Proceedings of the Sixth Annual U.S. Army Conference on Applied Statistics, 18-20 October 2000

Barry A. Bodt, Edward J. Wegman
EDITORS

Hosted by:
RICE UNIVERSITY

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 Sixth Annual U.S. Army Conference On Applied Statistics, 18-20 October 2000

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

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

Hosted by:
Rice University

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
SIXTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

Special Session on Digital Government

Luncheon Presentation

Clinical Session I

Contributed Session V

SIXTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

ABSTRACT

The Sixth U.S. Army Conference on Applied Statistics (ACAS) was hosted by Rice University in Houston, Texas during 18-20 October 2000. 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 of Standards and Technology (NIST). Several other organizations cooperated with planners to support the conference. Those organizations are RAND, Los Alamos National Laboratory (LANL), George Mason University (GMU), Office of Naval Research (ONR), and the Institute for Defense Analyses (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 the sixth conference and thirty-seven papers were presented. This document is a compilation of available papers offered at the conference.

FOREWORD

The Sixth U.S. Army Conference on Applied Statistics (ACAS) was hosted by Rice University in Houston, Texas during 18-20 October 2000. 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 of Standards and Technology (NIST). Several other organizations cooperated with planners to support the conference. Those organizations are RAND, Los Alamos National Laboratory (LANL), George Mason University (GMU), Office of Naval Research (ONR), and the Institute for Defense Analyses (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. The purpose of this conference is to promote the practice of statistics in the solution of these diverse Army problems.

The sixth ACAS boasted a balance of education, distinguished speakers, special topics, and contributed papers of DoD interest. The conference was preceded by a short course, "Data, Knowledge, and Information Integration to Support Decision Making," presented by the Statistics Group, LANL. Approximately forty-five students took part. Invited, general session addresses were heard from Nozer Singpurwalla (keynote), George Washington University; Stuart Geman, Brown University; Alan Agresti, University of Florida; Noel A. C. Cressie, Ohio State University; Emanuel Parzen, Texas A&M University; Donald Berry, University of Texas, MD Anderson Cancer Center; Naomi Oreskes, University of California, San Diego; and Matthew Caffrey, Air

Command and Staff College. Three especially relevant special sessions anchored the conference agenda. These sessions were Biological Warfare (organized by Marek Kimmel, Rice University), Digital Government (organized by Edward Wegman, GMU), and Reliability (organized by Arthur Fries, IDA). More than twenty contributed papers rounded out the program. An important event at the conference was the awarding of the Army Wilks Award to C.R. Rao of Penn State University for a distinguished career of service to the field of statistics and to the statistics research program of the ARO.

The Executive Board for the conference recognizes James R. Thompson and Diane Brown, Rice University, for hosting the conference; David Webb, ARL, and Edmund Baur, ARL, for assisting with web page advertisement; Edward Wegman, GMU, for fiscal oversight and the production of these proceedings; Jock Grynovicki, ARL, for chairing the Army Wilks Award Committee, and Barry Bodt, 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 Health Sciences*
Paul J. Deason *U.S.A. Training and Doctrine Command*	Eugene F. Dutoit *U.S. Army Infantry School*	Arthur Fries *Institute for Defense Analyses*
Jock O. Grynovicki *U.S. Army Research Laboratory*	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 Health Sciences*	James R. Thompson *Rice University*
Mark G. Vangel *National Institute Standards & Technology*	David W. Webb *U.S. Army Research Laboratory*	Edward J. Wegman *George Mason University*
Alyson G. Wilson *Los Alamos National Laboratory*		

SIXTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

SHORT COURSE

DATA, KNOWLEDGE, AND INFORMATION INTEGRATION TO SUPPORT DECISION MAKING

Presented by
The Statistics Group
Los Alamos National Laboratory

Abstract: This course will cover structured, quantitative approaches for combining data, knowledge, and information from multiple sources and in various forms (including both quantitative and qualitative) to support decision making. Techniques from various disciplines, including state-of-the-art expert elicitation, statistical and reliability analysis, and knowledge management, will be discussed and applied to develop formal methods for information integration. Possible DoD applications will be addressed.

Schedule:
Day 1 - The "Data" Day

(1) Welcome and Introduction (15 min)

(2) Course Overview, Methodology Development and Motivation (30 minutes)

(3) Problem and System Structuring, Decision Context (1.5 hours)
- Develop course example
- Defining problem objectives
- Developing decision contexts
- Major classes of representations and their strengths and weaknesses

(4) Capturing, Representing, and Quantifying Data and Expertise (2 hours)
- Sources of information and how to find them
- Information capture
- Information elicitation
- Quantifying expert judgment
- Defining communities of practice
- Obtaining structural and content knowledge from experts

(5) Creating Knowledge Systems (2 hours)
- Types of knowledge systems
- Organizing the contents of a knowledge system based on elicited structure
- Automating elicitation
- Knowledge access and protection

Day 2 - The “Now What?” Day

(6) Information Integration Analysis Issues and Tools (4 hours)
- Sources of uncertainty
- Reliability models
- Bayesian hierarchical models
- Propagation of information "up and down" the system structure

(7) Estimation, Planning, and Decision Making (2 hours)
- Development and structure of "what-if" analyses
- Optimization: experimental design basics, experimental design "new directions," sensitivity analysis
- Problem solutions in a decision context

(8) Conclusions and Resources (30 minutes)
- Decision context revisited
- Who do I hire?
- What do I buy?
- How much does it cost?
- How long does it take?

SIXTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

WEDNESDAY, OCTOBER 18

GENERAL SESSION I (0900 - 1030)

Warranty Contracts and Equilibrium Probabilities
Nozer Singpurwalla, George Washington University

The scenario of warranties is at the interface of philosophy, law, and probability. In this talk we describe a real life scenario involving litigation pertaining to a breach of warranty and discuss its ramifications from a statistical point of view. We claim that the three interpretations of probability, the objective, the subjective and the logical all come into play when designing an optimum warranty that is also just.

Variance and Invariance in Machine Vision
Stuart Geman, Brown University

I will propose a computer vision system based upon a collection of scale-invariant composition rules that define a part-whole hierarchy. I will make a connection to some striking invariance properties of natural images. I will suggest a coarse-to-fine computing engine for scene analysis within the compositional framework. I will discuss some experimental results and make some connections to biological vision systems.

SPECIAL SESSION ON BIOLOGICAL WARFARE (1100 - 1215)

Dispersal of Bacterial Pathogenes: Scenarios, Models and Accidents
Marek Kimmel, Rice University

Abstract unavailable

Use of Genomic Technologies to Decode Bacterial Biological Warfare Agents
George Weinstock, Department of Molecular Virology & Microbiology, Baylor College of Medicine

Bacteria contain a number of genes that contribute to their ability to cause human infections and to resist the action of antibiotics. However only a subset of bacteria within a given species contain these genes. Thus, some bacteria of a species contain only a few genes involved in infection while others contain many such genes. These pathogenicity and antibiotic resistance genes can be transferred between bacteria of the same species and often between bacteria of different species. In addition, these genes can be inserted into bacteria in the laboratory by recombinant DNA techniques, which facilitates the design and construction of bacteria that are highly virulent and also resistant to most antibiotics. Engineered bacteria of this type represent possible biological warfare or terrorism agents. Genomics technologies such as high-throughput sequencing and DNA chips are mechanisms by which the constellation of pathogenicity and antibiotic resistance genes of any bacterial isolate can rapidly be assessed.

CONTRIBUTED SESSION I (1330 - 1500)

AN INVESTIGATION IN THE USE OF CLUSTER ANALYSIS IN HELPING TO ESTABLISH AN OPERATIONAL SCENARIO FOR A COMBAT SIMULATION

Eugene Dutoit
Dismounted Battlespace Battle Lab, Fort Benning, GA

ABSTRACT

As part of an advanced concepts technology demonstration a simulation study was conducted to determine the operational effects, with respect to communications, on clearing a building in an urban setting. Prior to force-on-force simulations the study agency wanted to determine, in advance, the probable communications locations (or nodes). Subject matter experts, (SMEs), were provided floor diagrams and a description of the scenario. They were asked to identify locations where they believed that communications would occur. A clustering algorithm was used to investigate if the group of SMEs were reasonably consistent about the locations of the subjectively positioned communications nodes. This paper will present the results of this portion of the study and present a simple measure of subjective clustering consistency.

PURPOSE OF THIS PAPER

The purpose of the paper is to present a heuristic application of the K-means cluster analysis algorithm as a tool for developing a scenario that will be used for force-on-force simulation. As part of the military operations in urban terrain (MOUT) advanced concept technology demonstration (ACTD) a sequence of simulation studies were planned and conducted to determine the performance requirements for communications devices (radios) operating in a MOUT (city) environment. As a first step to address this question an initial simulation study was conducted to determine the operational sequences with respect to communications on a floor clearing operation in an urban building. That is; where are the likely places that communications will take place during a floor clearing operation? This information was used to develop the initial scenario. The results of the force-on-force modeling were then used as inputs to engineering / physics models to derive the physical requirements.

SCENARIO DEVELOPMENT

The Blue force was attacking and the Threat force was defending the floor. Prior to the force-on-force simulations, the study agency wanted to determine, in advance, the probable communications locations (or nodes) on the floor that the attacking force would use to coordinate the attack. There were four levels of Threat considered for the scenario; no soldiers on the floor (zero threat), three per floor (considered low threat), five per floor (considered medium threat) and seven per floor (considered high threat). In addition to these levels of Threat there were two levels of Blue communications capability. Limited communication was defined such that platoon leaders could communicate with all squad leaders using communications assets such as SINGARS and PRC-126 and the squad leaders were able to communicate with each other. Team leaders did not have radio assets and therefore all communications between squad leaders and team leaders were the use of voice and arm signals. Perfect communications was defined such that the platoon leaders could communicate with all squad leaders and team leader. In addition, the squad leaders could communicate with both team leaders beneath them as well as their two other squad leaders.

The study agency obtained a detailed engineering drawing of the floor plan which served as the "terrain" for the combat simulation. The drawing showed, in scaled detail, the location of offices, windows, stairwells, elevators and bathrooms. It was the floor plan as it actually existed in the real world. A grid-coordinate system was superimposed over the engineering drawing. The coordinate system was laid out so that the abscissa was scaled from 0 to 20 and the ordinate was scaled from 0 to 10. Keep these coordinate limits in mind because they will become important in a later discussion in this paper.

METHODOLOGY

Five subject matter experts (SMEs) indicated ,using the grid system cited above, the (X,Y) coordinates where communications would be most appropriate during the floor clearing operations. The SMEs were active Army officers with ranks of Captain to Lieutenant Colonel. Each SME was presented with the same floor plan and a concept of operation. Each of the eight combinations of "threat and level of communications" were presented to each of the five SMEs. Each SME plotted his subjective set of coordinates (communications nodes) for each threat and level of communications combination on the floor plan.

An Analytical Question (The Basis Of This Presentation)

Each of the five SMEs indicated where they would place the nodes of communication for each of the eight combinations of "threat and level of communication." After this was done the question was asked whether the various sets of plots could be examined to determine if there was some agreement to the locations of the subjectively positioned nodes? Stated in another way; is there some consistent structure in the positioning of the proposed communications nodes as they were presented on the floor plan? *An after-action discussion with the five SMEs indicated that there was, and would be, a high degree of consistency and agreement among the SMEs about the positioning of the communications nodes. Given the clearly defined mission of clearing the floor and the use of the appropriate tactics, techniques and procedures (TTPs) for a floor clearing operation, one would expect consistency of judgement from a set of trained Army officers.*

At this point in the study it was decided to investigate the use of cluster analysis to determine if it would provide some heuristic measure of agreement to complement the italicized statement in the above paragraph. It would be considered to be an objective confirmation of subjective input. In order to apply the clustering process the study leaders (not the SMEs) visually examined each of the eight sets of coordinate plots and subjectively grouped them into sets of clusters which appeared to be relatively independent. The K-means clustering algorithm was chosen for this task. *The algorithm requires that the number of clusters be specified in advance. These were subjectively determined by the study leaders.* The K-means clustering process can be described as ANOVA in reverse. The program starts with K specified random clusters, and then moves objects between those clusters with the goal to (1) minimize variability within clusters and (2) maximize variability between clusters. The ANOVA that is performed is a check on the efficiency of the algorithm and should be interpreted for descriptive purposes. The results should not be interpreted as tests of hypothesis that the cluster means are equal. Hence the *heuristic* interpretation.

All the data coordinates and the subjectively assigned clusters for each of the eight separate "threat/communications" combinations were analyzed separately. The SPSS algorithm uses a fast clustering procedure to group the data points into the required number of clusters. As a complementary indication of agreement the subjective visual clustering results provided by the study leaders were compared to the results of the fast clustering algorithm. A measure of agreement between the original subjective clustering and the results of the K-means clustering algorithm (objective clustering) was defined as:

(The number of computed cluster centers that are contained within the initial subjective clusters) /
(The number of subjective clusters)

Table 1 (given on the next page) shows the results of the clustering agreement for each of the eight "threat/communications" combinations expressed as a percentage.

TABLE 1
AGGREEMENT (PER CENT) FOR THE INITIAL SUBJECTIVE CLUSTERS

Scenario (Threat/communications combination)	Clustering sanity of the initial subjective clusters (%)
No threat/limited communications	88
No threat/perfect communications	89
Light threat/limited communications	94
Light threat/perfect communications	85
Medium threat/limited communications	80
Medium threat/perfect communications	86
Heavy threat/limited communications	83
Heavy threat perfect communications	88

The average value for the agreement measures as given in Table 1 is 86.6%. This was considered to be a respectable degree of consistency between the objectively determined clusters and the results of the study leaders. The following comments are offered.

Comments

a. The use of the K-means algorithm is not totally independent process. It is a function of the subjectively estimated number of clusters.
b. The study leaders built their clusters based on their knowledge of the geometry of the floors (walls, stairs, hallways, locations of rooms). Their clusters were also a function of their knowledge of TTPs and the terrain. The algorithm derives the clusters based on the proximity of the data points to each other and instructions on the number of clusters to be formed. The algorithm is not aware of the terrain or military doctrine.
c. In this case it was decided that the average value for the measure of agreement (86.6%) is a respectable degree of consistency of determining communication nodes and forming the subjective clusters.
d. There may be other applications of this procedure in the field of perceptual psychology and the ability of people to form groups/clusters based on a rough scatter of data or information and their prior knowledge about the subject matter to be clustered.

A Note Of Caution

Like all statistical or analytical tools or techniques, the use of clustering algorithms should be conducted with caution. The following example will illustrate what can happen if these techniques are applied blindly or without any knowledge of the field of application. Consider the description of the coordinate system described in the paragraph above (Scenario Development) and presented below for the convenience of the reader: "*A grid-coordinate system was superimposed over the engineering drawing. The coordinate system was laid out so that the abscissa was scaled from 0 to 20 and the ordinate was scaled from 0 to 10*" Assume that the data analyst knows nothing about the military tactics, techniques or procedures that are instrumental in establishing the initial/subjective set of communications nodes and their clusters. This naive analyst examines the plotted set of coordinates for each the eight threat/ communications combinations and decides that the floor can be divided into *two clusters* for each of the combinations. The ignorant analyst uses the k-means algorithm and specifies two clusters. By the nature of the clustering algorithm we would expect that the resultant cluster centers to fall somewhere in the middle of the two halves of the coordinate system described above with centers at (5,5) and (15,5) for each of the two specified clusters respectively. All measures of agreement would be 100% for each of the combinations and the analyst would be led to believe that the clusters represented the "true answer." This experiment was conducted for each of the eight threat/communications combinations and the results are shown in Table 2 below. Note that the coordinates of the computed clusters are all within the range of the expected values. However, from an operational point of view, the answer is wrong.

TABLE 2
COORDINATES OF CLUSTER CENTERS USING THE K-MEANS ALGORITHM FOR TWO SPECIFIED CLUSTERS

Scenario (Threat/communications combination)	Cluster 1	Cluster 2
No threat/limited communications	X = 4.6, Y = 4.8	X = 14.3, Y = 4.7
No threat/perfect communications	X = 4.6, Y = 4.9	X = 14.4, Y = 4.6
Light threat/limited communications	X = 14.3, Y = 4.8	X = 4.5, Y = 5.0
Light threat/perfect communications	X = 14.3, Y = 4.8	X = 4.5, Y = 5.1
Medium threat/limited communications	X = 14.4, Y = 4.9	X = 4.6, Y = 4.9
Medium threat/perfect communications	X = 14.3, Y = 4.8	X = 4.5, Y = 5.1
Heavy threat/limited communications	X = 14.3, Y = 4.9	X = 4.5, Y = 5.0
Heavy threat perfect communications	X = 14.3, Y = 4.8	X = 4.5, Y = 5.1

These results clearly show that ignorance of the subject matter field can lead to wrong conclusions and confirmations regardless of how sophisticated the analysis tool.

REFERENCES

Bayne, C., Beauchamp, J., Begovich, C., Kane, V. (1977), Monte Carlo Comparisons Of Selected Clustering Procedures. Paper presented at the annual meeting of the American Statistical Association.

SPSS Base 8.0 User's Guide, 1998, 315-321.

Army Research Laboratory (ARL)
Insight Report Notes

A Human Factors Evaluation of the Custom Tent Design used during the 1st Brigade Task Force Lanes Exercise

Fort Hood, Texas

Army Research Laboratory (ARL)
Human Research and Engineering Directorate (HRED)

Dr. Jock O. Grynovicki
Mr. Kragg P. Kysor
Dr. Madeline B. Swann

ACKNOWLEDGEMENT

The authors would like to acknowledge the contributions of Ronald Carty, of the U.S. Army Research Laboratory, to the research of alternative tactical operations center (TOC) shelter concepts. Mr. Carty provided outstanding photographic support, including audio-visual videography and digital still images, of the various shelter concepts and operational procedures involved in their employment. These images were used as sources of data in analyzing physical and human factors aspects of prototype equipment and military personnel tasks related to the TOC shelter designs. In addition, Mr. Carty provided instrumentation and obtained light measurements for representative locations within the TOC shelters that were used for evaluating the alternative design concepts. Mr. Carty's contributions were made with a high level of professionalism and were invaluable in providing needed data for the study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This paper summarizes insights developed during the 1st Brigade Task Force Lanes Exercise relative to the Custom Tent shelter used for the Combat Information Cell (CIC) of the 1st BCT Main CP. Through the TOC Summit venue, it was recognized that the currently fielded shelter system might be inadequate for future U.S. Army digitized TOCs. The commander of the Combined Arms Center (CAC) directed the TRADOC Analysis Center-Fort Leavenworth (TRAC-FLVN) to provide an analysis to inform the November 2000 committee's decision on the system to select for the U.S. Army's future division-level and brigade-level TOCs. Consequently, TRAC-FLVN requested ARL-HRED to provide a human factors evaluation of the Custom Tent Design to support TRAC-FLVN's overall analysis. This effort considered both the TOC layout and internal operations for personnel requirements, information flow, and current operations decision making. To fulfill this effort, ARL-HRED developed a survey instrument (see Appendix B) to assess human factors, battlefield management, staff collaboration, equipment modularity, mobility, and security issues.

Currently, the U.S. Army fields the Modular Command Post System (MCPS), formerly known as the Standard Integrated Command Post System-Extensions (SICPS-E), with its command post vehicles (CPVs). The MCPS consists of tents and bootwalls that connect the vehicles to the tents. Command Posts (CPs) at all echelons use this system to create a common workspace which allows commanders and staffs to perform C3 functions, fuse information, and the myriad of tasks required during military operations. Although the currently fielded MCPS allows for the establishment of a common workspace, it does not provide an open architecture within which staffs can better perform their functions. An open architecture allows for the uninterrupted view of CIC displays and unimpeded movement of personnel within the shelter.

Currently fielded systems are not adequate because of architectural support poles that break up the MCPS common workspace.

The objective of this study was to provide HF analyses regarding the Custom Tent Design as input to TRAC as the lead to inform the November 2000 decision committee on the form (platform or shelter) to select for the Army's future division-level and brigade-level TOCs. The scope of this analysis was on the Custom Tent shelter used for the Combat Information Cell (CIC) of a Battalion or Brigade Tactical Command Post.

Several major issues were considered for the improvement of TOC operations. They were concerned with the TOC's tactical mobility, execution of C2, interior environment, equipment and personnel arrangements, modularity, flexibility, and security. In addition, human factors considerations for the alternative TOC interior light levels and ambient noise were made.

Numerous manufacturers can provide custom-built tents. This study used two soft-walled, internal-framed custom tents built to user specifications by the Custom Canvas Manufacturing Company (Buffalo, NY). Because the shelter was built to user specifications, it was configurable for vehicle booting and expandable, limited only by the physical constraints of the framing system. Materials vary but are normally the same as those used in MCPS type systems. The manufacturers included flooring, ground covers, and repair material and equipment in addition to the actual tents.

Regarding tent installation, almost 64% percent of the staff that completed the survey felt that the Custom Tent design did not facilitated quick set up time. They felt that there were two many fasteners and wires. The Custom Tent set-up time was approximately 50 minutes per tent for the 1 BCT TOC. Extreme weather (e.g., wind, cold temperatures, & rain) and dark conditions increased the time as did the inexperience of all personnel who had only set up the tent two previous times. Although the time required to set up and boot four MCPS systems was not observed during this exercise, NCO's stated that the Custom Tent could be established much more quickly than four booted MCPS systems. Even though an individual MCPs takes between 20-30 minutes to set up, connecting four MCPS systems together and assuring that the gutters are water tight will take longer. The custom tent which equalls in size to four MCPS has no gutters.

Regarding tent only disassembly, 54 % of the surveyed staff felt that the Custom Tent design did not facilitated quick tear down while 33% felt that it did. Fifty eight percent of the staff surveyed rated the Toc components as hindering mobility.

It was observed that the 1 BCT Main CP required approximately 4.5 hours to march order (i.e., disassemble and prepare all the TOC digital and non digital equipment and Custom Tent for movement). This time appears to be long but is explainable considering the unit's current lack of experience. This was the first time that this digitized unit had conducted a march order. As a training exercise it emphasized the need for a march order SOP and the need for some staff training, task prioritization, and rehearsal. Some of the extra time required to march order was consumed by the unit's using three heavy expanded mobility tactical trucks (HEMTTs) from the FSB and moving floor boards, sandbags miscellaneous TOC equipment and the breakdown and packing of the 9 panel digital display.

The 1st Brigade layout provided more than adequate space for CP equipment and personnel during military operations. Over 95% of the staff surveyed felt that the custom tent provided adequate space for equipment and 83.3 % felt that it provided adequate space for the number of personnel required for effective TOC operations. Most of the time there appeared to be unneeded room in the MCPS shelters attached to the 1st BCT Main CP Custom Tents. These spaces were only used when TOC briefings were conducted. Possibly, the TOC with its attached vehicles could be reconfigured to eliminate the need for, at least, one of the two MCPSs used in this study.

The staffs' opinion on a quick establishment of an integrated LAN communication system was mixed. Only 37.5% of those surveyed rated the TOC system as facilitating a quick establishment while 41.7 % felt that it hindered a quick set up. The lack of an established and standardized TOC wiring diagram was felt strongly needed.. This brigade's 74B was working toward creating such a diagram.

The 1st Brigade Custom Tent with its open architecture configuration was regarded as facilitating the commander's ability to exercise C2 in the CIC to some degree. Eighty three percent of the staff felt that the design supported, facilitated or greatly facilitated the Commander’s ability to provide direction and management. Almost 80% felt that the design allowed the Commander to maintain an active command presence among the entire staff. The commander had the ability to rearrange functional elements to meet METT-TC requirements. The commander had access to each member of the staff throughout the exercise and his command presence was in clear view of the entire staff. All the staff teams could easily see and hear the commander.

The personnel line of sight view of the situational map and information displays varied.as did the staffs’ rating of the arrangement of equipment and personnel to facilitate access to information displays. This staff team working at the first row of tables had a clear view of the FBCB2 situational maps as well as the wall-mounted paper map with friendly and enemy updates and UAV information directly in front of them. The battle captain was able to supervise the efforts of the staff NCOs, conduct analyses and assessments of available information, assist in the review and dissemination of information from the other BFAs, and assist in monitoring the location and activity of friendly units. The battle major monitored and updated the information displayed on the electronic display screens and paper maps. However, the staff in the back row of tables of the CIC had some problems viewing the map boards and large screen displays. One suggestion was that the rear tables should have one or two computer monitors to view the Common Operational Picture (COP). The majority of the ABCS platforms were housed in the C2Vs and the operators had a “caved mentality” with little access to the battle staff.

The Custom Tent was relatively watertight when compared to the MCPS. The larger size of the Custom Tent reduced the need for gutters at the connecting points which often leaked. Consequently, though it rained intensely for many hours, no significant water was observed to leak from the ceiling attachments involving the Custom Tent Designs. However, it was noted that some connector pin assembly parts had broken which were used to connect the roof sections of the tents. Therefore, it is recommended that the connector parts be ruggedized to withstand the battlefield environment. If possible, the connector parts should remain attached to appropriate tent sections, even when the tent sections themselves are not connected, so that the connector parts will be available when needed and not get misplaced during frequent TOC relocations.

Another problem associated with the interior TOC environment and rain is the mud that can result on the ground space under the tent area. Though the Custom Tent design included a tarpaulin floor it did not prevent water getting into the TOC ground space and causing severe mud development on the floor. The mud was sticky and built up on the shoes of the military personnel which slowed their performance. The mud severely hindered the TOC displacement process as it took 8-10 soldiers an inordinate amount of time, ingenuity, and physical effort to manipulate and load the tarpaulin floor onto a truck.

Regarding safety, one potential TOC safety hazard that was noted involved personnel climbing on the Custom Tent to disassemble the camouflage netting. It was cumbersome and, perhaps, unsafe to climb on top of a tent that is 11 feet high and not intended to support the weight and movement of soldiers.

The soldiers were asked to rate the modularity of the Custom Tent design to allow for open TOC architecture to support the commander's layout preferences for the arrangement of equipment and personnel. Over 92% felt that it at least supported the Commanders layout preference. When asked to rate the adaptability of the TOC design to accommodate large screen displays and multiple displays, 83% of the soldiers rated the design as at least supported this type of display technology. Seventy five percent of the staff surveyed thought that the design permitted or enhanced the Commander's ability to observe the staff.

Eighty percent of the soldiers rated the flexibility and open architecture of this design as supporting the performance of tasks related to METT-T. In an interview, they stated that the ATCCS with the large screen displays allowed real time action for the commander and that the Jupiter gave the commander a versatile tool to manipulate and display data. It was also stated that this design provided an excellent ability to switch feeds from the various boxes through the DPV to display information on a large screen. The only negative comments were due to either power failure and equipment failure.

Only 45.8% of the soldiers rated the TOC system design able to support concealment and camouflage techniques while 50 % felt that the design hindered it. As confirmation, only 4.2% stated that it greatly facilitated. No light could be seen escaping from the TOCs at night. Fifty percent of the staff surveyed felt that it would be hard to take measures to prevent observation and detection. The respondents stated that the size and height of the TOC layout and the noise level from the numerous vehicles and generators would cause the TOC to be an easy target. This problem exists regardless of the TOC configuration used.

When asked how the did this TOC design affect the ability to control thermal signature, only 25% felt that it could be controlled. The numerous vehicles and generators added to the thermal signature problem.

Only 41.7%% of the soldiers indicated that this TOC design aided the ability to control physical TOC evidence (signature). None of the respondents stated that this design greatly facilitated their ability to control the signature. Once again, the size of the TOC layout and the noise level from vehicles and generators made the TOC an easy target.

The sources of illumination in the 1st brigade TOC "A" were fluorescent work lights located at a height of approximately 7 feet. The TOC CIC operations using the Custom Tent design were conducted in incident light levels ranging from 8-16 foot-candles. These levels of illumination are adequate for normal detail but not for prolonged periods of reading printed material. The levels of illumination in the adjacent Custom Tent design (i.e., left-hand side of the TOC), which support the CIC, ranged from 4-16 foot-candles. At the right-hand side of the TOC was a currently fielded tent, the MCPS, which had light levels ranging from 2-12 foot-candles. The TOC areas which supported the CIC had lower light levels than the CIC but the requirement for prolonged periods of reading printed material was also less.

A TOC at another location (i.e., TOC "B") had its fluorescent work lights mounted in the upper 4-foot section of the Custom Tent design ceiling. Overall, the lighting levels in work areas for TOC B were lower than for TOC A. The lighting level at any given location appeared to be most directly related to the distance from the light source. Consequently, the overall light levels in TOC B were lower than for TOC A.

