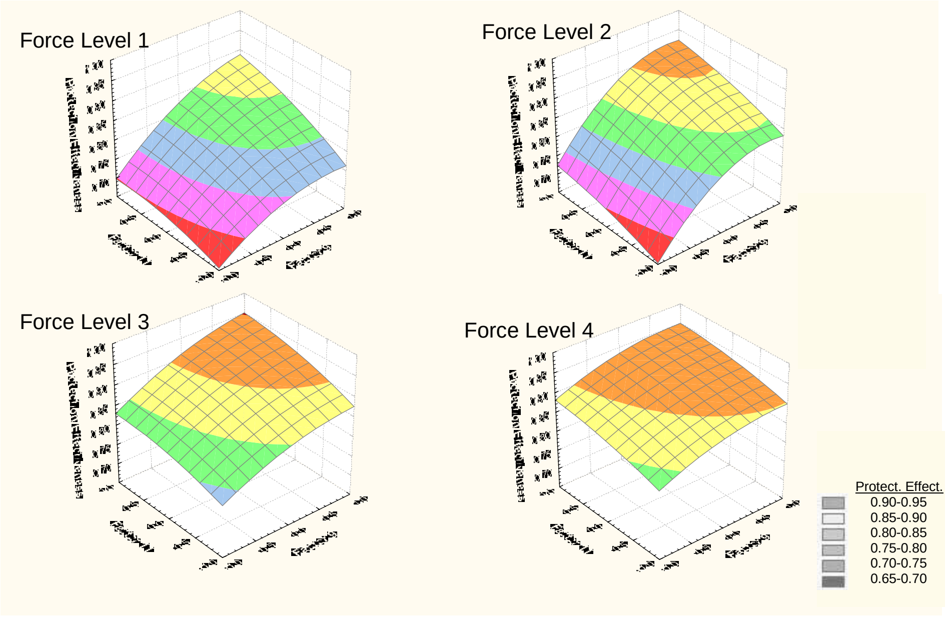


Figure 5. Protection Effectiveness Response Surfaces at Four Force Levels

4.0 Recommendations. The recommended experimental designs for the two steps in a sensitivity analysis are as follows.

1. Screening experiment: Use a two-level fractional factional design. If the number of factors is less than 32, use a Resolution V design. (If you can run more than 1,024 design points, the number of factors can be increased above 32 and Resolution V design can be used). Otherwise, use a Resolution IV design. To obtain some information on curvature, collect data at the center of design. Only one measurement at the center point is needed if the process is deteministic; if the process is stochastic, replicates are needed at the center point (10, 25, or 50 times, depending on the variability of the process). Even if curvature is indicated by an appropriate test, the factor causing the curvature cannot be identified.

2. Response Surface experiment: Use a Resolution V three-level fractional factorial design.

SAS is a registered trademark of the SAS Institute, Inc., Cary, NC in the U.S. and other countries.

References

Biles, W. E. (1979), "Experimental Design in Computer Simulation," in *Proceedings of the 1979 Winter Simulation Conference*, 3-9, Washington, DC: Institute of Electronics Engineers.

Box, G.E.P. and N. R. Draper. (1987), *Empirical Model Building and Response Surfaces*, New York: Wiley.

Box, G. E. P., W. G. Hunter and J. S. Hunter. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, New York: Wiley.

Daniel, C. (1962), "Sequences of Fractional Replicates in the 2^{p-q} Series,*" Journal of the American Statistical Association*, **57**:403-429.

Grier, J. B., T. G. Bailey and J. A. Jackson. (1999), "Response Surface Modeling of Campaign Objectives Using Factorial Analysis," *Military Operations Research*, **4**:2, 61-70.

Hunter, J. S. and T. H. Naylor. (1970), "Experimental Designs for Computer Simulation Experiments," *Management Science*, **16**:422-434.

Jacoby, J. E. and S. Harrison. (1962), "Multi-variate Experimentation and Simulation Models," *Naval Research Logistics Quarterly*, **9:**121-135.

Kleijnen, J. P. C. (1975), "Screening Designs for Poly-Factor Experimentation," *Technometrics*, **7:**487-493.

Law, A. M. and W. D. Kelton. (2000), *Simulation Modeling and Analysis (Third Edition),* New York: McGraw-Hill.

Montgomery, D. C. (2000), *Design and Analysis of Experiments (Fifth Edition),* New York: Wiley.

Weber, D. C. and J. H. Skillings. (2000), *A First Course in the Design of Experiments: A Linear Models Approach*, Florida: CRC Press.

CLEARANCE REQUEST FOR PUBLIC RELEASE OF DEPARTMENT OF DEFENSE INFORMATION

(See instructions on back.)

(This form is to be used in requesting review and clearance of DoD information proposed for public release in accordance with DoDD 5230.9.)

TO: Director, Freedom of Information & Security Review, Rm. 2C757, Pentagon

1. DOCUMENT DESCRIPTION

a. TYPE PAPER	b. TITLE Sensitivity Analysis Using Design of Experiments in Ballistic Missile Defense
c. PAGE COUNT 25	d. SUBJECT AREA Ballistic Missile Defense

2. AUTHOR/SPEAKER

a. NAME *(Last, First, Middle Initial)* TELFORD, JACQUELINE	b. RANK CIV	c. TITLE N/A
d. OFFICE N/A		e. AGENCY JHU/APL

3. PRESENTATION/PUBLICATION DATA *(Date, Place, Event)*

NEEDED FOR U.S. ARMY CONFERENCE ON APPLIED STATISTICS

CLEARED
FOR OPEN PUBLICATION

JAN 16 2002 14

DIRECTORATE FOR FREEDOM OF INFORMATION
AND SECURITY REVIEW
DEPARTMENT OF DEFENSE

4. POINT OF CONTACT

a. NAME *(Last, First, Middle Initial)* HOFF, SEWALL K.	b. TELEPHONE NO. *(Include Area Code)* (703) 697-8683

5. PRIOR COORDINATION

a. NAME *(Last, First, Middle Initial)*	b. OFFICE/AGENCY	c. TELEPHONE NO. *(Include Area Code)*
	BMDO/EA	

6. REMARKS

SRE LOG NUMBER: 01121901

COMPLETED

7. RECOMMENDATION OF SUBMITTING OFFICE/AGENCY

a. THE ATTACHED MATERIAL HAS DEPARTMENT/OFFICE/AGENCY APPROVAL FOR PUBLIC RELEASE *(qualifications, if any, are indicated in Remarks section)* AND CLEARANCE FOR OPEN PUBLICATION IS RECOMMENDED UNDER PROVISIONS OF DODD 5320.9. I AM AUTHORIZED TO MAKE THIS RECOMMENDATION FOR RELEASE ON BEHALF OF:

DIRECTOR, BMDO

b. CLEARANCE IS REQUESTED BY 02-01-11 *(YYYYMMDD)*.

c. NAME *(Last, First, Middle Initial)* Lt. Col Richard Lehner	d. TITLE Chief Media Relations
e. OFFICE EA	f. AGENCY BMDO
g. SIGNATURE	h. DATE SIGNED *(YYYYMMDD)* 01-12-19

02-S-0489

Model-based Methods for Biological Agent Identification in Mass Spectrometry

Fernando Pineda, Johns Hopkins University Applied Physics Laboratory

We describe a model-based pattern recognition approach for mass spectrometry. The approach is based on phenomenological probability density functions that characterize the distribution of molecular masses in the spectrum. In contrast to traditional template-based techniques, which rely on libraries of template mass spectra, our methods use phenomenological distributions derived from databases of protein sequence data. We describe a hypothesis testing approach that yields robust classification performance. If time permits we will also describe progress towards a Bayesian Belief Network approach. This approach potentially leads to intractable calculations that we partially mitigate by applying approximate algorithms based on statistical field theory approximations (i.e. saddle point).

Contributed Session VII

Identifying Storms in Noisy Radio Frequency Data via Satellite: an Application of Density Estimation and Cluster Analysis

Tom Burr, Angela Mielke, Abram Jacobson
Mail Stop E541 Los Alamos National Laboratory
Los Alamos, NM 87545

Abstract

The FORTE (Fast On-Orbit Recording of Transient Events) satellite collects records of radio frequency events that exceed a threshold. Here we consider data for the observed phenomena of "recurrent-emission storms." Each data point represents the total electron content (TEC) of the emanation of a 409.6 μ-s collection window. Some data records contain well-defined storm events which consist of many data points in a specialized cluster. We present a method involving noise rejection and cluster analysis to identify well-defined storms from the data records. We first remove noise using density estimation and then apply hierarchical clustering to the higher-density points. For each identified cluster of points, we fit TEC as a quadratic function of time (a quadratic shape is anticipated from atmospheric physics), and find more points that belong to the cluster using a careful extrapolation. The overall performance of finding each storm and identifying which points belong to which storm is assessed by comparing our results to test data produced by a human analyst. We also give results for three other mixture-fitting methods: (1) principal curve clustering, (2) a method using the integrated squared error norm, and (3) a method using the expectation-maximization algorithm.

1 Background and Data Description

The FORTE (Fast On-Orbit Recording of Transient Events) satellite collects records of radio frequency (RF) events that exceed a threshold ((Moore, 1995) and (Jacobson, 1999)). Here we consider data for the observed phenomena of "recurrent-emission storms." This data emanates from a single RF source that radiates while the satellite passes overhead. Each data point of a micro event represents the total electron content (TEC) of the emanation of a 409.6 μs collection window. Each data record contains approximately 100 to 400 micro events and is processed with a dechirping step that corrects for the frequency-dependent transit time through the atmosphere. The total record time is approximately 15 minutes which is the time for the satellite to pass over a region of the earth. A data record in our study contains 100 to 400 micro events ("points"). As defined here, "data records" were produced from a database query for at least 100 events and over a certain region of the earth.

Most data records from our query contain storms which consist of many data points (micro events) in a specialized cluster shaper. Atmospheric dispersion models suggest that a fully viewed storm will consist of data points in a bowl shape. Because most of the bowl is concave up, and some false positives are concave down, we will restrict attention to concave up feature clusters. We have empirically determined that a quadratic fit is adequate to describe most clusters of interest.

We present a method involving noise rejection and cluster analysis to identify well-defined storms from the data records. We first remove noise using density estimation and then apply hierarchical clustering to the higher-density micro events. For each identified cluster of micro events, we fit TEC as a quadratic function of time, and find more micro events that belong to the cluster using a careful extrapolation. Because some data records contain false alarms having many points generated over a very short time, a final inspection of the found clusters includes diagnostic checks of the time duration and of the residual variance of the points around the quadratic shape. We also reject concave down clusters. The overall performance of finding each storm and identifying which micro events belong to which storm is assessed by comparing our results to test data produced by a human analyst.

1.1 Examples

We plot examples using the four data sets D6, D7, D18 and D16 (of 30 analyzed) in Figure 1. Figure 1a is the easiest example in the sense that we expect to find the one cluster, probably without any false positives. There is only one storm in Figure 1a so the number of clusters is $K = 1$. In all cases, we use the following labeling convention. The first cluster (storm) is labeled with a 1, the second with a 2, etc., and the noise points are all labeled with the integer $K + 1$. Figure 1b is somewhat harder because there are two clusters quite close together. Figure 1c is harder still, and we might expect to have several false positives and/or false negatives in Figure 1d which is the hardest.

1.2 Performance Measures

Our goal is to find quadratic-shaped upward clusters and to do so with a small false positive rate. We can accept a relatively large false negative rate because we expect thousands of event records from which we need to extract well-characterized storms. The data from identified storms (clusters) will be linked to ground-based data for further analysis as described in Jacobson et. al. (1999) and to be extended in future work.

1.3 Our Approach

After some initial trial and error with several methods, we chose an iterative procedure that we call method 1 with the following steps.

1. Reject noise points (all rejected points are candidates for inclusion with a cluster later).

2. Get cluster result A with optimal values, and result B with near optimal values.

3. For each cluster, extrapolate using a quadratic fit and a "zone of ownership" to avoid ambiguous points. Each cluster center defines a zone of ownership using uncertainty bands that widen as the extrapolation distance increases. If a point "belongs" to two or more zones of ownership, then it is ambiguous and assigned to the noise class.

4. Compare A and B results. For each cluster in A that is confirmed in B, accept the cluster as a storm. We also apply diagnostic checks to reject found clusters with very short time duration or a concave down shape or a very large residual variance around the quadratic shape. For example, D10 contains many points over a very short time range that appear to any algorithm to be a cluster but are the result of calibration or data corruption.

Figure 2 illustrates method 1 without the initial noise removal for data sets D1, D3 and D8. Note that we do not find the 2 clusters in D1. Figure 3 illustrates method 1 with the initial noise removal for the same three data sets and note that we do find the 2 clusters in D1.

We now describe steps 1-4 in more detail.

1. Noise rejection. Do an initial noise rejection (all points removed are candidates to be added back in step 3) using the distance to the nearest $k_1, k_2, \ldots, k_5$ neighbors. We also tested a density estimation scheme that counted the number of neighbors within distances $r_1, r_2, \ldots, r_5$ of each point. Several composite measures for identifying noise were then implemented and tested. For example, we used the first 2 principal components (PCs) of the 10-dimensional data (5 k-nearest neighbor results and 5 density estimation results). However, the most basic k-nearest neighbor method with $k = 3$ to 5 worked nearly as well as more involved methods. For example, using $k = 5$, the false positive rates ranged from 0.03 to 0.33 for a range of thresholds while the associated false negative rates ranged from 0.79 to 0.36. A method based on the PCs of the 10-dimensional data had false positive rates ranging from 0.02 to 0.32 with associated false negative rates of 0.79 to 0.35. Because these rates are essentially the same, we use the distance to the 5th nearest neighbor to identify noise. In all cases, it was better to scale TEC and time to unit variance so we assume TEC and time have been rescaled in the remainder of our discussion.

2. Cluster results A and B. We used hierarchical clustering with single linkage (the distance between clusters is the minimum distance between a point in the first cluster and a point in the second cluster). This is called method = "connected" in Splus, and it favors long and thin clusters. We cut the hierarchical tree at a high percentile (approximately the 0.99 percentile) of all between-group distances to select the number of clusters. Six parameters are involved: 2 noise rejection thresholds (relative to the scene f_1 and absolute parameter f_2), 3 factors related to hierarchical clustering: cut the hierarchical tree at some high percentile f_3, reject clusters with small relative number of observations f_4, reject clusters with small relative number of observations f_5, and 1 factor f_6, related to extrapolation from original cluster. The factor f_6 is the fraction of the range of the original cluster to allow in extrapolation to identify new feature points. We identified good nominal values of these 6 parameters using data set D1. D1 is a good training example because it is moderately difficult, having 2 distinct but somewhat close clusters. We then used a low value of half the nominal and a high value of twice the nominal and searched over 3^6 runs using all 30 data sets to find good values. Good values had the lowest false negative rates subject to having a very low false positive rate. The result is that the optimal values over 30 data sets are approximately the same as those chosen from D1. Therefore, we use the values selected from exploratory analysis with D1 as the "result A" values and values that differ slightly from these are the "result B" values.

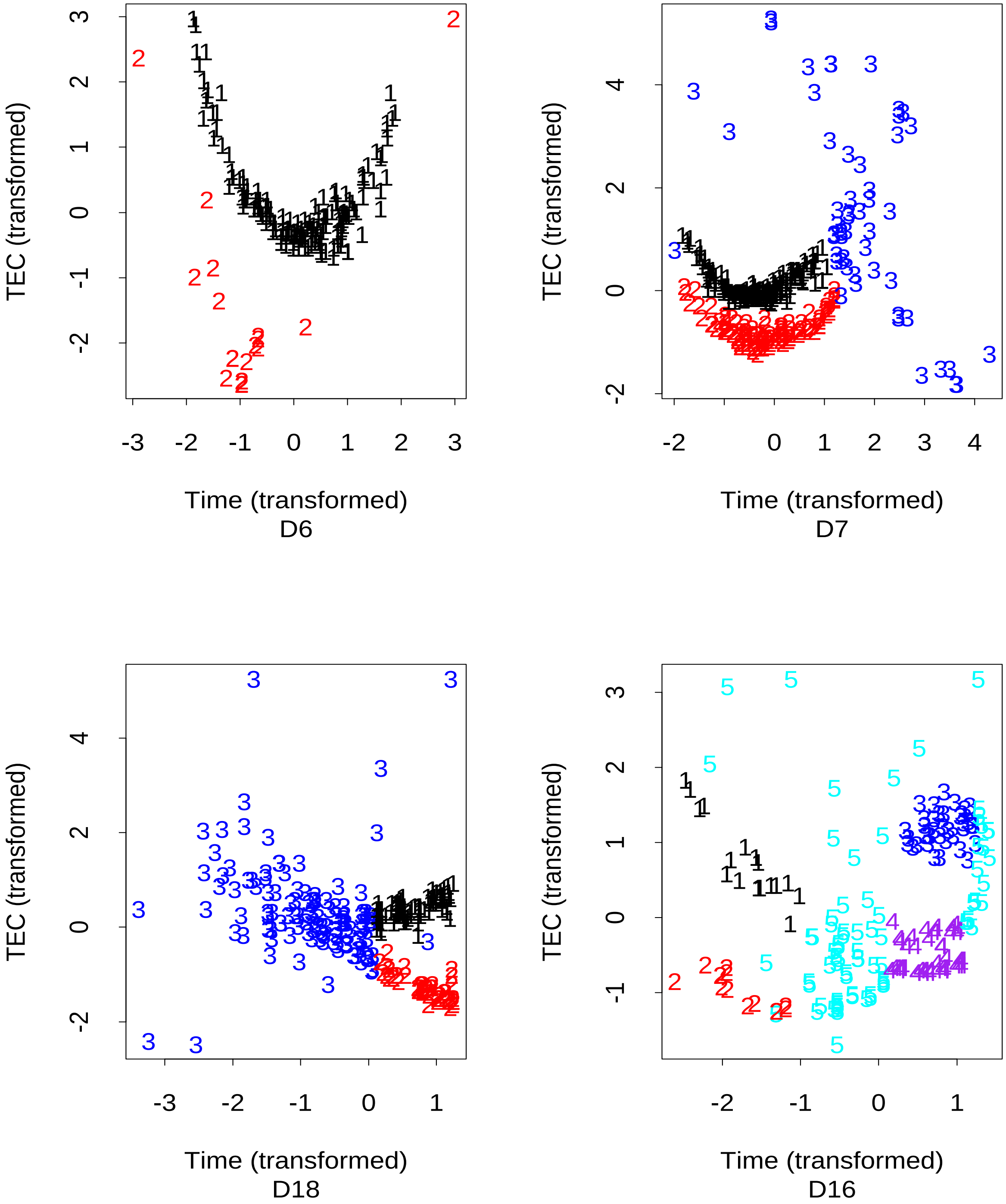


Figure 1. Examples in order of increasing difficulty, found in data sets D6, D7, D18, and D16.