The current 1st brigade TOC operations were conducted in average noise levels ranging from 67-78dB using the A weight and 79-83 dB using the C weight. This level of background noise was loud but did not exceed the steady state noise hazard requirement of 85 dB measured (using the A weight) as specified in Army Pam 40-501. The source of the loud background noise was due to the turbine engines of the C2V along with the primary power unit (PPU). The majority of the staff felt that they could not control the noise levels. The sound readings at the engine exhaust box were measured resulting in an average of 89 dB with the A weight and 92 dB with the C weight. Sound protection around these vehicles is required.

The majority of the staff surveyed felt that voice commands were easily heard throughout the TOC. The majority (75%) of the staff felt that the design promoted or at least supported efficient internal communication.

The consensus of the staff was that if the day and night shifts of the brigade or battalion staff were combined, then there was adequate manpower to set up or disassemble the CP that utilized the Custom Tent configuration. However, the majority of the staff (62.5%) surveyed expressed concern that there was not enough people to man the planning requirements of the TAC or complete all tasks that have increased because of digitization. As a consequence, TOC security was minimal.

Training and experience on the ABCS needs to be increased. Most of the staff was relatively new with only 2-3 months of experience. The TOC did have a couple of experienced ABCS staff members that could reconfigure and reactivate the ABCS workstations. However, the ABCS operators had to complete other TOC duties and could not devote their full attention to ABCS.

No additional personal skill identifier (ASI) or MOS was felt to be required for set up or disassembly of the Custom Tent configuration. Guidance for battlefield functional area layout was received from the commanders in both the battalion and brigade TOCs.

1. Introduction

This paper summarizes insights developed during the 1st Brigade Task Force Lanes Exercise relative to the Custom Tent shelter used for the Combat Information Cell (CIC) of the 1st BCT Main CP. Through the TOC Summit venue, it was recognized that the currently fielded shelter system might be inadequate for future U.S. Army digitized TOCs. The commander of the Combined Arms Center (CAC) directed the TRADOC Analysis Center-Fort Leavenworth (TRAC-FLVN) to provide an analysis to inform the November 2000 committee's decision on the system to select for the U.S. Army's future division-level and brigade-level TOCs. Consequently, TRAC-FLVN requested ARL-HRED to provide a human factors evaluation of the Custom Tent Design to support TRAC-FLVN's overall analysis. This effort considered both the TOC layout and internal operations for personnel requirements, information flow, and current operations decision making. To fulfill this effort, ARL-HRED developed a survey instrument (see Appendix B) to assess human factors, battlefield management, staff collaboration, equipment modularity, mobility, and security issues.

2. Background

The U.S. Army currently fields the Modular Command Post System (MCPS), formerly known as the Standard Integrated Command Post System-Extensions (SICPS-E), with its command post vehicles (CPVs). The MCPS consists of tents and bootwalls that connect the vehicles to the tents. Command Posts (CPs) at all echelons use this system to create a common workspace which allows commanders and staffs to perform C3 functions, fuse information, and the myriad of tasks required during military operations. Although the currently fielded MCPS allows for the establishment of a common workspace, it does not provide an open architecture within which staffs can better perform their functions. An open architecture allows for an uninterrupted view of CIC displays and unimpeded movement of personnel within the shelter. The MCPS is not adequate because of architectural support poles that break up the TOC's common workspace.

The U.S. Army doctrine specifies the types and functions of CPs at all echelons of command. The functionality offered by these various CPs has always focused on the warfighting doctrine of the Army. However, the doctrine allowed the commander to tailor the unit's CPs to meet the commander's needs and preferences as long as the doctrinal functionality was achieved. Consequently, CPs have always been as unique as the commanders operating in them. The wide variety of needs and preferences, coupled with changes in task organization and equipment, led to countless solutions for a CP's configuration. While the functionality found in CPs remained the same across units (coordinating and special staff activities), the physical organization and internal standing operating procedures (SOPs) were different based on the individual commander's requests. Organizations throughout the U.S. Army pursued TOC development programs to meet their needs. This led to redundant TOC development efforts that achieved varying results.

In early 1997, the Vice-Chief of Staff, Army (VCSA) established a policy to address TOC development. The intent of this policy was to focus all of the unique TOC development efforts that were known at that time. The goal for the U.S. Army of the future was a TOC compatible with all U.S. Army forces and interoperable with joint and combined forces. The purpose was not to impose a single standardized inflexible CP that would not meet the needs of all users. Rather, the intent was to leverage all of the financial and intellectual efforts from across many communities to focus on systems that would provide commanders and their staffs the facilities and information required for optimizing military decision-making processes. The “clearing house” task for focusing TOC development efforts was the Training and Doctrine Command (TRADOC) Program Integration Office-Army Battle Command System (TPIO-ABCS).

In early 1999, HQ TRADOC and the Army Digitization Office (ADO) requested that a high-level group address the issues that resulted from the complexity of TOC development. The result was a forum known as the "TOC Summit." The TOC Summit held an issue review board meeting in May 1999 and its first summit meeting on 3 June 1999. The TOC Summit continues today as the forum through which TOC development is focused. Through this venue, the shortcomings of the MCPS (e.g., lack of flooring, poles in the central work area, gaps in roof sections) led to the realization that the currently fielded system was inadequate for future Army TOCs. The commander of the Combined Arms Center (CAC) directed that the TOC program manager (PM-TOC) pursue a range of alternatives with a modular architecture and allowed the lower echelons (i.e., battalions & brigades) to use soft architectures and higher echelons (division level and above) to use hardened architectures. The CAC commander also tasked the TRAC-FLVN to provide analyses to inform the November 2000 decision committee on the system to select for the U.S. Army’s future division-level and brigade-level TOCs. Consequently, the TRAC-FLVN requested the ARL-HRED to provide human factor analysis support.

3. Study Objective.

To provide Human Factors analyses regarding the Custom Tent Design as input to TRAC to inform the November 2000 decision committee on the form (platform or shelter) to select for the Army’s future division-level and brigade-level TOCs.

4. Scope.

This analysis primarily focuses on the Custom Tent shelter used for the Combat Information Cell (CIC) of a Battalion or Brigade Tactical Command Post.

5. Assumptions.

The following assumptions were developed based on requirements identified in the ABCS Critical Requirements Document (CRD) and information from the ABCS users in the field:

a. <u>*An open architecture is required*</u>. There are two primary reasons this study assumes that a collaborative and open architectural environment is required. First, it recognizes that "virtual" TOCs and "C2 on the move (C2 OTM)" capabilities are not mature given today's technology. Physical collocation within TOCs is required until C2 OTM technology becomes more fully developed. Second, this study recognizes that during the conduct of military operations, commanders will want to collocate staff functions if the mission, enemy, terrain, troops, time available, and civilian affairs (METT-TC) conditions allow. Clearly, the ability to perform C2 OTM through the use of "virtual" TOCs will enhance the survivability of the CPs when threatened by enemy forces. During Stability and Support Operations (SASO) and low threat conditions, the commander may want to create an open architecture (collocate) to increase the sustainability of C2 functions and allow for human, face-to-face interaction.

b. <u>*The 1st Brigade Modified Table of Organization and Equipment (MTOE) serves as the standard*</u>. This study uses the 1st Brigade MTOE as the equipment and personnel "footprint" that the given Custom Tent Design alternative (sheltering system) must accommodate.

c. *The electronic signature of the equipment is the same across the alternatives*. This study contends that there are no significant differences between the given TOC alternatives regarding electronic signature because the same command, control, communications, and computer (C4) equipment is resident in each case. Consequently, an analysis of electronic signature is not necessary for distinguishing between alternatives. This does not imply that an electronic signature analysis is not required to answer other relevant survivability questions outside the purview of this study.

6. Issues and Configuration Tested

a. Major Issues.

Several major issues were considered for the improvement of TOC operations. They were concerned with the TOC's tactical mobility, execution of C2, interior environment, equipment and personnel arrangements, modularity, flexibility, and security. In addition, human factors considerations for the alternative TOC interior light levels and ambient noise were made.

b. Custom Tent Design.

Numerous manufacturers can provide custom-built tents. Figure 1 depicts soft-walled, internal-framed custom tents built to user specifications by the Custom Canvas Manufacturing Company (Buffalo, NY). Because the shelter is built to user specifications, it is configurable for vehicle booting and expandable, limited only by the physical constraints of the framing system. Materials vary but are normally the same as those used in MCPS type systems. Most manufacturers include flooring, ground covers, and repair material and equipment in addition to the actual tent.

Figure 1 depicts tents that the 3ID had built in its efforts to increase the deployability and reduce the signature of its DMAIN. The Custom Canvas Mfg. Co. built these tents based on unit specifications and material samples. These tents are built to MCPS specifications but provide the open architecture that the MCPS does not. Custom tents are currently in use by the XVIII Airborne Corps (82d Abn Div and 3 ID) and other military units. The tents provide blackout capability, can be heated and cooled, are man-portable, are connectable and expandable with currently fielded MCPS systems, and can be easily transported and erected.

Figure 1. External views of Custom Tent Design.

Common among all of the alternatives are the facts that they do provide 24-hour continuous operations functionality and facilitate the use of the Common Operating Picture (COP). However, the alternatives do not provide ballistic protection from direct or indirect fire, nor do they provide over-pressurization capabilities or High-Altitude Electromagnetic Pulse (HEMP) protection.

7. Methodology.

As the lead agency responsible for analyzing human factors aspects of major issues of the Custom Tent Design, ARL provided the following resources and assessment materials: (a) three HFE Subject Matter Expert (SME) observers and (b) an HF Questionnaire (see Appendix B-1) that was administered to the entire TOC staff.

Data collection consisted of SME observations, measurements of the physical layout of personnel and equipment, the responses from the Human Factors survey for 24 staff members, interviews, and information provided by the military unit. The duty positions surveyed are listed in Table 1. This data will be used in the analysis to provide human factors evaluation information to the November 2000 decision committee.

Table 1
Command Post Duty Position Surveyed

Position
1. Advanced Field Artillery Tactical Data System (AFATDS) Operator
2. FSE (Fire Support Element)
3. Brigade Targeting Officer
4. Brigade Fire Support Officer (BDE FSO)
5. G-2 DTAC Non-Commissioned Officer-in-Charge (NCOIC)
6. Military Intelligence (MI) ACT
7. S-2 (Day Shift) All Source Analysis System (ASAS) Operator
8. 1 BCT-4ID TF XXI
9. S-2 NCOIC
10. Battle CPT
11. Operations (OPN) SGT MAJOR
12. G-3 Operations Sergeant
13. S-3
14. Radio Telephone Operator(RTO) for S3
15. Operations (OPN) SGT
16. S-3 Engineer
17. S-6 BN Level
18. S-G (299 Eng S-3)
19. G-6
20. Tactical Automation Specialist (MOS 74B)
21. Executive Officer (XO)
22. S-6
23. BDE S-2
24. Military Intelligence (MI) ACT NCOIC

8. Results of Issue Considerations

A. Tactical Mobility

The study issue here is whether or not the TOC alternative is mobile. The Essential Elements of Analysis are as follows:

(1) What are the critical C2 times/events during CP displacement operations given the TOC alternative?

(2) Does the given TOC alternative facilitate employment (set-up) and march order (tear down) of the digitized division's CPs (short moves, whole CP as march unit)?

(3) Does the given TOC alternative facilitate continuous C2 operations during displacement of the echelons (Interim C2 OTM)?

(4) Does the given TOC alternative facilitate the digitized division's ability to establish fully operational CPs?

(5) Is the given TOC alternative able to sustain movement (rate and terrain) commensurate with the combat systems of the organization's MTOE?

(6) Does the given TOC alternative meet mobility objectives and/or thresholds identified in the ABCS CRD?

(7) Does the given TOC alternative allow for quick set-up and tear-down? (ARL)

Regarding tent installation, 37% percent of the staff that completed the survey felt that the Custom Tent design highly facilitated quick set up and all felt that it was as good or better for installation than the MCPS The Custom Tent set-up time was approximately 60 minutes per tent for the 1 BCT TOC. Extreme weather (e.g., wind, cold temperatures, & rain) increased the time as did the inexperience of all personnel. The military unit personnel believed that set-up time could be reduced to 20 minutes once they were trained. Although the time required to set up and boot four MCPS systems was not observed during this exercise, NCO's stated that the Custom Tent could be established much more quickly than four booted MCPS systems.

Regarding tent disassembly, 33% of the surveyed staff felt that the Custom Tent design highly facilitated quick tear down and all responded that it was as good or better in this respect than the MCPS. However, even though the potential for quick disassembly of the Custom Tent exists, the 1 BCT Main CP required approximately 4.5 hours to march order (i.e., disassemble and prepare the TOC equipment and Custom Tent for movement). This time appears to be long but is explainable considering the unit's current lack of training and experience. Better SOPs and more training are expected to reduce this time. Some of the extra time required to march order was consumed by the unit's using three heavy expanded mobility tactical trucks (HEMTTs) from the FSB and moving floor boards, sandbags, and miscellaneous TOC equipment. Some of this preparation time might be reduced by the shared knowledge of an SOP for loading plans and priorities of work.

The Custom Tent configuration did offer the potential for continuous C2 operations during the displacement of echelons because all the ABCS major components were housed in the C2Vs or "1068" vehicles. The staff was able to maintain continuous audio connectivity up to the time when the vehicles moved. One complaint was that the BFA operators remained inside the vehicles which inhibited task sharing and crosstalk but promoted a quicker march order.

The majority of the staff felt that the lack of portability of the Custom Tent design hindered (i.e., 5.3%) or only "borderline supported" (i.e., 20.8%) the mobility of the military unit (χ^2=8.93, $p<.25$). At present, the personnel in a designated vehicle in the TOC is responsible for breaking down, packing and transporting each MCPS. A Standard Operational Procedure (SOP) is needed to determine which vehicles will transport the various parts (e.g., roof, sides, poles, floor) of the Custom Tent design. The shared knowledge of such procedures will help to quicken the mobility of the military unit.

Several needed improvements were reported or noted during setup or breakdown. First, the Custom Tent poles need to be labeled and colored-coded to assure easier matching of horizontal with vertical poles. Secondly, the pins holding together the Custom Tent platform were easy to remove when compared to the nuts and bolts of the MCPS. The use of the pins facilitated quicker set up and disassembly. However , the pins need to be ruggedized because during a Custom Tent disassembly it was observed that several of the pins broke which delayed the frame poles from being taken apart. On another occasion, during setup, several of the pins were bent which delayed the erection of the Custom Tent frame. Thirdly, another improvement involves a system to facilitate the handing of cables and wires within the Custom Tent. This system may possibly consist of clips attached to the tent frame that will be readily available to hold and group the cables and wires in an organized manner.

B. 1st Brigade CIC Layout.

This study issue was whether or not this Custom Tent alternative allowed the digitized division to effectively exercise C2. Therefore, this section quantifies the ability of the Custom Tent used in the 1st Brigade CIC layout to provide adequate space for CP equipment and personnel and its ability to facilitate C2 operations. The EEAs are:

(1) Does the given TOC alternative facilitate C2, battle tracking, and info sharing?

(2) Does the given TOC alternative support user-friendly man-machine interfaces and standardized equipment requirements?

(3) Does the given TOC alternative meet functionality objectives and/or thresholds identified in the ABCS CRD?

(4) Does the given TOC alternative provide adequate space for CP equipment and personnel, and does it facilitate TOC operations?

(5) Does the given TOC alternative facilitate the commander's ability to exercise C2?

The 1st Brigade layout provided more than adequate space for CP equipment and personnel during military operations. Figure 2 shows the overall layout of the 1st Brigade TOC including the CIC area. Figure 3 shows a more detailed view of the CIC section. Based on the HF survey, 96% of the staff members ($\chi = 21.6$, $p<.01$) felt that the physical dimensions of the Custom Tent provided adequate space for digitized and non-digitized equipment. Also, 83% of the staff members ($\chi=10.7$, $p<.01$) stated that the physical dimensions of this design provided adequate space for the number of personnel required for effective TOC operations. Most of the time there appeared to be unneeded room in the MCPS shelters attached to the 1st BCT Main CP Custom Tents. These spaces were needed only when TOC briefings were conducted. Possibly, the TOC with its attached vehicles could be reconfigured to eliminate the need for, at least, one of the two MCPSs used in this study.

Personnel could easily change their locations to correspond to specific METT-TC operations at the discretion of the commander or battle captain. No BFA had to be dedicated to a specific table computer. Each BFA's laptop computer could be picked up and plugged into the local area network. Most of the wires were located at a 7-foot height where the walls joined the ceiling sections. The tables were not anchored which allowed for layout flexibility. In addition, the height of the command tables was approximately 28 inches which corresponds to conventional HF ergonomic guidelines. Most of the staff (63%, $\chi^2=2.25$, *N.S.*) felt that the Custom Tent design supported easy integration of ABCS and associated communication networks and nodes. The staffs' opinion on a quick establishment of an integrated communication system was mixed. Only 37.5% of those surveyed rated the TOC system as facilitating a quick establishment. The lack of an established and standardized TOC wiring diagram was felt strongly. This brigade's 74B was working toward creating such a diagram.

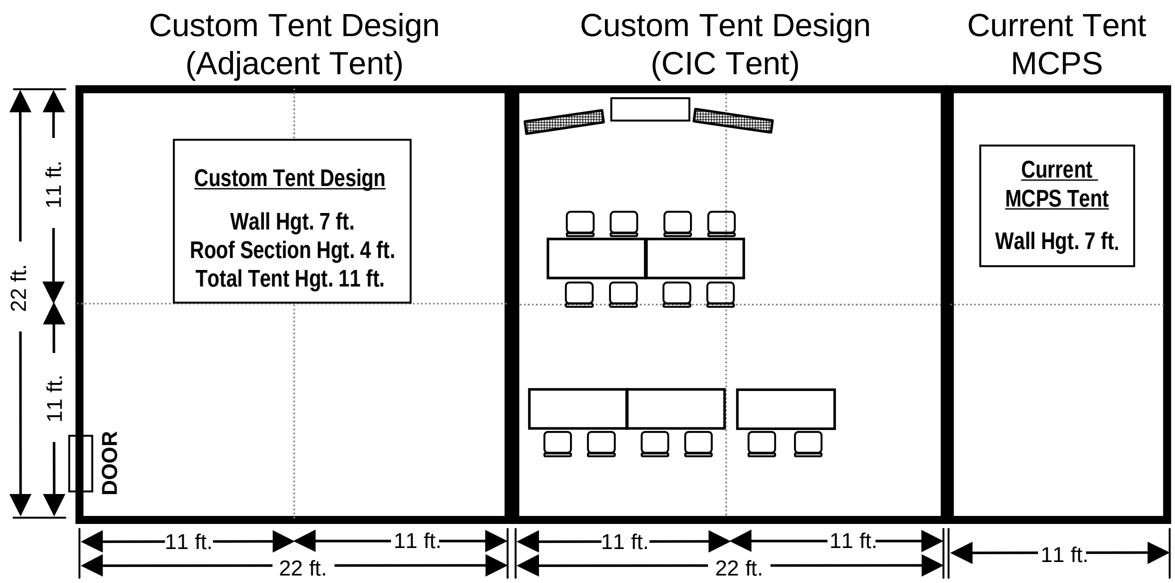


Figure 2. Layout of the DTAC during the 1st brigade task force exercise at Fort Hood, Texas using the two custom tent designs and the current MCPS tent (March 2000).

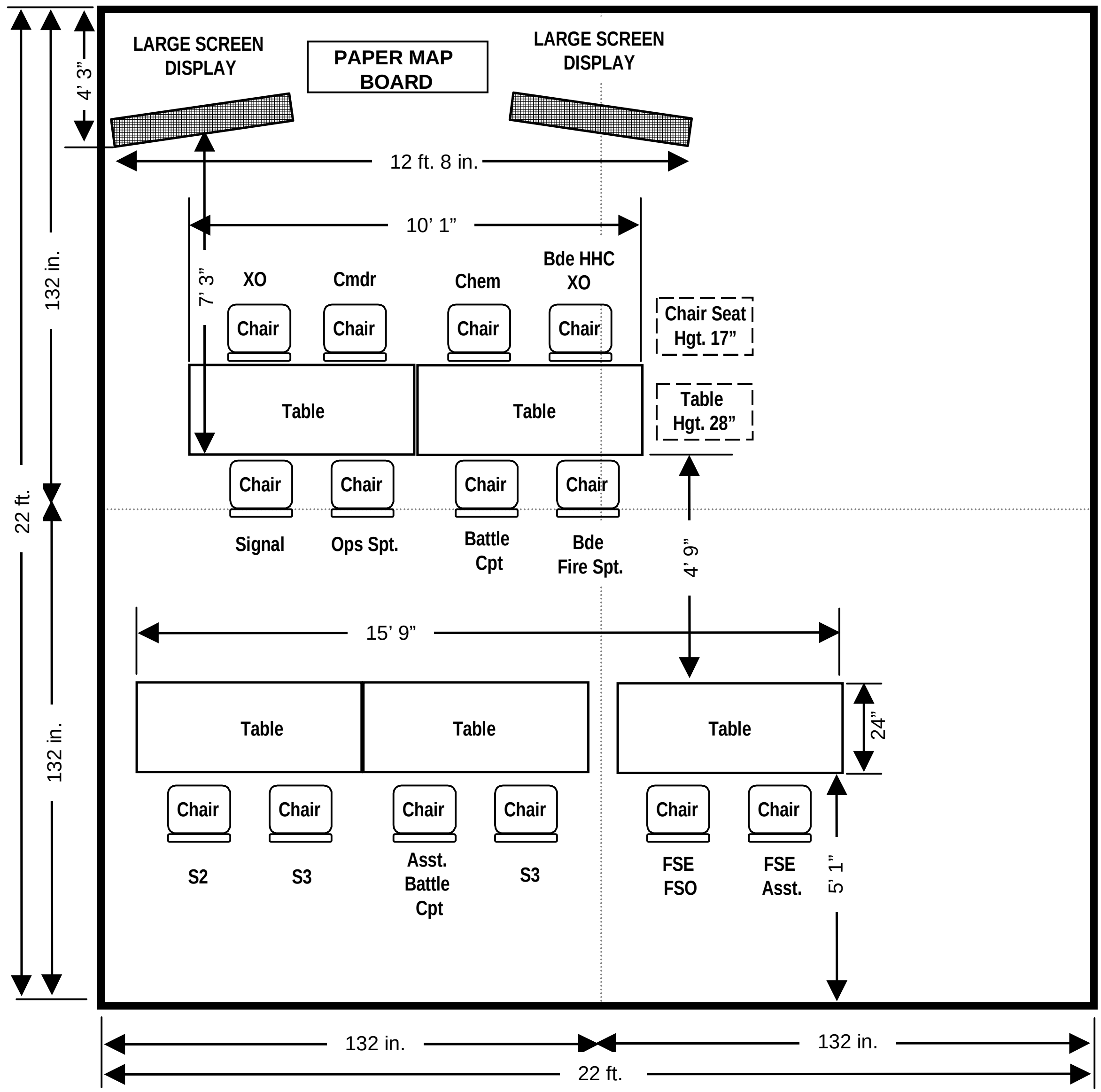


Figure 3. Layout of the combat information cell (CIC) using the custom tent design during the 1st brigade task force exercise at Fort Hood, Texas (March 2000).

The 1st Brigade Custom Tent with its open architecture configuration was regarded as facilitating the commander's ability to exercise C2 in the CIC to some degree. The commander had the ability to rearrange functional elements to meet METT-TC requirements. The commander had access to each member of the staff throughout the exercise and his command presence was in clear view of the entire staff. All the staff teams could easily see and hear the commander. The staff rated the commander's ability as "high" to (1) observe the staff (66.7%, $\chi^2 = 15.57$, $p<.01$), (2) implement risk management (58.3%, $\chi^2 = 18.07$, $p<.01$), (3) easily provide the staff with guidance and monitor activity (70.9%, $\chi^2=20.15$, $p<.01$), (4) focus the activity of the staff as desired (54.2%, $\chi^2=7.24$, p<.86), and (5) position himself in order to maintain an active command presence (79.2%, $\chi^2 = 20.99$, $p<.01$).

The 1st Brigade Custom Tent promoted efficient internal staff communications and total staff integration. The battle major, battle captain, G3, and operations officer (maneuver team) worked effectively as a team in providing the commander the maneuver information and screen displays he required to support his decision making. The staff surveyed felt that the modified tent design promoted BOS integration during the planning (54.2%,$\chi^2 = 17.5$, $p<.01$), preparation (58.3%, $\chi^2 = 24.3$, $p<.01$) and execution phase (58.3%,$\chi^2 = 20.99$, $p<.01$) of decision making. There was adequate room to promote team huddles which supported collaborative planning (79.1%,$\chi^2 = 27.7$, $p<.01$) and synchronization (83%, $\chi^2 = 14.32$, $p<.01$). The majority of the staff felt that the TOC design supported (20.8%) or promoted (58.3%) task sharing and teamwork within and among the staff leader teams ($\chi^2 = 11.83$, $p<.03$). Similarly, the open architecture promoted workload distribution among the staff (70.8, $\chi^2 = 14.32$, $p<.01$).

The personnel line of sight view of the situational map varied. This staff team working at the first row of tables in the CIC had a clear view of the FBCB2 situational map as well as the wall-mounted paper map with friendly and enemy updates and UAV information directly in front of them. The battle captain was able to supervise the efforts of the staff NCOs, conduct analyses and assessments of available information, assist in the review and dissemination of information from the other BFAs, and assist in monitoring the location and activity of friendly units. The battle major monitored and updated the information displayed on the electronic display screens and paper maps. However, the staff in the back row of tables of the CIC had some problems viewing the map boards and large screen displays. One suggestion was that the rear tables (see Figure 3) should have one or two computer monitors to view the Common Operational Picture (COP).

C. Interior Environment.

This study issue was whether or not the TOC alternative provided CPs that were interchangeable, expandable, and adaptable to meet changing mission needs. The EEAs are:

(1) Does the given TOC alternative CIC cool to 85^{o} F and heat to 50^{o} F and offer limited climate control elsewhere in the TOC?

(2) Is the given TOC alternative climate control consistent with currently fielded systems?

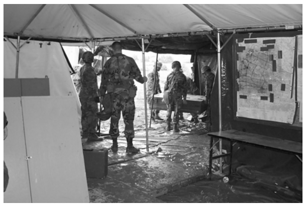
(3) Does the given TOC alternative prevent water/snow/wind from entering the TOC interior and interfering with CP operations in inclement weather?

(4) Is the given TOC alternative able to be integrated into/supported by the Army logistical support system and does it introduce any unique logistical support requirements?

(5) Does the given TOC alternative prevent typical battlefield conditions from adversely affecting CP operations (smoke, dust)?

The Custom Tent was relatively watertight when compared to the MCPS. The larger size of the Custom Tent reduced the need for gutters at the connecting points which often leaked. Consequently, though it rained intensely for many hours, no significant water was observed to leak from the ceiling attachments involving the Custom Tent Designs. However, it was noted that some connector pin assembly parts had broken which were used to connect the roof sections of the tents. Therefore, it is recommended that the connector parts be ruggedized to withstand the battlefield environment. If possible, the connector parts should remain attached to appropriate tent sections, even when the tent sections themselves are not connected, so that the connector parts will be available when needed and not get misplaced during frequent TOC relocations.

Another problem associated with the interior TOC environment and rain is the mud that can result on the ground space under the tent area (see Figure 4a). Though the Custom Tent design included a tarpaulin floor it did not prevent water getting into the TOC ground space and causing severe mud development on the floor. The mud was sticky and built up on the shoes of the TOC personnel which slowed their military performance. At one TOC location (i.e., TOC "A"), raised wooden floor sections, provided by local test support personnel, were successful in preventing mud problems. However, at another TOC location (i.e., TOC "B"), a large one-piece tarpaulin ground-level floor was provided that was not successful in preventing mud buildup. An additional problem occurred when the TOC B was disassembled and the military personnel tried to fold and place the single-piece floor material onto a truck (see Figure 4d). The mud on the tarpaulin caused the material to be so heavy and unmanageable that the personnel were initially unable to move it. After repeated futile attempts with many personnel and failed problem-solving ideas, the tarpaulin was loaded with difficulty onto the back of a truck (see Figure 4f). It was suggested that the tarpaulin floor be made in sections for easier handling. However, the fact that the ground-level tarpaulin did not prevent severe mud buildup suggests that a tarpaulin floor is not a good solution.



a.

b.

c.

d.

e.

f

Figure 4. Disassembly and loading of the Custom Tent for movement to another site.

Other environmental issues involved temperature, ventilation, and noise within the Custom Tent shelter. The environment was rated adequate or high (80%) for temperature (χ^2 =17.37, p<.01), (70.6%) for ventilation (χ^2 =10.27, p<.05), and (89.2%) for noise (χ^2 =21.54, p<.01).

Regarding safety, one potential TOC safety hazard that was noted involved personnel climbing on the Custom Tent to disassemble the camouflage netting. It was cumbersome and, perhaps, unsafe to climb on top of a tent that is 11 feet high and not intended to support the weight and movement of soldiers.

D. Modularity.

This study issue was whether or not the TOC alternative is extensible to future concepts and provides modularity to meet the needs of commanders and considerations of METT-TC. The EEAs are:

(1) Is the given TOC alternative extensible to future operational concepts identified in TRADOC Pam (TP) 525-66, ATDs, ACTD, and other future concept efforts?

(2) Can the given TOC alternative physically and adequately integrate joint, multi-national, and coalition forces?

(3) Does the given TOC provide CPs that are interchangeable, compatible, expandable, and adaptable to meet changing missions and needs?

(4) Does the given TOC alternative meet adaptability objectives and/or thresholds identified in the ABCS CRD?

The Custom Tent meets the above criteria. It can be transported by aircraft, ships, trains, and trucks and moved by material handling equipment.

The soldiers were asked to rate the modularity of the system design to allow for open TOC architecture to support the Commander's layout preferences for the arrangement of equipment and personnel. In other words, could the TOC be constructed with the standardized units and dimensions but have enough flexibility in construction to support the Commander's layout preference? When asked to rate the adaptability of the TOC design to accommodate large screen displays and multiple displays, 83.4% of the soldiers stated that it was adaptable. Only 12.5% rated the modularity of the system design as hindering the ability to construct an open architecture design (χ^2=20.29, p <.01).

The majority of the staff surveyed highly rated the ability of the modified tent design to accommodate large screen displays (70.9%, χ^2 =11.0, p<.05), and multiple displays in horizontal or vertical configurations (68.8%, χ^2 =30.84, p<.01).

E. Flexibility.

This study issue is related to the flexibility and open architecture of the TOC design to affect the performance of tasks related to mission, enemy, terrain, troops, and time available (METT-TC). Open architecture is essential because during military operations commanders will want to collocate staff functions to enhance the performance of tasks related to METT-TC. This will increase the sustainability of C2 functions and allow for face-to-face interactions. As a result of the wide variety of commander needs and preferences as well as changes in task organization and equipment, the physical organization and the internal standing operating procedures of the TOCs change based on commander preferences. Therefore, the TOC must be flexible in order to meet the needs of all users, including all Army forces ad interoperable with Joint and Combined Forces. In other words, the TOC must have the ability to be tailored to meet the commander's needs and preferences while maintaining the functionality required by doctrine.

Ninety-six percent of the soldiers ($\chi^2 = 20.15$, $p < .01$) rated the flexibility and open architecture of this design as supporting the performance of tasks. They stated that ATCCS with the large screen displays allowed real time action for the commander and that the Jupiter gave the commander a versatile tool to manipulate and display this data. It was also stated that this design provided excellent ability to switch feeds from the various boxes through the DPV to display information on a large screen. The only negative comments were due to either power failure and equipment failure.

F. Security.

Although a detailed technical analysis of electronic signature was beyond the scope of this study, questions about security were asked. The EEAs are:

(1) Does the given TOC alternative enhance the ability to employ concealment techniques and camouflage?

(2) Does the given TOC alternative allow the employment of noise, light, thermal, and physical evidence control?

(3) Does the given TOC alternative facilitate protection from surprise, observation, detection, interference, espionage, terrorism, and sabotage?

Only 45.8% of the soldiers rated the TOC system design able to support concealment and camouflage techniques (χ^2 =4.04, *p*=N.S.). As confirmation, only 4.2% stated that it greatly facilitated. No light could be seen escaping from the TOCs at night. Fifty percent (χ^2 =12.4, $p<.03$) of the staff surveyed felt that it would be hard to take measures to prevent observation and detection. This was the first time that the Custom Tent configuration was camouflaged. The respondents stated that the size of the TOC layout and the noise level from the numerous vehicles and generators would cause the TOC to be an easy target. This problem exists regardless of the TOC configuration used.

When asked how the did this TOC design affect the ability to control thermal signature, only 25% (χ^2 =14.74, $p<.01$) felt that it could be controlled. The numerous vehicles and generators added to the thermal signature problem.

Only 41.7%% (χ^2 =7.49, *p*=N.S.) of the soldiers indicated that this TOC design aided the ability to control physical TOC evidence (signature). None of the respondents stated that this design greatly facilitated their ability to control the signature. Once again, the size of the TOC layout and the noise level from vehicles and generators made the TOC an easy target.

When asked how well did the TOC system design allow the staff to take measures to protect from surprise, observation, and detection, 45.8% (χ^2 =12.37, $p<.025$) stated that it either borderline supported or facilitated measures of protection. However, no one stated that it greatly facilitated. It was stated that with thermals, the TOC could be seen and heard. This is due to its size (i.e., numerous connected military vehicles and generators). There was concerned expressed that the signature is too large to be within FM (radio) range of forward BCTs.

Sixty-six percent of the soldiers (χ^2 =12.9, $p<.025$) stated that the TOC design would allow the staff to take measures to protect from espionage, terrorism or sabotage. However, only 8.3% of the respondents stated that it greatly facilitated their ability. The TOC did not provide ballistic protection from direct or indirect fire. In addition, the TOC does not provide over-pressurization capabilities or high-altitude electromagnetic pulse (HEMP) protection. Also, it was stated that there were no weapons except personal weapons for protection.

G. Interior Light Levels.

It was desired to know what the general illumination levels were at representative locations within the 1st brigade TOCs. The light levels were measured by using a Gossen Luna-Pro light meter placed with the sensor in an upward position to record the ambient incident light at representative work locations indicated in the Figures 5 and 6. Table 2 shows the scale number indicated by the light meter and the equivalent readings in foot-candles and Lux (see Appendix E).

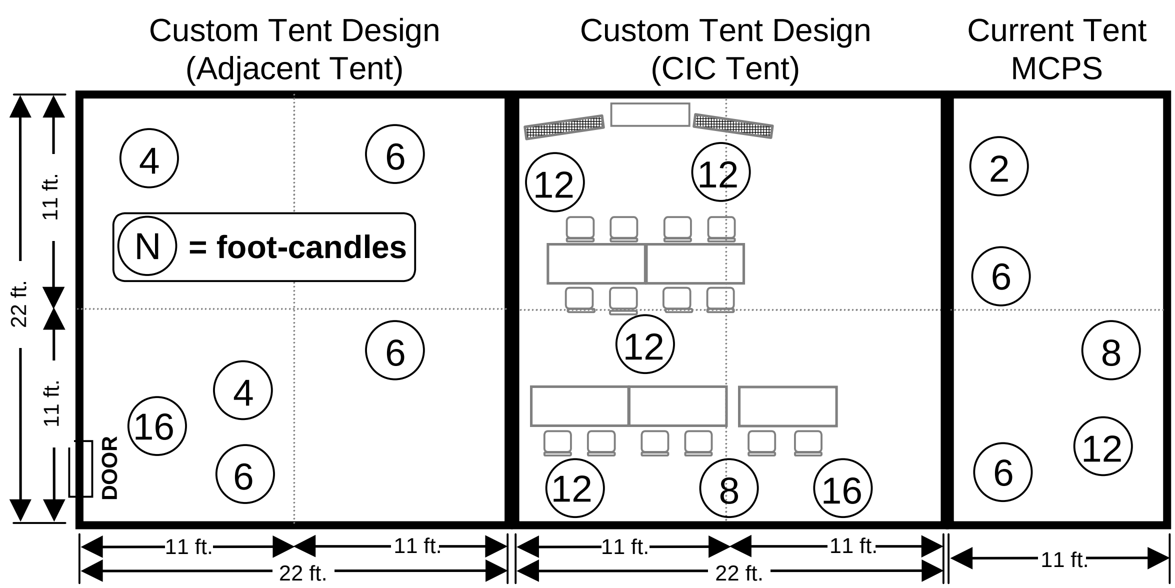


Figure 5. Diagram showing the sources of illumination at TOC A.

The sources of illumination in the 1st brigade TOC "A" (see Figure 5) were fluorescent work lights located at a height of approximately 7 feet. The TOC CIC operations using the Custom Tent design were conducted in incident light levels ranging from 8-16 foot-candles. These levels of illumination are adequate for normal detail but not for prolonged periods of reading printed material (see Appendix F). The levels of illumination in the adjacent Custom Tent design, which support the CIC, ranged from 4-16 foot-candles. At the opposite side of the TOC was a current tent design, the MCPS, which had light levels ranging from 2-12 foot-candles. The TOC areas which supported the CIC had lower light levels than the CIC but the requirement for prolonged periods of reading printed material was also less.

Based on the HF survey, 71% of the military respondents felt that the brightness of the light in the TOC was "adequate" and an additional 21% felt that the brightness was "excellent" ($\chi^2 = 41.65$, $p < .01$). When asked how the ambient lighting affected C4IS operations, 71% of the respondents felt that these operations were "somewhat facilitated" and an additional 12% felt they were "greatly facilitated" ($\chi^2 = 41.97$, $p < .01$). Regarding how the TOC design affected the ability to control lighting, 67% of the respondents indicated that lighting control was "somewhat facilitated" and an additional 8% felt that this lighting control was "greatly facilitated." However, 17% felt that lighting control was "somewhat" or "seriously hindered."

The respondents made suggestions to improve the usefulness of the CP: (1) provide support in tents for lights, (2) provide readily available hangers to mount the lights, and (3) provide improved light sets.

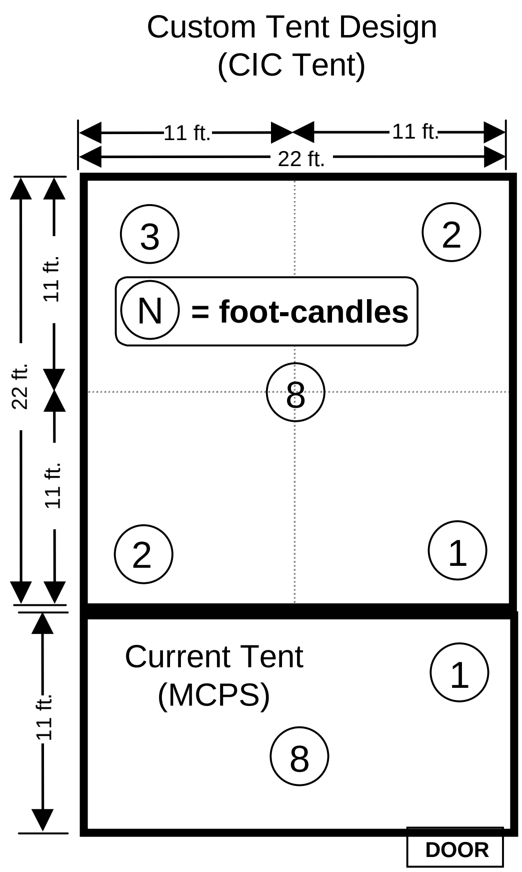


Figure 6. Diagram showing the sources of illumination at TOC B.

A TOC at another location (i.e., TOC "B") had its fluorescent work lights mounted in the upper 4-foot section of the Custom Tent design ceiling (see Figure 6). Thus, the source of illumination for the TOC B was several feet higher than existed for the TOC A. The resulting work level lighting for the higher mounted lights ranged from 1-8 foot-candles in the Custom Tent. The light levels in the adjacent MCPS tent also ranged from 1-8 foot-candles. Overall, the lighting levels in work areas for TOC B were lower than for TOC A. The lighting level at any given location appeared to be most directly related to the distance from the light source. Consequently, the overall light levels in TOC B, with the higher mounted light sources, resulted in lower light levels in the work areas than for TOC A.

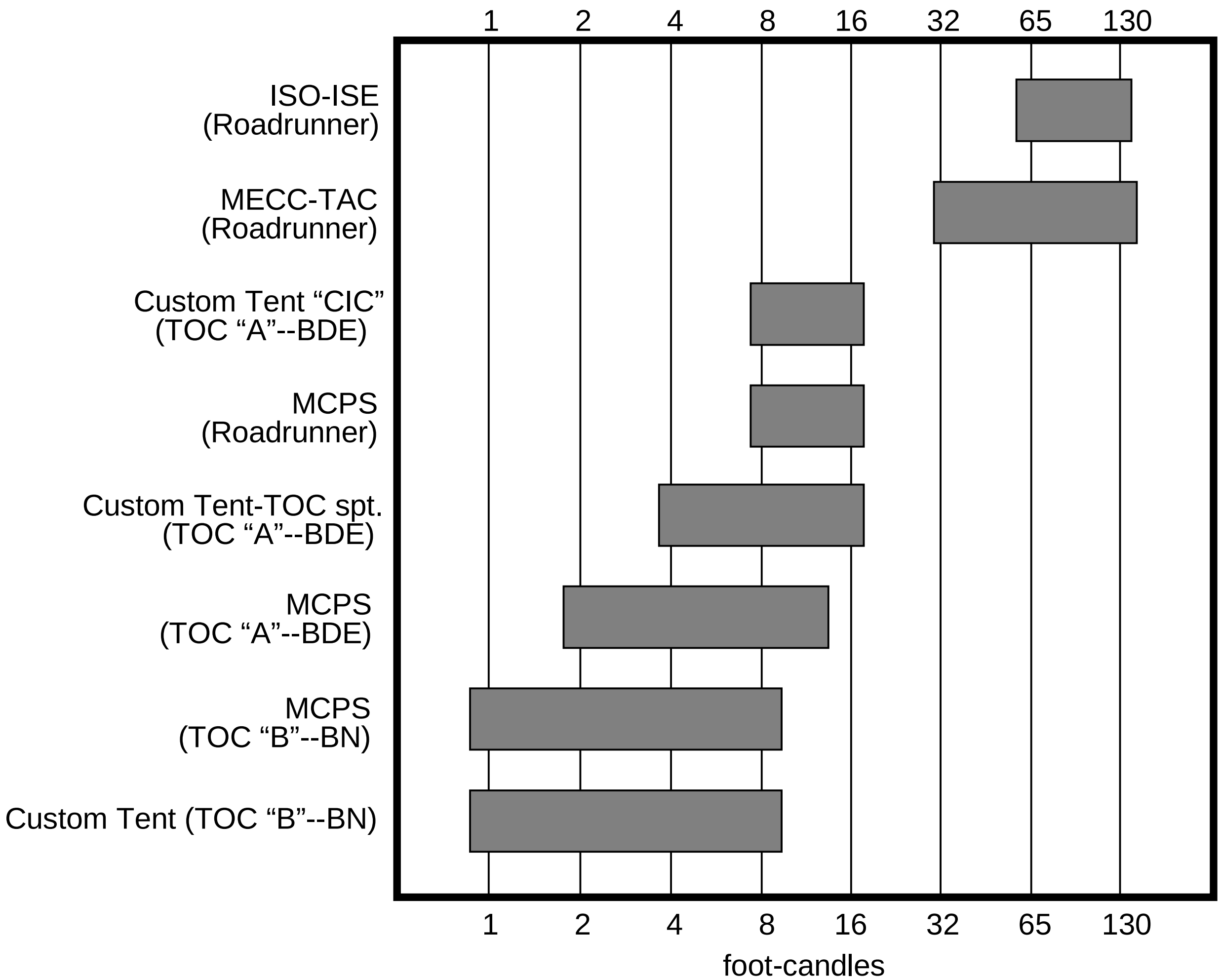


Figure 7. A comparison of light levels for various concepts of tactical operations centers(TOCs).

A comparison was made for the light levels of the TOCs observed in this study and a previous study (i.e., the 4th Infantry Division's III Corps Roadrunner and Iron Horse Sprint exercises). The light level ranges are presented in Figure 7. Overall, it can be seen that the light levels (i.e., 32-130 foot-candles) were higher when using the Mobile Expandable Container Configuration

(MECC) shelters for the CIC and the Information Support Element (ISE). However, a lower amount of ambient light (i.e., 10-20 foot-candles) may be more desirable for TOC operations that consist of viewing ABCS computer and large screen displays (see guidelines in Appendix F).

.

Table 2

Incident Light Levels

Scale No.	Foot-Candles	Lux
5	0.26	2.8
6	0.50	5.5
7	1.00	11.0
8	2.00	22.0
9	4.00	44.0
10	8.00	88.0
11	16.00	175.0
12	32.00	350.0
13	65.00	700.0
14	130.00	1400.0

H. Sound and Noise Levels.

It was desired to know what the general sound or noise levels were at representative locations within the Brigade TOC using the custom tent design.. The sound levels were measured using a sound level meter placed with the sensor in a horizontal position to record the sound level at specific work locations indicated in the Figures 5-6. The accuracy of the meter at 144 dB was ±2 dB. When set on the "A" weighting, the meter measures frequencies in the 500-10,000 Hz range which is the area of greatest sensitivity to the human ear. When set to the "C" weighting, the meter measures uniformly over the frequency range from 32-10,000 Hz, giving an indication of the sound level at a wider range.

The current 1st brigade TOC operations were conducted in average noise levels ranging from 67-78dB using the A weight and 79-83 dB using the C weight. This level of background noise was loud but did not exceed the steady state noise hazard requirement of 85 dB measured (using the A weight) as specified in Army Pam 40-501. The source of the loud background noise was due to the turbine engines of the C2V along with the primary power unit (PPU). The majority of the staff felt that they could not control the noise levels (67%, $\chi 2 = 17.66$, $p < .01$). The sound readings at the engine exhaust box were measured resulting in an average of 89 dB with the A weight and 92 dB with the C weight. Sound protection around these vehicles is required. The noise levels at various positions can be found in Table 3.

Table 3

Noise levels (dB) obtained within the digitized Brigade TOC using a Custom Tent Design

Position	Weighting	Average (dB)	Peak (dB)
S2	A	71	75
	C	79	81
Battle Captain	A	74	78
	C	80	82
FSE	A	67	68
	C	82	83
S3	A	69	73
	C	80	82
Commander	A	69	72
	C	83	84

The majority (62.5%) of the staff surveyed ($\chi 2 = 10.57$, $p < .05$) felt that voice commands were easily heard throughout the TOC. Twenty-five percent of the staff surveyed felt that the design hindered the ability to control noise ($\chi 2 = 17.66$, $p < .01$). However, the majority of the staff felt that the design promoted or at least supported efficient internal communication. (75%, $\chi 2$ =21.4%, p<.01).

I. Manpower, Personnel, and Training.

The consensus of the staff was that if the day and night shifts of the brigade or battalion staff were combined, then there was adequate manpower to set up or disassemble the CP that utilized the Custom Tent configuration. However, the majority of the staff (62.5%) surveyed expressed concern that there was not enough people to man the planning requirements of the TAC or complete all tasks that have increased because of digitization (χ^2 =2.08, p= *.62*). Consequently, TOC security was minimal.

Training and experience on the ABCS needs to be increased. Most of the staff was relatively new with only 2-3 months of experience. The TOC did have a couple of experienced ABCS staff members that could reconfigure and reactivate the ABCS workstations. However, the ABCS operators had to complete other TOC duties and could not devote their full attention to ABCS.

The majority of the staff surveyed (83.3%) felt that no additional personal skill identifier (ASI) or MOS was felt to be required for set up or disassembly of the Custom Tent configuration (χ^2 =10.7, p<.01). Guidance for battlefield functional area layout was received from the commanders in both the battalion and brigade TOCs.

APPENDIX A

REFERENCES

3rd Infantry Division (Mechanized), DMain Redesign Briefing, Fall 1998

Army Field Manual No. (FM) 101-5-1, *Operational Terms and Graphics,* Headquarters, Department of the Army, 30 September 1997.

Deployable Rapid Assembly Shelter (DRASH), "Shelter Products", [http://www.drash.com/shelters.html].

English, H.B. & English, A.C. (1958*). A Comprehensive Dictionary of Psychological and Psychoanalytical Terms.* N.Y.: David McKay Company

Gichner Shelter Systems – Military Shelters -MERWS, "Military Shelter Systems MERWS", [http://www.gichner.com/military/merws.html].

Logicon – A Northrop-Grumman Company, "Standard Integrated Command Post System", [http://www.logicon.com/shelter/sicps.html].

Natick Research Development and Engineer Center, "LSS", [http://www.sbccom.army.mil/hooah/pubs.htm]

Nilsson, T. H. (1981). Reference light source for luminance and illuminance photometry. *Behavior Research Methods & Instrumentation,* Vol. 13(1), 18-19.

TRADOC Program Integration Office-Army Battle Command System, *Capstone Requirements Document*, *Revision 1b, Annex D, Final Draft*, 30 July 1999.

Training and Doctrine Command Pamphlet (TRADOC Pam) 525-66, *Military Operation Future Operational Capability (FOC),* Headquarters, Department of the Army, 1 May 1997.

Van Cott, H.P. & Kinkade, R.G. (1972). *Human Engineering Guide to Equipment Design.* Washington, D.C.: American Institutes for Research

Weatherhaven, "MECC", [http://www.weatherhaven.com/photos/product08.htm].

APPENDIX B

Tactical Operations Center
HUMAN FACTORS QUESTIONNAIRE

ARMY RESEARCH LABORATORY
HUMAN RESEARCH AND ENGINEERING DIRECTORATE

Privacy Act Statement

Authority: 5 USC § 301, Authority for the Secretary of the Army to Issue Army Regulations; AR 73-1, Test and Evaluation Policy. **Principal Purpose:** The data to be collected with this form are to be used for research and evaluation purposes only. **Routine Uses:** This is an experimental data collection questionnaire developed by the Test and Experimentation Command pursuant to its research and testing mission as prescribed in AR 73-1. When identifier (name and social security number) is requested they are to be used for administrative and statistical control purposes only. Full confidentiality of the responses will be maintained in the processing of these data. **Disclosure:** Completion of this questionnaire is required for this test. You are encouraged to provide complete and accurate information in the interests of research and testing, but there will be no effect on individuals for not providing all or part of the information.

Instructions

The purpose of this questionnaire is to record HUMAN FACTORS data on CIC Command Post (CP) designs. Your answers will not be given to or shown to anyone except those who are assessing CIC for the Army. (For example, none of your information will be given to your chain of command or put in your personnel file.) Your answers will be treated confidentially. Please fill out the questionnaire carefully. If you need additional space to answer a question, indicate by an arrow (→) and continue on the back of the page. Be sure to number the item on the back of the page. If you have any questions concerning this questionnaire, please contact an ARL or TRAC team representative for help.

Thank you for your help

1) Name: Last ______________________________
First ______________________________
Middle Initial _______________________

2) User PIN: __________________________
(First initial of last name and last four digits of social security number)

3) TOC Configuration: a. STCPS ____ b. DRASH ____
c. CUSTOM TENT ____ d. MECC ____ e. Large SICPS ____

4) Command Post Duty Position:
______________________________________(e.g., G-2, G-3, Operations Sgt, Commander)

5) Shift: a) Day: _________________ b) Night:____________

I. MANPOWER/PERSONNEL

1) Were there an adequate number of personnel available in your Command Post (CP) to perform operations including setup and breakdown?

☐ Yes ☐ Do Not Know ☐ No ☞ Please explain: ________________________

2) Did you require personnel augmentation in order to perform CP operations including setup, breakdown, camouflage and security?

☐ No ☐ Do Not Know ☐ Yes ☞ To perform which functions ☐ (Day) ☐ (Night)?

3) Were there enough ABCS personnel (manpower) in your CP to reconfigure and reactivate workstations? ☐ Yes ☐ Do Not Know ☐ No ☞ Please explain:

4) Were there enough personnel in your CP to prevent mission delays?

☐ Yes ☐ Do Not Know ☐ No ☞ Please explain:

5) Do the personnel skills required to setup/breakdown the TOC necessitate a specific Additional Skill Identifier (ASI) or unique MOS?

☐ Yes ☐ Don't know ☐ No ☞ please explain:

II. TRAINING

1) Did you receive formal training on this TOC configuration regarding setup and breakdown.

☐ No ☐ Yes ☞ When: ____________month ____________year

2) Did you receive formal training on internal TOC layout ?

☐ No ☐ Yes ☞ When: ____________month ____________year

3) Did you receive instruction from your commander regarding interior layout for your specific Battlefield Functional Area?

☐ Yes ☐ No ☞ please explain:

III. HUMAN FACTORS

1) The noise level of your work area while operating within the CP was:
☐ Very Low ☐ Low ☐ Borderline ☐ High ☐ Very High

2) The intensity (brightness/darkness) of light in the TOC was:
☐ Excellent ☐ Adequate ☐ Borderline ☐ Inadequate ☐ Very Poor

3) The view of the situational map from your work area was :
☐ Excellent ☐ Adequate ☐ Borderline ☐ Inadequate ☐ Very Poor

4) Does the system design facilitate quick setup?

☐ Yes ☐ No ☞ please explain: ______________________________

5) Does the system design facilitate quick breakdown?
☐ Yes ☐ No ☞ please explain: ______________________________

6) Rating of how the portability of the components of the TOC design system affects the mobility of the military unit:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

7) Did the physical dimensions of the system design provide adequate space for digital and non-digital equipment:

☐ Yes ☐ No ☞ please explain: ______________________________

8) Did the physical dimensions of this TOC design provide adequate space for the number of personnel required for effective TOC operations:

☐ Yes ☐ No ☞ please explain: ______________________________

9) Rating of this TOC design to allow for optimal arrangements of equipment and personnel that facilitated the ease of access to information displays:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

10) Rating of the how the physical dimensions of this TOC design affected the efficiency of communications among personnel:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

11) Indicate how each of the following physical conditions affected C4IS operations: (mark one ☒ per row)

	Seriously Hindered	Hindered	Supported (Adequate)	Facilitated	Greatly Facilitated
a. Ambient lighting					
b. Temperature					
c. Noise					
d. Ventilation					

12) Rating of the modularity of the system design to allow for open TOC architecture to support the Commander's layout preferences:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

13) Rating of the adaptability of the TOC system design to accommodate large screen displays and multiple displays:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

14) Rating of the adaptability of the TOC system design to accommodate horizontal and vertical map boards:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

15) Rating of the flexibility and open architecture of the TOC design to affect the performance of tasks related to mission, enemy, terrain, troops, and time available (METT-TC):

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

16) Indicate how this TOC design impacted on effective C4IS operations related to METT-TC: (mark one ☒ per row)

	Seriously Hindered	Hindered	Borderline	Facilitated	Greatly Facilitated
a. Mission					
b. Enemy					

c. Troops					
d. Terrain					
e. Time Available					

17) Rating of the physical characteristics of equipment and personnel arrangements in the TOC system to contribute to a safe working environment:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

18) Rating of the TOC system design to employ concealment and camouflage techniques:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

19) Indicate how this TOC design affected the ability to control :
(mark one ☒ per row)

	Seriously Hindered	Hindered	Borderline	Facilitated	Greatly Facilitated
a. Noise					
b. Lighting					
c. Thermal					
d. Physical TOC Evidence (Signature)					
e. Time Available					

20) Rating of the TOC system design to allow the staff to take measures to protect from surprise, observation, detection:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

☞ please explain: ____

____

____

____

21) Rating of the TOC system design to allow the staff to take measures to protect from espionage, terrorism or sabotage:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

22) Rating of the TOC system design to allow the Commander to provide directions and management for the TOC staff:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

23) Rating of the TOC system design to allow the Commander to obtain information on the mission, enemy forces, friendly forces, terrain, weather:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

☞ please explain: ____

____

____

24) Rating of the TOC design to allow C4I accessibility from all TOC locations:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

25) Rating of the TOC system design to ensure that voice commands are easily heard throughout the TOC:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

26) Rating of the TOC system design to ensure easy integration of ABCS and associated communication networks and nodes (e.g., LANs, WAN's, satellites, data facsimiles, etc.):

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

27) Rating of the TOC system design to ensure the easy visual scanning of map boards at all locations:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

28) Rating of the TOC system design to ensure the easy and quick establishment of an integrated communication system:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

29) Rating of the TOC design to enhance the Commander's ability to observe the staff:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

30) Rating of the TOC design to ensure that the Commander can observe the entire staff and implement risk management:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

31) Rating of the TOC design to create an environment for the Commander to provide the staff with guidance and monitor activities:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

32) Rating of the TOC design to allow the Commander to focus the activities of his staff:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

33) Rating of the TOC design to allow the Commander to always position himself in order to maintain an active command presence among the entire staff:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

34) Rating of the TOC design to promote task sharing and teamwork among the battlestaff:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

☞ please explain: ____

____

____

____

35) Rating of the TOC design to promote workload distribution among the battlestaff:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

36) Rating of the TOC design to promote prioritizing actions:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

37) Rating of the TOC design configuration to permit teamwork and reduced workload:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

38) Rating of the TOC to promote staff workload sharing and collaboration:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

39) Rating of the TOC design to allow BFA's adequate space to advise and assist the Commander:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

40) Rating of the TOC system design to promote collaborative planning:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

41) Rating of the TOC design to promote synchronization:

(1) Seriously Hindered	(2) Somewhat Hindered	(3) Borderline Support	(4) Somewhat Facilitated	(5) Greatly Facilitated

42) Rating of the TOC design to permit easy visual and auditory communication to integrate TOC personnel for planning and execution:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

43) Rating of the TOC design to provide adequate space for the assembly of staff members during wargaming:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

44) Rating of the TOC design to provide adequate space for supervision of staff during wargaming?