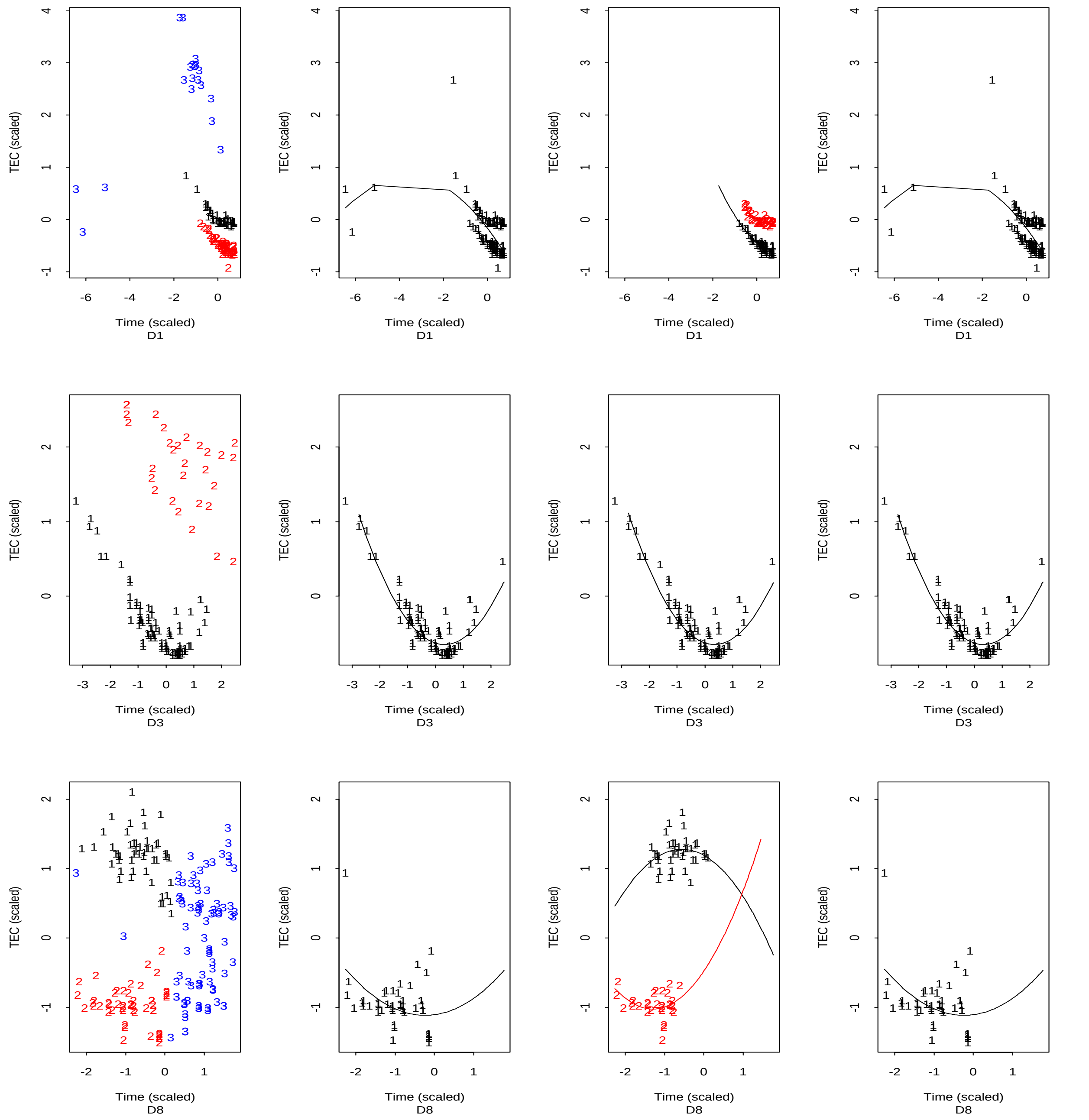


Figure 2. Example results using Method 1 without noise removal. The correctly labeled data is in column 1. The results using parameters A are in column 2. The results using parameters B are in column 3, and the final result is in column 4.

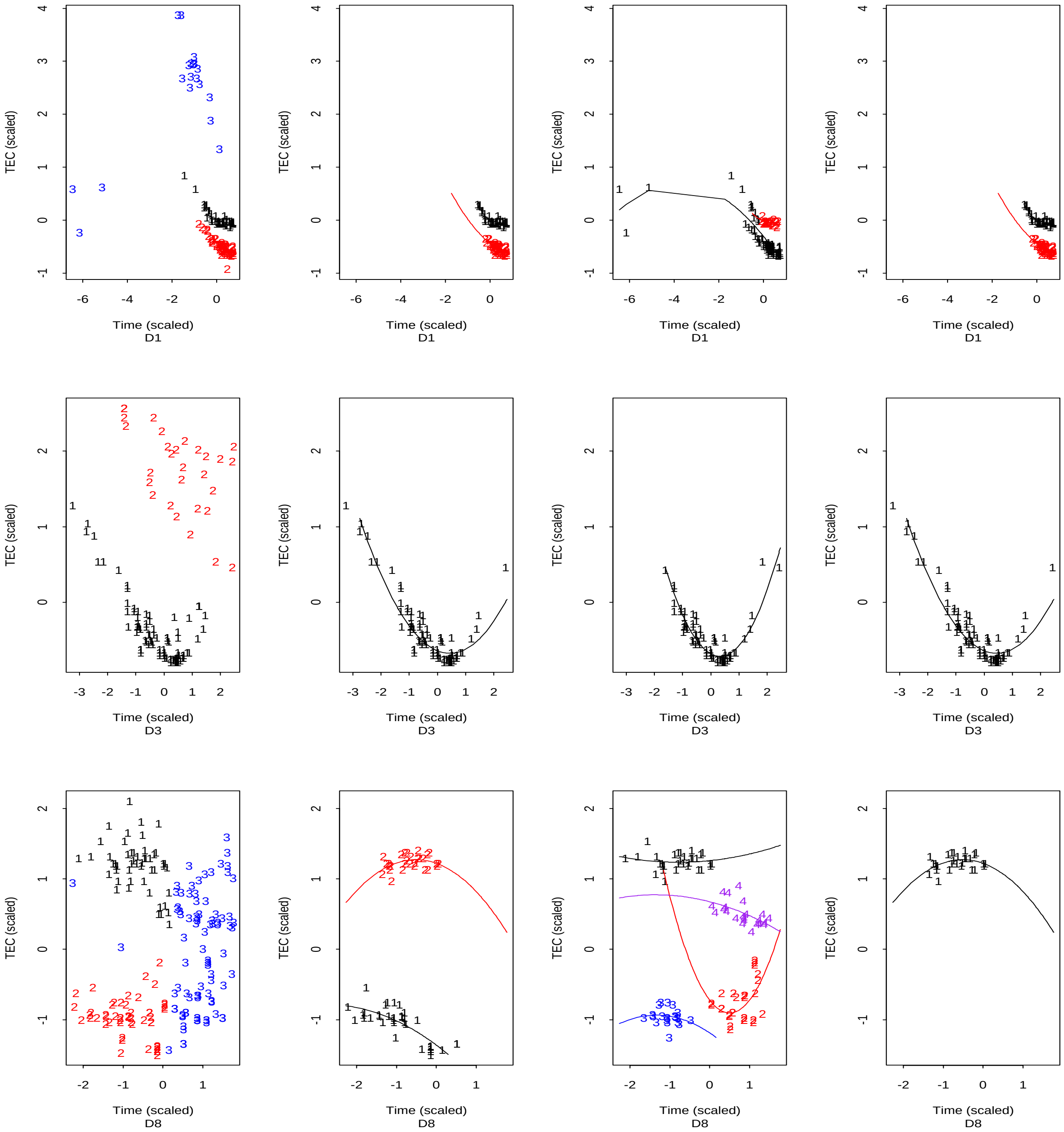


Figure 3. Same as Figure 2, except with noise removal.

3. Extrapolate outward from each cluster to identify new points. Some of the clusters have gaps so we anticipated the need to extrapolate outward from each central cluster in order to identify points that were clearly part of the storm. To do so, we fit TEC (scaled) as a quadratic function of time (scaled), and used the least squares estimates of uncertainty in the fitted function (the variance/covariance matrix of the fitted intercept, linear, and quadratic terms) to choose the "zone of ownership" of a cluster. If two or more clusters appeared to "own" a point then the point was ambiguous and labeled as a noise point. To control the false positive (FP) rate, we did not allow extrapolation to extend too far as a function of the range of the original core cluster. Otherwise, small, vague clusters would "own" too many noise points leading to a high FP rate.

4. Cluster stability check. We used two clustering results (A and B) to check for cluster stability and provide a final control on the FP rate. If and only if a cluster was found in both A and B and passed the three diagnostic checks (time duration, direction of curvature, and residual variance) would the final result include that cluster.

2 Results

We queried archived FORTE records and manually created 30 training cases, each having 0 to 5 clusters. Most cases had 1 or 2 clusters, but a few had 3, 4, or 5, and one case had 0 clusters.

We must define when a cluster has been found, which is somewhat subjective. Also, for each found cluster we define the false positive and false negative rates (fpr and fnr) as the number of false positive points divided by the true number of points in the class and the number of false negatives to be the number of missed points divided by the true number of points in the class.

Concerning whether a cluster has been found, consider the simulated data case in Figure 4. This simulated example is very challenging because of the nature of the overlap of the two clusters. The results from each of four methods (the other three methods are described in Section 3) are shown in Figure 5.

This simulated example raises two questions related to performance measures:

(A) Should we define the effective number of clusters by using some notion of cluster separation?

(B) Should we consider our performance with method 1 in the simulated example to be one false positive and two false negatives?

Because the guessed cluster using method 1 is a hybrid of the two clusters, in our application, a better performance would be to find no clusters in this example. That is, large contamination of a cluster with members of other clusters is undesired, so we prefer to strongly penalize this performance. Our current performance measure works as follows. For the first true cluster (we order by true cluster size, so the first is the largest true cluster) define the guessed cluster by "majority rule," which means that the most common guess is defined to be "the guess." For the second largest true cluster, again define the guessed cluster by "majority rule," but only using previously unassigned cluster labels. Continue for all true clusters. This defines the number of false negatives. If the "majority rule" implies that the guess is the noise cluster, then accept the second-to-largest frequency class as the guessed class. This is a somewhat liberal choice. Reverse the roles of "guess" and "true" to define the number of false positives. Therefore, in Figure 5, method 1 and the L2E method both resulted in one false positive and one false negative with a large false positive rate for the one found cluster. The mixreg and princurve methods each have 2 false negatives.

To define a measure of difficulty for each case, we use the ratio of the within-to-between feature distances. We have experimented with using the median, mean, and minimum ratio for each pair of features and found that the median ratio is as good as any other choice and sometimes is much better. Therefore, currently the median ratio of the within-to-between feature distances is our measure of difficulty. We do not yet try to define a measure of difficulty for a case with no features.

The results on the 30 real data sets for method 1 using the false positive and negative definitions described above are:

false positives: 0 (found 24 of 59); fpr: 0.16, fnr:0.15; false negatives: 35/59.

The false negatives tended to appear for the more difficult cases as defined by the median ratio of the within-to-between class distances. We also evaluated three other methods, each described in the next section.

3 Other Approaches

Here we describe three other approaches and give results from each. These three methods treat this as a mixture fitting problem but the details differ among the three methods. Mixture fitting is known