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

45) Rating of the TOC design to promote the development of an accurate common operating picture (COP):

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

46) Rating of the TOC design to provide the staff an unobstructed field of view of situational maps:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

47) Rating of the TOC design to ensure easy information flow between BFA's and synchronization:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

48) Rating of how the TOC design promoted BOS integration throughout the:
(mark one ☒ per row)

	Seriously Hindered	Hindered	Borderline	Facilitated	Greatly Facilitated
Planning					
Preparation					
Execution					

49) Rating of how the system design promoted efficient internal communication:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

50) Rating of how the TOC system design promoted efficient total staff integration:

(1) Seriously (2) Somewhat (3) Borderline (4) Somewhat (5) Greatly

Hindered Hindered Support Facilitated Facilitated

51) Rating of the TOC system design to allow easy assimilation of audio data by the entire staff:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

52) Rating of the TOC system design to allow easy assimilation of visual data by the entire staff:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

53) Rating of the TOC system design to facilitate information being sent to the appropriate destination:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

54) Rating of the TOC system design to allow the entire staff to monitor operational situations:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

55) Rating of the ability for you to complete your tasks using automated tools:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

56) Rating of the ability for you to maintain better situational awareness using the available automated tools compared to analog methods:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

57) Rating of the ability for you to exchange information with personnel you need to colaborate with using the available automated tools:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

58) Rating of the ability for you to exchange information with personnel you need to collaborate with based on the current TOC design:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

59) Rating of the ability of digitization to provide adequate features for maintaining a COP with other personnel, even when face to face communications were not possible:

(1) Seriously Hindered (2) Somewhat Hindered (3) Borderline Support (4) Somewhat Facilitated (5) Greatly Facilitated

55) What are your top three suggestions to improve the usefulness of the CP:

a) ______________________________

b) ______________________________

c) ____

IV. SAFETY

1) Did you identify any TOC safety hazards (shortcomings)? [example: electrical hazards, equipment configuration/design, non secured items (loose), sharp edges]

☐ No ☐ Yes ☞ Please explain:______________________________________

____

____

____

V. HEALTH HAZARDS

1) Please identify any health hazards (example: noise level, exposure to chemicals, oxygen deficiency, heat/cold stress, work stress) associated with the CP (includes ABCS).

☐ None or Comments: ____

____

____

____

VI. SAFETY / HEALTH HAZARDS -- COMMENTS

1) What are your comments concerning safety and health hazards while operating within CP.

☐ None or Comments: ____

____

____

APPENDIX C
Tables Of Questionnaire Results

Table C1:
Manpower/Personnel and Training

N=24

	Yes	No	Do Not Know	Chi -Square	Sig
Were there an adequate number of personnel available in your Command Post (CP) to perform operations including setup and breakdown?	33.3%	62.5%	4.2%	2.13	.79
Did you require personnel augmentation in order to perform CP operations including setup, breakdown, camouflage and security?	41.7%	41.7%	16.7%	0.06	.83
Were there enough ABCS personnel (manpower) in your CP to reconfigure and reactivate the workstation?	37.5%	41.7%	20.8%	0.05	.87
Were there enough personnel in your CP to prevent mission delays?	54.2%	25.0%	20.8%	2.58	.13
Do the personnel skills required to setup/breakdown the TOC necessitate a specific Additional Skill Identifier (ASI) or unique MOS?	16.7%	83.3%	0.0%	10.66	<.01
Did you receive formal training on this TOC configuration regarding setup and breakdown?	16.7%	70.8%	0.0%	8.05	<.01
Did you receive formal training on internal TOC layout?	16.7%	75.0%	0.0%	8.03	<.01
Did you receive instruction from your commander regarding interior layout for your specific Battlefield Functional Area?	41.7%	58.3%	0.0%	0.67	.78

Table C1:
Human Factors Summary
N=24

	Very Low	Low	Borderline	High	Very High	No Response	Chi-Square	Sig
The noise level of your work area while operating within the CP was:	0.0%	37.5%	45.8%	12.5%	4.2%	0.0%	20.24	<.01
	Very Poor	Inadequat	Borderline	Adequate	Excellent	No Response	Chi-Square	Sig
The intensity (brightness/darkness) of light in the TOC was:	0.0%	4.2%	4.2%	70.8%	20.8%	0.0%	41.65	<.01
	Excellent	Adequate	Borderline	Inadequate	Very Poor	No Response	Chi-Square	Sig
The view of the situational map from your work area was:	8.3%	29.2%	16.7%	25.0%	20.8%	0.0%	3.08	.47
	Yes	No				No Response	Chi-Square	Sig
Does the system design facilitate quick setup?	37.5%	63.5%					1.5	0.20
Does the system design facilitate quick breakdown?	33.3%	54.2%				12.5%	1.41	.22
	Seriously Hindered	Somewhat Hindered	Borderline Support	Somewhat Facilitated	Greatly Facilitated	No Response		
Rating of how the portability of the components of the TOC design system affects the mobility of the military	20.8%	37.5%	20.8%	16.7%	0.0%	4.2%	8.63	..07

unit:								
	Yes	No						
Did the physical dimensions of the system design provide adequate space for digital and non-digital equipment:	95.8%	4.2%					21.6	<.01
Did the physical dimensions of this TOC design provide adequate space for the number of personnel required for effective TOC operations?	83.3%	16.7%					10.7	<.01
	Seriously Hindered	Somewhat Hindered	Borderline Support	Somewhat Facilitated	Greatly Facilitated	No Response		
Rating of this TOC design to allow for optimal arrangements of equipment and personnel that facilitated the ease of access to information displays:	4.2%	33.3%	12.5%	45.8%	4.2%	0.0%	16.82	<.01
Rating of how the physical dimensions of this TOC design affected the efficiency of communications among personnel:	8.3%	20.8%	33.3%	29.2%	8.3%	0.0%	6.41	..17
Indicate how each of the following physical conditions affected C4IS operations: Ambient lighting	0.0%	0.0%	12.5%	70.8%	12.5%	4.2%	41.97	<.01
Indicate how each of the following physical conditions affected C4IS operations: Temperature	8.3%	16.7%	20.8%	50.0%	0.0%	4.2%	17.37	<.01
Indicate how each of the following physical conditions affected C4IS operations: Noise	0.0%	16.7%	29.2%	50.0%	0.0%	4.2%	22.54	<.01
Indicate how each of the following physical conditions affected C4IS operations: Ventilation	4.2%	20.8%	33.3%	33.3%	4.2%	4.2%	10.27	<.05
Rating of the modularity of the system design to allow for open TOC architecture to support the Commander's layout preferences:	0.0%	12.5%	29.2%	50.0%	4.2%	4.2%	20.29	<.01
Rating of the adaptability of the TOC system design to accommodate large screen displays and multiple displays:	4.2%	12.5%	12.5%	41.7%	29.2%	0.0%	11.00	<.05
Rating of the adaptability of the TOC design to accommodate horizontal and vertical map boards:	4.2%	12.5%	12.5%	62.5%	8.3%	0.0%	30.84	<.01
Rating of the flexibility and open architecture of the TOC design to affect the performance of tasks related to mission, enemy, terrain, troops, and time	4.2%	16.7%	20.8%	54.2%	4.2%	0.0%	20.15	<.01

available (METT-TC):								
Indicate how this TOC design impacted on effective C4IS operations related to METT-TC: Mission	0.0%	0.0%	20.8%	58.3%	12.5%	8.3%	27.63	<.01
Indicate how this TOC design impacted on effective C4IS operations related to METT-TC: Enemy	0.0%	4.2%	33.3%	45.8%	8.3%	8.3%	19.57	<.01
Indicate how this TOC design impacted on effective C4IS operations related to METT-TC: Troops	4.2%	4.2%	29.2	50.0%	4.2%	8.3%	20.48	<.01
Indicate how this TOC design impacted on effective C4IS operations related to METT-TC: Terrain	4.2%	20.8%	20.8%	41.7%	4.2%	8.3%	11.65	<.025
Indicate how this TOC design impacted on effective C4IS operations related to METT-TC: Time Available	12.5%	20.8%	16.7%	33.3%	8.2%	8.3%	4.58	.55
Rating of the physical characteristics of equipment and personnel arrangements in the TOC system to contribute to a safe working environment:	4.2%	4.2%	37.5%	41.7%	12.5%	0.0%	15.99	<.01
Rating of the TOC system design to employ concealment and camouflage techniques:	20.8	29.2%	20.8%	20.8%	4.2%	4.2%	4.04	.48
Indicate how this TOC design affected the ability to control: Noise	0.0%	25.0%	41.7%	33.3%	0.0%	0.0%	17.66	<.01
Indicate how this TOC design affected the ability to control: Lighting	4.2%	12.5%	8.3%	66.7%	8.3%	0.0%	33.07	<.01
Indicate how this TOC design affected the ability to control: Thermal	4.2%	29.2%	41.7%	25.0%	0.0%	0.0%	14.74	<.01
Indicate how this TOC design affected the ability to control: Physical TOC Evidence (Signature)	16.7%	33.3%	25.0%	16.7%	0.0%	8.3%	7.49	.14
Indicate how this TOC design affected the ability to control: Time Available	16.7%	25.0%	29.2%	25.0%	0.0%	4.2%	6.21	..19
Rating of the TOC system design to allow the staff to take measures to protect from surprise, observation, detection:	33.3%	16.7%	37.5%	8.3%	0.0%	4.2%	12.37	<.025
Rating of the TOC system design to allow the staff to take measures to protect form espionage, terrorism or sabotage:	16.7	8.3%	45.8%	16.7%	4.2%	8.3%	12.9	<.025
Rating of the TOC system design to allow the Commander to provide directions and management for the TOC	0.0%	12.5%	16.7%	50.0%	16.7%	4.2%	16.54	<.01

staff:								
Rating of the TOC system design to allow the Commander to obtain information on the mission, enemy forces, friendly forces, terrain, weather:	8.3%	8.3%	8.3%	45.8%	29.2%	0.0%	13.58	<.01
Rating of the TOC design to allow C4I accessibility from all TOC locations:	0.0%	4.2%	50.0%	25.0%	8.3%	12.5%	20.53	<.01
Ratings of the TOC system design to ensure that voice commands are easily heard throughout the TOC:	4.2%	25.0%	8.3%	41.7%	20.8%	0.0%	10.57	<.05
Rating of the TOC system design to ensure easy integration of ABCS and associated communication networks and nodes (e.g., LANs, WAN's, satellites, data facsimiles, etc.):	25.0%	12.5%	16.7%	29.2%	16.7%	0.0%	2.25	.96
Rating of the TOC system design to ensure the easy visual scanning of map boards at all locations:	20.8%	25.0%	12.5%	29.2%	12.5%	0.0%	2.68	.85
Rating of the TOC system design to ensure the easy and quick establishment of an integrated communication system:	4.2%	37.5%	20.8%	25.0%	12.5%	0.0%	7.67	.31
Rating of the TOC design to enhance the Commander's ability to observe the staff:	4.2%	20.8%	8.3%	50.0%	16.7%	0.0%	15.57	<.01
Rating of the TOC design to ensure that the Commander can observe the entire staff and implement risk management:	4.2%	29.2%	8.3%	50.0%	8.3%	0.0%	18.07	<.01
Rating of the TOC design to create an environment for the Commander to provide the staff with guidance and monitor activities:	4.2%	20.8%	4.2%	54.2%	16.7%	0.0%	20.15	<.01
Rating of the TOC design to allow the Commander to focus the activities of his staff:	4.2%	16.7%	25.0%	37.5%	16.7%	0.0%	7.24	.39
Rating of the TOC design to allow the Commander to always position himself in order to maintain an active command presence among the entire staff:	4.2%	4.2%	12.5%	54.2%	25.0%	0.0%	20.99	<.01
Rating of the TOC design to promote task sharing and teamwork among the battlestaff:	4.2%	16.7%	20.8%	45.8%	12.5%	0.0%	11.83	.08
Rating of the TOC design to promote workload distribution among the battlestaff:	4.2%	25.0%	20.8%	45.8%	4.2%	0.0%	14.32	<.05
Rating of the TOC design to promote prioritizing	0.0%	4.2%	41.7%	54.2%	0.0%	0.0%	32.15	<.01

actions:								
Rating of the TOC design configuration to permit teamwork and reduce workload:	20.8%	16.7%	25.0%	33.3%	4.2%	0.0%	5.57	.21
Rating of the TOC to promote staff workload sharing and collaboration:	8.3%	20.8%	29.2%	37.5%	4.2%	0.0%	9.63	<.0
Rating of the TOC design to allow BFA's adequate space to advise and assist the Commander:	4.2%	12.5%	12.5%	62.5%	8.3%	0.0%	27.67	<.01
Rating of the TOC system design to promote collaborative planning:	4.2%	16.7%	20.8%	50.0%	8.3%	0.0%	15.57	<.01
Rating of the TOC design to promote synchronization:	0.0%	16.7%	41.7%	33.3%	8.3%	0.0%	14.32	<.01
Rating of the TOC design to permit easy visual and auditory communication to integrate TOC personnel for planning and execution:	4.2%	20.8%	20.8%	37.5%	12.5%	4.2%	7.38	.21
Rating of the TOC design to provide adequate space for the assembly of staff members during wargaming:	16.7%	12.5%	25.0%	20.8%	16.7%	8.3%	1.25	.95
Rating of the TOC design to provide adequate space for supervision of staff during wargaming?	8.3%	12.5%	12.5%	45.8%	12.5%	8.3%	11.68	<.025
Rating of the TOC design to promote the development of an accurate common operating picture:	8.3%	4.2%	29.2%	50.0%	8.3%	0.0%	18.07	<.01
Rating of the TOC design to provide the staff an unobstructed field of view of situational maps:	4.2%	33.3%	20.8%	29.2%	12.5%	0.0%	9.50	.08
Rating of the TOC design to ensure easy information flow between BFA's and synchronization:	8.3%	12.5%	41.7%	25.0%	12.5%	0.0%	8.92	.18
Rating of how the design promoted BOS integration throughout the: Planning	4.2%	12.5%	20.8%	50.0%	4.2%	8.3%	17.49	<.01
Rating of how the design promoted BOS integration throughout the: Preparation	8.3%	12.5%	20.8%	58.3%	0.0%	0.0%	24.26	<.01
Rating of how the design promoted BOS integration throughout the: Execution	4.2%	20.8%	16.7%	50.0%	8.3%	0.0%	15.57	<.01
Rating of how the system design promoted efficient internal communication:	0.0%	25.0%	12.5%	54.2%	8.3%	0.0%	21.42	<.01
Rating of how the TOC system design promoted efficient total staff integration:	0.0%	25.0%	29.2%	41.7%	4.2%	0.0%	14.74	<.01
Rating of the ROC system	4.2%	16.7%	37.5%	37.5%	4.2%	0.0%	13.49	<.01

design to allow easy assimilation of audio data by the entire staff:								
Rating of the TOC system design to allow easy assimilation of visual data by the entire staff:	8.3%	12.5%	33.3%	45.8%	0.0%	0.0%	17.25	<.01
Rating of the TOC system design to facilitate information being sent to the appropriate destination:	4.2%	16.7%	25.0%	50.0%	4.2%	0.0%	17.23	<.01
Rating of the TOC system design to allow the entire staff to monitor operational situations:	4.2%	16.7%	29.2%	37.5%	12.5%	0.0%	8.50	.20
Rating of the ability for you to complete your tasks using the automated tools:	25.0%	20.8%	12.5%	33.3%	8.3%	0.0%	4.75	.57
Rating of the ability for you to maintain better situational awareness using the available automated tools compared to analog methods:	20.8%	16.7%	8.3%	45.8%	8.3%	0.0%	11.41	<.025
Rating of the ability for you to exchange information with personnel you need to collaborate with using the available automated tools:	20.8%	8.3%	33.3%	20.8%	12.5%	4.2%	4.46	.51
Rating of the ability for you to exchange information with personnel you need to collaborate with based on the current TOC design:	12.5%	8.3%	33.3%	37.5%	8.3%	0.0%	9.75	<.05
Rating of the ability of digitization to provide adequate features for maintaining a COP with other personnel, even when face to face communications were not possible:	20.8%	12.5%	16.7%	41.7%	8.3%	0.0%	8.08	.07
	Response	No Response						
What are your top three suggestions to improve the usefulness of the CP?	79.2%	20.8%						
	No	Yes						
Did you identify any TOC safety hazards (shortcomings)? [example: electrical hazards, equipment configuration/design, non secured items (loose), sharp edges]	58.3%	41.7%						

APPENDIX D

Questionnaire Comments

Soldier ID#	*Question#*	*Comment*
	6	**Were there an adequate number of personnel available in your Command Post (CP) perform operations including setup and breakdown?**
1	6	Too many things to do. We can't cover it all.
5	6	Our MTDE does not provide personnel to man the Planning Requirements of the TAC,
5	6	of TOC operations including the security requirements.
8		There were plenty of people to break down a normal brigade CP but there was too much
8		that contractors brought and we had to load.
10		In comparison to last year's NTC rotation, we lack in personnel and experience.
11	6	S-3 needs more soldiers. Make reserve slots and more active duty slots.
12	6	We are forced to argue with TOC staff, personnel from plans. Even then we are barely
12	6	properly.
14	6	We are short people and have been for a long time. Sometimes we have to do a lot more
14	6	kills morale and mission.
16		Require three persons to man all systems (FBCB2J, MCS, SINGARS) plus other
20		Need more authorization t brigade level or do not count reservist slot against active duty
22	6	Not close.
	7	**Did you require personnel augmentation in order to perform CP operations setup, breakdown, camouflage and security?**
5	7	All duties were conducted by the personnel assigned to the TOC.
8	7	We had all cells and all shifts up for days putting up and fixing camouflage and security.
10	7	Normally dayshift would setup TOC and nightshift sleeping quarters, however due to
10	7	had to pull from the night crew to setup TOC and cover down on security as well as
10	7	fundamental precautions.
12	7	There is no TOC security.
16	7	We needed it but didn't receive it.
20	7	Did not receive personnel.
22	7	No, we didn't get them. Often people worked two shifts.
23		Not manned to do guard, KP, camo net repair/adjustments and monitor radio, operate
23		senior NCOs ended up being RTD's operators instead of planning and supervising.
	8	**Were there enough ABCS personnel (manpower) in your CP to reconfigure and workstations?**
5		We have 3 personnel trained to operate two boxes.
8		The workstations are unstable and go down all the time. The contractors are only here we can't go into superuser and fix problems all night.
10		Most of the people in the TOC under this command are new but there are a couple of
10		assisted in reactivation. However, these were extra people to handle the particular
10		procedures.
12	8	We have a handful (2) who are proficient in running the systems. There is no depth.
13		Short one ABCS operator.
14		Sometimes when you needed them, they were not around.
18		Not enough operators available for work stations.
20		Not enough for personnel.
21		We didn't have enough trained operators for night and day shifts.
22	8	Often lost operators to work details, guard duty, and KP.
23	8	Not enough contractor support. Operators not allowed to reconfigure systems.

	9	**Were there enough personnel in your CP to prevent mission delays?**
3		Must stop in the setup to answer radios and questions.
5		But to run full up operation, it required our full MTDE allocation of personnel. Two at
5		Bde FSE two RTOs. Two AFATDS operators, one NCDIC, and Bde FSO.
10		But there are still the inexperienced that create a great time defect in comparison to
14		Well, yes, if you suck it up and drive on knowing what you got is what you got even
14		People to do your work.
16	9	As I stated in previous comment, the moment “sidebar” mission developed we were
16	9	people short. At times, an officer manned the SINGARS.
17	9	At times.
23	9	Skeleton crew because of guard/KP daily tasks.
24	9	We were not part of this last test.
	10	**Do the personnel skills required to setup/breakdown the TOC necessitate a specific Additional Skill Identifier (ASI) or unique MOS?**
4		Setup and breakdown of the TOC only requires a soldier’s technical skills.
6		MOS not needed.
8		New people are forced to learn everything on the fly which is hard and creates stress.
8		TOC training. COMMS setup, camouflage, how to run wire and ground systems, how
8		affiliate DNVTS, what additional elements go into a CP, how to manage personnel with
12		There is a lot of training that operators must complete to be proficient.
13		No ASI or unique MOS needed to setup/breakdown TOC.
14		Not really but you can’t be a rock if you know what I mean.
15		No, because we have not set it up enough to be quick. Once the TOC is setup a couple
15		quick.
16		The setup is simple and principal. Each section knows what needs to be done. Training
16		helped to bring all personnel to speed.
17	10	Yes and no. It is not a necessity but it would greatly improve TOC setup/breakdown
18	10	Everyone needs the training for TOC setup/breakdown.
19	10	As long as everyone knows what they are doing ahead of time and as long as there are
19	10	(one month qualified) to take care of the computers and electronics. On the other
19	10	(or group of people) knew how to do all the required tasks, then the setup/breakdown
19	10	smoother.
20	10	Each TOC setup will be different.
23	10	Basic NCA’s and soldiering tasks for those in TOC.
	13	**Did you receive instruction from your commander regarding interior layout for your specific Battlefield Functional Area?**
3	13	Did the layout as the CDR would have wanted it.
5	13	As we developed the layout of the TOC, the Bde CDR and XO clearly laid out the Form
5	13	each BOSS would interact within the TOC.
6	13	All we have is our vehicle, table and three cars.
7	13	Our interior layout consists of only a table.
8	13	We are told how to pack the vehicles but as far as radios and tables are concerned we
9		I was lied to by my chief commander. I was told I was only going to out here 5 days and
9		computer.
10		BFA was laid out from previous experience.
12		The setup is constantly changing.
14		But as we got out, we go no equipment.
15		Almost too much.
20		Hide excess equipment.
21		There are too many ports.
22	13	Hide excess equipment.

	17	**Does the system design facilitate quick setup?**
3	17	Too many cables or fasteners running through or over the floorboards.
6	17	We have SICPS and circus tents (they are not meant to be put up together).
7	17	With setting up two tents (custom tents) plus two SICPS, plus gutters/"Custom Tents" and
7	17	gutters leak when it rains.
8	17	We have to put down floor boards then setup a tent that nobody had seen and only goes
8	17	special way with inadequate labeling.
9	17	There is too much useless stuff mostly concerning appearance. In a combat situation
9	17	jump in a hurry, maybe then sucking up to the VIPS will take less time.
10	17	Any TOC system design should facilitate quick setup. We will get there with time
12	17	There is a lot of integrate wiring and delicate handling for certain equipment.
13	17	There is a lot of cable to be run in addition to normal setup.
14	17	With the proper number of people, I think so.
16	17	The large setups were quick because it was as big as regular four setups.
17		Takes longer then SICPS but is bigger.
20	17	Too many parts(fasteners) , small and large.
23	17	Plenty of room.
	18	**Does the system design facilitate quick breakdown?**
3	18	Only certain number of personnel trained to breakdown certain system or systems.
5	18	There are a number of cables that are required to run from each vehicle to connect intra-
5	18	inter-TOC connectivity. Also a number of power and audio/video/LAN cables required
5	18	feeds to the CIC and intra-TOC communication system.
6	18	Not sure – have not done breakdown yet.
7	18	We haven't taken down the "custom tents" yet.
8	18	It is not modular, each section is designed to pack itself up and move and help everyone
8	18	done. Right now we have to wait on the slowest man to bring down the tents and floor
9	18	As I explained above, there are too many extras.
12	18	It is very complicated.
16	18	Ask me at the end of April.
17	18	A lot of wires and common stuff takes times.
20		Too many parts, small and large.
23		TOC is too big.
5	18	Same reason as setup. A LOT of cables to recover.
	20	**Did the physical dimensions of the system provide adequate space for digital and digital equipment?**
8	20	The tents are plenty big enough but they need clips or bars to run wires.
9	20	Equipped.
10	20	But is very awkward to install internal wiring.
14	20	There is enough room.
17	20	There was plenty of space for all of the equipment.
	21	**Did the physical dimensions of this TOC design provide adequate space for the of personnel required for effective TOC operations?**
6	21	Not for the BUB (Battle Update Briefings).
7	21	For a BDE TOC, no one with the amount of personnel present at the BUB.
9	21	There is plenty of room for normal operations, however, when giving tours of the TOC to
9	21	from world university, some which are enemy countries, things did get a little tight.
14	21	Yes.
16	21	We are using too large 4F tracks in a briefing area. There are two regular SICPS on the
17		TOC was big enough to hold all personnel.
22	21	C2Vs and 1068s are too cramped.

27 How did ventilation affect C4IS operations?

8 Temperature is something soldiers deal with. Lighting is okay but if a TOC is quiet that
8 22 isn't getting around like it is supposed to.

44 What was the rating of the design to allow the staff to take measures to protect
26 **observation, and detection?**

2 The size makes it easier to see.
3 **27** Too large and takes up plenty of time setup and tear down.
5 Availability of personnel assigned to the TOC. Not enough people to perform TOC
5 Boxes and radios) and provide adequate security.
7 We have a SCAR but with all of the contractors we can't control it completely and without
7 our systems will not work.
8 Well, the new tents have a higher profile which is bad and we have to put up camo net
8 which stands a chance of blowing over easier. Not to mention, the corner dimension
8 light which is bad for discipline.
9 Camo net takes too long to setup.
10 Only three windows are present. TOC concealed light during evening Ops but during day
10 say borderline support.
12 The TOC is very resource intensive.
13 ALL BFA sections within the TOC are setup in a way to assist each other and share info.
14 Well, we have C2Vs that would probably be fine vecause they were made for being
14 this big camo net that in this that we are in, hinders movement and jumping time which
14 A camo net in this age is useless with today's technology.
15 It is a very high TOC. It could be lower and still work fine.
16 The size is difficult to conceal which makes it impossible to movement of area in case of a
16 attack. Laying automated systems out of the tracks makes it virtually impossible to
18 Not enough personnel for different types of guards.
19 Anytime you have an element this big (TOC), the enemy will be able to identify it unless
more natural cover and concealment.
21 The TOC is so big that concealment is almost impossible.
22 Two-toned colors.
23 Again, TOC is too big. Size of camo net makes it obvious it's a large building division CP.
23 Generators, easy to move up on to observe.

47 What was the rating of the TOC system design to allow the Commander to obtain the mission, enemy forces, friendly forces, terrain, and weather?

3 Plenty of new equipment (toys) but still relying on old method of obtaining info., radios.
6 The systems are down or cannot communicater with each other.
7 Need ATTCS systems to function properly before deployment not during.
8 Well, we have a huge bridge and a map. There isn't much info that we can't fit up there.
12 The large selections are a great tool for monitoring information from all ATCCS systems.
14 It's good.
19 The command and control is made easier by the layout in technology.
20 Systems unreliable as of right now.
21 Had a good view of all systems on LSO and could get the information he needed.
23 Large enough to accommodate maps, large screen TVs, etc.
24 The one map load gets crowded.

52 What was the rating of the TOC design to ensure the easy and quick establishment establsihment of an integrated communication system?

5 Relates to FHMUX

58 What was the rating of the TOC design to promote task sharing and teamwork battlestaff?

3 Battlestaff positioned further away from sections (ATTCS devices) to maintain positive
3 Lane as well as helping key players.
5 Biggest drawback is the FHMUX . With this we push up to 4 radios through one antenna.
5 reduces the range of both voice and digital communications.
8 All the TOC cells are shoved away in C2Vs. This means nobody can see us and we can't
8 commander is situated with his back to us. So we get no safety words and we are all
8 vehicles.
9 Although teams work in a big part of the TOC operations, it seems to be more forced on
9 rather than designed that way.
15 It's very large.
16 The CDR has not spent a lot of time at this loca. The Battle CPT seems to relay info
16 Battle NCO ha an understanding of this task.
18 Not enough personnel from each of the sections. Personnel had different missions to
20 Not enough room for all of staff and current TOC layout.
21 All staff is together so they work together.
22 Good vertical movement but not good horizontal movement.
23 BFAs are too spread out- ATTC's operators inside C2Vs have "cave mentality." Tends to
23 cross talk among BFAs.