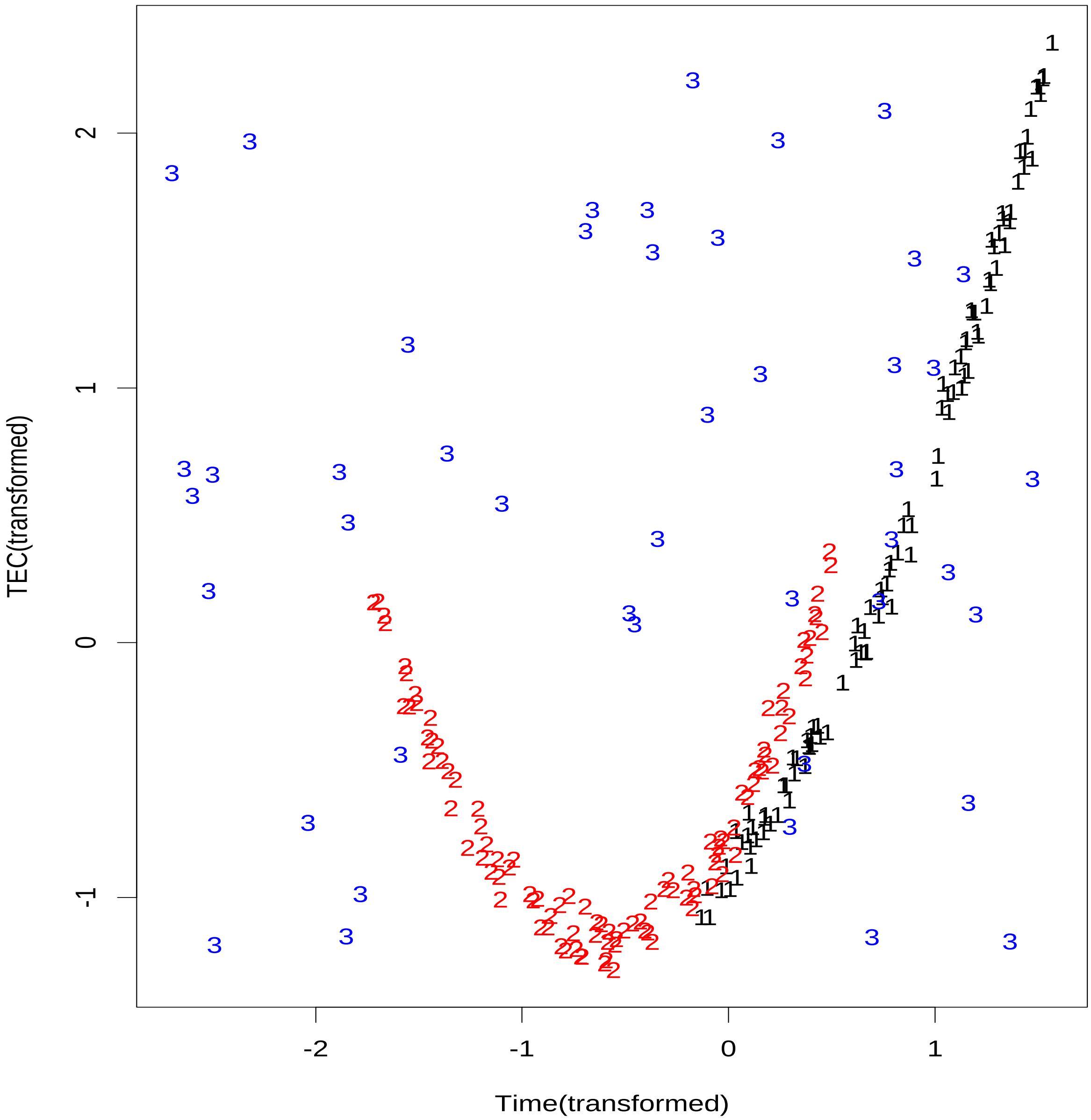


Figure 4. Simulated example with overlapping features

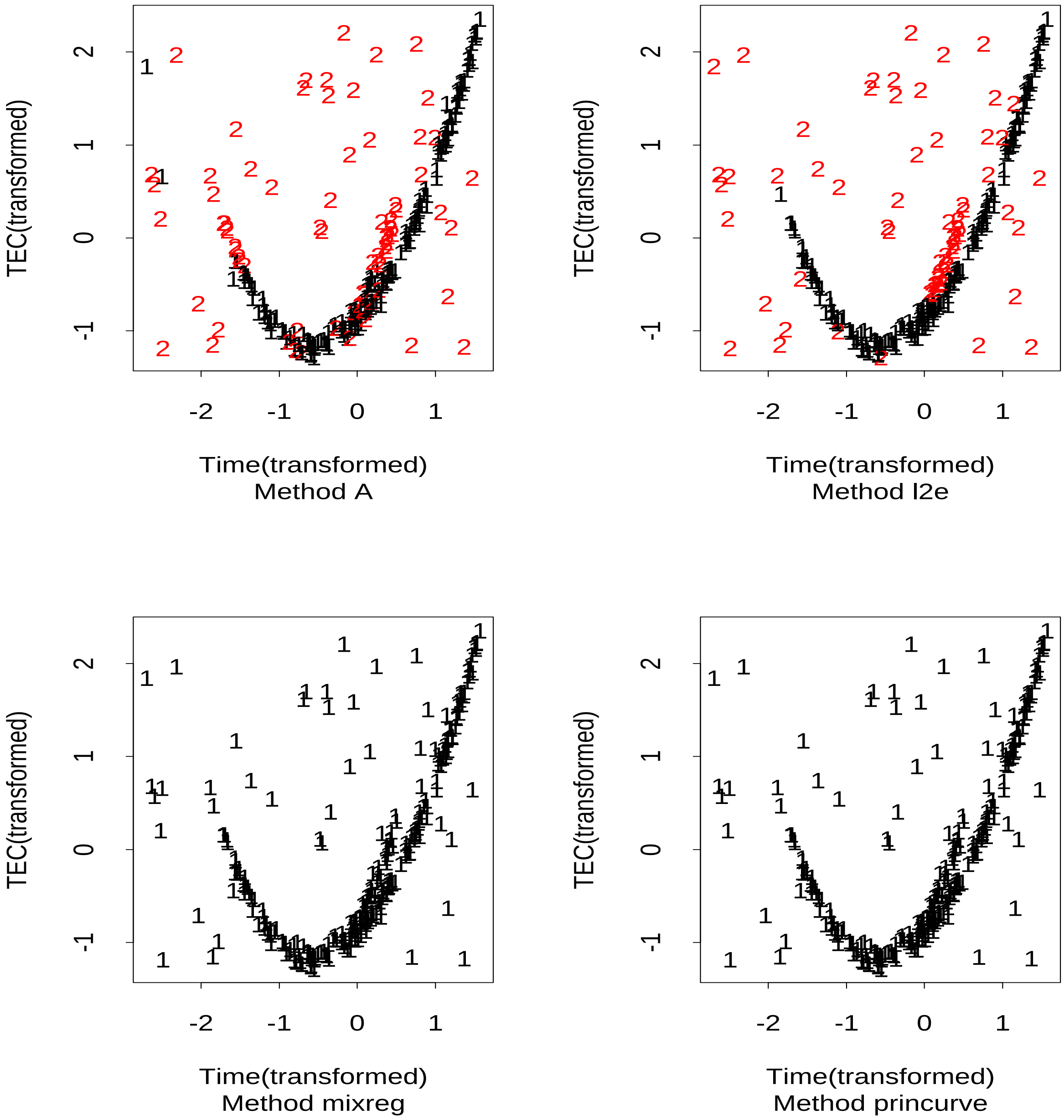


Figure 5. Results with each of the 4 methods for the simulated data in Figure 4.

to be difficult. In all cases the mixture model for the cluster classes assumes that the errors around the quadratic fit are Gaussian. The stochastic component of each model for each observation x is therefore some version of $\sum_{k=1}^{K} \pi_k \phi(x|\mu_k \sigma_k^2)$, where π_k are the relative fractions in each cluster, μ_k is the mean of cluster k and σ_k^2 is the variance of cluster k, and ϕ is the Gaussian probability density. We use the notation of the cited references for each method below, so both the approach and the notation differ among the three methods.

Principal curve clustering

Principal curve clustering with noise was developed by Stanford and Raftery (2000) to locate principal curves in noisy spatial point process data. Although our data is time series, the technique is a viable candidate when we view the predictor as time and the response as TEC. A principal curve is a smooth curvilinear summary of p-dimensional data (Hastie and Stuetzle, 1989). It is a nonlinear generalization of the first principal component line observations. Stanford and Raftery developed an algorithm that first uses hierarchical principal curve clustering (HPCC, which is a hierarchical and agglomerative clustering method) and next uses an iterative relocation (reassign points to new clusters) based on the classification estimation-maximization (CEM) algorithm. A probability model uses the principal curve probability model for the feature clusters and a homogeneous Poisson process model for the noise cluster. Because our features are approximately quadratic in shape, a principal curve with 2 to 5 degrees of freedom should be an adequate fit. Future work will consider using a probability model that uses a quadratic fit using HPCC, although we do not anticipate that this will have much impact.

The noise in most cases appears to be a non homogeneous Poisson process with a strong tendency for the noise cluster to appear in one or a few regions of the scene. Therefore, we modified the noise model by assuming Poisson-distributed noise only within the range of the observed noise. Because the procedure must identify noise, this implies that we should adaptively choose the noise regions of the scene, and also perhaps the noise model. At present, we use a noise model from the initially identified noise points.

Let X denote the set of observations, $x_1, x_2, \ldots, x_n$ and C be a partition considering of clusters $C_1, C_2 \ldots, C_{K+1}$ where the cluster C_j contains n_j points. The noise cluster is C_{K+1} and assume feature points are distributed uniformly along the true underlying feature so their projections onto the feature's principal curve are randomly drawn from a uniform $U(0, \nu_j)$ distribution, where ν_j is the length of the jth curve. An approximation to the probability for $0, 1, \ldots, K$ clusters is available from the Bayesian Information Criterion (BIC), which is defined as $BIC = 2\log(L(X|\theta) - M\log(n)$, where L is the likelihood of the data X, and M is the number of fitted parameters, so $M = K(DF + 2) + K + 1$ For each of K features we fit 2 parameters (ν_j and σ_j defined below) and a curve having DF degrees of freedom. There are K mixing proportions (π_j defined below) and the estimate of scene area is used to estimate the noise density. The likelihood L satisfies $L(X|\theta) = \prod_{i=1}^{n} L(x_i|\theta)$, where $L(x_i|\theta) = \sum_{j=0}^{K} \pi_i L(x_i|\theta, x_i \in C_j)$ is the mixture likelihood (π_j is the probability that point i belongs to feature j), $L(x_i|\theta, x_i \in C_j) = (1/\nu_j)(1/\sqrt{2\pi}\sigma_j)\exp(-\frac{||x_i - f(\lambda_{ij})||^2}{2\sigma_j^2})$ for the feature clusters, and $||x_i - f(\lambda_{ij})||$ is the Euclidean distance from x_i to its projection point $f(\lambda_{ij})$ on curve j, and $L(x_i|\theta, x_i \in C_j) = 1/Area$ for the noise cluster.

Briefly, the HPCC-CEM steps are:

(1) Make initial estimate of noise points and remove then;

(2) Form an initial clustering with at least seven points in each cluster (we use clara for fast clustering in R);

(3) Fit a principal curve to each cluster;

(4) Calculate a clustering criterion $V = V_{About} + \alpha V_{Along}$, where $V_{About} = \sum_{j=1}^{n} ||x_i - f(\lambda_{ij})||^2$ and $V_{Along} = 1/2 \sum_{j=1}^{n} ||\epsilon_j - \overline{\epsilon}||^2$ and $\epsilon_j = f(\lambda_j) - f(\lambda_{j+1})$, so that V_{About} measures the orthogonal distances to the curve ("residual error sum of squares") and V_{Along} measures the variance in arc length distances between projection points on the curve. Minimizing V (the sum is over all clusters) will lead to clusters with regularly spaced points along the curve, and tightly grouped around it. Large values of α will cause the method to avoid clusters with gaps and small values of α favor thinner clusters. Clustering (merging clusters) continues until V stops decreasing.

Because we believe that data set D1 is a good training set, we evaluated the BIC for D1 for DF ranging from 2 to 6, a range of candidate numbers of clusters, and a range for the parameter α (around the nominal value of 0.4). Conventionally, differences of 2-6 between BIC values represent positive evidence for the model having larger BIC. The goal was to find the α and DF values that caused the BIC to maximize at the correct number of clusters. Unfortunately, with D1 the "maximize BIC" algorithm always chose 3 or 4 clusters for any values of α and DF. Therefore, we altered the probability model for the noise cluster slightly to reflect the fact that the noise tended to locate near the features rather than be randomly scattered throughout the scene. We did this by adjusting the area for the noise cluster to agree closely with the observed area of the noise. The results remained the same which indicates either that the BIC approximation is not particularly effective or that the groups have been inconsistently defined by the human expert. In a similar experiment, we calculated the BIC for the correctly partitioned data and for an estimate of the partition. BIC was maximum for the correct labels in only 10 of the 30 cases with the scene-wide Poisson noise model and in only 11 of the 30 cases with the local Poisson noise model. It might be possible to use the BIC to define a degree of difficulty of each case. For example, if the BIC for the true labels is very close to the BIC for a fairly major rearrangement of the labels, then the case would be considered difficult. Recall, we currently define the measure of difficulty as the median ratio of ratio of the within-to-between feature distances.

The results for the HPCC-CEM method are:

false positives: 4; fpr: 0.11, fnr: 0.08; false negatives: 34/59.

We also modified the results of the HPCC-CEM method by including our diagnostics (direction of curvature, residual variance for each cluster, and duration of each cluster) and the strategy of choosing two sets of slightly different parameters (each set nearly optimal for D1) to get two clustering results, and then accepting only those clusters that were found by both clustering results.

The results for this modified HPCC-CEM method are

false positives: 1; fpr: 0.02, fnr: 0.04; false negatives: 50/59.

We note that this modification results in a very large false negative rate: We detect only 9 of 59 clusters and find one false positive. Although the "per found cluster" false positive and negative rates are impressively low, it appears that this modification is too cautious in conjunction with the principal curve clustering method.

Minimum Integrated Square Error (ISE) Method

The minimum integrated squared error (or L2 distance) appears to be a good approach to fit mixture models, including mixtures of regression models. (Scott, 2002 and Scott and Szewczyk, 2002). The minimum L2 distance method tries to find the largest portion of the data that matches the model. In this case, we seek feature 1 having the most points, regard the remaining points as noise, remove the feature and then repeat the procedure in search of the feature 2, and so on until a stop criterion is reached. It should also be possible to estimate the number of components in the mixture in the first evaluation of the data but we will not consider that approach here. To motivate the form of the L2 estimator, consider estimating the smoothing parameter h in density estimation with $\hat{h} = \text{arg min} \hat{\int}[f_h(x) - f(x)]^2 dx$. By interpreting terms and ignoring which does not depend on h, we find that $\hat{h}$ should satisfy $\hat{h} = \text{arg min}_h[\hat{\int}\hat{f}_h^2(x)dx - 2\hat{\int}\hat{f}_h(x)f(x)dx]$. The key quantity to estimate is the second term, which is the expected height of the density estimate (the first integral can be evaluated exactly for any h so it does not require estimation). It follows that a reasonable estimator is $\hat{h} = \text{arg min}_h[\frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_k \nu_k^2]$, where ν_k is the bin count of bin $B_k = (x_0 + kh, x_0 + (k+1)h)$ and x_0 is the bin origin. Methods to estimate h arose from nonparametric density estimation but Scott (2002) has shown that in the parametric setting with model $f(x|\theta)$, instead of estimating h we estimate θ using $\text{arg min}_\theta \hat{\int}[f_h(x|\theta) - f(x|\theta_0)]^2 dx$ where the true parameter θ_0 is unknown. Again the key quantity to estimate is the expected height of the density, $\int f(x|\theta)f(x|\theta_0)dx$ and again by interpreting terms it follows that a reasonable estimator minimizing the parametric ISE criterion is $\hat{\theta}_{L2E} = \text{arg min}_\theta[\int f(x|\theta)^2 dx - 2/n\sum_{i=1}^n f(x_i|\theta)]$.

This assumes that the correct parametric family is used. To achieve robustness the concept can be extended to include the case in which the assumed parametric form is incorrect. If we focus on the distribution of the residuals in regression then the L2E criterion can fit mixtures of regression models which is a good model for our data. Our data is a mixture of 0 to 5 features plus feature noise and scene noise. Each feature consists approximately of a quadratic feature for which regression of TEC on time and time^2 should provide a good fit. David Scott kindly provided Splus code for fitting a concave up quadratic using the L2E method. The results of this method on data set D1 and D3 are shown in figures 6 and 7 respectively. The top left plot is the original data D1 (transformed). The top right plot is the first feature found ("feature 1"). All points in feature 1 are removed and the procedure is repeated. The middle left plot is the data with feature 1 removed. The middle right plot is the next feature found on the remaining data ("feature 2"). The bottom left plot is the remaining data with both features removed. We chose stop criteria (involving the number of points in each feature and residual sum of squares in each feature) that correctly stopped finding features after these 2 features. The bottom right plot indicates the final result which is qualitatively in good agreement with the human analyst's labels (both features were found). However, not all feature points agreed with those of the human analyst. The results of this same procedure for data set D3 (figure 7) indicate that feature 1 was judged to be 2 features. Therefore, the qualitative performance is not as good with data set D3. Overall, using stop criteria that were optimized for D1, we found 33 of 59 clusters with 26 false positives.

The results for the L2E method are:

false positives: 26; fpr: 0.33, fnr: 0.17; false negatives: 26/59.

Because of this high false positive rate, we also implemented the L2E method by insisting on finding at most 2 or at most 1 cluster.

The results for L2E with "at most 2" clusters imposed are:

false positives: 17; fpr: 0.35, fnr: 0.16; false negatives: 26/59.

The results for L2E with "at most 1" cluster imposed are:

false positives: 1; fpr: 0.34, fnr: 0.16; false negatives: 32/59.

We also modified our results for L2E with "at most 1" cluster imposed, using the same diagnostics and "comparison of two cluster results" as in the other two methods. The modified results are:

false positives: 0; fpr: 0.26, fnr: 0.18; false negatives: 40/59.

Although we found zero false positives, we detect only 19 of 59 clusters. It appears that this modification is too cautious (false negative rate is too high) in conjunction with the L2E method. However, L2E does well at finding the single largest cluster.

Mixreg method

Rolf Turner (2000) implemented a method to handle a mixture of regression models via the well-known EM (estimation - maximization) algorithm (Dempster, Laird, and Rubin, 1977) The function mixreg is available from statlib (http://lib.stat.cmu.edu) for use in Splus (1999).

The likelihood L satisfies $L(X|\theta) = \prod_{i=1}^{n} L(x_i|\theta)$, where $L(x_i|\theta) = \sum_{j=1}^{K} \pi_i L(x_i|\theta, x_i \in C_j)$ is the mixture likelihood (π_j is the probability that point i belongs to feature j) and $L(x_i|\theta, x_i \in C_j) = f_{ij} = (1/\sigma_j)\phi(\frac{y_i - X_i\beta_j}{\sigma_j})$ where ϕ is the standard Gaussian distribution.

Introduce indicator variable z_i of which component of the mixture generated observation y_i and iteratively maximize $Q = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \ln(f_{ik})$ with respect to θ where θ is the complete parameter set of the model consisting of vectors of regression coefficients, the variances σ_i^2, and the mixing probabilities π_j for each class. The γ_{ik} satisfy $\gamma_{ik} = \pi_k f_{ik} / \sum_{i=1}^{K} \pi_k f_{ik}$. Then Q is maximized with respect to θ by weighted regression of y_i on $x_1, \ldots, x_n$ with weights γ_{ik} and each σ_k^2 is given by
$\sigma_k^2 = \sum_{i=1}^{n} \gamma_{ik}(y_i - x_i\beta_k)^2 / \sum_{i=1}^{n} \gamma_{ik}$. In practice the components of θ come from the value of θ in the previous iteration and $\pi_k = 1/n \sum_{i=1}^{n} \gamma_{ik}$.

A difficulty in choosing the number of components in the mixture, each with unknown mixing probability, is that the likelihood ratio statistic has an unknown distribution. Therefore, the mixreg function suite includes a bootstrap strategy (implemented in the function bootcomp) to choose between 1 and 2 components, between 2 and 3, etc. The strategy to choose between K and $K+1$ components is: (a) calculate the log-likelihood ratio statistic Q for a model having K and for a model having $K+1$ components; (b) simulate data from the fitted K component model; (c) fit the K and $K+1$ component models to each simulated data set and calculate the corresponding bootstrap log-likelihood ratio statistic Q^*; (d) compute the p-value for Q as the $p = 1/n \sum_{i=1}^{n} I\{Q \geq Q^*\}$.

The bootcomp function actually implements a semi parametric bootstrap described in Turner (2000). However, bootcomp requires a very long run time and did not lead to an improvement over a slightly unfair strategy involving using the correct data partition in the initial guess for mixreg.

We also include mixreg results using the correct partition with fitted quadratic relations as starting values. Such excellent starting values are not available in practice but allow us to assess how well this version of mixreg could be in the best case.

The mixreg results (with "excellent" starting values) are:

false positives: 4; fpr: 0.20, fnr: 0.06; false negatives: 35/59.

4 Summary

The results presented suggest that either of the 4 approaches presented can find the single dominant feature (storm) in most examples with a low false positive rate. However, all methods have difficulty finding the other features, if any, in most examples. Our current application can accept a relatively high false negative rate; therefore, we have been at least partially successful. Although we had zero false positives among the 29 real test cases (we regard D1 as a training case), as discussed, we had a false

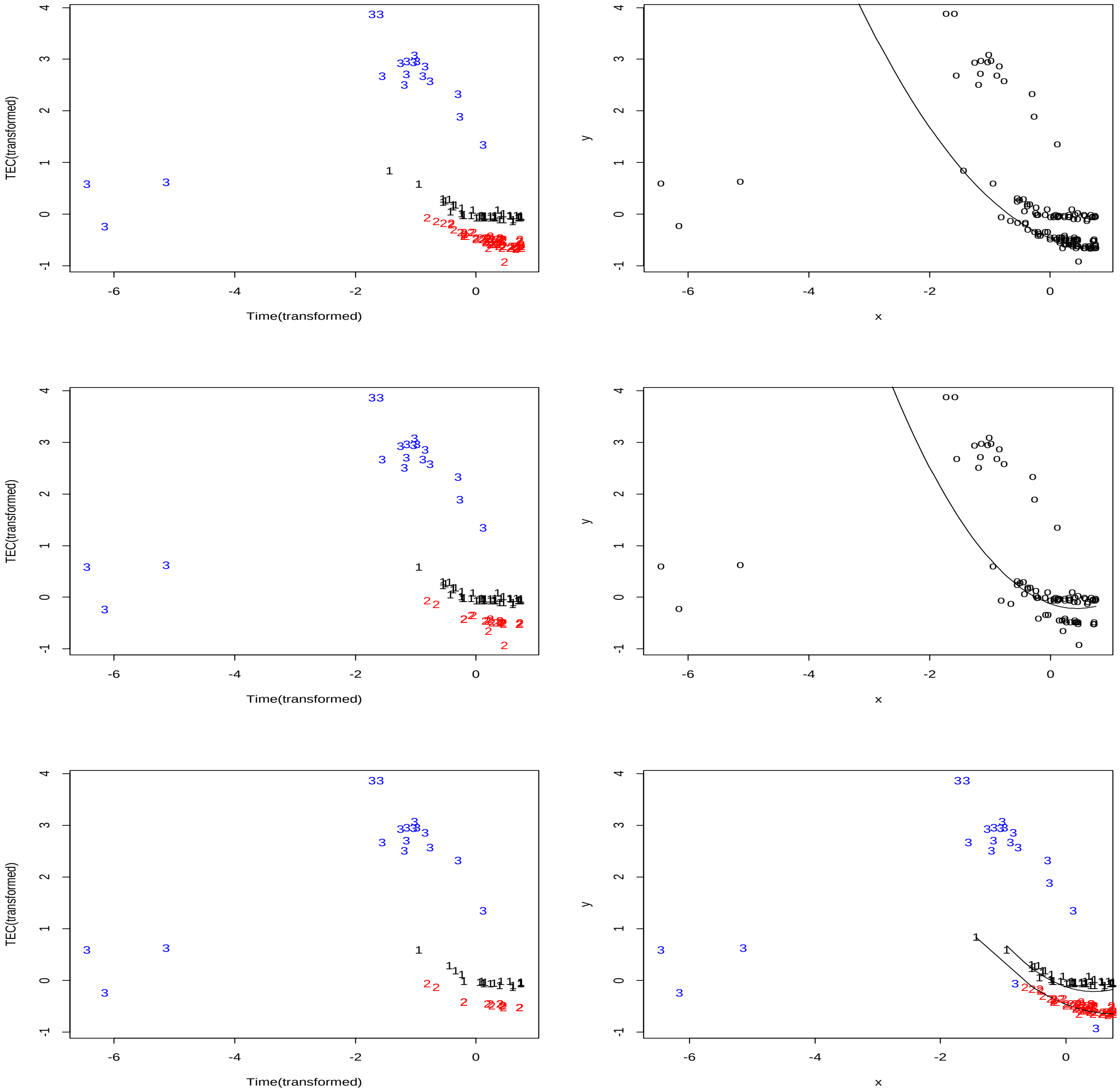


Figure 6. L2E method results on D1.

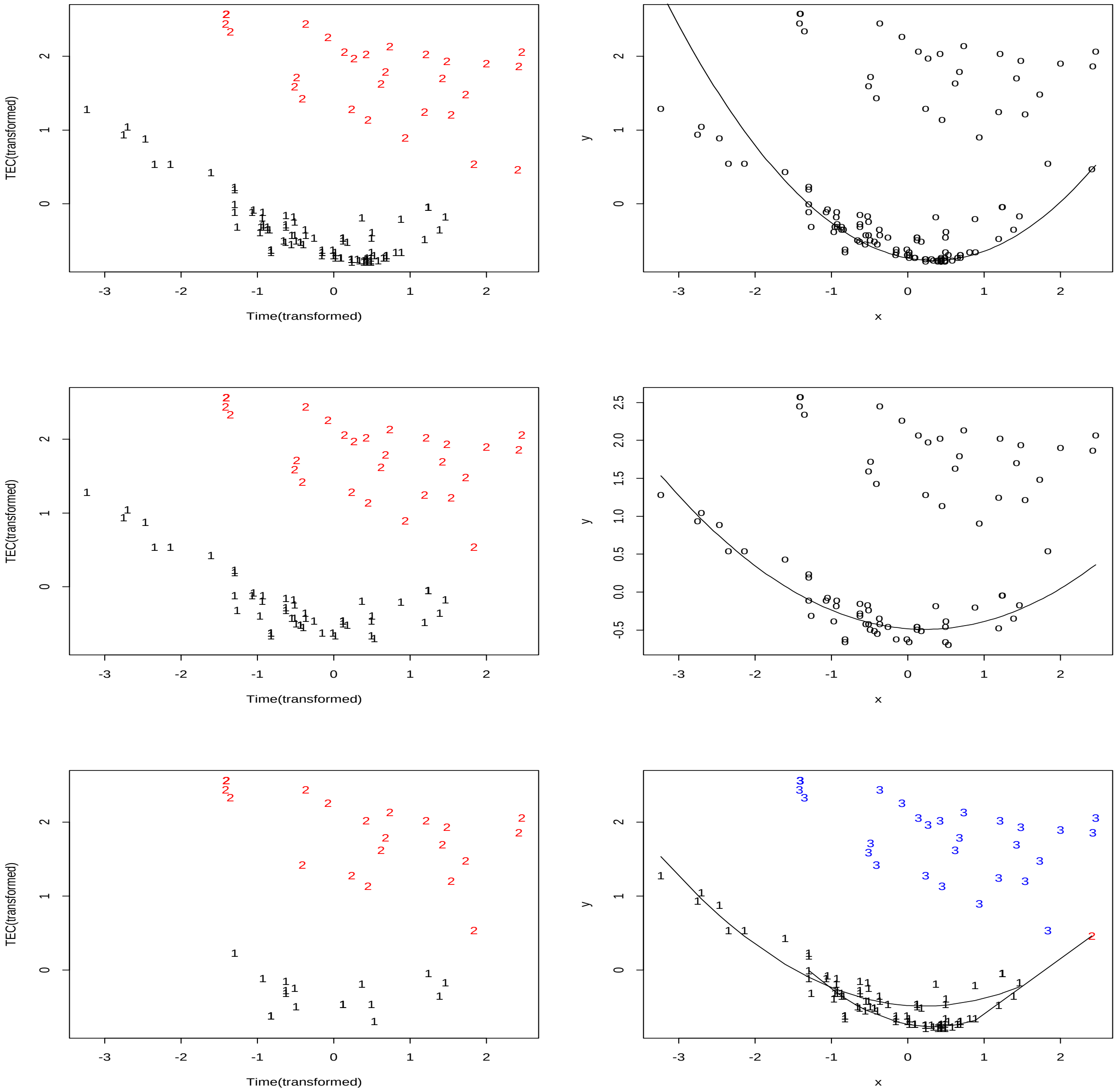


Figure 7. L2E method results on D3.

positive in a simulated case (if we use one particular definition of false positive) or at least a high false positive rate in the found cluster. This is undesirable in our intended use for the identified storms.

We believe that this data set provides rich opportunity for evaluating methods for fitting mixture models. It could be argued that because the BIC criterion did not indicate a clear signal in favor of the "correct" model that there is some inherent ambiguity in the analyst's labels. Therefore we think a useful next step would be to consider the data partition (into features and noise) having the highest posterior probability to be the "true model." This also has drawbacks because there is no guarantee that the highest posterior probability would belong (in practice) to the true model.

At present, we plan to use our conservative method 1 (low false positive rate method) involving noise rejection, hierarchical clustering, and possible relabeling of some noise points as feature points if they fall in a cluster's zone of ownership. However, we also plan a more extensive evaluation of the three alternatives presented here because we anticipate that some improvements could be made with modifications to the basic approach. Perhaps a blended approach using some information from each of the four methods applied to each new case would be effective.

5 References

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood for Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Soc Series B* 39, 1-38.

Hastie, T., and Stuetzle W. (1989), "Principal Curves," *Journal of the American Statistical Assoc* 84, 502-516.

Jacobson, A., Knox, S., Franz R., and Enemark, D. (1999), "FORTE Observations of Lightning Radio-Frequency Signatures: Capabilities and Basic Results," *Radio Sci*, 34(2), 337-354. See also http://nis-www.lanl.gov/nis-projects/forte.

Moore, K., Blain, P., Briles, S., and Jones, R. (1995), "Classification of RF Transients Using Digital Signal Processing and Neural Network Techniques," *Applications and Science of Neural Networks*, *Proceedings SPIE*, 2492, 995-1006.

Stanford, D., and Raftery, A. (2000), "Finding Curvilinear Features in Spatial Point Patterns: Principal Curve Clustering with Noise," *IEEE Trans on Pattern Analysis and Machine Intelligence*, 22(6) 601-609.

Splus5 for Linux, Insightful Corporation, Seattle Washington, 1999.

Scott, D. (2002), "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics* 43(3), 274-285.

Scott, D., and Szewczyk, W. (2002), "From Kernels to Mixtures," *Technometrics* 43(3), 323-335.

Turner, R. (2000), "Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions," *Applied Statistics* 49, part 3, 371-384.

The Moment Preservation Method of Cluster Analysis

Bernard Harris, University of Wisconsin, Madison

Abstract Unavailable

Rank Adapted Kernel Density Estimation

David B. Kim
Department of Mathematical Sciences
U.S. Military Academy
West Point, NY 10996

October, 2001

Abstract

We consider adapting bandwidths of a kernel density estimator according to the ranks of observations. The specifics of bandwidth selection is motivated by a deterministic decomposition of a density into densities of order statistics and their asymptotic behaviors. The resulting estimator has a local bandwidth similar to that of Abramson (1982) and Breiman et al. (1977) with a new feature of rank correction. We investigate its properties and demonstrate that not only it can smooth out the bumps in the tails while maintaining interesting features in data-rich region but also that it can reduce the boundary bias when the support of the target density is compact.

1 Introduction

Kernel density estimation is one of the more intuitive and widely used methods of nonparametric density estimation. Most of the research activities have been focused on the selection of the bandwidth, or smoothing parameter. The simplest implementation of the method uses only one bandwidth over the entire data set. This, however, has many shortcomings that motivated the various strands of researches on adaptive bandwidth selection.

Starting with the works by Breiman et al. (1977) and Abramson (1982), various ways of determining bandwidths depending on the local behavior of the density have been investigated. What one would like to accomplish in adaptive bandwidth selection is to be able to sharply delineate the boundary of a compact support and to be able to smooth out bumps in the tails while depicting interesting features in the density near the mode at the same time. Both Breiman et al. (1977) and Abramson (1982) attempt the latter by adjusting the bandwidth by a factor of a positive power of the reciprocal of the density. Since a larger bandwidth corresponds to a smoother estimate, this approach aims to use larger bandwidths where the density is smaller, and vice versa. Of course, the true underlying density is to be unknown in most cases, so in using this type of approach, one first finds a pilot estimate of the density and use it instead of the true underlying density.

Many researchers have also looked at the former problem using boundary kernels and other approaches (see for example, Müller (1991)). In this report, we look at a new adaptive bandwidth selection scheme where both sharp delineation of the boundary and more smoothing in the region of sparse density can be accomplished at the same time.

2 New Estimator: Motivation and Definition

In this section, we define the new kernel density estimator whose bandwidths are adapted according to the ranks of observations, and we investigate its properties. Let $X_1, \cdots, X_n$ be *iid* observations from an unknown density $f(x)$. Recall that a Kernel density estimator $\hat{f}(x)$ with a bandwidth h identical for all data points can be written as

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}K\left(\frac{x-X_j}{h}\right), \tag{1}$$

where $K(\cdot)$ is a Kernel function. Kernel functions can be classified by its order k, where the k the order kernel $K(t)$ is such that, for positive integers j,

$$\int t^j K(t)\,dt = 0,\; j < k,\; \text{ and} \int t^k K(t)\,dt \neq 0,$$

where we note that we will assume the integration is over the entire real line $\mathbb{R}$ unless otherwise specified. Higher order kernel functions usually lead to estimators with smaller asymptotic biases, but the kernel functions of order higher than two can no longer be probability density functions (*pdfs*). Having a kernel function that is a *pdf* itself guarantees that the resulting kernel density estimator will again be a *pdf.* For this reason, we shall restrict our attention to the second order kernel which itself is a *pdf.*

The adaptive scheme presented in this paper takes its motivation from the following identity:

$$f(x) = \frac{1}{n}\sum_{j=1}^{n} f_{(j)}(x) \tag{2}$$

where $f_{(j)}(x)$ is the *pdf* of the jth order statistic, $X_{(j)}$. Kim (1999) showed that using the densities of the putative asymptotic densities instead of $f_{(j)}$'s gives a convergent approximation to $f(x)$. The asymptotic normality of the central order statistics, viz,

$$\sqrt{n}\left(X_{(j)} - \xi_{p_j}\right) \overset{\text{a}}{\sim} N\left(0, \frac{p_j(1-p_j)}{f^2(\xi_{p_j})}\right),$$

where $p_j = j/n$ and $\xi_{p_j} = F^{-1}(p_j)$. This can be seen as a consequence of the convergence in *pdfs* (Rao 1973, p. 422), specifically,

$$f_{(j)}(x) \approx \frac{\sqrt{n}}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{n(x-\xi_{p_j})^2}{2\sigma_j^2}\right), \tag{3}$$

where we let $\sigma_j^2 = \frac{p_j(1-p_j)}{f^2(\xi_{p_j})}$.

The second order kernel functions are bona fide densities, and a kernel density estimate using the second order kernels is the arithmetic average of n *pdf*'s centered at n data points:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right), \tag{4}$$

where $K(\cdot)$ is a bona fide density. Silverman (1986) defines an adaptive kernel estimate $\hat{f}_a(x)$ as follows:

$$\hat{f}_a(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h\lambda_i} K\left(\frac{x - X_i}{h\lambda_i}\right), \tag{5}$$

where λ_i is a local bandwidth factor adapted to each observation X_i. Comparing Eq.(2) and Eq.(5) suggests that one way of choosing the local bandwidth is to refer to the standard deviation in Eq.(3). Clearly, we do not know a priori the population quantile, ξ_{p_j}, not to mention the true density f. But $X_{(j)}$ is a consistent point estimator of ξ_{p_j}, so using $X_{(j)}$ in its stead we get the following estimator:

$$\hat{f}_1(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h\lambda_i} K\left(\frac{x - X_{(i)}}{h\lambda_i}\right), \tag{6}$$

where $\lambda_i = \sqrt{\frac{p_i(1-p_i)}{f^2(X_{(i)})}}$. Even though this estimator still depends on the unknown true density, the dependence is reminiscent of the estimators of Abramson (1982) and Breiman et al. (1977). Still there is one new feature in the local bandwidth not found in the previous estimators: the factor of $\sqrt{p_i(1-p_i)}$, which is presumably responsible for a surprising and unique adaptability which will be shown in the next section. The choice of the global bandwidth h may be made as in using the estimator of Abramson (1982) after choosing a suitable pilot estimator.

3 Examples

In this section, the new rank adapted estimator will be compared to other estimators. Comparison is among rank adapted, Abramson, fixed bandwidth estimators, and the dashed line for the *true density*. The global bandwidth h for each of the estimators was determined using the unbiased cross validation (UCV) criterion. For the rank adapted and Abramson estimators, the pilot estimator using Silverman (1986)'s "rule of thumb" $\hat{h}$. All the computation and the generation of the figures were done in R.

It is well known that the estimators of Abramson (1982) and Breiman et al. (1977) were motivated by the need to smooth out bumps in the tails while depicting interesting features in the density near the mode at the same time, which cannot be achieved with a single uniform bandwidth. The similarity of the new rank adapted estimator and Abramson's estimator leads one to expect the both should behave similarly. In the case of a sample of size 500 from the standard normal distribution, seen in Fig. 1, one indeed observes that.

Fig. 2 shows the comparison of the local bandwidth factors λ_i for Abramson's estimator and the new rank adapted estimator. One observes the attenuation of the growth of the local bandwidth factor in the tails for the new rank adapted estimator–the attenuation is due to the rank correction factor. This is an encouraging sign since the rapid growth of the local bandwidth factor in Abramson's estimator has been shown to be responsible for its rather surprisingly poor performance for a very large sample (Terrell and Scott 1992).