81 What are your top three suggestions to improve the usefulness of the CP?

1 No KP. No camo net.
3 Improved light sets. Too many small pieces of equipment (cotterpins) are easy to loose.
3 Ceiling too high.
5 Get the CTP function to work. Provide ability to share large amounts of data (OPORD) to
5 a paper and a visual. Reduce the cable requirements to provide feeds into CIC, Intra-
5 Inter-TOC.
8 Quit putting so much stuff in at one time (its overloaded). Train more (soldiers don't know
8 Ask the soldiers what they need. Don't buy something that a salesman told you was
8 fix it to fit soldiers. Have them design it usefully.
9 Concentrate on the ability to jump quicker. Leadership using common sense would help
9 Sufficient amount of personnel to operate the TOC and all the other tasks, i.e., guards,
9 other data that take away from operation personnel.
10 Needs more convenient transport/portability is very cumbersome, better support for
10 ventilation, better sealing, i.e., waterproofing.
11 Get rid of camo net, get rid of TAVs, get rid of SICCPS.
12 Simplify the wiring, a common wire, improve compatibility of ATTCS; soft light.
13 Gutters need to be longer and add snaps to match the roof, the doorwalls to match tent,
13 (up and down) or ends for water run off.
14 Required people, contractors when you need them.
15 Make the side poles adjustable, camo the outside of tent, pad in light hangers.
16 Increase personnel strength at (MTOC), increase training for personnel/cross
16 Automated systems throughout.
17 Adequate number of personnel.
18 More personnel/more operators for work stations, assigned specific duties, more
18 hand.
20 Silencer on generators, large tent to accommodate more personnel for battle update
21 Make smaller, quicker to set up tents, keep as many things in vehicles, less cabling.
22 Open up the central floor more-like a pit, bleacher seats on each side-better flow in the
22 actually located. 60K generator needed. We need more power for this much draw.
22 also 60K is quicker.
23 Maker smaller. Move Engineer BN TOC out of Bde TOC. Get operators out of the C2Vs-

23 analyst to lead interface.
24 Listen to the lower enlisted. They good ideas. Stick with one TOC layout until after NTC.

82 Did you identify any TOC safety hazards (shortcomings)?

2 38 Walking on tents to put up camouflage. Some soldiers slipping could have been hurt.
5 38 The number of cables; power, communications, and data required to establish has a
5 38 when introduced into inclement weather.
6 38 People tripping over the floorboards and broken palettes outside the TOC.
7 Without level ground the use of floorboards and pallets is a hazard. Also, putting camo
8 To put a camouflage net on the TOC the soldiers had to climb on the tents which were
8 hold shifting weight.
9 Loose floors, electrical wires spanning long distances, rain leaking in on electric
9 being assembled at night, fumes inside the TOC.
16 Wiring is all over the AO. This causes two hazards: electrical and obstacles.
20 Electrical cords need to be kept on floor to prevent shorts when it rains. Not enough
20 lights.
21 To set up camo you have to walk on the tent.
22 Power lines could be setup so as not to run across the TOC.
24 Personnel on top of the circus tent too help put como net.

83 Were there any health hazards associated with the CP (including ABCS)?

2 The exhaust from C2V being side by side makes it hard to refuel. The exhaust blows on
2 driver's door.
3 Cables, some power- running back, across, over and on sides of TOC walls.
5 Carbon monoxide created by all power generation devices required to power the TOC.
7 Without proper ventilation in the TOC, it is a health hazard.
8 Work stress: the ABCS does not work. They constantly break or don't do what the
8 When they breakdown, the soldiers get blamed and chewed out but know one cares
8 In the TOC because they are all totally fed up with "BS" systems that don't work and
8 a contractor every 15 minutes.
9 Stress can build up, especially for lower enlisted who when things go wrong, they seem to
9 the leaders.
10 Only complaint is that tent holds both heat and cold. If you can create a portable device
10 wonderful.
12 High noise level, bright lights, TOC exhaust fumes.
14 Have to due with shortage of personnel, too many missions and not enough people – ask
16 Just the usual that comes from working in the military. Stress level is higher because of
16 BDECDR is determined to go to NTC and win the fight digitally.
19 Noise level could be lowered.
20 Noise level, exhaust fumes from C2V flow into TOC, not enough ventilation.
21 The carbon dioxide from the vehicles stayed trapped in the TOC due to lack of ventilation.
22 Noise level equals high, oxygen is not good, ventilation bad.
23 Exhaust froc C2Vs PPU gets into TOC.

90 What are your comments concerning safety and health hazards while operating

5 Also, Elctro-magnetic signatures generated by the power cables, added screens, etc.
6 Floor boards and palettes.
7 Fumes from gas and fire hazards.
15 Need lightning rods.
20 Dust from cleaning the floors to keep TOC neat.
21 Implement ventilation system.
22 Fumes prevalent.

APPENDIX E

Definition of Terms (English & English, 1958)

Scale--unit of light measurement as indicated on the GOSSEN LUNA-PRO light meter.

Candle (candle power)--the unit of luminous intensity of a source of light. It was originally measured by comparison with a standard international candle. Since 1948, a candle is one-sixtieth of the luminance per square centimeter of a complete radiator at the temperature of solidification of platinum.

Foot-Candle--a unit of illuminance or illumination equal to that produced by a uniform point source of one standard candle on a surface every point of which is one foot away from the source.

Lux (meter candle)--the illuminance of a surface one square meter in area receiving uniformly distributed flux of one lumen; or the illuminance produced at the surface of a sphere having a radius of one meter by a uniform point source of one international candle situated at its center.

Lumen--the unit of luminous flux. It is equal to the flux through a unit solid angle from a uniform point source of one candle, or to the flux on a unit surface all points of which are at unit distance from a uniform point source of one candle. It is the strength of the light energy. -- Symbol, L.

Four related terms may be compared:

1. *Luminance*--the light energy emitted, reflected, or transmitted; the luminous flux emitted per unit solid angle and unit projected area of source. This was formerly called photometric brightness. It may be measured in lamberts of millilamberts.

2. *Illuminance*--is the strength of light arriving at, or incident to, a surface; it is what the layman calls the illumination of the surface. Its measurements are in plane geometry terms.

3. *Luminosity*--the brightness-producing capacity of light. Luminosity is not a function of the physical intensity of the light (i.e., of luminance) but of that light under all the prevailing physical conditions (distance, grain of the light surface, translucence of the medium, etc.). It is luminosity, not luminance, which is the physical correlate of brightness. It is measured by the ratio of photometric quantity to radiometric quantity, e.g., lumens (photometric) per watt (radiometric).

4. *Brightness*--is the psychological attribute of color or light as it is perceived. Its physical correlate is luminosity.

NOTE: If a light meter is not available that measures foot-candles or Lux then an ordinary photographic light meter can be used by converting a combination of (1) an arbitrary film sensitivity (ASA rating) with its associated (2) aperature (f-number) and (3) shutter speed to light measurement in Lumen/Meter^2. These measurements can be computed using a formula presented by Nilsson (1981) (i.e., Lumen/meter^2 = (215.3 x f number^2) / ASA x shutter time in seconds).

APPENDIX F

General Illumination Levels and Types of Illumination for Different Task Conditions and Types of Tasks (Van Cott & Kincaide, 1972)

Task condition	Type of task or area	Illuminance level (Ft.-c)	Type of illumination
Small detail, low contrast, prolonged periods, high speed, extreme accuracy.	Sewing, inspecting dark meterials, etc.	100	General plus supplementary, e.g., desk lamp.
Small detail, fair contrast, speed not essential.	Machining, detail drafting, watch repairing, inspecting medium materials, etc.	50-100	General plus supplementary.
Normal detail, prolonged periods.	Reading, parts assembly; general office and laboratory work.	20-50	General, e.g., overhead ceiling fixture.
Normal detail, no prolonged periods.	Washrooms, power plants, waiting rooms, kitchens.	10-20	General, e.g., random natural or artificial light.
Good contrast, fairly large objects.	Recreational facilities.	5-10	General.
Large objects.	Restaurants, stairways, bulk-supply warehouses.	2-5	General.

A Quantitative Method for Evaluating Machine Translation Systems

Barbara D. Broome, Ann E. M. Brodeen, Frederick S. Brundick
U.S. Army Research Laboratory

Malcolm S. Taylor
OAO Corporation

Abstract

The development of machine translation of human language has been impeded, initially by over-expectation, and subsequently by a lack of impartial and quantitative methods to measure its effectiveness. In this paper we describe a pilot study designed to explore the use of standardized reading comprehension tests in the evaluation of machine translation systems. In an investigation of the value added by a French-to-English machine translation system, we find a substantial enhancement of ability for non-French-trained participants. Our conclusions are supported by a rigorous statistical analysis. An approach of this type, which we advance as a general methodology, provides a structured method for systems evaluation, and one that is adaptable to more complex evaluation issues.

Introduction and Background

Machine Translation (MT), a computer-based application that seeks to convert the content of a passage provided in one human language to another, preserving as much of the original meaning as possible, was one of the first large applications in Artificial Intelligence (AI) funded by the U.S. government. It received considerable attention until the mid-sixties when, following a negative assessment by the Automatic Language Processing Advisory Committee, government funding was severely curtailed [1]. The over-optimism associated with AI in general, and with MT in particular, subsided, and the translation effort has continued on a much smaller scale with gradual, but marked, advances.

As international interaction has increased, both commercially and politically, the need for translation has increased in tandem. The Army in particular, with its land operations in foreign countries and its use of coalition forces, stands to benefit from translation tools. At the same time, our understanding of what can and cannot be automated has made our expectations of the contribution of AI more realistic. As a result, machine translation, with its capabilities and its limitations, has recently received more favorable attention from both the commercial and defense communities. There are a variety of translation systems on the market, ranging from quick word-to-word translators to deep syntactic and semantic analyzers. Defense efforts like the Army's Forward Area Language Converter (FALCon) program [2] and DARPA's Trans-lingual Information Detection, Extraction and Summarization (TIDES) program [3] are examples of ongoing efforts to leverage and develop translation tools to support the soldier at all echelons.

The availability of sound MT system evaluation techniques, however, continues to be an issue [4]. The number of languages, the differing technical approaches, and the very complexity of human language itself, all present barriers to assessing the operational effectiveness of MT systems. In earlier studies, a variety of evaluation techniques were employed. Most involved subjective assessments of translation quality, emphasizing the correctness of the output syntax, morphology, and semantics.

In 1993 DARPA funded a three-component evaluation of MT technology, addressing *adequacy*, *fluency*, and *informativeness*. As in previous evaluations, each component was measured subjectively. The third, however, presented an interesting alternative concept. A paragraph was machine translated to English and the participant was given a series of SAT-like questions and asked to assess to what degree the answers could be found in the MT output [5]. Earlier studies concluded that *informativeness* was the least useful of the three components [6]. We, however, find it potentially the most useful, given one critical change: rather than subjectively assess the degree to which information can be found in the translation, simply have the participants actually answer questions based on the MT output. That is, give participants a standard foreign language reading comprehension test that is converted to English via MT.

The strengths of this comprehension-based approach are twofold. First, it focuses on the system as a pairing of the MT tool with its user. Instead of asking how good the machine-produced translation is, it focuses on the more tractable issue of how much better a translation task can be performed when given access to machine translation. Second, it moves evaluation from the subjective to the objective realm, using the number of correctly answered questions on the reading comprehension test as a measure of the system's operational effectiveness.

Our research further explores the use of reading comprehension tests to establish a baseline in the evaluation of MT technology. We consider the specific concerns identified in previous studies, refining the approach to establish a formal evaluation method. Next, we describe the results of a pilot study in which we employ a standardized reading comprehension test to assess the effectiveness of a French MT system.

The Pilot Study

Applying formal statistical techniques to the results of the pilot study, we address three research questions.

Are test scores for untrained participants without access to MT equivalent to random guessing?

If we hope to measure the effect of a French MT in improving reading comprehension for non-French-readers, we need to establish their level of performance absent MT support. While one might assume that participants without MT are randomly guessing at answers, the similarities between French and English may provide cues (or possibly even miscues) that negate this assumption. In our pilot test, we provide questions in both French and French-to-English-via-MT formats. Future testing would be simplified if we could verify the randomness assumption, eliminating the need to include the original French questions. Using a binomial hypothesis test, we verify that test scores for untrained participants without access to MT are, indeed, indistinguishable from random guessing.

Twenty participants, with no formal training in French, took the reading comprehension portion of the 1995 SAT II: French Subject Test [7].* The reading comprehension portion consisted of 29 multiple-choice questions, each with 4 possible answers. The participants answered a combined total of 151 questions correctly.

If the responses were the result of random guessing, the number of correct responses follows a binomial distribution

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, \ldots, n \tag{1}$$

with parameters $n = 20 \times 29 = 580$ and $p = \frac{1}{4}$.

To test the hypotheses

H_0: The results—number of questions answered correctly—do not exceed what one would expect from guessing,
H_a: The results exceed what one would expect from guessing,

it is sufficient to evaluate the sum

$$\sum_{x=151}^{580} b(x; 580, \tfrac{1}{4}). \tag{2}$$

This sum is the probability of observing 151 or more correct responses if the participants are guessing. Evaluation of expression (2) yields a p-value of 0.30, far too large to reject the null hypothesis. The distribution of $b(x; 580, \frac{1}{4})$ shown in Figure 1 suggests the adequacy of a normal approximation. We chose to evaluate expression (2) directly due to its ease of computation.

How do test scores for participants without access to MT compare to their test scores when MT is available?

This question provides insight into the critical evaluation issue of whether or not a selected MT system provides value. Using the Wilcoxon Signed-Rank Test, we find that test scores for untrained participants are significantly higher when given access to MT.

Upon completion of the exercise detailed above, the participants then proceeded to respond to the identical test after it had been submitted to SYSTRAN MT software [8] for translation into English, their native language. At the conclusion of this phase, we had at our disposal a set of 20 paired observations, $(x_i, y_i), i = 1, 2, \ldots, 20$, the number of questions answered correctly before- and after-MT.

*Copyright ©1998 by College Entrance Examination Board. All rights reserved.

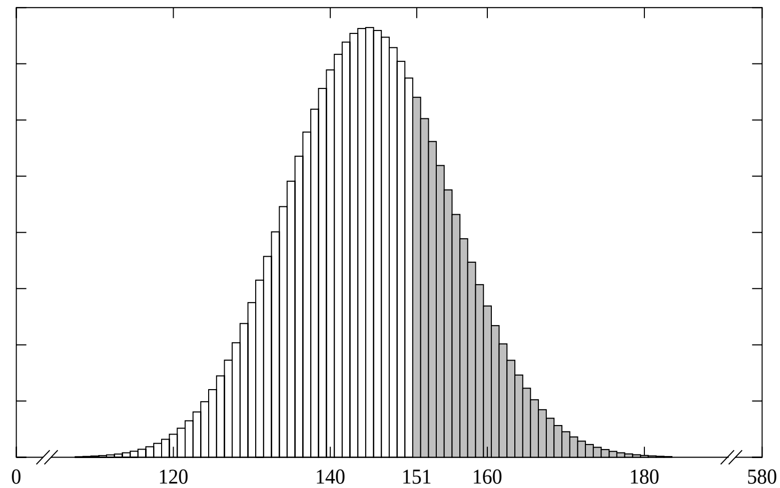


Figure 1. The Distribution Of $b(x; 580, \frac{1}{4})$.

To investigate the impact of the MT system we chose the Wilcoxon signed-rank test with hypotheses

H_0: The MT system has no discernible impact on the ability of the participants to respond correctly,
H_a: The MT system has a discernible impact on the ability of the participants to respond correctly.

Formally, the hypotheses involve comparison of the before-mean $E(X)$ and after-mean $E(Y)$.

H_0: $E(Y) \leq E(X)$
H_a: $E(Y) > E(X)$.

For these data, presented in Table 1, we have $x_i < y_i \; \forall \; i$, with one exception. In that instance, a tie occurred. Rejection of the null hypothesis is thus assured *a priori*. The p-value for the Wilcoxon test applied to these data was determined to be less than 0.005.

How do test scores for untrained participants using MT compare with those of participants who have two or more years of strong French language study?

At this point, we take full advantage of the fact that we have constructed the pilot study from the 1995 SAT II: French Subject Test, for which detailed statistics are available on the test scores of over 3,000 students. The recommended preparation for taking the French SAT is 3–4 years of French language study in high school, or two years of strong preparation. Using a two-sample t-test, we find that test scores for untrained participants with access to MT are significantly higher than those of French-trained participants that took the same French SAT test in 1995.

To compare the level of performance of the test participants with that of students taking this College Entrance Board exam, we chose a two-sample t-test with

Table 1. Number of Questions Answered Correctly

Participant	Without MT	With MT
1	10	21
2	6	14
3	9	18
4	8	21
5	6	21
6	4	17
7	7	19
8	10	21
9	6	17
10	7	15
11	11	17
12	6	22
13	9	9
14	7	21
15	9	23
16	10	20
17	6	18
18	9	20
19	4	14
20	8	9

population variances unknown. We were able to extract from information provided by the testing service an estimate of the mean and variance of student test scores based on a sample of well over 3,000 applicants, more than adequate to support a normal population assumption. The test score for an i^{th} participant may be modeled as $x_{i\cdot} = \sum_{j=1}^{29} x_{ij}$, where x_{ij} is a Bernoulli random variable with parameter p_j. The number of summands, 29, is sufficiently large to support a second normal assumption (under the Central Limit Theorem) for the conceptual population of untrained participants.

We established the hypotheses

H_0: The participants' MT-enhanced performance did not exceed that of students taking the college entrance board exam.
H_a: The participants' MT-enhanced performance exceeded that of students taking the college entrance board exam.

Formally, the hypotheses involve comparison of means of the two populations

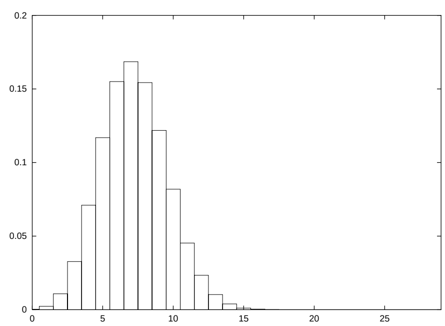
H_0: $E(X) \leq E(Y)$
H_a: $E(X) > E(Y)$

where $E(X)$ denotes the mean of the hypothetical population of participant scores and $E(Y)$ the mean of the population of student scores.

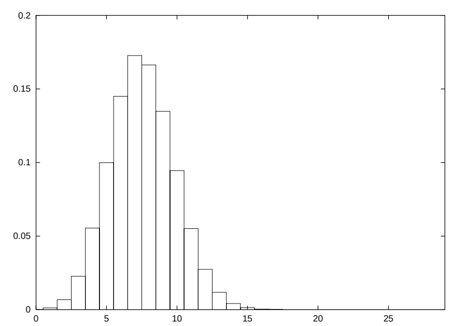
Upon evaluation of the t-statistic with 19 degrees of freedom, the null hypothesis H_0 was rejected with an associated p-value of less than 0.005. These data suggest that untrained participants, with the aid of an MT device, not only rise to the level of, but exceed, the performance of students taking the College Entrance Board exam.

A graphical summary of the three research questions appears in Figure 2.† For the statistician, this additional detail might be superfluous, but for the audience we hope to persuade, it is more likely essential. Someone for whom binomial, Wilcoxon, and t-tests are unfathomable jargon, let alone hypothesis testing and p-values, can still appreciate the striking similarity of Figures 2a and 2b (*question one*); that the histogram in Figure 2d lies to the right of that in Figure 2b (*question two*); and that the histogram in Figure 2d is to the right of that in Figure 2c (*question three*). The raw data underlying Figure 2 is presented in Table 2.

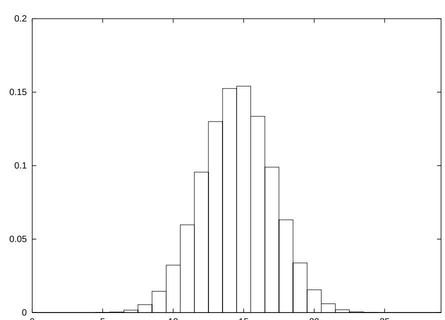
The three tests—the binomial, Wilcoxon, and Student's-t—are *a priori* nonorthogonal contrasts whose significance levels (p-values) must be accounted for at the experimental level. We chose to address this by applying a Bonferroni correction for the hypotheses tested. The Wilcoxon and the t-test, by virtue of their exceedingly small p-values, remain highly significant.



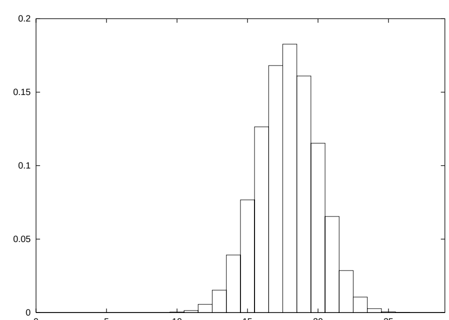
a) Random Guessing; $b(x; 580, \frac{1}{4})$



b) Without Translation



c) SAT Test Scores



d) Machine Translation

Figure 2. Graphical Summary of Pilot Study

†In panels b–d the raw data are simulation enhanced.

Table 2. Number of Participants Answering Each Question Correctly

Question Number	Without MT Count	Without MT %	With MT Count	With MT %	SAT %
1	5	25	18	90	88
2	5	25	7	35	75
3	7	35	15	75	73
4	8	40	7	35	38
5	9	45	5	25	41
6	3	15	18	90	69
7	3	15	12	60	53
8	3	15	19	95	49
9	5	25	18	90	63
10	5	25	16	80	57
11	6	30	13	65	61
12	6	30	17	85	57
13	10	50	16	80	63
14	9	45	2	10	17
15	4	20	14	70	39
16	2	10	14	70	55
17	3	15	19	95	67
18	1	5	17	85	31
19	4	20	4	20	47
20	8	40	3	15	28
21	7	35	16	80	57
22	5	25	7	35	52
23	5	25	14	70	20
24	4	20	18	90	51
25	2	10	5	25	26
26	5	25	4	20	54
27	7	35	9	45	26
28	4	20	15	75	44
29	6	30	15	75	53

Conclusion

In conclusion, we submit the method outlined in this study as a general MT evaluation technique. Its emphasis on quantitative measures of effectiveness provides a greatly needed structure for impartial and statistically sound assessments of the effectiveness of MT systems. Further, the approach is flexible enough to expand and adapt to more complex issues, such as evaluating improvements to a single system under development, comparing widely diverse commercial and research systems, and assessing various user characteristics on MT effectiveness.

Acknowledgment

Ms. Broome gratefully acknowledges the helpful support and advice of LTC James Bass, Dr. Guisseppe Forgionne, Dr. Ana Schwartz, Mr. Iftikhar Sikder and Ms. Dawn French in designing the pilot study, which was initiated as a class project at the University of Maryland Baltimore County.

References

[1] Kay, M. "Machine Translation," http://www.lsadc.org/Kay.html.

[2] DeHart, J., C. Schlesiger and M. Holland. "Issues in Optical Character Recognition for Army Machine Translation," Symposium on Document Image Understanding Technology, Annapolis, MD, April 14–16, 1999.

[3] Hovy, E., N. Ide, R. Frederking, J. Mariani, A. Zampolli. "Multilingual Information Management: Current Levels and Future Abilities," a study commissioned by the National Science Foundation in 1998, http://www.cs.cmu.edu/~ref/mlim.

[4] Taylor, K. and J. White. "Predicting What MT is Good for: User Judgements and Task Performance," Proceedings AMTA, October 1998.

[5] White, J. S. and K. B. Taylor. "A Task-Oriented Evaluation Metric for Machine Translation," http://www.cst.ku.dk/projects/eagles2/2ndworkshop/annprog2.html.

[6] O'Connell, T. A., F. E. O'Mara, and K. B. Taylor. "Sensitivity, Portability and Economy in the ARPA Machine Translation Evaluation Methodology," http://ursula.georgetown.edu/mt_web/3Q94FR.html.

[7] Real SAT II: Subject Tests (French, French with Listening), The College Entrance Board and Educational Testing Services, Princeton, NJ, 1998.

[8] SYSTRAN Software, Inc. SYSTRAN PROfessional for Windows. La Jolla, CA, 1998.

CONTRIBUTED SESSION II (1330 - 1500)

Passive Unicast Network Tomography based on TCP Monitoring

Yolanda Tsang, Mark Coates and Robert Nowak [1]
Department of Electrical and Computer Engineering, Rice University
6100 South Main Street, Houston, TX 77005–1892

Abstract

Network tomography is a promising new technique for studying the (internal) behavior of large-scale networks based solely on end-to-end measurements. Most network loss tomography methods utilize active probing. While such measurement schemes are efficient, the probing burden may become prohibitive for large-scale networks. As an alternative, we propose a new, completely passive measurement framework based on sampling of existing TCP traffic flows.

The new passive unicast network tomography methodology we propose shows considerable promise. We demonstrate its performance using extensive ns-2 simulations. We observe that we are able to accurately estimate the losses experienced by existing TCP flows. As these can differ substantially from losses suffered by other forms of traffic, we surmise that in some situations inference from active probing may offer a poor reflection of existing TCP loss rates.

I. Introduction

A. Background and Motivation

Accurately characterizing the performance of a large-scale network is essential if it is to be successfully managed and controlled. The characterization must extend further than path-level behavior; it is necessary to acquire information about internal (link-level) performance. One way to achieve this is to gather statistics at as many internal routers as possible, but the collection and compilation of such statistics is an onerous and expensive task. Often it proves impossible for any one organization or individual to collect relevant statistics when various parts of a network are administered by different parties.

An attractive alternative is to infer the internal performance from end-to-end measurements that are comparatively easy to make. A number of authors have proposed methods for achieving this task; the general problem has been termed "network tomography". One of the most promising proposals is MINC (Multicast Inference of Network Characteristics) [1], which proposes strategies for estimating loss, delay distributions and variances, and topology [6, 11, 5]. All of these techniques use active multicast probing, then exploit the inherent correlation between the losses/delays observed by multicast receivers. The performance of these algorithms is impressive [7], but there are two serious deficiencies in the methodology. Firstly, multicast protocols are not supported by significant portions of the Internet. Secondly, the internal performance measured by active multicast probes often

[1]This work was supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0349, the Office of Naval Research, grant no. N00014-00-1-0390, and Texas Instruments.

differs significantly from that encountered by unicast packets, which comprise by far the most substantial component of Internet traffic [12].

Recently, strategies have been proposed to avoid these limitations [9, 10, 12]. These strategies involve the use of unicast packets to acquire the end-to-end statistics required for network tomography. The difficulty encountered by these techniques is that unicast packets do not obey the same well-behaved correlations as multicast packets. Network probing using back-to-back packets has been proposed in a number of measurement schemes [3, 4, 8, 15]. A network tomography procedure based on measurements of back-to-back (closely time-spaced) unicast packets was first proposed in [9]. To address the problem of imperfect correlations, the authors used probabilistic modeling methods and likelihood maximization techniques (Expectation-Maximization (EM) and exact inference on factor graphs). The maximum bias of estimates was characterized in terms of the observed path-level correlations. The authors in [12] proposed an alternative strategy based on sending multiple-packet probes to improve the observed correlations, and then applied the multicast-based algorithms of the MINC project for loss estimation.

B. Contribution

Although the unicast network tomography proposal [9] suggested that the collection of statistics is possible using both active probing and passive sampling of existing traffic, the experiments and performances reported in previous work [9, 10, 12] use active probing strategies. Moreover, a number of issues encountered in performing passive sampling have not been addressed. In this paper, we concentrate on the inference of internal link losses from passive unicast end-to-end measurement. The motivation for passive inference is two-fold. Firstly, there is a risk that for a substantial network the insertion of the large number of probes required for accurate estimates might significantly impact the performance of the network (perturbing the very quantity to be estimated). Secondly, the sampling of existing flows potentially offers an opportunity to estimate the losses experienced by the existing flows (which can be substantially different from those experienced by inserted probes). We propose a new, truly passive methodology for unicast network tomography, and we assess the feasibility and performance our approach through `ns-2` [2] simulation experiments.

We use extensive `ns-2` simulations to explore various aspects of our new methodology's performance. We assess whether sufficient back-to-back packet statistics can be extracted from existing TCP [14] flows between the source and receivers. We quantify the performance of the inference algorithm in terms of mean absolute error.

The paper is organized as follows. In Section II, we review the basic unicast network tomography problem and the technical issues involved. In Section III, we describe the new passive measurement framework, focusing on the collection of statistics from TCP flows. In Section IV, we apply the loss modeling and likelihood analysis techniques devised in [9] to our new approach. In Section V, we describe and discuss the results of `ns-2` simulations exploring the performance of the passive measurement and inference methodology. In Section VI, we discuss some of the limitations of the passive scheme, and propose some potential remedies. Conclusions are made in Section VII.