An additional role the rank correction factor plays is demonstrated in the next example where a random sample was drawn from the uniform distribution on (0,1), which has a density with a compact support. Fig. 3 shows that the new rank adapted estimator is more adept at delineating the sharp boundary of the uniform density than other estimators shown (it shows the steepest descent at the boundaries).

In the next example, where the random sample of size 500 is drawn from an exponential distribution (with mean=1), which has a long tail on one end and a sharp boundary on the other end, the dual capabilities of the new rank adapted estimator is demonstrated.

The final two examples show the comparison of the estimators for famous real life data sets (see Silverman (1986) for more details on the data sets). The performance of the new estimator on the suicide data set (Fig. 5), which has similar features to the exponential example seen before, is very encouraging. It demonstrates that the new estimator does an adequate job of sharply delineating the boundary and smoothing in the region of sparse density. Fig. 6 shows the comparison of the estimators for the Yellowstone geyser data set. Again, one observes the competitive performance of the new estimator along with its novel capability where both sharp delineation of the boundary and more smoothing in the region of sparse density can be accomplished at the same time.

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates – A square root law. *The Annals of Statistics 10*, 1217–1223.

Breiman, L., W. Meisel, and E. Purcell (1977). Variable kernel esitmates of multivariate densities. *Technometrics 19*, 135–144.

Kim, D. B. (1999). *Quantile Decomposition of a Density*. Ph. D. thesis, University of California, Santa Barbara.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika 78*(3), 521–530.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (Second ed.). John Wiley & Sons, New York.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* London: Chapman & Hall.

Terrell, G. R. and D. W. Scott (1992). Variable kernel density estimation. *Ann. Statist. 20*(3), 1236–1265.

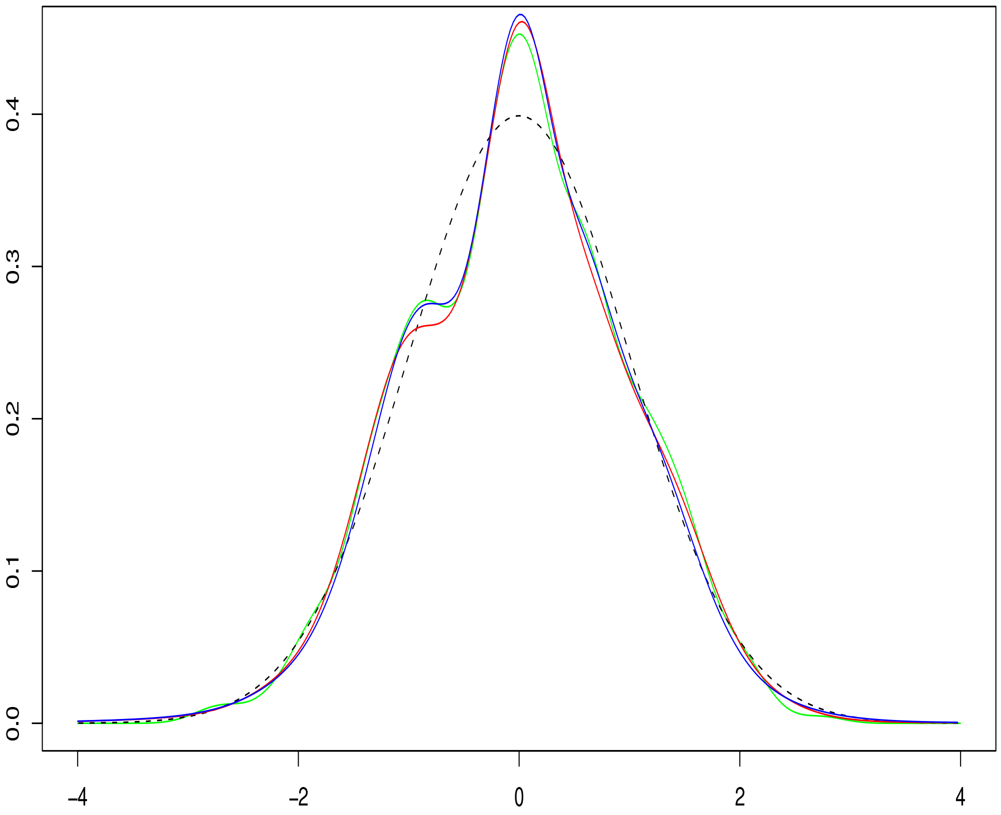


Figure 1: Comparison of the three estimators for a random sample of size 500 from the standard normal distribution.

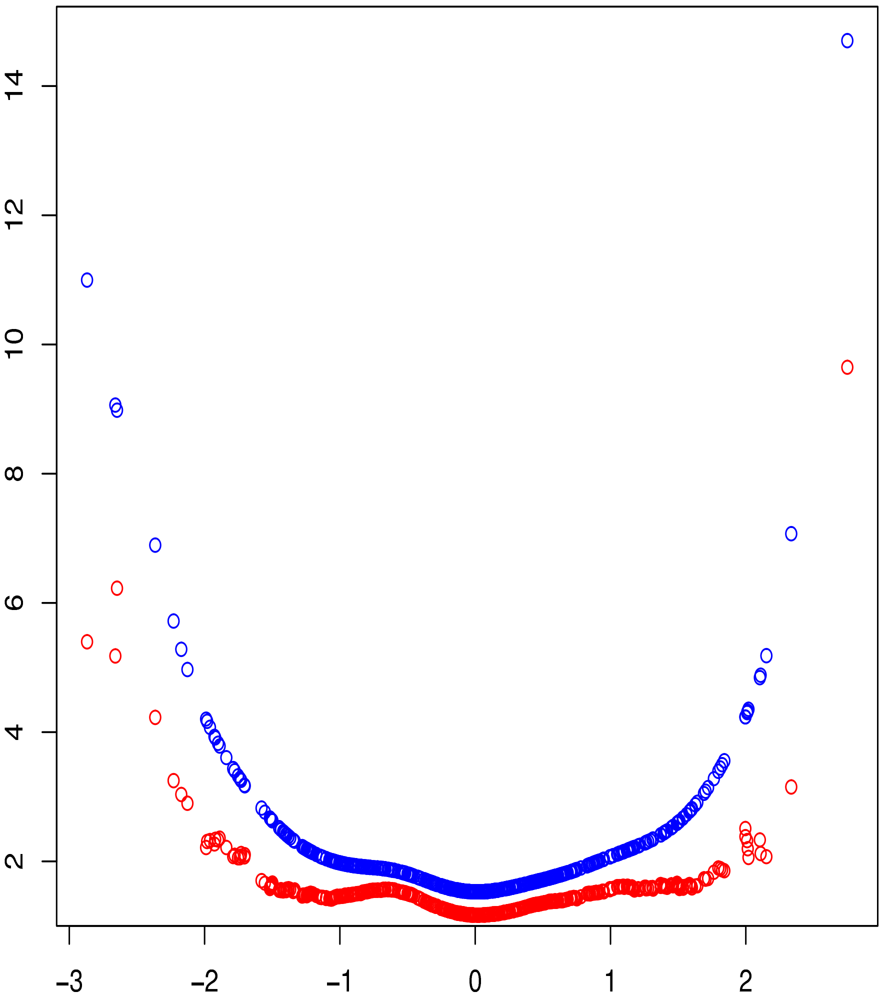


Figure 2: Comparison of the local bandwidth factors λ_i for a random sample of size 500 from the standard normal distribution.

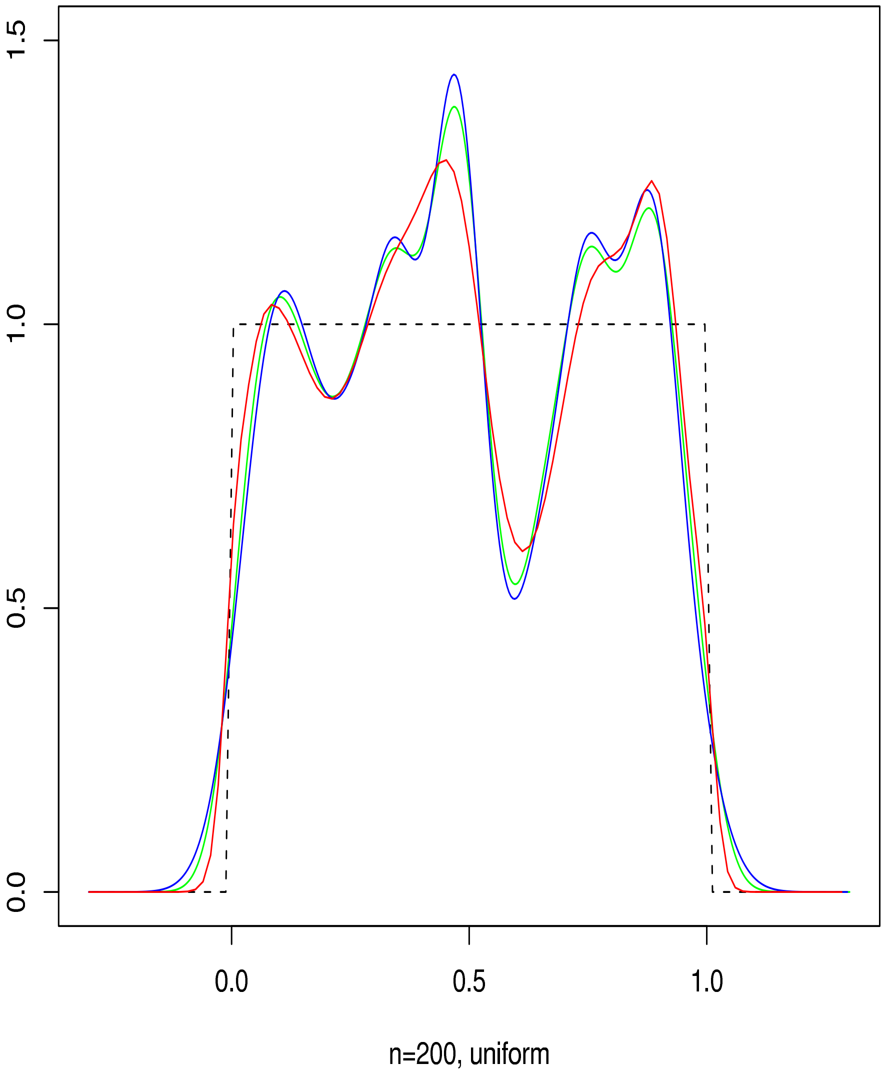


Figure 3: Comparison of the three estimators for a random sample of size 200 from the uniform distribution on (0,1).

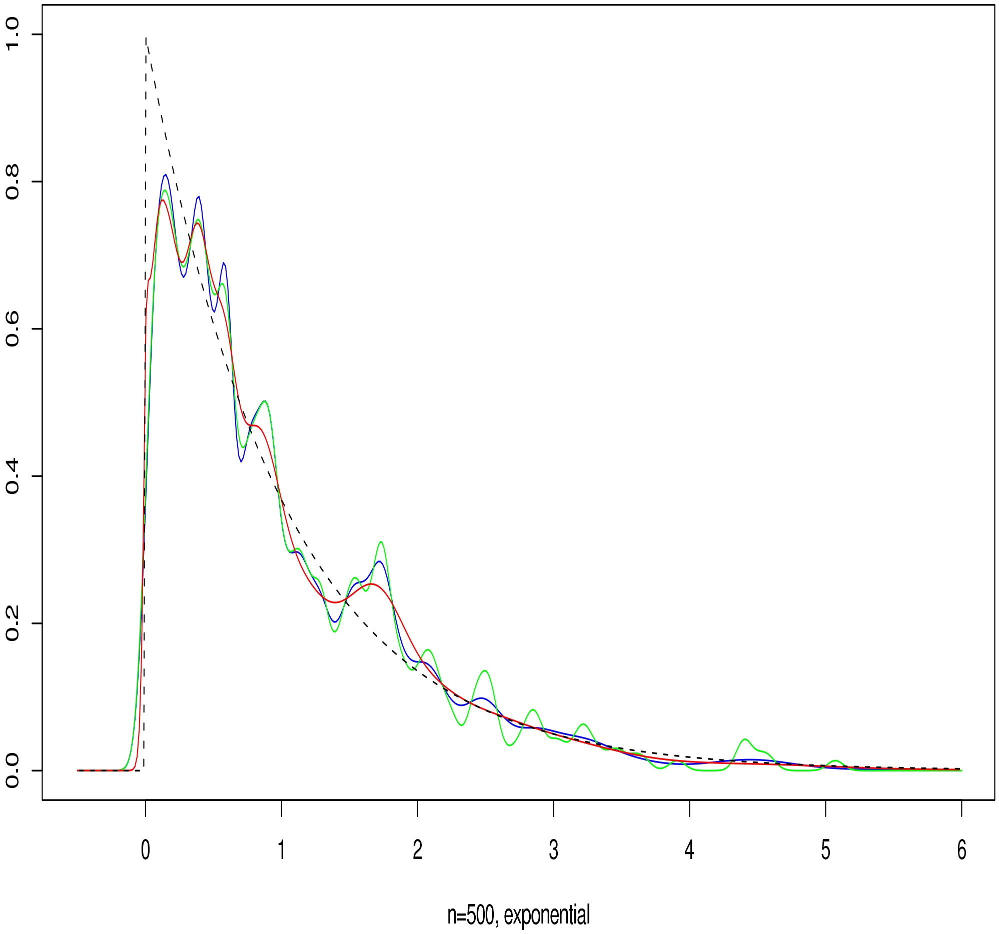


Figure 4: Comparison of the three estimators for a random sample of size 500 from the exponential distribution with mean=1.

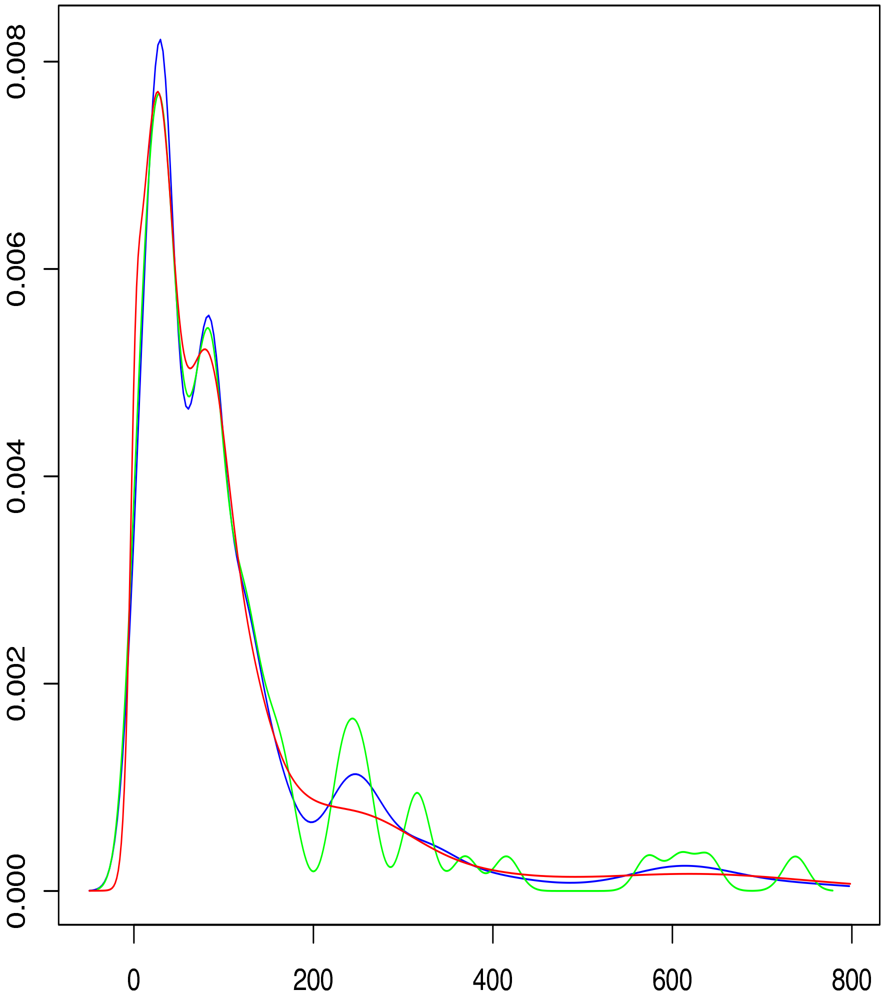


Figure 5: Comparison of the three estimators for the suicide data set.

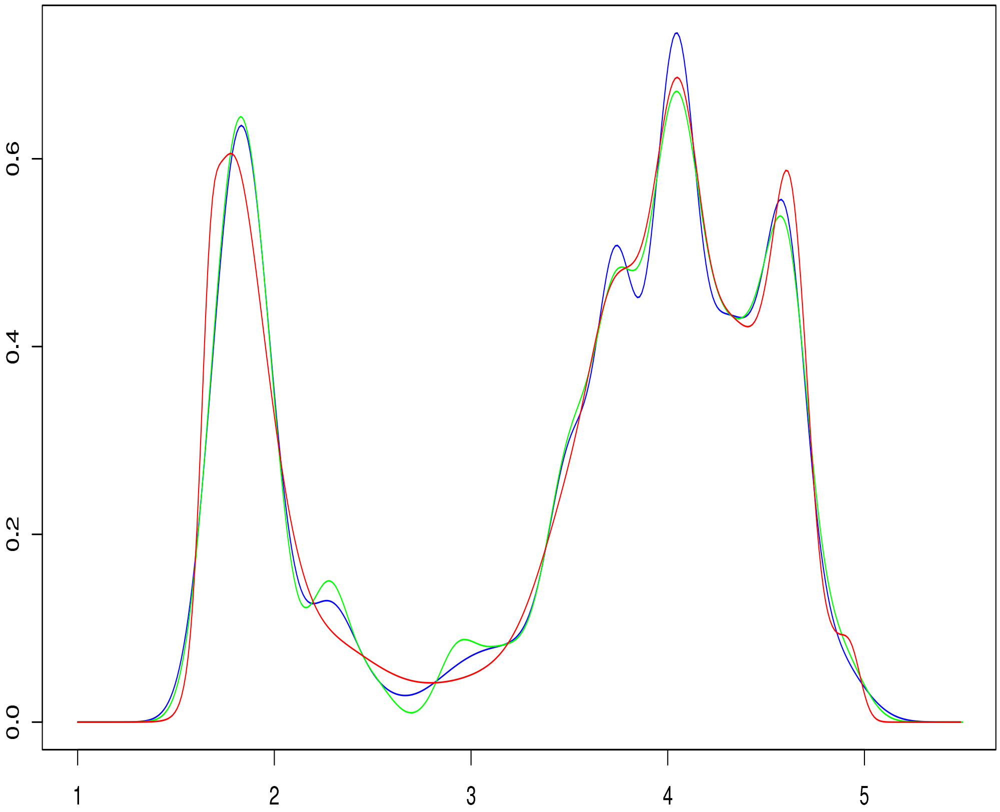


Figure 6: Comparison of the three estimators for the geyser data set.

Contributed Session VIII

THE ROLE OF EXPERT KNOWLEDGE IN UNCERTAINTY QUANTIFICATION (ARE WE ADDING MORE UNCERTAINTY OR MORE UNDERSTANDING?)

Jane M. Booker
Los Alamos National Laboratory
MS P946 Los Alamos, NM 87545

Mark C. Anderson
Los Alamos National Laboratory
MS D411 Los Alamos, NM 87545

Mary A. Meyer
Los Alamos National Laboratory
MS F600 Los Alamos, NM 87545

Uncertainty quantification can be broadly defined as the process of characterizing, estimating, propagating, and analyzing various kinds of uncertainty for a complex decision problem. In the realm of complex computer and physical models, it is more focused upon computational and modeling uncertainties, e.g., sensitivities of outputs to input values or verification and validation. In either case, sources of uncertainty (including variability) can arise from the two broad categories: epistemic and aleatory. We begin by presenting a brief taxonomy of uncertainties involved in uncertainty quantification of physical and simulation models. Included in this exercise will be some definitions, which remain somewhat non-standard within and between various communities (e.g., artificial intelligence). While listing these sources is relatively straightforward, evaluating uncertainties in complex computer codes presents huge obstacles. In addition, the human enters into the quantification of uncertainty process in much the same way as in any decision process, and that entrée presents both additional obstacles and some solutions. We will discuss how the use of expert judgment and expertise is involved in uncertainty quantification.

1. Definition of Terms. The first term in the title is expert knowledge which is defined as what is known by qualified individuals, responding to complex, difficult (technical) questions, obtained through formal expert elicitation (Meyer and Booker, 2001). It is a snapshot of the expert's state of knowledge at the time, and may be expressed in either qualitative and quantitative form.

Knowledge can be elicited in two distinctive forms: expertise and expert judgment. Expertise refers to that information from experts about the definition and structure of a complex problem. How experts organize and represent their problem solving knowledge and how information flows within a problem are part of expertise. When experts identify relevant data and information sources, including models, experimental results and numerical methods, they are providing expertise. Identification of uncertainties associated with these is also part of expertise. Expertise is used extensively in making everyday decisions; it may not be elicited and documented as such.

Examples of expertise include:

Decisions about what variables to enter into a statistical analysis.
Decision about which data sets should be analyzed.

Assumptions used in model or method selection.

Decisions concerning which forms of uncertainty are appropriate to use (e.g., using a probability density function to represent aleatory uncertainty).

Descriptions of experts' thinking, problem solving, and information sources used in arriving at any of the above.

Expert judgment refers to the contents of the expert's knowledge. When experts provide estimates of phenomena (qualitative or quantitative) or the uncertainties associated with those estimates, they are providing expert judgment. Any assumptions, heuristics, cues, and historical information that experts use in providing estimates is also considered part of expert judgment. Examples of expert judgment include: estimating the occurrence of an event, estimating the uncertainty in a parameter, and predicting the performance of a new product.

Uncertainty quantification is the other major term in the title. In the broadest sense, it refers to the process of characterizing, estimating, propagating and analyzing various types of uncertainty (including variability) for a complex decision or physical problem. In more specific modeling complex and physical communities it focuses upon measurement, computational, parameter (including sensitivities of outputs to input values), and modeling uncertainties leading to verification and validation of the computation and modeling.

Generally speaking, there are two different categories of uncertainty, aleatory and epistemic. Aleatory refers to uncertainty due to random variation or inherent variation. It is irreducible and includes the basic statistical concepts of variability and the definition of probability as describing the uncertainty associated with the outcome of an experiment or event. By contrast, epistemic uncertainty is reducible and stems from a lack of knowledge. There is disagreement on how to classify *error*, which could stem from numerical methods, the process of discretization or simply mistakes. We maintain that error could be either aleatory or epistemic in nature.

2. Uncertainties in the Modeling Process. Using the modeling definition of uncertainty quantification, let us examine the steps involved in modeling and the uncertainties associated with those steps. A complex physical or decision model usually begins with observations of nature. Next models for capturing the observed behavior are conceptualized and then mathematically formulated into computational models. At this point, issues emerge regarding the computational process. Numerical models may be employed to represent the physical model (if one exists). Numerical implementation and evaluation of the model is then implemented. Finally, if physical models and/or numerical models are lacking, surrogate models may be used, such as a statistical response function or a neural network. Implementing and evaluating surrogate models are necessary steps in the surrogate modeling process.

Associated with these modeling steps are various kinds of uncertainties, some of which are extremely difficult to estimate. Regardless of the difficulty, decisions are made every day by humans, whose mental intervention now enters the modeling process as potentially important sources of uncertainty. Uncertainties of measurement include noise, resolution and processing, many of which could be classified as aleatory. Mathematical modeling uncertainties are found in choices and uses of equations, boundary conditions, initial conditions and inputs. Many of these could be classified as epistemic. For numerical modeling, uncertainties arise in the use of weak formulations, in the choices of discretizations for mesh sizing and time steps, in the use of approximate solution algorithms, and with issues of truncation and roundoff. Statistical or surrogate modeling is listed as distinctly different from physical or mathematical modeling in that interpretability of these models is one step removed from the real world. Neural networks and other dimension reduction models fall into this category. Uncertainties also emerge from errors of approximation, interpolation and extrapolation.

In general modeling uncertainties (whether from mathematical, numerical or surrogate models) are extremely difficult to characterize, understand and estimate. They can be both epistemic and aleatory in nature. We list the issue of model parameters and the uncertainties and sensitivities of inputs to outputs separately from model uncertainty because significant progress has been made in this area (McKay et al., 1999). Beyond modeling issues are uncertainties that might be termed as scenario uncertainties. These would encompass the application realm and the choices made regarding the problem definition (Oberkampf et al., 2001).

Regardless of the source or type of uncertainty involved, decisions made in the modeling process also contain uncertainties. Therefore the "human in the loop" is an additional source of uncertainty, whether specifically acknowledged or not.

3. Uncertainties from Humans. The human decision making contribution to the overall uncertainties in the modeling process can originate in the cognitive and motivational biases that affect human thinking and judgment. By *bias,* we do not imply statistical bias (or shift of the mean value) but instead refer to a skewing from a standard or reference point that can degrade the quality of the information and contribute to uncertainty.

Among the list of various cognitive biases, the most prominent contributor to uncertainty is the *underestimation of uncertainty bias* (or false precision bias). Humans tend to believe and think about the world as having more precision that it really does. Other cognitive biases contributing to uncertainty are:

- *Availability*—how humans account for rare events depends upon whether they have experienced them or not.
- *Anchoring*—humans cannot move from preconceptions, but instead anchor to them even in light of new data/information.
- *Inconsistency*—humans forget what has preceded and hence produce inconsistent conclusions.

The most noted motivational biases that contribute to uncertainty are:

- *Group Think*—following the leader, regardless of the consequence.
- *Impression Management*—being politically correct.
- *Wishful Thinking*—wanting something makes it a reality.
- *Misimpression*—poor, incorrect or bad translation of information.

To top this list of biases as contributors to uncertainty in human thinking and judgment is the well-studied phenomena that humans are poor probability thinkers (Meyer and Booker, 2001). It is so easy to contradict the axioms of probability, even if the person is an expert in probability theory. Human thinking is just not conducive to probabilistic thinking. Therefore asking expert to provide probability estimates for uncertainties is quite dangerous.

4. Countering Human Contributions to Uncertainty. One would hope that experts could offer assistance in understanding and estimating uncertainties in the modeling process, without contributing additional uncertainty. The human brain has a tremendous capacity to integrate complexities, such as uncertainty. Experts should be able to identify sources of uncertainty, provide estimates (quantitative or qualitative) for these, be able to update estimates as new information becomes available and suggest methods of how to propagate uncertainties throughout the modeling process. With the use of some recently developed tools and technologies, it is possible to counter the human contributions to uncertainty enabling us to take advantage of the knowledge that experts are capable of providing.

Formal, structured elicitation of expertise and expert judgment is designed to counter the common biases arising from human cognition and behavior. These techniques (Meyer and Booker, 2001) draw from cognitive psychology, decision analysis, statistics, cultural anthropology and knowledge acquisition. They add rigor and defensibility, and increase the ability to update judgments in light of new knowledge.

In addition to providing bias minimization, formal elicitation provides documentation and utilizes the way people think, work and solve problems. For example, an expert unfamiliar with probability would never be asked to express uncertainties in the form of a probability density function. Instead he might be accustomed to thinking about uncertainty in terms of a range of possible values; therefore, the range would be elicited.

A second major advancement to counter human contributions to uncertainty is the development of alternative mathematical theories for handling different kinds of uncertainty, such as ambiguity and vagueness. Because humans have difficulty with consistent thinking that preserves the axioms of probability, these other theories offer alternatives for characterizing uncertainties which may be more consistent with human thinking. These theories include (Oberkampf, et al. 2001):

Possibility Theory (for crisp or fuzzy sets).
Fuzzy Sets
Dempster-Schafer (Evidence) Theory
Choquet Capacities
Upper and Lower Probabilities
Convex Sets
Interval Analysis Theories
Information Gap Decision Theory (non measure based) (Ben-Haim, 2001)

While most of these theories are measure based, all are set based, using either crisp (classical) or fuzzy set theory. They are all axiomatic and have a calculus (or algebra) with rules for combining sets and implementing the axioms. They are internally consistent and coherent such that one cannot be caught up in a situation of "heads I win and tails you lose." With modern computational methods and computers, they are computationally practical. As seen in Section 5, many have or are in the process of developing metrics for uncertainty.

Having these choices is an advantage, but there may be different kinds of uncertainties within a modeling problem such that these different theories would apply. For combining all uncertainties within a problem, we need bridges and linkages between the various theories. To our knowledge the only developed linkage has been between probability theory and membership functions in fuzzy set theory (Singpurwalla and Booker, 2002). Hierarchical relationships between them have been developed as shown in Figure 1, with more specific theories at the bottom and more general ones at the top.

As noted in Figure 1, probability theory can have more than one interpretation. Historically speaking there are as many as eleven different interpretations of probability, all consistent with its axioms and calculus (Bement, et al, 2002). Two competing interpretations today are Frequentist and Subjective (or Personalistic). The latter encompasses the ever expanding set of Bayesian analysis methods, and it is the interpretation that permits linkage between probability and fuzzy logic (Singpurwalla and Booker, 2002).

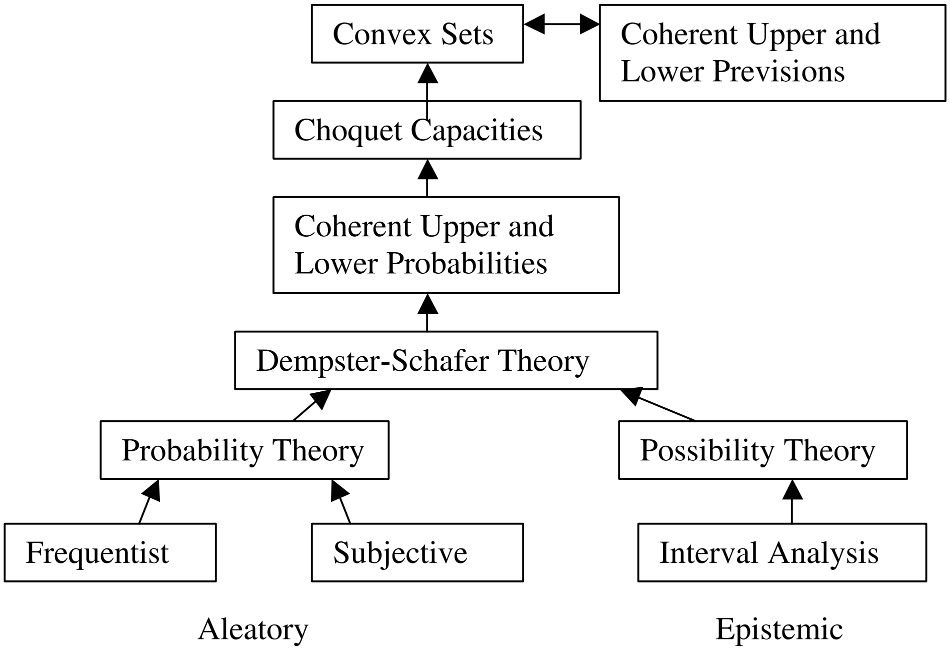


Figure 1. Hierarchy of Theories for Crisp Sets

5. Some Measure Theoretic Approaches to Uncertainty. In this section, we provide brief comparison of three of the measure based, crisp set approaches in Figure 1: Probability Theory, Dempster-Schafer Theory and Possibility Theory.

Probability is based upon a single measure function, *Pr*, and has the property of additivity as noted in equation (5.2) below.

(5.1) $Pr: 2^X \to [0,1] \quad Pr(\varnothing) = 0 \quad Pr(X) = 1$

(5.2) $Pr(\cup_i A_i) = \Sigma_i Pr(A_i) - \Sigma_{j>k} Pr(A_j \cap A_k) + \ldots + (-1)^{n+1} Pr(\cap_i A_i)$

$Pr(\cap_i A_i) = \Sigma_i Pr(A_i) - \Sigma_{j>k} Pr(A_j \cup A_k) + \ldots + (-1)^{n+1} Pr(\cup_i A_i)$

Dempster-Schafer is based upon two measure functions, belief (*Bel*) and plausibility (*Pl*). Non additivity is illustrated in equation (5.4) below.

(5.3) $Bel: 2^X \to [0,1] \quad Bel(\varnothing) = 0 \quad Bel(X) = 1$

$Pl: 2^X \to [0,1] \quad Pl(\varnothing) = 0 \quad Pl(X) = 1$

(5.4) $Bel(\cup_i A_i) = \Sigma_i Bel(A_i) - \Sigma_{j>k} Bel(A_j \cap A_k) + \ldots + (-1)^{n+1} Bel(\cap_i A_i)$

$Pl(\cap_i A_i) = \Sigma_i Pl(A_i) - \Sigma_{j>k} Pl(A_j \cup A_k) + \ldots + (-1)^{n+1} Pl(\cup_i A_i)$

Possibility is also based upon two measure functions, possibility (*Pos*) and necessity (*Nec*). Non additivity is illustrated in equation (5.6) below.

(5.5) $Pos: 2^X \to [0,1] \quad Pos(\varnothing) = 0 \quad Pos(X) = 1$

$Nec: 2^X \to [0,1] \quad Nec(\varnothing) = 0 \quad Nec(X) = 1$

(5.6) $Pos((_{i}A_i) = \sup_i Pos(A_i)$
$Nec('_{i}A_i) = \inf_i Nec(A_i)$

Applications of Dempster-Schafer are conspicuously lacking in the literature, especially regarding the important issue of how this theory might be useful for capturing how some experts think about uncertainties. Perhaps the most applicability can be found using fuzzy membership functions and possibility theory (Ross, 1995).

Some metrics for uncertainty have been developed under these alternative theories. Information theory based concepts such as entropy provide the foundation for some, as noted below.

Hartley measure for nonspecificity:

$H(A) = \log_2|A|$, where $|A|$ is the cardinality of A.

Generalized Hartley measure for nonspecificity in Dempster-Schafer:

$N(m) = \Sigma_{A\in 2\mathrm{X}}\, m(A) \log_2|A|$, where m: $2^{\mathrm{X}} \top\ [0,1]$, $m(\varnothing) = 0$, and $\Sigma_{A\in 2\mathrm{X}}\, m(A) = 1$.