C. Related Work

The proposals and performance analysis presented in this paper are important extensions of the work presented in [9, 10, 12], but the basic idea of exploiting correlations of closely-spaced packets remains the same. The authors of [7] presented extensive performance analysis of the multicast loss inference technique proposed in [6]; in this paper, we aim to perform a similar analysis for passive unicast inference. In [9, 10] the nature of the probing (active vs. passive) was not specifically addressed, and the methodology proposed in [12] is based on active probing.

In [7, 13], the authors also applied packet pairs in detecting shared losses using unicast probes. In a two-receiver tree-structured network, they were interested in detecting whether or not losses in two flows suffering similar losses are occurring on the shared path. The authors in [13] mentioned the possibility of passively recording the statistics but they applied only active probing to their simulations. In our work, we estimate individual link statistics, and thereby the congestion level of all links in a larger network.

II. Unicast Loss Tomography

In this section, we provide a brief overview of the unicast loss inference problem, describe the back-to-back measurement framework proposed in [9], and review the loss modeling framework. We illustrate the problem of unicast loss inference by considering the simple case in which a single source sends packets to multiple receivers. The problem and methodology are readily extended to the multiple-source case. Figure 1 depicts an example of this form of topology; the network appears to the source as a tree. The nodes of the tree correspond to the source (node 0), internal routers (nodes 1–4) and receivers (nodes 5–11). We define a 'link' as the connection between any two adjacent nodes in the tree, deem the set of links connecting a source and any receiver a 'path', and a subset of connected links in a path is referred to as a 'subpath'. The tree of Figure 1 does not necessarily depict all routers encountered by packets traveling from the source to receivers. It is possible that a number of routers are passed as a packet travels from node 1 to node 3, for example.

We consider the situation where measurements can only be made at the edge of the network and assume that the routing (and thus the topology) table is fixed for the duration of the measurement. The goal of passive unicast loss inference is to estimate the loss rates on the internal links of the network using solely passive sampling of the existing traffic. Estimating the loss rates is equivalent to estimating success rates, and henceforth we shall speak solely of success rates, since they simplify mathematical expressions in the proposed framework. (The success rate is simply one minus the loss rate).

It is straightforward to estimate *path* success rates, but, unfortunately, there is no unique mapping of the path success rates to the success rates on *individual* links in the path. To overcome this difficulty, the authors in [9] propose a methodology based on measurements of back-to-back packet pairs. These measurements provide

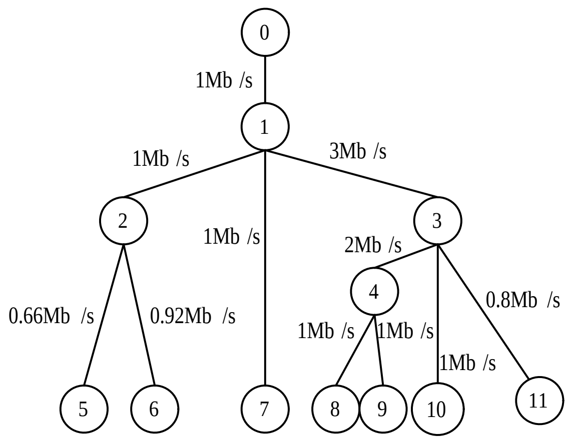


Figure 1: An example network topology with a single source (node 0), 4 internal routers, and 7 receivers.

an opportunity to collect more informative statistics that can help to resolve the links.

Back-to-back packet pairs have been utilized for inferring a number of network performance metrics [3, 4, 8, 15]. A back-to-back packet pair refers to two closely time-spaced packets, possibly destined for different receivers, but sharing a common set of links in their paths. If two back-to-back packets are sent across a link and the one of the pair was received, then it is highly likely that the other packet was also received. In other words, we expect that the conditional success probability of one packet (given that the other was received) may often be close to one. This observation has been verified experimentally in real networks [15].

A unicast network tomography method that exploits the correlation between back-to-back packet losses is developed in [9]. The basic idea is quite straightforward. Suppose two back-to-back packets are sent to two different receivers. The paths to these receivers share a common set of links from the source but later diverge. If one of the packets is dropped and the other successfully received, then (assuming strong correlation of losses on common links) one can infer that the packet was probably dropped on one of the unshared links. This enables the resolution of losses on individual links.

III. TCP-Based Measurement Framework

Earlier unicast network tomography schemes have focused on active probing [9, 10, 12] (insertion of additional packets whose sole purpose is for measurement), although the possibility of passive measurement (collecting measurements directly from existing traffic) was mentioned, but not investigated, in [9]. Although active probing allows one to control the timing and nature of the measurements, thus assuring that sufficient measurements are made, it imposes an extra burden on the network which may make it impractical in many situations. Furthermore, active probing will disrupt the transmission of normal traffic, which may lead to biased estimates of the true losses (experienced in the absence of probe packets). Also, probe traffic may experience significantly different losses than normal traffic (e.g., TCP) since the temporal structure of the probing is specified by the experimenter, and is generally different than that of normal traffic. For example, TCP often results in clusters of packets separated by a significant (round-trip) time, contrary to a uniform or Poisson probing scheme.

We propose a passive traffic monitoring/sampling scheme in order to circumvent the problematic issues surrounding active probing. We focus on TCP-based measurement, because we are interested in estimating the link-level losses experienced by TCP connections flowing from the source to the receivers. When estimating TCP losses, the spacing of measurements should clearly not be uniform or exponentially distributed. Where the TCP traffic to a particular connection is dense, we should make many measurements; where light, only a few. The measurement process should ideally be a subsampled version of the true traffic. This is impossible to achieve exactly, because we have other constraints. For example, we need to make back-to-back measurements and we need to ensure that measurements are sufficiently spaced to provide the approximate inter-pair temporal independence assumed in our statistical model.

The situation we consider is one where the source has numerous contemporaneous TCP connections with a number of receivers. We want to extract as many informative measurements as we can from the existing TCP traffic. We concentrate first on extracting the important packet-pair measurements, which are less common than the isolated packet measurements. We first inspect the sending times of the TCP traffic at the source and decide that two packets form a packet-pair if their time-spacing is less than a threshold δ_t seconds. This threshold is dependent on the sending rate of the source. Commencing at the start of the measurement period, we sweep forward in time, seeking and identifying the first packet-pair. We then step forward by a fixed time interval $\Delta_t \gg \delta_t$ and begin to search for the next pair. In this way, we ensure that the pairs we include in our analysis are separated by a reasonable time interval, making the assumption of statistical independence between pairs more realistic.

Following the collection of these pairs, we include any isolated packets that do not violate the time separation requirement. In general, we observe the number of such packets to be significantly larger than the number of pairs. We are now in a position to define the following statistics. For the purpose of anti-causal conditioning (see Section V), let $n_{i,j}$ denote the number of pairs wherein the first packet is sent to receiver i and the second to receiver j AND the second packet is successfully received. Let $m_{i,j}$ denote the number of these pairs in which both packets are successful. Let n_i denote the number of isolated packets sent to receiver i and let m_i denote the number successfully received. Collecting all the measurements, define:

$$\mathcal{M} \equiv \{m_i\} \cup \{m_{i,j}\} \qquad \text{and} \qquad \mathcal{N} \equiv \{n_i\} \cup \{n_{i,j}\},$$

where the index i alone runs over all receivers and the indices i, j run over all pairwise combinations of receivers in the network.

IV. Loss Modeling and Likelihood Analysis

The modeling frameworks of [6, 9, 10, 12] assume a simple Bernoulli loss model for each link for individual packet transmissions. The *unconditional* success probability of link i (the link into node i) is defined as

$$\alpha_i \;\equiv\; \Pr(\text{packet successfully transmitted from } \rho(i) \text{ to } i),$$

where $\rho(i)$ denotes the index of the parent node of node i (the node above i-th node in the tree; *e.g.,* referring to Figure 1, $\rho(1) = 0$). A packet is successfully sent from $\rho(i)$ to i with probability α_i and is dropped with probability $1 - \alpha_i$. Loss processes on separate links are modeled as mutually independent.

If a back-to-back packet pair is sent from node $\rho(i)$ to node i, then we define conditional success probability of link i as:

$$\gamma_i \;\equiv\; \Pr(\text{1st packet } \rho(i) \to i \mid \text{2nd packet } \rho(i) \to i\ ),$$

where 1st and 2nd refer to the temporal order of the two packets, and $\rho(i) \to i$ is shorthand notation denoting the successful transmission of a packet from $\rho(i)$ to i.

The sampled TCP flows and corresponding traffic statistics lead to following likelihood function. We denote the collections of the unconditional and conditional link success probabilities as $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively. The *joint* likelihood of all measurements is given by

$$l(\mathcal{M} \,|\, \mathcal{N}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \;=\; \prod_i \mathcal{B}i\,(m_i \,|\, n_i, p_i(\boldsymbol{\alpha})) \;\times\; \prod_{i,j} \mathcal{B}i(m_{i,j} \,|\, n_{i,j}, p_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\gamma})),$$

where $\mathcal{B}i(m|n,p) \equiv \binom{n}{m} p^m (1-p)^{n-m}$, the binomial likelihood function, $p_i(\boldsymbol{\alpha})$ is a product the unconditional success probabilities in the path from the source to receiver i, and $p_{i,j}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ is a product of conditional success probabilities (on the common links in the paths to receivers i and j) and unconditional success probabilities on the links to j not shared in the path to i.

The EM Algorithm developed in [9] can be used to compute maximum likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Beginning with an initial guess for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the algorithm is iterative and alternates between two steps until convergence. The Expectation (E) Step computes the conditional expected value of the unobserved packet losses at internal nodes given the observed data, under the probability law induced by the current estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. The Maximization (M) Step combines the observed (path) losses and expected unobserved (internal) losses to compute new estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Each iteration of the EM Algorithm is close to $O(NL)$ in complexity, where N is the number of possible measurements that we can make and L is the number of levels involved. The exact complexity depends on the topology of the network. Our `ns-2` experiments have shown that the algorithm typically converges in a small number of iterations (typically 5-15). Moreover, it can be shown that the original (observed data only) likelihood function is monotonically increased at each iteration of the algorithm, and the algorithm converges to a local maximum of the likelihood function.

V. `ns-2` Simulation Experiments

We evaluated the passive loss inference framework using the `ns-2` simulation environment. In the simulations that we perform, we strive to investigate a number of issues. We gauge the performance of the combined EM loss inference algorithm and passive measurement framework under a variety of traffic conditions and queueing policies. We also explore the measurement period required to collect a sufficient number of data for accurate inference in a passive framework.

A. Simulation Framework

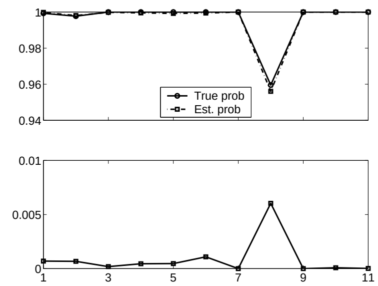
Network Topology: We use the same 12-node network topology in all experiments (see Figure 1). This topology is intended to reflect (to some extent) the nature of many networks — a slower entry link from the source, a fast internal backbone, and then slower exit links to the receivers. The chosen topology gives us the flexibility to explore the effects of having receivers at different distances from the source, and to examine the effect of varying fan-outs. We fix the size of all queues to be 35 packets. We consider four different traffic scenarios and perform all experiments using the droptail queuing policy throughout the network.

Traffic Generation and Statistics Collection: In all experiments, we assume that there are TCP connections to the receivers that last for the extent of the measurement period. In addition, we set up a variety of short-duration TCP sessions, both from source to receiver and as cross-traffic on internal links, as well as exponential on/off traffic sources traversing various paths. In total there are approximately thirty TCP connections and thirty UDP connections operating within the network at any one time. The average utilization of the network is in all cases relatively high; otherwise, we experience very drops and loss estimation is of little interest.

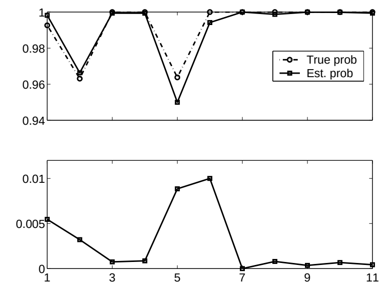
We utilize all the TCP connections flowing from the source to the receivers when collecting statistics using the procedure discussed in Section III. We set the maximum time-spacing between packets within a pair to $\delta_t = 1$ ms and the minimum spacing between pairs to $\Delta_t = 10$ ms. We collect measurements over a 300 second interval. The four traffic scenarios are described as follows.

Traffic Scenarios 1–3 (heavy losses at 1 or 2 links): In these three traffic scenarios, we strive to ascertain the capability of our passive framework to discern where significant losses are occurring within the network. We assess its ability to determine how extensive the heavy losses are and to provide accurate estimates of loss rates on the better performing links. In each case, we establish each heavy loss link by adding substantial exponential on-off and short-duration TCP session cross-traffic flows to the link traffic. In Scenario 1, link 4-8 experiences heavy losses (enabling assessment of the framework's ability to localize losses at links near the receivers). In Scenario 2, links 1-2 and 2-5 experience substantial losses (testing the framework's capacity to separate cascaded losses). In Scenario 3, links 1-2 and 4-8 experience substantial loss. This last scenario tests the ability to resolve distributed losses in different branches of the network.

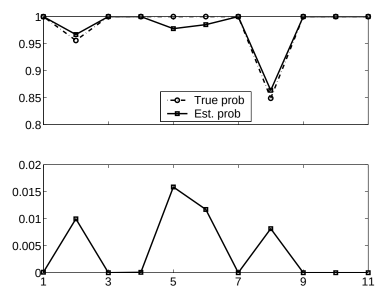
Traffic Scenario 4 (mixed traffic with medium losses): In this last scenario, we introduce many on-off UDP and on-off TCP connections throughout the topology and insert extra links (not depicted in Figure 1) connecting to the internal nodes. These links allow us to develop TCP cross-flows that have a range of different round-trip-times.



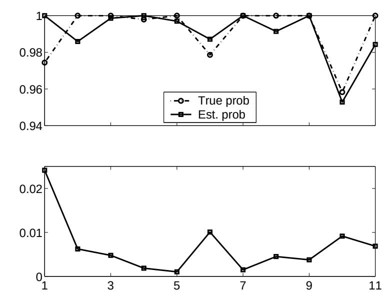
(a) Heavy losses on link 4-8



(b) Heavy losses on links 1-2 and 2-5



(c) Heavy losses on links 1-2 and 4-8



(d) Traffic mixture - medium losses

Figure 2: Simulation Results. True and estimated link-level success rates of TCP flows from source to receivers for all traffic scenarios. In each subfigure, the three panels display for each link 1-11 (horizontal axis): (top) true and estimated success rates using drop-tail queues, and (bottom) mean absolute error for each link.

B. Simulation Results

We conduct ten independent simulations of each traffic scenario and queuing policy over a measurement period of 300 seconds. Figure 2 displays the results of our simulations for each of the different traffic scenarios. We see that the estimated success rates are in good agreement with the true TCP success rates (based on direct counts of total losses on each link). In the heavy-traffic scenarios, we see that the worst-case mean absolute error is about one percent. The passive framework is thus capable of identifying where heavy losses occur.

In Figure 3, we examine the relationship between the average mean absolute error and the measurement period. As expected, the error decreases as the measurement period increases. Note that even for a 60 second measurement period, the averaged mean absolute error is less than 0.6 percent.

Finally, we draw attention to the fact that the losses we measure for the TCP flows can be very different from those of the UDP traffic. For example, in the mixed traffic scenario, we observed average TCP losses of approximately 3 percent

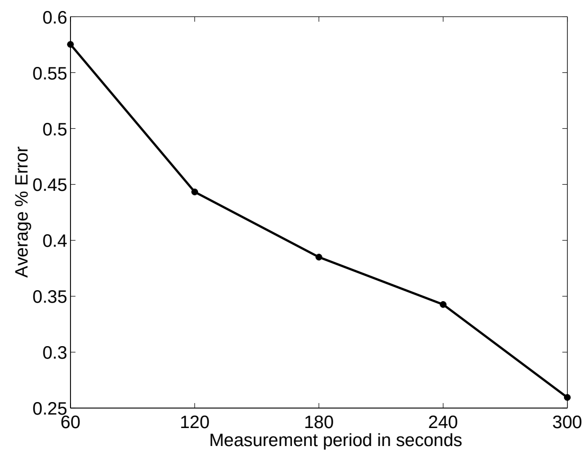


Figure 3: The performance error (mean absolute error averaged over all links) versus measurement period.

on links 2-6 and 3-10, whereas the on-off exponential traffic experienced losses of nearly 20 percent. This is a strong indication that active probing may provide a poor indication of losses in the existing TCP traffic.

VI. Discussion

The sampling (or perhaps more aptly "mining") of the TCP traffic flows for packet-pair events is a crucial step in our methodology. As discussed in Section III, the simplest approach is simply to scan the traffic flows, locating packet-pair occurrences in a sequential fashion (beginning at the start of the measurement period). When a packet pair is located, we skip ahead Δ_t seconds and resume the scan. This results in a fast extraction algorithm, with the amount of measurements involving each receiver dependent on the throughput to that receiver. The disadvantage of this simple algorithm is that some pairs are more informative than others. Due to the nature of TCP, there are large numbers of back-to-back pairs to the same receiver (arising whenever the source receives notification that it can send a group of packets equal to the current window size). The pairs involving different receivers are rarer, occurring when the source switches from one connection to another. Because they are fewer in number, they provide more information for inference.

A potentially more effective alternative algorithm is to consider the reduced set of back-to-back "cross-pairs" that involve different receivers. We begin by locating and including all cross-pairs. The time intervals between cross-pairs are often sufficiently large so that we can include all of them without violating the requirement of a Δ_t separation. If not, we scan through the set; when we have to decide which of a set of closely-spaced cross-pairs to eliminate, we eliminate all but the pair with the least representation in the current set of included pairs. After the set of cross-pairs has been finalized, we incorporate "auto-pairs" (back-to-back packets destined for the same receiver), excepting those that violate the Δ_t time-separation criterion with the pairs already included. If we discover that this algorithm has resulted in the under-representation of some type of auto-pair, then we start again, including

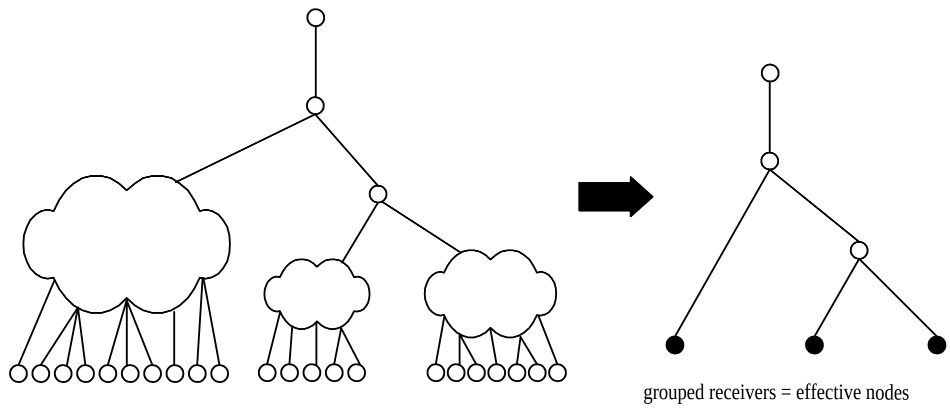


Figure 4: An example of "grouping" receivers to reduce network complexity.

the under-represented pair in the initial set. Finally, single (isolated) packets are included, again maintaining the Δ_t time-separation. In this way, we can extract more informative statistics from a given set of traffic data, which should lead to improved estimates. Conversely, a more effective sampling strategy of this nature should reduce the measurement time duration required for accurate loss inferences.

A more aggressive approach to obtain more informative data could involve alternative servicing strategies at the source. For example, instead of a basic round-robin service strategy, the source could employ a scheme that would enhance the chances of cross-pairs occurrences, without necessarily deviating from a TCP format. We are currently investigating this possibility as well as the more sophisticated data collection process described above.

One final point of discussion is the issue of scalability. The EM algorithm itself is reasonably scalable (approximately linear in the number of nodes), but as the number of receivers grows, the potential for a sufficient number of cross-pairs may diminish. Alternative data collection and servicing strategies, like those mentioned proposal above, could mitigate this problem. Nonetheless, the possibility of "data-starvation" may limit one's ability to passively estimate all link-level losses in very large networks. This may not be as bad as it seems. For example, rather than estimating all link-level loss rates, in many practical situations it may be sufficient to determine loss rates on a few of the first links along the paths from the source to the receivers. By grouping receivers into clusters, the actual network can be abstracted into a smaller network with "effective nodes" replacing the original clusters. This is illustrated in Figure 4, where we use a "cloud" to denote the corresponding aggregation of links (subnetwork) to the clustered receivers. Auto-pairs and cross-pairs can be shared among clustered receivers, and we should be able to reliably estimate the loss rates on the upper links (close to the source) as well as average loss rates for the subnetworks associated with the clustered receivers. We are currently investigating this approach through theoretical analysis and `ns` simulations.

VII. Conclusions

The new passive unicast network tomography methodology we have proposed shows considerable promise. We have demonstrated using extensive ns-2 simulations that sufficient data can be collected using passive sampling to perform accurate loss inference, even for relatively short measurement periods. Moreover, we have observed that we are able to accurately estimate the losses experienced by existing TCP flows. As these can differ substantially from losses suffered by other forms of traffic, we surmise that in some situations inference from active probing may offer a poor reflection of existing TCP loss rates.

Recognizing some of the potential limitations of the passive scheme (data starvation, scalability problems), we have proposed alternative data-mining and servicing strategies at the source that may provide more informative data. We also discuss a method for clustering receivers to reduce the effective complexity of the network, thus allowing us to focus on identifying losses on a subset of interesting links. Both of these are topics under current investigation.

References

[1] Multicast-based inference of network-internal characteristics (MINC). See gaia.cs.umass.edu/minc.

[2] The network simulator-2. For more information, see http://www.isi.edu/nsnam/ns/.

[3] M. Allman and V. Paxson. On estimating end-to-end network path properties. In *Proc. Sigcomm*, 1999.

[4] J-C. Bolot. End-to-end packet delay and loss behaviour in the Internet. In *Proc. ACM Sigcomm '93*, pages 289–298, Sept. 1993.

[5] R. Cáceres, N. Duffield, J. Horowitz, F. Lo Presti, and D. Towsley. Loss-based inference of multicast network topology. In *Proc. IEEE Conf. Decision and Control*, Dec. 1999.

[6] R. Cáceres, N. Duffield, J. Horowitz, and D. Towsley. Multicast-based inference of network-internal loss characteristics. *IEEE Trans. Info. Theory*, 45(7):2462–2480, November 1999.

[7] R. Cáceres, N. Duffield, J. Horowitz, D. Towsley, and T. Bu. Multicast-based inference of network-internal characteristics: Accuracy of packet loss estimation. In *Proc. IEEE Infocom'99*, March 1999.

[8] R. Carter and M. Crovella. Measuring bottleneck link speed in packet-switched networks. Technical Report BU-CS-96-006, Computer Science Dept., Boston University, Mar. 1996.

[9] M. Coates and R. Nowak. Network loss inference using unicast end-to-end measurement. In *Proc. ITC Seminar on IP Traffic, Measurement and Modelling*, pages 28–1–28–9, Monterey, CA, Sep. 2000.

[10] M. Coates and R. Nowak. Networks for networks: Internet analysis using Bayesian graphical models. *IEEE Neural Network for Signal Processing Workshop*, Dec. 2000.

[11] N. Duffield and F. Lo Presti. Multicast inference of packet delay variance at interior network links. In *Proc. IEEE Infocom*, Mar. 2000.

[12] N.G. Duffield, F. Lo Presti, V. Paxson, and D. Towsley. Inferring link loss using striped unicast probes. *Proc. IEEE Infocom'01*, April 2001. Available as http://www.research.att.com/projects/minc/dlpt00.ps.

[13] K. Harfoush, A. Bestavros, and J. Byers. Robust identification of shared losses using end-to-end unicast probes, November 2000. Errata to this publication available as BUCS Technical Report 2001-001.

[14] J. Kurose and K. Ross. *Computer Networking: A top-down approach featuring the Internet*. Addison Wesley, 2001.

[15] V. Paxson. End-to-end Internet packet dynamics. *IEEE/ACM Trans. Networking*, 7(3):277–292, June 1999.

Modeling Transmission Loss in a Network with a Large Number of Nodes

Jayaram Sethuraman, Florida State University

Suppose that a signal with an initial strength from an originating node is transmitted through a network with a large number of intermediate nodes. There will be dissipation as well as some boosting of the signal between nodes. We will explore a general probabilistic model for the total loss in transmission, i.e. for the final strength of signal after passing though a large number of nodes.

To make the problem more mathematical, we assume that a signal has strength X_0 at the originating node 0 and it is transmitted through a path consisting of nodes i=1,2,...,n. Denote the strength of the signal at node i by X_i, i=1,2,...,n. The nodes themselves do not have to be on a straight line, they are the nodes along a certain path. The ratio $p_i=X_i/X_{(i-1)}$ represents the loss/boost factor at node i; p_i <1 means that there was a loss and p_i >1 means that there was a boost to the signal at node i. We are interested in strength X_n of the final signal after it comes through node n, or more particularly the final loss/boost factor $Z_n = X_n/X_0$. We present a probabilistic model for the loss/boost factors $p_1,...,p_n$ and obtain simple limiting distributions for the final loss/boost factor Z_n. In some models, the mean of the final loss/boost factor is 1 indicating that on the average there is no loss. In these cases, one can examine the variance, which we obtain, to devise systems with tolerable amounts of fluctuations while the same time there is no loss of strength on the average. In other probabilistic models, there will be a loss or boost in the strength of the final signal. This information can be useful in designing robust systems.

Monte Carlo Filters and Its Applications in Target Tracking and Wireless Communications [1]

Rong Chen
Department of Information and Decision Science,
University of Illinois at Chicago,
Chicago, IL 60607.
rongchen@uic.edu

1 Introduction

Stochastic systems are routinely used in science, engineering and economics. Many of these systems have a natural dynamic structure; others can often be built up dynamically. Except for a few special cases such as the linear Gaussian models or the discrete hidden Markov models, statistical analysis of these systems still present major challenges to researchers. The Monte Carlo filter (or sequential Monte Carlo) approach recently emerged in the fields of statistics and engineering shows a great promise in solving a large class of nonlinear filtering/prediction problems and general optimization problems, opening up new frontiers for cross-fertilization between statistical science and a wide spectrum of application areas such as telecommunications, bioinfomatics, and business data analysis.

Monte Carlo Filters (MCF) can be loosely defined as a family of methodologies that use Monte Carlo simulation to solve *on-line* estimation and prediction problems in dynamic systems. By recursively generating Monte Carlo samples of the state variables or some other latent variables, these methods can easily adapt to the dynamics of the underlying stochastic systems. Although the basic principle behind MCF dates back to the "growth Monte Carlo" method (Rosenbluth & Rosenbluth, 1955) known in molecular physics in the 50's, a complete theoretical framework for the MCF only appeared recently (Liu & Chen, 1998).

Efficient MCF has been designed for a number of problems including blind deconvolution, target tracking problems, and digital signal extraction in fading channels. It has also been shown that MCF complements MCMC and the two can be fruitfully combined (MacEachern et al. 1999). It has been used successfully in many problems such as energy minimization in molecular modeling and combinatorial optimization (Grassberger 1997; Wong & Liang 1997), speech recognition (Rabiner, 1989), target tracking problem (Gordon et al. 1993, Avitzour 1995), computer vision (Isard and Blake, 1996), economical time series (Pitt & Shephard, 1997); and DNA and protein sequence

[1] This work was supported in part by the U.S. National Science Foundation (NSF) under grants CCR-9980599 and DMS-0073601.

analysis (Churchill 1989; Krough et al. 1994; Liu et al. 1999), probabilistic expert systems (Spiegelhalter & Lauritzen 1990, Kong et al. 1994), simulating protein structures (Vasquez & Scherago 1985) and genetics (Irwing et al. 1994).

In section 2 we introduce the general stochastic dynamic systems, and one of its most commonly seen forms – the state space model. Several applications including target tracking and wireless communications will be introduced. In section 3 we detail the framework of MCF and its implementation issues. Section 4 contains an efficient variation of the MCF, the mixture Kalman filter. In section 5 we present several approaches to improve the efficiency of the MCF. Application examples are presented in Section 6.

2 Stochastic Dynamic Systems

A *stochastic dynamic system* (SDS) is defined as a sequence of evolving probability distributions, $\pi_t(\boldsymbol{x}_t)$, $t = 0, 1, 2, \ldots$, where the dimensionality of the state variable $\boldsymbol{x}_t$ often increases as the system evolves, i.e., $\boldsymbol{x}_t = (\boldsymbol{x}_{t-1}, x_t)$. In many scientific problems, it is of interest to evaluate expectations of a function of the state variable with respect to π_t at any time t.