U-uncertainty measure for nonspecificity in possibility theory:

$U(r) = \Sigma_{j=2}\, (r_j - r_{j+1})\log_2 j$, where $r(x) = Pos(\{x\})$, for $r_j = r_{j+1}$, for all j.

Shannon entropy for total uncertainty in probability theory:

$S(p) = -\,\Sigma_{x\in \mathrm{X}}\, p(x)\log_2 p(x)$

Generalized Shannon entropy for total uncertainty in Demspter-Schafer theory:

$AU(Bel) = \max_{p_x} (-\Sigma_{x\in \mathrm{X}}\, p_x \log 2 p_x)$, where $Bel(A) = \Sigma_{x\in \mathrm{X}}\, p(x)$, for all $A \in 2^{\mathrm{X}}$.

Hamming distance for fuzzy sets:

$f(A) = \Sigma_{x\in \mathrm{X}}\, [1-|2A(x)-1|]$, where $A(x)$ is a membership function.

6. Role of Expert Knowledge in Uncertainty Quantification. Experts and the knowledge they provide can be valuable in the processes of understanding, estimating and propagating different kinds of uncertainties within a complex model. However, the experts are human, subjected to the cognitive and motivational biases which also contribute to the overall uncertainties within that problem. There is hope for minimizing these contributions by utilizing bias minimization elicitation techniques and by offering experts alternative (to probability) theories for handling uncertainties.

Two major areas of research are necessary to fulfill the usefulness of these alternative theories. First, continued study is required to be able to link the various theories for handling different types of uncertainties within the same modeling problem. Second, more study is needed to determine the usefulness of these alternative theories as better methods (than probability) for capturing the way experts think and problem solve.

REFERENCES

Ben-Haim, Y. (2001). *Information Gap Decision Theory*. Academic Press, New York.

Bement, T.R., Booker, J.M., Keller-McNulty, S., Singpurwalla, N.D. (2002). Testing the Untestable; Reliability in the 21st Century. To appear *IEEE Transactions on Reliability*.

McKay, M. D., Morrison, J.D., Upton, S.C. (1999). Evaluating Prediction Uncertainty in Simulation Models. *Computer Physics Communications* **117** 44-51.

Meyer, M.A. and Booker, J.M. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. SIAM, Philadelphia, PA.

Oberkampf, W., Helton, J., Sentz, K. (2001). Mathematical Representation of Uncertainty. *AIAA Proceedings of Non-Deterministic Approaches Forum*, Reston, VA.

Ross, T.J. (1995). *Fuzzy Logic with Engineering Applications.* McGraw Hill, New York.

Singpurwalla, N.D. and Booker, J.M. (2002). Membership Functions and Probability Measures of Fuzzy Sets. Los Alamos National Laboratory report, LA-UR-02-0032.

Optimal Reliability Apportionment

Nozer D. Singpurwalla, George Washington University

Abstract Unavailable

On Computing and Comparing the Reliability of Competing Networks [1]

P. J. Boland
Department of Statistics
National University of Ireland, Dublin
Dublin 4, Ireland

F. J. Samaniego
Department of Statistics
University of California
One Sheilds Avenue,
Davis, CA 95616

E. M. Vestrup
Dept. of the Mathematical Sciences
DePaul University
2230 N. Kenmore
Chicago, IL 60615

Abstract

Comparing the reliability of two networks of more than modest size can be a computationally intensive exercise. In this paper, domination theory and the notion of the signature of a network, and their respective roles in calculating the reliability of a network, are briefly reviewed. The computational advantages of the former, and the interpretive richness of the latter, beg the question: how are the two related? The exact functional relationship between the signature vector and the vector of signed dominations is obtained. A detailed example is given in which the connection between these two concepts is usefully exploited.

A network G with v vertices and n edges is typically denoted by the symbol $G(v,n)$. We follow the usual convention that postulates that nodes cannot fail, but that edges can be in either a functioning or a failed state. In communication networks, as in many other types of networks, the primary quality characteristic of interest is connectivity. A two-terminal network is connected if there is at least one set of functioning edges providing a path from one terminal to the other. We will restrict attention to networks that are coherent, that is, to networks with the property that every edge is relevant and that all supersets of path sets are also path sets.

The network characteristics upon which we will be focusing are defined in terms of edges whose lifetimes are treated as independent and identically distributed random variables. We will be concerned with the distribution of T, the failure time of the network and, in particular, with the probability that it is connected at a given time t_0. In the latter instance, we'll treat the states of edges (i.e. working or failed states) as independent Bernoulli variables. For concreteness, all references in the sequel to the reliability of a network pertain to two-terminal reliability.

It is known, of course, that the reliability of the network $G(v,n)$ in i.i.d. edges can be expressed as a polynomial $h(p)$ of order n, that is, as

$$h(p) = \sum_{r=1}^{n} d_r p^r, \tag{1}$$

where p is the common success probability for the edges. Satyarananaya and Prabhakar [10] showed that the coefficients in (1) could be obtained as the signed dominations associated with the network. Indeed, domination theory, described in 1984 as a breakthrough by Agrawal and Barlow [1] among computational tools in network reliability, continues to be a widely used algorithmic vehicle for calculating the reliability polynomial. We review the concept of dominations in Section 2. We note that, as useful as domination theory has proven to be in simplifying the computation of the reliability of a network, it has not been found particularly useful in

[1] Work supported in part by grant DAAD 19-99-1-1082 from the U.S. Army Research Office.

comparing one network design with another as is required, for example, in searching for universally optimal networks of a given size (v, n).

A quite different tool was introduced by Samaniego [9] for studying the performance properties of coherent systems. The concept of signature applies equally well to network reliability. The signature of a network is a probability vector $\boldsymbol{s}$ whose components are simply the respective probabilities that the first, second, $\cdots$, and nth edge failures (ordered by time of occurrence) are fatal to the network. Assuming, again, i.i.d. edge states at a fixed time t_0, the reliability polynomial of a network can be expressed in terms of the network's signature vector. Unlike the domination vector, the properties of the signature vector are readily interpretable and have a close relationship to the failure time T of the network itself. We review the notion of signature, and some of the problems to which it has been applied, in Section 3.

The main goal of this paper is to identify the exact relationship between the vector of signed dominations $\boldsymbol{d}$ and the signature vector $\boldsymbol{s}$. This is accomplished in Section 4. Because dominations are central to the computation of the reliability of a network, and signatures are rich in interpretation regarding the relative performance of competing networks, the exact linkage of the two through the functional relationship $\boldsymbol{s} = f(\boldsymbol{d})$ established here enables one to exploit the benefits of both. Our closing example demonstrates the utility of this linkage.

2. A Brief Look at Domination Theory

The notion of dominations was discovered in the process of seeking a reduction in the complexity of the well-known inclusion-exclusion formula (see [6]) for calculating the probability that all edges are functioning in at least one of a given network's minimal path sets. The inclusion-exclusion rule applies to the union of any m sets, and may be written as

$$\begin{aligned} P\left(\bigcup_{i=1}^{m} A_i\right) &= \sum_1 P(A_i) - \sum_2 P(A_i \cap A_j) + \sum_3 P(A_i \cap A_j \cap A_k) \\ &- \cdots + (-1)^{m+1} P\left(\bigcap_{i=1}^{m} A_i\right), \end{aligned} \tag{2}$$

where $\sum_i$ represents a sum over all i-fold intersections. If A_i in (2) represents the event that all edges in the ith minimal path set are working, and there are m minimal path sets in all, then the formula in (2) provides the probability that the network will function.

Suppose that one has a list of minimal path sets of a given network in n i.i.d. edges. A formation is defined as a union of minimal path sets. A *formation* is thus the union of the edges in a fixed collection of minimal path sets. Finally, an *i-formation* is a union of the components in a set of i minimal path sets. For example, the union $\{1, 2, 3, 4\}$ of the minimal path sets $\{1, 2\}$, $\{2, 3\}$, and $\{3, 4\}$ would be an example of a formation that is both a 2-formation and a 3-formation. We will refer to a particular formation as even if it is the union of an even number of minimal path sets and as odd if it is the union of an odd number of minimal path sets. It can, of course, be both at the same time.

The minimal path sets of this network are the sets of edges $\{1, 4\}$, $\{2, 5\}$, $\{1, 3, 5\}$, and $\{2, 3, 4\}$. The *signed domination* of a given union of minimal path sets is simply the difference between the number of even dominations and the number of odd dominations for that union. Satyarananaya and his co-workers showed that is a large variety of network reliability settings, the awkward expression for network reliability in (2) could be replaced by the simple form of the reliability polynomial in (1), where $(d_1 ... d_n)$ is the vector of signed dominations.

3. A Brief Look at Signatures.

The signature of a network of order n (that is, having n edges) is defined as the probability distribution $\boldsymbol{s}$ on the integers $\{1, 2, \cdots, n\}$ for which

$$s_i = P(T = X_{(i)}), \qquad i = 1, 2, \cdots, n, \tag{3}$$

where $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics from a random (i.i.d.) sample drawn from the (arbitrary) continuous lifetime distribution F, and T is the lifetime of the network.

The fact that the signature $\boldsymbol{s}$ depends only on the network design, and not on the distribution F, is a consequence of the fact that each of the $n!$ orderings of the failure times $X_1, X_2, \cdots, X_n$ of the n edges is equally likely to occur under the i.i.d. assumption. Thus, the probability that the ith edge failure is fatal to the network is solely dependent on the likelihood that the last working edge in some minimal cut set is the ith edge to fail overall. In other words, calculating s_i is simply a matter of examining minimal cut sets and counting how many among the equally likely permutations of $X_1, X_2, \cdots, X_n$ coincide precisely with a particular minimal cut set failing, before any other, upon the occurrence of $X_{(i)}$, the (ordered) ith edge failure time.

As shown in Samaniego [9] (see also [7]), the survival function of a network's lifetime T can be written as a simple function of $\boldsymbol{s}$ and F. When focusing on the reliability of the network at a fixed time t_0, where $P(X_j > t_0) = p$ for all j, this representation reduces to the reliability polynomial in pq-form, that is, in the form

$$h(p) = \sum_{j=1}^{n} \left(\sum_{i=n-j+1}^{n} s_i \right) \binom{n}{j} p^j q^{n-j}. \tag{4}$$

The tail probabilities of the signature vector $\boldsymbol{s}$ have an interpretation through the concept of path set. This connection was noted by Boland [4] and exploited in his study of indirect majority systems. As is apparent from (4), the coefficient of $p^j q^{n-j}$ in the reliability polynomial in pq-form can be interpreted as the number of path sets of order j, as it is precisely those sets, among the collection of $\binom{n}{j}$ sets with exactly j working components, that each contribute the positive probability $p^j q^{n-j}$ to the reliability polynomial. If we let a_j stand for the proportion of path sets among the $\binom{n}{j}$ sets of j working components (with the complementary components non-working), then we see that the reliability polynomial can be written as

$$h(p) = \sum_{j=1}^{n} a_j \binom{n}{j} p^j q^{n-j}. \tag{5}$$

It follows that the vector $\boldsymbol{a}$, which is fundamentally related to path sets, and the vector $\boldsymbol{s}$, which is fundamentally related to cut sets, are related to each other through the system of equations

$$a_j = \sum_{i=n-j+1}^{n} s_i, \qquad j = 1, \cdots, n, \tag{6}$$

For future reference, the linear relationship between the vectors $\boldsymbol{a}$ and $\boldsymbol{s}$ will be denoted as $\boldsymbol{a} = \boldsymbol{P}\boldsymbol{s}$.

In the introduction, we alluded to the fact that signatures are rich in interpretation and are particularly useful in the comparison of competing networks. We summarize here a collection of results that support this remark. The random variables

X_1 and X_2, discrete or continuous, are stochastically ordered (i.e., $X_1 \leq_{st} X_2$) if the survival functions $S_i(x) = P(X_i > x)$ are suitably ordered, that is, if $S_1(x) \leq S_2(x)$ for all x. We say that X_1 is smaller than X_2 in the hazard rate (or uniform stochastic) ordering if the ratio of survival functions $S_2(x)/S_1(x)$ is nondecreasing in x. This ordering will be denoted by $X_1 \leq_{hr} X_2$. Finally, X_1 is said to be smaller than X_2 in the likelihood ratio ordering ($X_1 \leq_{lr} X_2$) if the ratio $f_2(x)/f_1(x)$ is nondecreasing in x, where f_i represents the density or probability mass function of X_i. As is well known, stochastic order is the weakest of these three relations; indeed, it is easy to verify that $lr \Rightarrow hr \Rightarrow st$. With the notation established above, we may now restate results from Kochar, Mukerjee and Samaniego [7] relating properties of signatures to properties of network lifetimes.

Theorem 1 *Let $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ be the signatures of two networks with n i.i.d. edges, and let T_1 and T_2 be their respective lifetimes. If $\boldsymbol{s}_1 \leq_{st} \boldsymbol{s}_2$ or $\boldsymbol{s}_1 \leq_{hr} \boldsymbol{s}_2$ or $\boldsymbol{s}_1 \leq_{lr} \boldsymbol{s}_2$, then $T_1 \leq_{st} T_2$ or $T_1 \leq_{hr} T_2$ or $T_1 \leq_{lr} T_2$, respectively.*

The results above have been applied with profit to stochastic comparisons of k-out-of-n structures with system-wise or component-wise redundancy (see [7]), to indirect majority systems of varying design (see [4]) and to consecutive k-out-of-n systems with varying n (see [5]). We will apply these results again in an example in the concluding section.

4. The Linkage Between Dominations and Signatures.

As noted above, the reliability polynomial of a network with signature $\boldsymbol{s}$ may be written as

$$h(p) = \sum_{j=1}^{n} \left(\sum_{i=n-j+1}^{n} s_i \right) \binom{n}{j} p^j q^{n-j} \tag{7}$$

or equivalently as

$$h = (p) \sum_{r=1}^{n} \left(\sum_{j=1}^{r} a_j \binom{n}{j} \binom{n-j}{r-j} (-1)^{r-j} \right) p^r, \tag{8}$$

where $\boldsymbol{a}$ is as in (6). Examining the expressions in (1) and (8), we see that the vectors $\boldsymbol{d}$ and $\boldsymbol{a}$ are related via the equations

$$d_r = \sum_{j=1}^{r} a_j \binom{n}{j} \binom{n-j}{r-j} (-1)^{r-j}, \qquad r = 1, \cdots n. \tag{9}$$

Alternatively, the components of the domination and signature vectors satisfy the relationships

$$d_r = \sum_{j=1}^{r} \left(\sum_{i=n-j+1}^{n} s_i \right) \binom{n}{j} \binom{n-j}{r-j} (-1)^{r-j}, \qquad r = 1, \cdots, n. \tag{10}$$

If we denote the linear relationship between $\boldsymbol{d}$ and $\boldsymbol{a}$ in (9) as $\boldsymbol{d} = \boldsymbol{M}a$ and if we denote the linear relationship between $\boldsymbol{a}$ and $\boldsymbol{s}$ in (6) as $\boldsymbol{a} = \boldsymbol{P}s$, then we may write the relationship of interest as $\boldsymbol{s} = \boldsymbol{P}^{-1}\boldsymbol{M}^{-1}\boldsymbol{d}$.

The following result identifies this latter relationship explicitly.

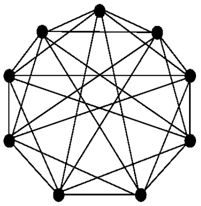
Theorem 2 *Let $\boldsymbol{d}$ and $\boldsymbol{s}$ denote the domination and signature vectors for a given network of order n. Then for $i = 1, \cdots, n$ we have*

$$s_i = \sum_{j=1}^{n-i} \frac{-(n-i)_j + (n-i+1)_j}{(n)_j} d_j + \frac{(n-i+1)_{n-i+1}}{(n)_{n-i+1}} d_{n-i+1}, \tag{11}$$

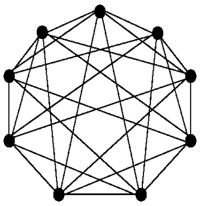
where $(k)_j$ denotes the number of ways of selecting without replacement j items from k items, accounting for the order of selection. For a proof of this result, see Boland, Samaniego and Vestrup [6].

We suggested in Section 1 that the comparison of networks via their vector of dominations was unintuitive and, for complex networks, quite difficult. The reason for this is that the difference of two polynomials in standard form (that is, in the form displayed by (1)) is another polynomial in standard form. For two complex networks, the difference polynomial $\sum(d_{2r} - d_{1r})p^r$ will typically be of quite high degree. Because of the requirement $\sum d_r = 1$ on the domination vector of an arbitrary network, it can never be the case that all coefficients of the difference polynomial will have the same sign. Thus, determining whether one reliability polynomial is uniformly larger than another for all $0 < p < 1$ is a task equivalent to finding the roots of a high degree polynomial. But that algebraic problem is a quite famous one, a problem that was dramatically resolved by Evariste Galois. Finding roots of polynomials of degree greater than 4 is a problem that is not "solvable by radicals"; thus, closed-form expressions for the solutions of such problems are not possible in general.

Transforming this problem into the world of signatures changes things substantially. To see this more graphically, let us consider the comparison between the two G(9,27) networks of order, pictured below.