Consider the *generalized state space model* of the form

$$\begin{array}{rlcl} \text{(state equation):} & x_t = \phi_t(\boldsymbol{x}_{t-1}, \varepsilon_t) & \text{or} & x_t \sim f_t(\cdot \mid \boldsymbol{x}_{t-1}) \\ \text{(observation equation):} & y_t = \gamma_t(\boldsymbol{x}_t, e_t) & \text{or} & y_t \sim g_t(\cdot \mid \boldsymbol{x}_t), \end{array} \qquad (1)$$

where $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ is the unobserved state variable and $\boldsymbol{y}_t = (y_1, \ldots, y_t)$ is the observed information available up to time t, and ε_t and e_t are noises which lead to the conditional densities f_t and g_t. When x_t is discrete and f_t is Markovian, the model is often termed the *hidden Markov Model* and is widely used in speech recognition and biological sequence analysis. When x_t is continuous, the model is widely used in engineering and time series analysis. This model can also take the form of a "dynamic Bayesian network" (Boyen & Koller, 1998). Important problems with many of these systems are (a) the on-line estimation of the "true state characteristics," e.g. $E[h(x_t) \mid \boldsymbol{y}_t]$ (b) the prediction of a future behavior, e.g. $E[h(x_{t+1}) \mid \boldsymbol{y}_t]$, and (c) the revision of previous states when given new information (smoothing), e.g. $E[h(x_{t-s}) \mid \boldsymbol{y}_t]$. Thus, because of the Bayes theorem, one is most interested in the computation of expectations with respect to the *a posteriori* SDS, $\pi_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t \mid \boldsymbol{y}_t)$, a task that often eludes all theoretical attempts.

The state space modeling has a long history and the list of its applications is endless. When the system is linear and Gaussian, the posterior distribution $p(x_t \mid \boldsymbol{y}_t)$ is Gaussian and can be obtained recursively through the Kalman filter; when x_t only takes on a few discrete values, the problems can also be solved by a forward-backward approach. Otherwise, either a suboptimal solution or a crude approximation to the Bayes estimator has to be employed. Previously proposed methods include the extended Kalman filer (e.g. Anderson & Moore 1979, Gelb 1974), Gaussian sum filters (Anderson & Moore 1979), the iterated extended Kalman Filter (Jazwinski 1970) and many others. Recently, researchers in statistics, engineering, and AI communities began to turn their attention to Monte Carlo-based filtering algorithms.

2.1 Target Tracking

Designing a sophisticated target tracking algorithm is an important task for both civilian and military surveillance systems, particularly when a radar, sonar, or optical sensor is operated in the present of clutter or when innovations are non-Gaussian (Bar-Shalom and Fortmann, 1988). We show three examples of target tracking using the MKF: (a) targets in the presence of random interference (clutter); (b) targets with non-Gaussian innovations; and (c) targets with maneuvering.

2.1.1 Random (Gaussian) accelerated target in clutter

Suppose the target follows a linear and Gaussian state space model:

$$\begin{cases} x_t &= Hx_{t-1} + Ww_t \\ y_t &= Gx_t + Vv_t \end{cases} \tag{2}$$

where x_t is the state variable (location and velocity) of the target and w_t, v_t are white noises with identity covariance matrix. For targets moving on a straight line, we have $x_t = (s_t, \nu_t)$ where s_t is the true target location and ν_t is its current velocity. In this case

$$H = \begin{pmatrix} 1 & T \\ 0 & T \end{pmatrix}; \quad W = \sigma_w^2 \begin{pmatrix} T/2 \\ 1 \end{pmatrix}; \quad G = (1, 0) \quad \text{and} \quad V = \sigma_v^2, \tag{3}$$

where T is the time duration between two observations and the random acceleration is assumed to be constant in the period, with rate $\sigma_w^2 w_t / T$. For targets moving in two (three) dimensional space, the state variable becomes $x_t = (\boldsymbol{s}_t, \boldsymbol{v}_t)$ with $\boldsymbol{s}_t$ and $\boldsymbol{v}_t$ being two (three) dimensional vectors. The corresponding matrixes can be expanded similarly.

In a clutter environment, we observe m_t signals $\{z_{t1}, \ldots, z_{tm_t}\}$ at time t, with

$$m_t \sim \text{Bernoulli}(p_d) + \text{Poisson}(\lambda\Delta),$$

where p_d is the probability of a true signal y_t being detected, λ is the rate of a Poisson random field, and Δ is the area of the surveillance region. In words, at time t the true signal has probability p_d to be detected, together with false signals, such as deceiving objects, electro-magnetic interferences, etc., which are distributed as a Poisson point process in the space. The problem is to track the real target on line in real time.

2.1.2 Random (Non-Gaussian) accelerated target in a clean environment

Consider again model (2), but with non-Gaussian errors w_t and v_t. Here we analyze the case when $w_t \sim t_{k_1}$ and $v_t \sim t_{k_2}$, but the same approach can be applied to other mixed-Gaussian settings. By defining $\Lambda_t = (\Lambda_{t1}, \Lambda_{t2})$, where $\Lambda_{ti} \sim \chi^2_{k_i}$ independently, we can rewrite model (2) as:

$$\begin{cases} x_t &= Hx_{t-1} + (\sqrt{k_1}/\sqrt{\lambda_1})We_t \\ y_t &= Gx_t + (\sqrt{k_2}/\sqrt{\lambda_2})V\varepsilon_t \end{cases} \quad \text{if} \quad (\Lambda_{t1}, \Lambda_{t2}) = (\lambda_1, \lambda_2)$$

with $e_t \sim N(0, I)$ and $\varepsilon_t \sim N(0, I)$. Again, it is a nonlinear/nonGaussian state space model.

2.1.3 Maneuvered target in a clean environment:

This situation can be modeled as follows:

$$\begin{aligned} x_t &= Hx_{t-1} + Fu_t + Ww_t \\ y_t &= Gx_t + Vv_t \end{aligned}$$

where u_t is the maneuvering acceleration. Here we consider an example of Bar-Shalom and Fortmann (1988) in which a two-dimensional target's position is sampled every $T = 10s$. The target moves in a plane with constant course and speed until $k = 40$ when it starts a slow 90^o turn which is completed in 20 sampling periods. A second, fast, 90^o turn starts at $k = 61$ and is completed in 5 sampling times. Figure 1 shows the trajectory of the target and its x-direction and y-direction velocity in one simulated run. Consequently, the matrices in this example are

$$H = \begin{pmatrix} 1 & 0 & 10 & 0 \\ 0 & 1 & 0 & 10 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}; G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}; F = \begin{pmatrix} 5 & 0 \\ 0 & 5 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}; W = \sigma_w^2 \begin{pmatrix} 5 & 0 \\ 0 & 5 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}; V = \sigma_v^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix};$$

The slow turn is the result of acceleration inputs $u_t^x = u_t^y = 0.075$ $(40 < t \leq 60)$, and the fast turn is from $u_t^x = -u_t^y = -0.3$ $(61 < t \leq 65)$. Other u_t's are zero (i.e. no maneuvering). We will analyze this example in later sessions.

2.2 Fading Channel in Wireless Communications

Many mobile communication channels can be modeled as Rayleigh flat-fading channels, which have the following form:

$$\text{State Equations:} \quad \begin{cases} \boldsymbol{x}_t &= F\boldsymbol{x}_{t-1} + Ww_t \\ \alpha_t &= G\boldsymbol{x}_t \\ s_t &\sim p(\cdot \mid s_{t-1}) \end{cases}$$

$$\text{Observation Equation:} \quad y_t = \alpha_t s_t + Vv_t$$

where s_t are the input digital signals (symbols), y_t are the received complex signals, and α_t are the unobserved (changing) fading coefficients. Both w_t and v_t are complex Gaussian with identity covariance matrices. Given the input signals s_t, the system is linear in $\boldsymbol{x}_t$ and y_t.

3 The General Form of Sequential Monte Carlo

One of the key components of MCF is importance sampling. Suppose a set of Monte Carlo samples, $\{\boldsymbol{x}^{(j)}, j = 1, \ldots, m\}$, has been generated from a trial distribution q. In order to compute the expectation, $\mu = E_\pi h(\boldsymbol{x})$, say, under distribution π, we can make an 'adjustment' by giving $\boldsymbol{x}^{(j)}$ a weight $w^{(j)} \propto \pi(\boldsymbol{x}^{(j)})/q(\boldsymbol{x}^{(j)})$. Then μ can be estimated by a weighted average of the $h(\boldsymbol{x}^{(j)})$. Since the weights are independent of function $h(\cdot)$, in a practical sense we can think of π as being approximated by a discrete distribution supported on the $\boldsymbol{x}^{(j)}$ with probabilities proportional to $w^{(j)}$. Based on this weighted-sample principle, Liu and Chen (1998), formulated the following MCF framework for applying Monte Carlo methods to a SDS.

Suppose $\{\pi_t(\boldsymbol{x}_t), t = 0, 1, \ldots\}$ is the SDS of interest. Let $S_t = \{\boldsymbol{x}_t^{(j)}, j = 1, \ldots, m\}$ be a set of Monte Carlo samples at time t with a corresponding set of weights $W_t = \{w_t^{(j)}, j = 1, \ldots, m\}$. We call $\{\boldsymbol{x}_t^{(j)}, w_t^{(j)}\}_{j=1}^m$ a properly weighted sample of π_t if $E_{\pi_t} h(\boldsymbol{x}_t)$, for any $h(\cdot)$, can be estimated by a weighted average of the $\boldsymbol{x}_t^{(j)}$ using the $w_t^{(j)}$. When the system evolves from stage t to $t+1$, we can produce a properly weighted sample of π_{t+1} by the following *Sequential Importance Sampling (SIS)* step:

(A) For each j, $j = 1, \ldots, m$, generate a $x_{t+1}^{(j)}$ (or multiple of them) from a trial distribution $q_{t+1}(x_{t+1} \mid \boldsymbol{x}_t^{(j)})$; attach it to $\boldsymbol{x}_t^{(j)}$ to form $\boldsymbol{x}_{t+1}^{(j)} = (\boldsymbol{x}_t^{(j)}, x_{t+1}^{(j)})$.

(B) Compute the "incremental weight"

$$u_{t+1}^{(j)} = \frac{\pi_{t+1}(\boldsymbol{x}_{t+1}^{(j)})}{\pi_t(\boldsymbol{x}_t^{(j)}) q_{t+1}(x_{t+1}^{(j)} \mid \boldsymbol{x}_t^{(j)})}; \quad \text{and let} \quad w_{t+1}^{(j)} = u_{t+1}^{(j)} w_t^{(j)}. \tag{4}$$

MCF is achieved by recursively applying the SIS step to a SDS. For a Markovian state space model, SIS only needs to keep a record of $x_t^{(j)}$, instead of the whole path $\boldsymbol{x}_t^{(j)}$, when proceeding from t to $t+1$. In a simplest form of MCF for this case, the "bootstrap filter" (or "particle filter") (Gorden et al. 1993; Kitagawa, 1996), one starts with equally weighted samples at time t and uses the state equation, $f_{t+1}(\cdot \mid x_t)$ as a trial distribution. The incremental weight is then proportional to the likelihood of the new observation, i.e., $g_{t+1}(\boldsymbol{y}_{t+1} \mid x_{t+1}^{(j)})$. Finally, one can resample so as to obtain a set of equally weighted sample for time $t+1$. In many applications, however, a more careful design of the MCF procedure can yield a much improved algorithm. The following issues are important for the design of MCF procedures.

(1). Trial sampling distribution: Choosing good trial distributions q_t is the critical first step in designing a good MCF scheme. Kong et al. (1994) and Liu & Chen (1995, 1998) suggested using $q_{t+1}(x_{t+1} \mid \boldsymbol{x}_t) = \pi_{t+1}(x_{t+1} \mid \boldsymbol{x}_t)$. In the state-space model, this is just the "local posterior distribution"

$$q_{t+1}(x_{t+1} \mid \boldsymbol{x}_t, y_{t+1}) = p(x_{t+1} \mid \boldsymbol{x}_t, y_{t+1}) \propto f_{t+1}(y_{t+1} \mid x_{t+1}) g_{t+1}(x_{t+1} \mid \boldsymbol{x}_t) \tag{5}$$

with incremental weight $u_{t+1} = \int f_{t+1}(y_{t+1} \mid \boldsymbol{x}_{t+1}) g_{t+1}(x_{t+1} \mid \boldsymbol{x}_t) dx_{t+1}$. This, if achievable, is apparently better than the one used by the bootstrap filter. More sophisticated choices of q_t will be discussed in later sections. When sampling from (5) is infeasible, methods such MCMC or Gaussian approximation can be used. See Liu & Chen (1998) for a summary.

(2). Resampling: Resampling is an indispensable component of MCF. Suppose $S_t = \{\boldsymbol{x}_t^{(j)}, j = 1, \ldots, m\}$ is properly weighted by $W_t = \{w_t^{(j)}, j = 1, \ldots m\}$ with respect to π_t. Instead of carrying the weight W_t as the system evolves, it is legitimate, and sometimes preferable (Liu & Chen 1995), to insert a resampling/reallocation step between SIS recursions in order to stabilize the weight distribution. The following scheme is typical: (i) draw (or systematically allocate) a new set of samples (denoted as S_t') from S_t with probability proportional to $w_t^{(j)}$; and (ii) assign equal weights to the samples in S_t'. Some theoretical and heuristic arguments of resampling are given in Liu & Chen (1995). Note that if the weights $w_t^{(j)}$ are nearly constant, resampling only reduces the number of distinctive samples and incurs extra Monte Carlo variation. However, when the weights become very skewed, carrying many samples with very small weights in an SIS setting is apparently wasteful. Resampling can provide chances for the good (i.e., "important") samples to amplify themselves and hence "rejuvenate" the sampler to produce better samples for the future states.

(3). Marginalization: When implementing Monte Carlo strategies, it is often a good practice to carry out as much analytical computation as possible (Hammersley & Hanscomb 1965). In importance sampling, it can be easily shown that the algorithm is more efficient after some components of the system are integrated out (marginalization). Liu et al. (1994) show that marginalization is also beneficial for Gibbs sampling. MacEachern et al. (1999) demonstrate that marginalization can greatly improve a MCF algorithm in a nonparametric Bayes problem.

4 The Mixture Kalman Filter

Many dynamic systems belong to the class of conditional dynamic linear models (CDLM) of the form

$$\boldsymbol{x}_t = \boldsymbol{F}_{\lambda_t} \boldsymbol{x}_{t-1} + \boldsymbol{G}_{\lambda_t} \boldsymbol{u}_t, \tag{6}$$

$$\boldsymbol{y}_t = \boldsymbol{H}_{\lambda_t} \boldsymbol{x}_t + \boldsymbol{K}_{\lambda_t} \boldsymbol{v}_t, \tag{7}$$

where $\boldsymbol{u}_t \sim N(0, \boldsymbol{I})$ and $\boldsymbol{v}_t \sim N(0, \boldsymbol{I})$ are the state and observation noise, respectively; and λ_t is a sequence of random indicator variables which may form a Markov chain, but are independent of $\boldsymbol{u}_t$ and $\boldsymbol{v}_t$ and the past $\boldsymbol{x}_s$ and $\boldsymbol{y}_s$, $s < t$. The matrices $\boldsymbol{F}_{\lambda_t}$, $\boldsymbol{G}_{\lambda_t}$, $\boldsymbol{H}_{\lambda_t}$ and $\boldsymbol{K}_{\lambda_t}$ are known given λ_t.

We observe that for a given trajectory of the indicator λ_t in a CDLM, the system is both linear and Gaussian, for which the Kalman filter provides a complete statistical characterization of the system dynamics. We proposed a novel sequential Monte Carlo method, the mixture Kalman filter (MKF) for on-line filtering and prediction of CDLM's; it exploits the conditional Gaussian property and utilizes a marginalization operation to improve the algorithmic efficiency. Instead of dealing with both $\boldsymbol{x}_t$ and λ_t, the MKF draws Monte Carlo samples only in the indicator space and uses a mixture of Gaussian distributions to approximate the target distribution. Compared with the generic sequential Monte Carlo method, the MKF is substantially more efficient (e.g., it produces more accurate results with the same computing resources).

Let $\boldsymbol{Y}_t = (\boldsymbol{y}_0, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_t)$ and $\boldsymbol{\Lambda}_t = (\lambda_0, \lambda_1, \cdots, \lambda_t)$. By recursively generating a set of properly weighted random samples $\left\{ \left(\boldsymbol{\Lambda}_t^{(j)}, w_t^{(j)} \right) \right\}_{j=1}^m$ to represent $p(\boldsymbol{\Lambda}_t | \boldsymbol{Y}_t)$, the MKF approximates the target distribution $p(\boldsymbol{x}_t | \boldsymbol{Y}_t)$ by a random mixture of Gaussian distributions

$$\sum_{j=1}^{m} w_t^{(j)} \mathcal{N}_c \left(\boldsymbol{\mu}_t^{(j)}, \boldsymbol{\Sigma}_t^{(j)} \right),$$

where $\boldsymbol{\mu}_t^{(j)} = \boldsymbol{\mu}_t \left(\boldsymbol{\Lambda}_t^{(j)} \right)$ and $\boldsymbol{\Sigma}_t^{(j)} = \boldsymbol{\Sigma}_t \left(\boldsymbol{\Lambda}_t^{(j)} \right)$ are obtained with a Kalman filter on the system (6)-(7) for the given indicator trajectory $\boldsymbol{\Lambda}_t^{(j)}$. Denote $\kappa_t^{(j)} \triangleq \left[\boldsymbol{\mu}_t^{(j)}, \, \boldsymbol{\Sigma}_t^{(j)} \right]$. Thus, a key step in the MKF is the production at time t of the weighted samples of indicators, $\left\{ \left(\boldsymbol{\Lambda}_t^{(j)}, \kappa_t^{(j)}, w_t^{(j)} \right) \right\}_{j=1}^m$, based on the set of samples, $\left\{ \left(\boldsymbol{\Lambda}_{t-1}^{(j)}, \kappa_{t-1}^{(j)}, w_{t-1}^{(j)} \right) \right\}_{j=1}^m$, at the previous time $(t-1)$. For details, see Chen & Liu (2000).

5 Delayed-sampling Method

Dynamic systems often possess strong short term or long term "memory,", i.e., future observations can reveal substantial information on the current state. In the case of sudden noise spike and system malfunction, future observations become critical in combating temporary loss of information or mis-information. It is a good practice to have a "waiting" period, i.e., a buffer, in which information accumulates before being used for generating new Monte Carlo samples. However, a MCF scheme usually does not go back to "regenerate" past samples in view of new information (it is possible to design strategies for a MCF scheme to revise previous draws), although the past estimations can be adjusted by using the new importance weights. Hence, if the samples are generated by using false information (i.e. from a sampling distribution that is far away from the truth), MCF can quickly loose track of the state variable. A natural method to overcome this difficulty is the *delayed-sample method*, which uses future information in generating sample of the current state.

To achieve this end for the state space model (1), we can define the target SDS as $\{\pi_t(\boldsymbol{x}_t) = p(\boldsymbol{x}_t \mid \boldsymbol{y}_{t+\Delta}), t = 1, 2, \ldots\}$ for some $\Delta \geq 0$. Suppose at time $t + \Delta$ we have a set of properly weighted samples $\{(\boldsymbol{x}_t^{(j)}, w_t^{(j)})\}_{j=1}^m$ of the new SDS. Then at time $t + \Delta + 1$, we can use $q_{t+1}(x_{t+1}^{(j)} \mid \boldsymbol{x}_t^{(j)}, \boldsymbol{y}_{t+\Delta+1}) = p(x_{t+1}^{(j)} \mid \boldsymbol{x}_t^{(j)}, \boldsymbol{y}_{t+\Delta+1})$, to generate Monte Carlo samples of x_{t+1}. This involves computing

$$p(x_{t+1} \mid \boldsymbol{x}_t, \boldsymbol{y}_{t+\Delta+1}) \propto \int p(\boldsymbol{x}_{t+\Delta+1}, \boldsymbol{y}_{t+\Delta+1}) dx_{t+2} \cdots dx_{t+\Delta+1}. \tag{8}$$

The weight can be computed as $w_{t+1}^{(j)} \propto w_{t-1} p(y_{t+\Delta+1}, \boldsymbol{y}_{t+\Delta} \mid \boldsymbol{x}_t^{(j)}) / p(\boldsymbol{y}_{t+\Delta} \mid \boldsymbol{x}_t^{(j)})$. A main difficulty with this approach is that one needs to evaluate the multidimensional integrals such as (8). Here we discuss two approaches: the straightforward exact evaluation approach and a pilot evaluation approach.

(1). Exact Sampling and Evaluation: When the state variable takes values in a discrete set $\mathcal{A} = \{a_1 \ldots, a_J\}$ (e.g., when the state variable represents the transmitted signal in digital communications), the aforementioned integration can be done by exhausting all possible combinations of the $x_{t+1}, \ldots, x_{t+\Delta+1}$ string, which involves growing/trimming a tree with $J^{\Delta+1}$ branches. Specifically, the sampling distribution is

$$p(x_{t+1} \mid \boldsymbol{x}_t^{(j)}, \boldsymbol{y}_{t+\Delta+1}) \propto \sum_{x_{t+1}^{t+\Delta+1} \in J^{\Delta+1}} \prod_{d=1}^{\Delta+1} p(y_{t+d} \mid \boldsymbol{y}_{t+d-1}, x_{t+1}^{t+d}, \boldsymbol{x}_t^{(j)})$$

and the weight is

$$w_{t+1} = w_t \frac{\sum_{x_{t+1}^{t+\Delta+1} \in J^{\Delta+1}} \prod_{d=1}^{\Delta+1} p(y_{t+d} \mid \boldsymbol{y}_{t+d-1}, x_{t+1}^{t+d}, \boldsymbol{x}_t^{(j)})}{\sum_{x_{t+1}^{t+\Delta} \in J^{\Delta}} \prod_{d=0}^{\Delta} p(y_{t+d} \mid \boldsymbol{y}_{t+d-1}, x_{t+1}^{t+d}, \boldsymbol{x}_t^{(j)})},$$

where $x_{t+1}^{t+d} = (x_{t+1}, \ldots, x_{t+d})$. This approach has the difficulty that the complexity increases exponentially with J and Δ.

(2). Pilot-delayed Sampling and Evaluation:

When the cardinality J of $\mathcal{A}$ is large or when the state variable is continuous, the previous approach is infeasible. However, (8) can be approximated by using a relative small number of pilot streams. For example, in the discrete case we can send out J groups of pilots, each with k members. Every individual in the j-th group starts with the same value a_j of x_{t+1}, and propagates Δ SIS steps to time $t + \Delta + 1$. The sum of the weights of every pilot in j-th group, called "pilot weight for a_j", is given to value a_j. Finally we draw $x_{t+1} = a_j$ with probability proportional to the pilot weight. To correct the bias introduced by pilot approximation, we need to multiply the usual incremental weight (4) by the inverse of the pilot weight. When the state variable consists of an "important" but simple (say, binary) component and an "unimportant" one, the pilot streams can be generated to cover all possible combinations of the "important" components but only cover a random portion of the "unimportant" part. When the state variable x_t is continuous, one can generate for each $\boldsymbol{x}_t^{(j)}$ a pilot group of x_{t+1} from $g_{t+1}(\cdot|\boldsymbol{x}_t^{(j)})$. Each individual in the group is then processed with SIS recursions to $t+\Delta+1$ and obtains a pilot weight. The weight of each individual member can then be used to resample a x_{t+1} from the starting pilot group. The final weight can be adjusted similarly.

6 Some Numerical Examples

In this section we provide some numerical results for the examples given in section 2.

6.1 Target tracking

6.1.1 Random (Gaussian) accelerated target in clutter

By letting Λ_t be the identifier of the target, Liu and Chen (1998) formulated the problem stated in section 2.1 into a CDLM. More precisely, they let $\Lambda_t = 0$ if the target is not observed, and $\Lambda_t = i$ if the i-th observed object is the signal generated from the true target, i.e., $y_t = z_{ti}$. Then the system is linear and Gaussian with given Λ_t, and the remaining z signals bear no information. Some of their results are shown in Figure 2 (a) and (b), which reveal the tracking errors (the differences between the estimated and true target locations) of 50 simulated runs of the tracking model, with $r^2 = 1.0, q^2 = 1.0, p_d = 0.9$ and $\lambda = 0.1$. Five hundred Monte Carlo samples were used for both the MKF and a standard Monte Carlo filter (i.e. an SIS with resampling applied to the state variable x_t). Here we also tested the split-track filter (Figure 2 (c)), which, at each step, kept 500 trajectories with the highest likelihood values (recursively). The MKF performed much better than the other two algorithms in this problem.

6.1.2 Random (Non-Gaussian) accelerated target in a clean environment

Simulations were carried out with the matrices (3) with $T = 1$ and no interference and $w_t \sim t_3$ and $v_t \sim t_3$. The following table shows a comparison of the MKF and a standard Monte Carlo filter in terms of the number of times the target was lost ($|x_t - \hat{x}_t| > 1200$) and the cpu time for one hundred simulated runs.

noise variance	MC size (m)	MC Filter		MKF	
		cpu time	# miss	cpu time	# miss
	20	9.49843	72	19.4277	1
	50	20.1622	20	51.6061	1
σ_w^2= 16.00	200	80.3340	7	181.751	1
σ_v^2= 1600	500	273.369	4	500.157	1
	1500	1063.36	3	2184.67	1

Figure 3 shows the tracking mean squared error, after the lost tracks are eliminated. We observe that although it takes about twice as much CPU time as the standard Monte Carlo filter with the same m, the MKF performs much more efficiently in the same CPU time.

We also tested the idea of using a finite mixture of Gaussian distributions to approximate the t distribution, i.e. approximating t_3 with $\sum_{i=1}^{k} p_i N(0, \sigma_i^2)$. Similar results were obtained. The advantage of this approach is that a more efficient MKF can be used for discrete indicators. But on the other hand, the approximation causes some biases.

6.1.3 Maneuvered target in a clean environment:

To apply the MKF to this application, we need to specify prior structure of u_t. First, we assume that maneuvering can be classified into several categories, indicated by an indicator. In particular, we assume a three level model, $I_t = 0$ indicates no maneuvering ($u_t = 0$), and $I_t = 1$ and 2 indicate slow and fast maneuvering, respectively, ($u_t \sim N(0, \sigma_i^2)$, $\sigma_1^2 < \sigma_2^2$). In this study we used $\sigma_1^2 = 1$ and $\sigma_2^2 = 36$. We also specify transition probabilities $P(I_t = j \mid I_{t-1} = i) = p_{ij}$ for the maneuvering status. Specifically, we assume $p_{ii} = 0.8$ and $p_{ij} = 0.1$ for $i \neq j$ (i.e. it is more likely to stay in a particular maneuvering state than to change the maneuvering state). Second, there are different ways of modeling the serial correlation of the u_t. Here we assume a multi-level white noise model, as in Bar-Shalom and Fortmann (1988), where the u_t are assumed independent, given the indicator. This is the easiest but not a very realistic model. Other possible models are currently under investigation.

In Figure 4 we present the root mean square errors of the MKF estimates of the target position for 50 simulated runs. Comparing our result with that of Bar-Shalom and Fortmann (1988, pp 143) who used the traditional detection-and-switching method, we see a clear advantage of the proposed MKF.

6.2 Digital Signal Extraction in Fading Channels

In this section, we provide some computer simulation examples to demonstrate the performance of the MCF in fading channels. The fading process is modeled by the output of a Butterworth filter of order $r = 3$ driven by a complex white Gaussian noise process. The cutoff frequency of this filter is 0.05, corresponding to a normalized Doppler frequency (with respect to the symbol rate $\frac{1}{T}$) $f_d T = 0.05$, which is a fast fading scenario. Specifically, the fading coefficients $\{\alpha_t\}$ is modeled by the following ARMA(3,3) process:

$$\begin{aligned} & \alpha_t - 2.37409\alpha_{t-1} + 1.92936\alpha_{t-2} - 0.53208\alpha_{t-3} \\ = \; & 10^{-2}(0.89409u_t + 2.68227u_{t-1} + 2.68227u_{t-2} + 0.89409u_{t-3}), \end{aligned} \tag{9}$$

where $u_t \sim \mathcal{N}_c(0, 1)$. The filter coefficients in (9) are chosen such that $\mathrm{Var}\{\alpha_t\} = 1$. It is assumed that BPSK modulation is employed, i.e., the transmitted symbols $s_t \in \{+1, -1\}$.