AX3



AX4

In standard form, the reliability polynomials for these two networks are found to be

$$\begin{aligned} h_{AX3}(p) = \quad & 419904p^{27} - 6021144p^{26} + 41705280p^{25} - 18489826p^{24} \\ & +586821717p^{23} - 1413876060p^{22} + 2677774329p^{21} \\ & -4074363810p^{20} + 5048856414p^{19} - 5135792742p^{18} \\ & +4303029693p^{17} - 2967712776p^{16} + 1676975886p^{15} \\ & -769265910p^{14} - 282176568p^{13} + 80853282p^{12} \\ & +17445456p^{11} - 2667060p^{10} + 257634p^{9} - 11828p^{8} \\ h_{AX3}(p) = \quad & 414720p^{27} - 5934288p^{26} + 41015964p^{25} - 181453380p^{24} \\ & +574666025p^{23} - 1381692972p^{22} + 2611463517p^{21} \\ & -3965536554p^{20} - 4904464002p^{19} + 4979513718p^{18} \\ & +4164454729p^{17} - 2867022480p^{16} + 1617256842p^{15} \\ & -740601350p^{14} - 271201476p^{13} + 77576922p^{12} \\ & +16709916p^{11} - 2550156p^{10} + 245898p^{9} - 11268p^{8} \end{aligned}$$

While the uniform superiority of one of these networks over the other is certainly not obvious by inspection, one could by numerical means show that $h_{AX3}(p) \geq h_{AX4}(p)$ for all $p \in [0, 1]$. However, the comparison of the two signatures immediately yields this same conclusion, and in addition, a stronger one. From table 1 below, it is apparent that the signature of network AX3 is stochasitcally larger than that of AX4, yielding the uniform domination of AX3 over AX4 alluded to above. However, the comparison of the two signatures vectors yields an additional new insight. The ratios of the two survival functions displayed in the last column of table 1 shows that AX3 dominates AX4 in the hazard rate ordering as well. We can thus rightly say that AX3 is not only better than AX4, it is actually quite a bit better!

TABLE 1: Signature Tail Probabilities $S(x) = \sum_{i=x}^{27} s_i$ And Their Ratios

x	$S_{AX3}(x)$	$S_{AX4}(x)$	$S_{AX3}(x)/S_{AX3}(x)$
1	1.0	1.0	1.0
2	1.0	1.0	1.0
3	1.0	1.0	1.0
4	1.0	1.0	1.0
5	1.0	1.0	1.0
6	1.0	1.0	1.0
6	1.0	1.0	1.0
7	0.999970	.0999970	1.0
8	0.999787	.0999787	1.0
9	0.999149	0.999149	1.0
10	0.997367	0.997367	1.0
11	0.993612	0.993612	1.0
12	0.985922	0.985922	1.0
13	0.971744	0.971743	1.0000005
14	0.947220	0.947214	1.0000063
15	0.906907	0.906867	1.0000442
16	0.843421	0.843240	1.0002148
17	0.747317	0.746717	1.0008024
18	0.607883	0.606416	1.0024183
19	0.417560	0.415077	1.0059834
20	0.189140	0.186804	1.0125000
21	0.0	0.0	–
22	0.0	0.0	–
23	0.0	0.0	–
24	0.0	0.0	–
25	0.0	0.0	–
26	0.0	0.0	–
27	0.0	0.0	–

The main thesis of this note can be summarized quite succinctly: Domination theory is a useful tool in making network reliability calculations. However, for the purpose of comparing the performance characteristics of two competing networks, reliabilities expressed in terms of signed dominations will tend to have little intuitive content and may be of limited use (except for the possibility of brute force computation). The utility of signatures in the comparison of networks of the same size immediately raises questions about the exact relationship between the domination and signature vectors. The functional relationship linking the signature vector with the vector of dominations is displayed above. This linkage allows one to combine the computational advantages of domination theory with the intuitive and interpretive qualities of signatures for the purpose of making comparisons among networks. The growing but still inconclusive literature on the existence, uniqueness and identification of uniformly optimal networks of a given size (see, for example, [2], [3], [8], and [12]) should benefit from the application of these linked tools.

5. References.

[1] Agrawal, A. and Barlow, R.E., A survey of network reliability and domination theory, **Operations Research** (1984), 32, 478-92.

[2] Y. Ath and M. Sobel, Some conjectured uniformly optimal reliable networks, **Probability in the Engineering and Informational Sciences** (2000), 14, 375-83.

[3] F. T. Boesch, X. Li, and C. Suffel, On the existence of uniformly optimally reliable networks. **Networks** (1991) 21, 181-91.

[4] P. Boland, Signatures of Indirect Majority Systems, **Journal of Applied Probability** (2001), 38, (to appear)

[5] P.J. Boland and F.J. Samaniego, A Note on Stochastic Ordering for Consecutive k-out-of-m Systems, Technical Report #395, (2001), Department of Statistics, University of California, Davis

[6] P.J. Boland, F.J. Samaniego, and E.M. Vestrup, Linking Dominations and Signatures In Network Reliability Theory, Technical Report #394, (2001), Department of Statistics, University of California, Davis

[7] W. Feller, **Introduction to Probability Theory and its Applications** (1968), 3rd edition, New York: Wiley and Sons

[8] S. Kochar, H. Mukerjee and F.J. Samaniego The signature of a coherent system and its application to comparisons among systems, **Naval Research Logistics**, 46 (1999) 507-23.

[9] W. Myrvold, K.H. Cheung, L.B. Page, and J.E Perry, Uniformly most reliable networks do not always exist. **Networks** (1991), 21, 417-19.

[10] F.J. Samaniego, On Closure of the IFR Class Under Formation of Coherent Systems, **IEEE Transactions on Reliability Theory**, R-34 (1985), 69-72.

[11] A. Satyarananaya, and A. Prabhakar, A New Topological Formula and Rapid Algorithm for Reliability Analysis of Complex Networks, **IEEE Transactions on Reliability Theory** (1978), R-30, 82-100.

[12] M. Shaked and J.G. Shanthikumar, **Stochastic Orders and Their Applications** (1994), San Diego: Academic Press.

[13] G. Wang, A proof of Boesch's conjecture. **Networks** (1994), 24, 277-84.

Statistical Methods in Software Engineering for Defense Systems: Summary of a Workshop

Jesse Poore, University of Tennessee,
Siddhartha Dalal, Telcordia

Since the major defense system acquisition programs are essentially all software-intensive, their development into mature systems involves software engineering. Of late, a large percentage of these acquisition programs are either late, over budget, or provide less than their total functionality due to problems with the embedded software. To address this, the Committee on National Statistics and the Committee on Applied and Theoretical Statistics of the National Academy of Sciences organized a workshop July 19-20, 2001 entitled "Statistical Methods in Software Engineering for Defense Systems. At this workshop, experts from academia and industry suggested methods that have demonstrated their value in analogous industrial applications, with defense software experts providing opinions as to their applicability for defense systems.

We present an overview of the primary themes supported by the workshop presentations. These are the value of model-based testing, exemplified by markov chain usage models and automated efficient test generation system, in addressing the astronomical space of sequences of user commands, and the benefits of test automation and its requisites, especially the need for rigor in software requirements specification. In addition, the speakers will touch on the following issues that also received some attention: (a) the benefits gained from proper approaches to system architecture, (b) new approaches to treating interoperability problems, (c) the necessary resources to institute statistical software engineering methods in the defense operational test agencies, and (d) additional methods in areas such as defect analysis, models of software aging, measurement of software risk, and modeling costs of software development.

Special Session III

Network Evaluation Via Activity Vector Clustering

Jeff Solka, U.S. Naval Surface Warfare Center

Evaluation of the various activities on a multi-platform network can be a difficult task. This talk will examine the application of some rudimentary statistical procedures to the characterization of activities on a moderately sized network. The application of agglomerative clustering and visualization methods will be illustrated as applied to network activity vectors.

Audit Data Analysis and Mining

Ningning Wu, University of Arkansas, Little Rock

The need for intrusion detection systems that are able to monitor large amounts of audit trail data, detect intrusions quickly, and generate few false alarms is pressing. This presentation introduces a network anomaly detection system, Audit Data Analysis and Mining (ADAM), which uses an application of association rules and classification techniques to detect attacks using the network audit trail data. ADAM is able to detect network intrusions in real time with very low false alarm rate. One of its advantages is the ability to detect novel attacks without dependency on the training data of attacks, due to a novel application of the pseudo-Bayes estimators technique.

Mining a Data Stream to Understand User Behavior

Diane Lambert, Bell Labs, Lucent Technologies

Data on transactions, like credit card purchases, calls or web accesses, arrive in a fast, never-ending stream. These data contain an enormous amount of information about how users and customers behave, but extracting up-to-date information on users and analyzing it in real time -- faster than the data arrive -- is a huge challenge. This talk describes an automated, statistically principled way to design short, accurate summaries of high dimensional user behavior that can be kept current with a stream of transactions and can be used both for automated real-time analysis (e.g., fraud detection) and less formal exploratory data analysis. An example involving the calls of a set of 96,000 customers who made about 18 million wireless calls over a three-month period will be presented.

General Session III

Accelerated Testing: Obtaining Reliability Information Quickly

William Q. Meeker, Iowa State University

Accelerated tests are used to obtain timely information on products reliability. Changes in technology, the calls for rapid product development, and the need to continuously improve product reliability have combined to increase the need for developing improved methods for accelerated testing. Laboratory tests with increased use rates or higher than usual levels of accelerating variables like temperature or voltage are used to accelerate failure mechanism. Then the results are used to make predictions about product life or performance over time at use or design conditions. The predictions involve extrapolation in several dimensions. Interesting statistical problems arise in modeling physical phenomena, use of engineering/physical information, planning accelerated tests, and quantifying uncertainty. This talk reviews the basic physical and statistical models and methods used in accelerated testing. Current research in this area will be outlined and areas for future research will be described.

Contributed Session IX

A Special Topic in Risk Analysis
Bernie Harris, University of Wisconsin, Madison

Abstract unavailable

Exact Moments of the 2 x 2 x 2 Distribution
Robert Launer, U.S. Army Research Laboratory, Army Research Office

In the 1974 Army Design of Experiments Conference, the author proposed an exact model for the distribution of the 2X2 contingency table chi-square statistic under the alternate hypothesis. The exact moments for the asymptotic distribution of the resulting statistic under the alternate hypothesis were calculated as functions of the distributional parameters. That information was used to obtain easily computed type II errors for moderate to large sample analyses. In this talk, that analysis is extended to the 2X2X2 statistic.

Mixed Model Inference for Army Test and Evaluation
Thomas Mathew, University of Maryland, Baltimore County
David Webb, U.S. Army Research Laboratory

Mixed effects and random effects models are widely used for analyzing Army Test and Evaluation data. In particular, such models are used for investigating gun tube accuracy under a wide array of firing conditions. The study of tube-to-tube variability is a major concern in this context. The problems that arise in this situation turn out to be somewhat different from the traditional problems encountered in mixed and random effects models. Furthermore, the development of finite sample inference is a primary concern, the reason being that there is a high cost per observation associated with Army Test data. In the talk, I will introduce some of the mixed effects models, and the relevant hypotheses testing problems that arise in the study of gun tube accuracy. The concept of a generalized p-value will be used to provide solutions to such testing problems. Comparisons with some approximate tests will also be discussed.

General Session IV

Visual Data Mining of Remote Sensing Data

Jürgen Symanzik, Utah State University

In the first part of this talk, we look at techniques (e.g., linked brushing, grand tour, parallel coordinates) and software tools (e.g., XGobi, ExplorN, and the ArcView/XGobi link) that are useful for visual data mining. We focus on the visual exploration of satellite images that consist of multiple spectral bands that have been remotely sensed by earth observation satellites.

In one of our examples, the area of interest is the greater Atlanta, GA, region. Eighteen satellite images from January 1997 through December 1997 form the basis of our analysis. Visual data mining techniques provide us with valuable insights into this spatial/temporal data set. Other examples of visual data mining of remote sensing data are given as well.

Author Index

Rec'd 17 MAR 04

Quality Assurance and OPSEC Review

This form is an approval record for ARL generated information to be presented or disseminated external to ARL. Note: Submit all manuscripts in electronic format or camera ready copy. See attached instructions. If more space is needed, use reverse of form *(include block numbers)*.

A. General Information

1. Present Date
02//11/04

2. Unclassified Title
Proceedings of the Seventh Annual U.S. Army Conference on Applied Statistics

3. Author*(s)*
Barry A. Bodt
Edward J. Wegman

4. Office Symbol*(s)*
AMSRD-ARL-CI-CT
George Mason University

5. Telephone Nr*(s)*
(410) 278-6659
(703) 993-1691

6. Contractor generated ☒ No □ Yes
If yes, enter Contract No. and ARL COR

7. Type: ☒ Report □ Abstract □ Publication □ Presentation *(speech, briefing, video clip, poster, etc)* □ Book □ Book Chapter □ Web

8. Key Words
Data Mining, Statistics, Experimental Design

9. Distribution Statement *(required)* Is manuscript subject to export control? ☒ No □ Yes

Circle appropriate letter and number. *(see instructions for statement text)*
☒ B C D E F X 1 2 3 4 5 6 7 8 9 10 11

10. Security Classification
Unclassified

B. Reports

11. Series
TR

12. Type
FINAL

13. No. of pages
175

14. Project No.
P611102.H48

15. Period Covered
24-26 Oct 01

16. Sponsor

C. Publications

17. Is MS an invited paper? □ No □ Yes

18. Publication is a refereed journal?
□ No □ Yes

19. Material will be submitted for publication in

Journal ______ Country ______

D. Presentations

20. Conference Name/Location

21. Sponsor

22. Conference Date

23. Due Date

24. Conference is □ Open to general public □ Unclassified/controlled access
□ Classified

25. For nonpublic meetings: Will foreign nationals be attending?
□ No □ Yes *(If yes, list countries and identify International Agreement(s))*
□ Don't know

26. Material will be
□ Oral presented only □ Oral presented and published in

(If published, complete block 18 and 19, Section C.)

E. Authors Statement:

27. All authors have concurred in the technical content and the sequence of authors. All authors have made a substantial contribution to the manuscript and all authors who have made a substantial contribution are identified in Block 3.

Barry A. Bodt Barry A. Bodt
ARL Lead Author or COR

2/18/04
Date

F. Approvals

28. First line Supervisor of Senior ARL Author or COR
Patricia H. Jones Patricia H. Jones
Name

2/18/04
Date

29. Reviewer(s) (Technical/Editorial/NA)
(Tech)
Name*(s)*

13 Feb 2004
Date

30. Limited distribution information for release to foreign nationals

Foreign Disclosure ______ Date ______

31. Classified Information

Classified by ______

Declassified on ______

Command Security Manager ______ Date ______

OPSEC REVIEW CHECKLIST

OPSEC POC: Complete and explain any positive responses in block 9.
Note: ARL must be the proponent of the proposed information for release.

1. Does this material contain Sensitive Information? ☐ YES ☒ NO

2. Does this information contain state-of-the-art, breakthrough technology? ☐ YES ☒ NO

3. Does the United States hold a significant lead time in this technology? ☐ YES ☒ NO

4. Does this information reveal aspects of reverse engineering? ☐ YES ☒ NO

5. Does this material reveal any security practices or procedures? ☐ YES ☒ NO

6. Does this information reveal any security practices or procedures? ☐ YES ☒ NO

7. Would release of this information be of economic benefit to a foreign entity, adversary, or allow for the development of countermeasures to the system or technology? ☐ YES ☒ NO

8. Does this material contain:

a. Any contract proposals, bids, and/or proprietary information? ☐ YES ☒ NO

b. Any information on inventions/patent application for which patent secrecy orders have been issued? ☐ YES ☒ NO

c. Any weapon systems/component test results? ☐ YES ☒ NO

d. Any ARL-originated studies or after action reports containing advice and recommendations? ☐ YES ☒ NO

e. Weakness and/or vulnerability information? ☐ YES ☒ NO

f. Any information on countermeasures? ☐ YES ☒ NO

g. Any fielding/test schedule information? ☐ YES ☒ NO

h. Any Force Protection, Homeland Defense *(security)* information? ☐ YES ☒ NO

i. Information on subjects of potential controversy among military services or other federal agencies? ☐ YES ☒ NO

j. Information on military applications in space, nuclear chemical or biological efforts: high energy laser information; particle beam technology; etc? ☐ YES ☒ NO

k. Contain information with foreign policy or foreign relations implications? ☐ YES ☒ NO

OPSEC Approval Statement

I, the undersigned, am aware of the adversary's interest in DOD publications and in the subject matter of this material and that, to the best of my knowledge, the net benefit of this release out weights the potential damage to the essential security of all ARL, AMC, Army, or other DOD programs of which I am aware.

FRED S. BRUNDICK Fred S Brundick
OPSEC Reviewer *(Printed name/signature)*

11 Feb 2004
Date

9. Space for explanations/continuations/OPSEC review comments

Final Release Clearances

32. Public/Limited release information

a. Material has been reviewed for OPSEC policy.

Douglas Kingsley
ARL OPSEC Officer

17 MAR 04
Date

b. The information contained in this material is ☒ / is not ☐ approved for public release/ has received appropriate tech/editorial review.

Division Chief

10 Mar 04
Date

c. This information is accepted for public release.

Douglas Kingsley
Public Affairs Office

17 MAR 04
Date