In order to demonstrate the high performance of the proposed adaptive receiver, in the following simulation examples we compare the performance (in terms of bit error rate) of the proposed sequential Monte Carlo receivers with that of the following three receiver schemes:

- *Known channel lower bound:* In this case, we assume that the fading coefficients $\{\alpha_t\}$ are known to the receiver.

- *Genie-aided lower bound:* In this case, we assume that a genie provides the receiver with an observation of the modulation-free channel coefficient corrupted by additive noise with the same variance, i.e., $\tilde{y}_t = \alpha_t + \tilde{n}_t$, where $\tilde{n}_t \sim \mathcal{N}_c(0, \sigma^2)$. The receiver then uses a Kalman filter to track the fading process based on the information provided by the genie; The transmitted symbols are then demodulated. It is clear that such a genie-aided bound is lower bounded by the known channel bound. It should also be noted that the genie is used only for calculating the lower bound. Our proposed algorithms estimate the channel and the symbols simultaneously with no help from the genie.

- *Differential detector:* In this case, no attempt is made to estimate the fading channel. Instead the receiver detects the phase difference in two consecutively transmitted bits by using the simple rule of differential detection: $\widehat{b_t b}_{t-1} = \mathrm{sign}\left(\Re\left\{y_t^* y_{t-1}\right\}\right)$.

The differential encoding and decoding are employed to resolve the phase ambiguity. The adaptive receiver implements the MKF algorithm described in Section 4. The number of Monte Carlo samples drawn at each time was empirically set as $m = 50$. Simulation results showed that the performance did not improve much when m was increased to 100, while it degraded notably when m was reduced to 20. The resampling procedure discussed in Section 3 was employed to maintain the efficiency of the algorithm, in which the effective sample size threshold is $\bar{m}_t = m/10$. The *delayed-weight* method discussed in Section 5 was used to extract further information from future received signals, which resulted in an improved performance compared with concurrent estimation. In each simulation, the sequential Monte Carlo algorithm was run on 10000 symbols, (i.e., $t = 1, \cdots, 10000$). In counting the symbol detection errors, the first 50 symbols were discarded to allow the algorithm to reach the steady state. In Figure 5, the bit error rate (BER) performance versus the signal-to-noise ratio (defined as $\text{Var}\{\alpha_t\}/\text{Var}\{n_t\}$) corresponding to delay values $\delta = 0$ (concurrent estimate), $\delta = 1$, and $\delta = 2$ is plotted. In the same figure, we also plot the known channel lower bound, the genie-aided lower bound, and the BER curve of the differential detector. From this figure it is seen that, with only a small amount of delay the performance of the MKF can be significantly improved by the delayed-weight method compared with the concurrent estimate. Even with the concurrent estimate, the MKF does not exhibit an error floor, as does the differential detector. Moreover, with a delay δ=2, the MKF essentially achieves the genie-aided lower bound. We have also implemented the delayed-sample method for this case and found that it offers little improvement over the delayed-weight method.

REFERENCES

Ackerson, G.A. and Fu, K.S. (1970) On state estimation in switching environments. *IEEE Trans. Autom. Control*, **AC-15**, 10-17.

Akashi, H. and Kumamoto, H. (1977) Random sampling approach to state estimation in switching environments. *Automatica*, **13**, 429-434.

Anderson, B.D.O. and Moore, J.B. (1979). *Optimal Filtering.* New Jersey: Prentice-Hall.

Avitzour, D. (1995) A stochastic simulation Bayesian approach to multitarget tracking. *IEE Proc. Radar, Sonar and Navigation*, **142**, 41-44.

Bar-Shalom, Y. and Fortmann, T.E. (1988) *Tracking and Data Association*, Academic Press: Boston

Berzuini, C., Best, N.G., Gilks, W.R., and Larizza, C. (1997) Dynamic conditional independence models and Markov chain Monte Carlo methods. *J. Amer. Statist. Assoc.*, **92**, 1403-1412.

Boyen, X. and Koller, D. (1998). Tractable inference for complex stochastic processes. *Technical Report, Computer Sci. Dept., Stanford Univ.*

Carlin, B.P, Polson, N. G. and Stoffer, D. S. (1992) A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Amer. Statist. Assoc.*, **87** 493-500.

Carpenter, J., Clifford, P. and Fearnhead, P. (1997) An improved particle filter for non-linear problems. *Technical Report*, Oxford University.

Carter, C.K. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541-553.

Chen, R. and Liu, J.S. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society, Series B*, **62**, 493-508

Chen, R., Wang, X. and Liu, J.S. (2000). Adaptive joint detection and decoding in flat-fading channels via mixture Kalman filtering. *IEEE trans. Information Theory*, **46**, 2079-2094

Churchill, G.A. (1989), Stochastic Models for Heterogeneous DNA Sequences, *Bulletin of Mathematical Biology* **51**, 79-94.

Doucet, A., Godsill, S.J. and Andrieu, C. (1999) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, in press.

Gelb, A. (1974) *Applied Optimal Estimation*. Boston: MIT press

Gordon, N.J., Salmon, D.J. and Smith, A.F.M. (1993) A novel approach to nonlinear/non Gaussian Bayesian state estimation. *IEE Proc. Radar Signal Process.*, **140**, 107-113.

Grassberger (1997). Pruned-enriched rosenbluth method: simulations of q polymers of chain length up to 1000000.. *Physical Review E*, **56**, 3682–3693.

Hammersley, J.M. and Handscomb, D.C. (1965). *Monte Carlo Methods*. Chapman & Hall.

Hürzeler, M. and Künsch, H.R. (1998) Monte Carlo approximations for general state-space models. *J. Comp. Graph. Statist.*, **7**, 175-193.

Irwing, M., Cox, N. and Kong, A. (1994). Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Science, USA*, **91**, 11684-11688.

Isard, M. and Blake, A. (1996). Contour tracking by stochastic propagation of conditional density. In *Computer Vision - ECCV' 96*, B. Buxton and R. Cipolla (eds), Springer: New York.

Jazwinski, A. (1970) *Stochastic Processess and Filtering Theory*. New York: Academic Press.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. *J. Basic Engineering*, **82**, 35-45

Kitagawa, G. (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comp. Graph. Statist.*, **5**, 1-25.

Kong, A., Liu, J.S., and Wong, W.H. (1994) Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.*, **89**, 278-288.

Krogh A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994). Hidden Markov models in computational biology - Applications to protein modeling. *J. Mol. Biol.*, **235**, 1501-1531.

Liu, J.S. and Chen, R. (1995) Blind deconvolution via sequential imputations. *J. Amer. Statist. Assoc.*, **90**, 567-576.

— (1998) Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, **93**, 1032-1044

Liu, J.S., Chen, R., and Wong, W.H. (1998) Rejection control for sequential importance sampling. *J. Amer. Statist. Assoc.*, **93**, 1022-1031.

Liu, J.S., Wong, W.H., and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27-40.

MacEachern, S.N., Clyde, M.A., and Liu, J.S. (1999) Sequential importance sampling for non-parametric Bayes models: the next generation. *Canadian J. Statist.*, **27**, 251-267.

Pitt, M.K. and Shephard, N. (1999) Filtering via simulation: auxiliary particle filters. *J. Amer. Statist. Assoc.*, **94**, 590-599.

Rabiner, L.R. (1989), A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,*Proceedings of the IEEE*, **77**, 257-286.

Rosenbluth, M.N. and Rosenbluth, A.W. (1955). Monte Carlo calculation of the average extension of molecular chains. *J. Chem. Phys.*, **23**, 356-359.

Rubin, D.B.(1987) A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. *J. Amer. Statist. Assoc.*, **52**, 543-546.

Rubinstein, R.Y. (1981) *Simulation and the Monte Carlo Method*. New York: Wiley.

Shephard, N. (1994) Partial non-Gaussian state space. *Biometrika*, **81**, 115-131.

Spiegelhalter, D.J. and Lauritzen, S.L. (1990), Sequential Updating of Conditional Probabilities on Directed Graphical Structures, *Network*, **20**, 579-605.

Smith, M.C. and Winter, E.M. (1978) On the detection of target trajectories in a multi-target environment. *Proc. 17th (1978) IEEE Conf. on Decision & Control.*, San Diego, CA.

Svetnik, V.B. (1986) Applying the Monte Carlo method for optimum estimation in systems with random structure. *Auto. Remo. Cont.*, **47**, 818-827.

Tanizaki, H. (1996) *Nonlinear Filters: Estimation and Applications.* New York: Springer.

Tugnait, J.K. (1982) Detection and estimation for abruptly changing systems. *Automatica*, **18**, 607-615.

West (1992) Mixture models, Monte Carlo, Bayesian updating and dynamic models. *Computer Science and Statistics*, **24**, 325-333.

West, M. and Harrison, J. (1989) *Bayesian Forecasting and Dynamic Models.* New York: Springer.

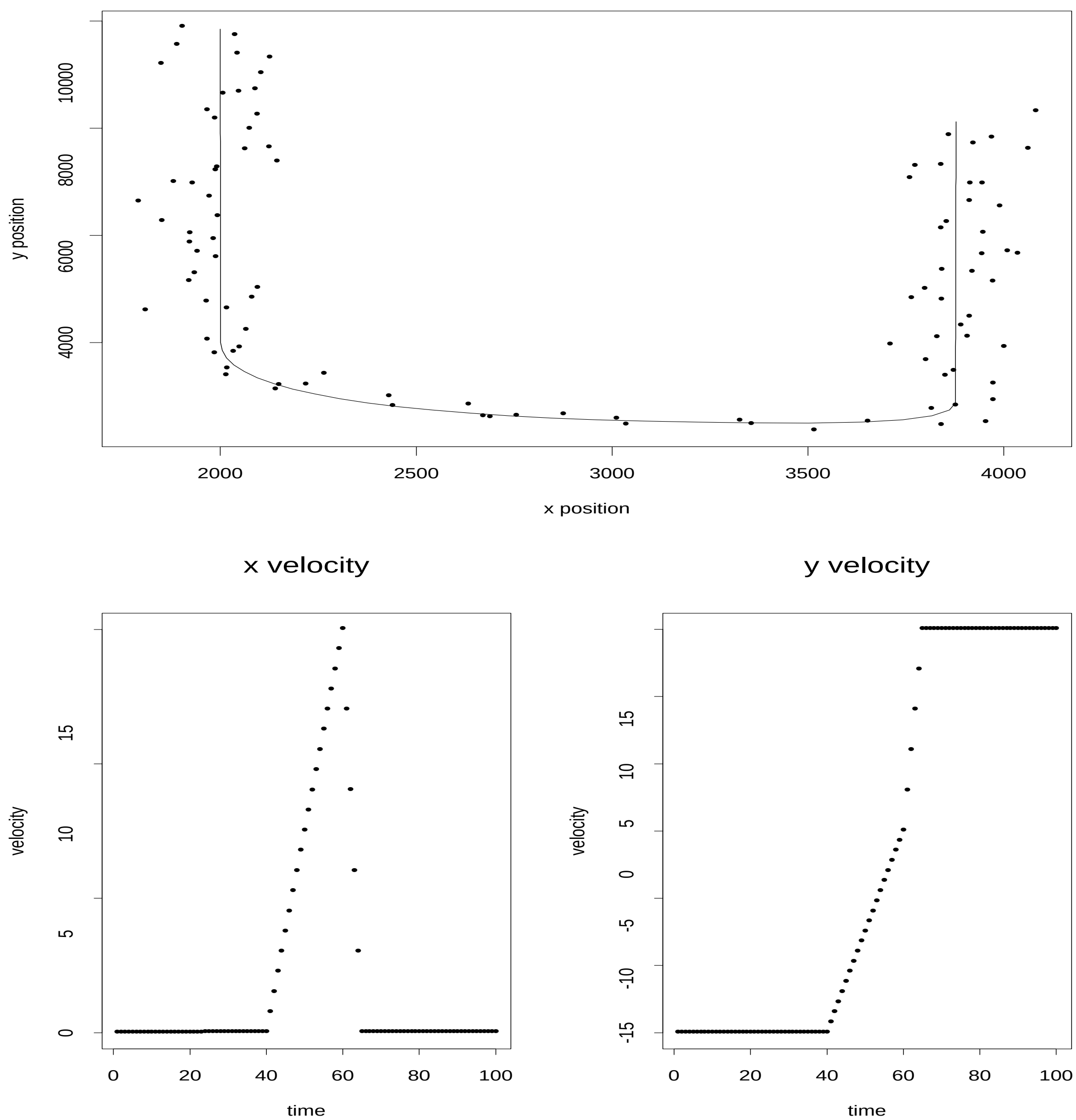


Figure 1: The position and velocity of a simulated 2 dimensional maneuvering target. (Top) Position. (Bottom) Velocity.

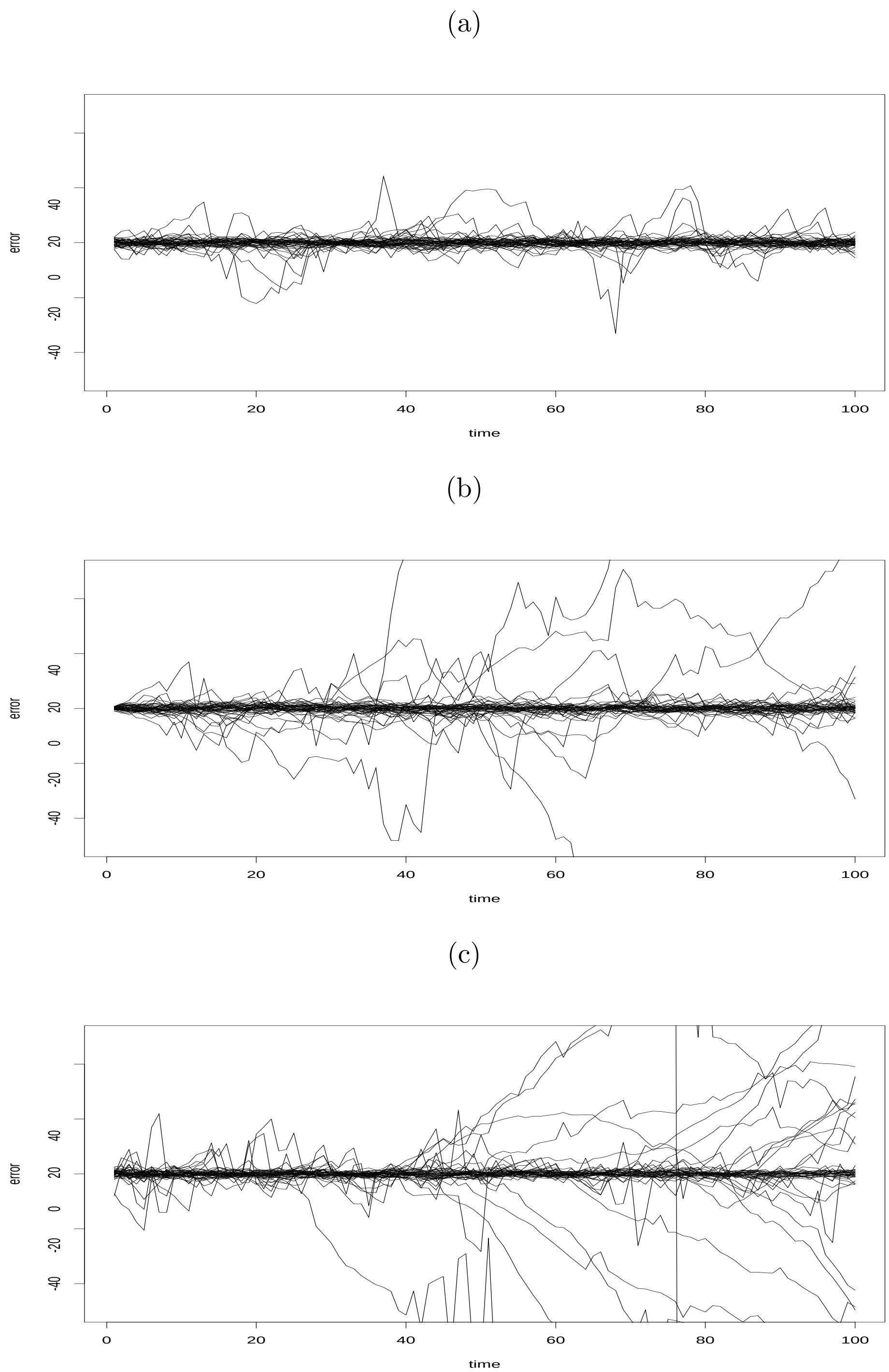


Figure 2: The tracking errors of 50 runs of the MKF (a), a standard Monte Carlo filter (b), and the split-track filter (c) for a simulated one-dimensional target moving system.

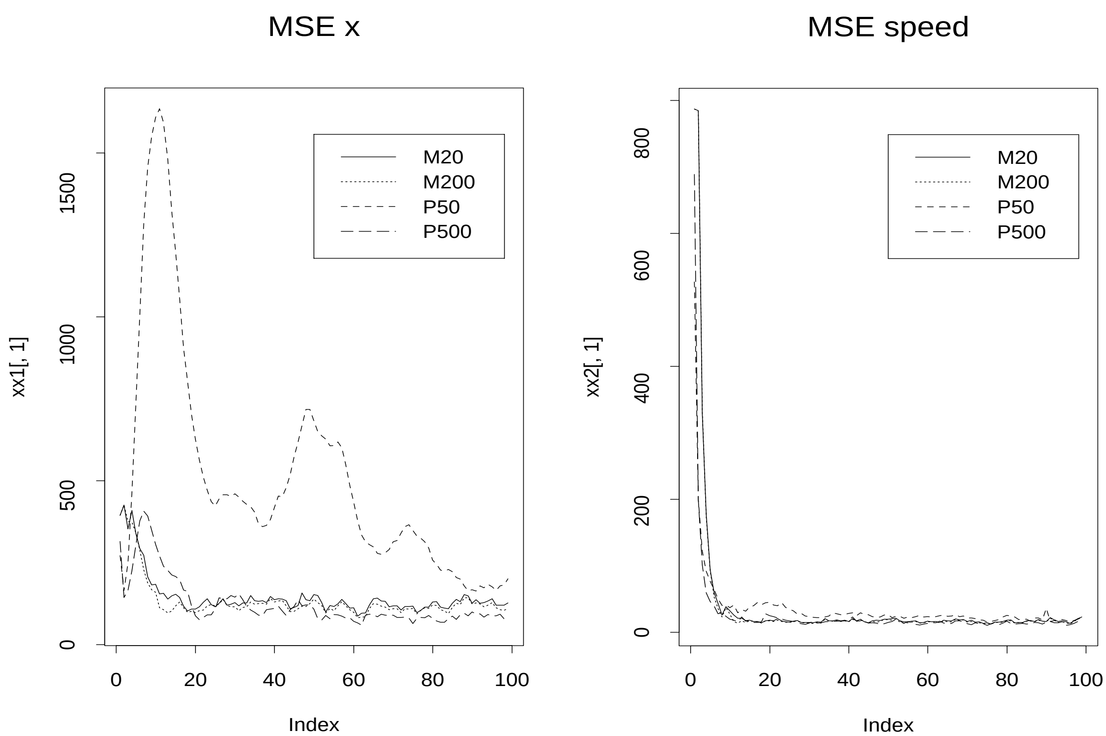


Figure 3: The MSE's of location and speed of 50 runs of the MKF and a standard MC filter for a simulated one-dimensional target moving system using different Monte Carlo sample sizes. 'MKF20' and 'MKF200' are the MSE's of the MKF with Monte Carlo sample size 20 and 200 respectively. 'MC50' and 'MC500' are the MSE's of a standard MC filter with Monte Carlo sample size 50 and 500 respectively.

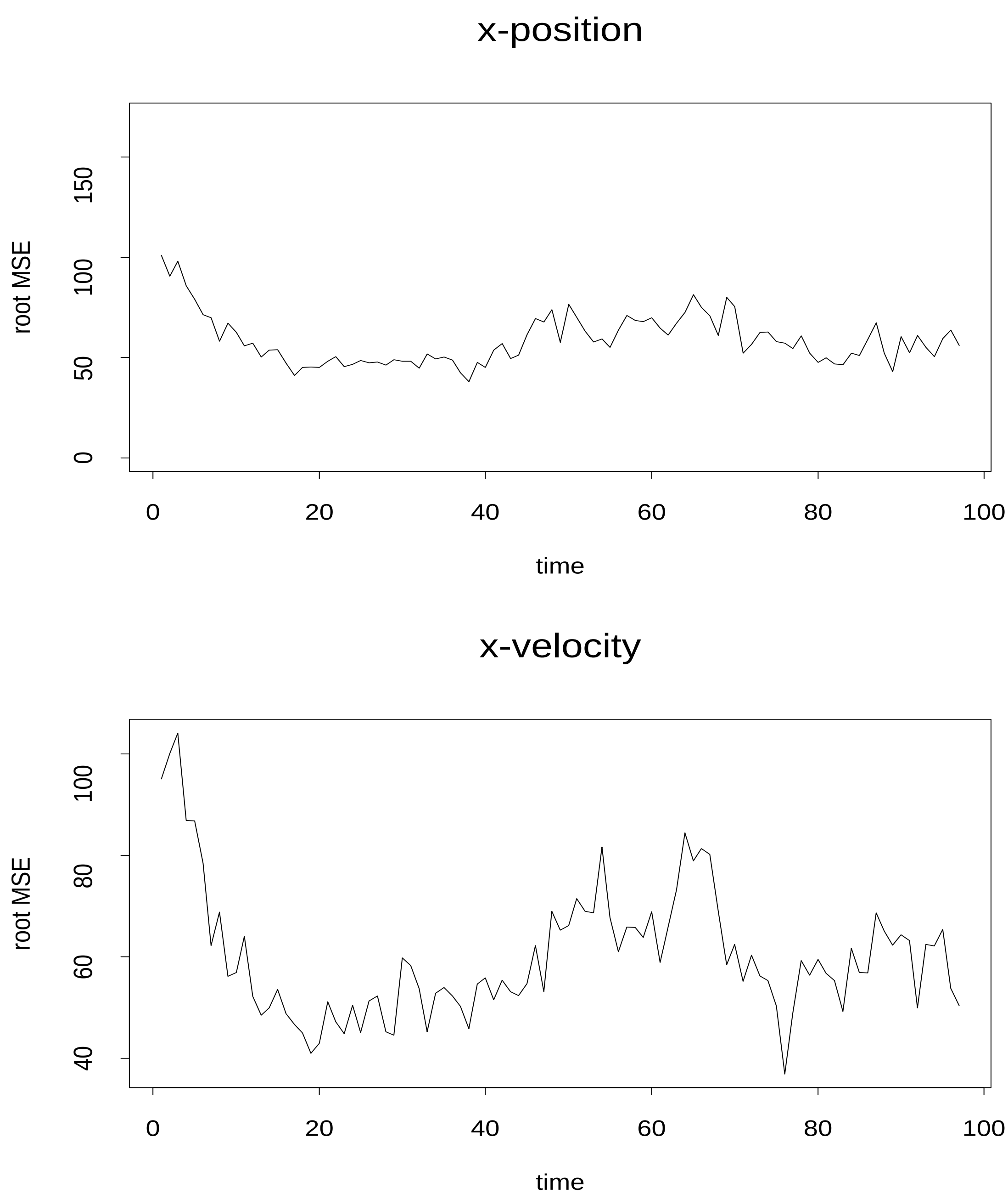


Figure 4: The root MSE's of the x-positiion and x-direction velocity of 50 runs of the MKF for a simulated two-dimensional target moving system with maneuvering.

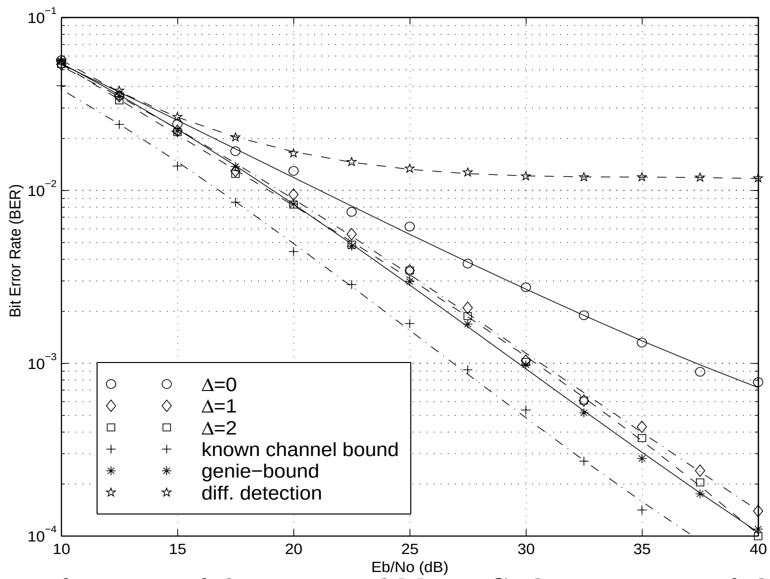


Figure 5: BER performance of the sequential Monte Carlo receiver in a fading channel with Gaussian noise and without coding. The delayed-weight method is used. The BER curves corresponding to delays $\delta = 0$, $\delta = 1$ and $\delta = 2$ are shown. Also shown in the same figure are the BER curves for the known channel lower bound, the genie-aided lower bound and the differential detector.

CONTRIBUTED SESSION III (1530 - 1700)

Clustering and Partial Mixture Estimation
David Scott, Rice University

The use of density estimation to find clusters in data is supplementing ad hoc hierarchical methodology. Examples include finding high-density regions, finding modes in a kernel density estimator, and the mode tree. Alternatively, a mixture model may be fit and the mixture components associated with individual clusters. Fitting a high-dimensional mixture model with many components is difficult to estimate in practice. Here, we describe a new algorithm that estimates a subset of the complete model. In particular, we demonstrate how to fit one component at a time and how the fits may be organized to reveal the complete clustering model.

Estimating Parameters in a Bimodal Distribution
Douglas Frank, Indiana University, Pennsylvania

The problem is we have data from a mixture of two populations with unknown means. The source of each datum cannot be identified. We assume the fraction of data from one population is an unknown parameter p. We show methods of estimating the parameter p as well as the means and variances of the mixed populations. The problem is phrased in terms of fitting bimodal test scores but has several possible military applications. For instance we may be receiving fire from two enemy weapon systems with differing rates of fire or kill ratios. We can estimate the number of each type of weapon as well as its capabilities with this procedure.

DEPARTMENT OF STATISTICS
University of Wisconsin
1210 W. Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1036

March 2001

ACCURATE LOWER TOLERANCE LIMITS FOR THE NORMAL RANDOM EFFECTS MODEL

by

Bernard Harris
Department of Statistics
University of Wisconsin-Madison

and

Shun-Yi Chen
Tamkang University
Tamsui, Taipei
Taiwan, ROC 251

Supported in part by the Army Research Office under Grant Number DAAD19-99-1-0185.

1 Introduction and Summary

In this paper we obtain lower tolerance limits for the balanced normal random effects model. Government certification and qualification requirements as well as contract specifications are frequently stated in terms of lower tolerance limits. Thus, the efficient determination of such tolerance limits is important in technological application of statistics.

This study was motivated by the need for qualification of polymer composite for aircraft construction. The fabricator prepares material in rolls, each of which are several hundred feet in length (when unrolled). The various rolls correspond to the batches in the prior discussion.

Thus let

$$X_{ij} = \mu + b_i + e_{ij}, \quad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J \tag{1}$$

where X_{ij} is the jth observation from the ith batch. The b_i's and e_{ij}'s are mutually independent normally distributed random variables with $E(b_i) = E(e_{ij}) = 0$ and variances σ_b^2 and σ_w^2 respectively.

Let

$$\hat{\mu} = \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij}/IJ, \tag{2}$$

$$\bar{X}_{i\cdot} = \sum_{j=1}^{J} X_{ij}/J, \tag{3}$$

$$SS_B = \sum_{i=1}^{I} J(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2, \tag{4}$$

$$SS_w = \sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{i\cdot})^2. \tag{5}$$

Further, define

$$s_B^2 = SS_B/(I-1), \tag{6}$$

and

$$s_w^2 = SS_w/I(J-1). \tag{7}$$

Let P and γ, $0 < P, \gamma < 1$ be given. Then $\hat{u}(P,\gamma)$ is a $100\gamma\%$ lower tolerance level for the $100(1-P)^{th}$ percentile, u_p, of the random variable X if

$$P\{\hat{u}(P,\gamma) \leq u_p\} \geq \gamma. \tag{8}$$

Clearly, if X is a normally distributed random variable with mean μ and variance σ^2, (8) may be written

$$P_{\bar{X},s}\{\bar{X} - ks \leq \mu - K_p\sigma\} \geq \gamma; \tag{9}$$

where $K_p = 100P^{th}$ percentile of the standard normal distribution, $k = k(n,\gamma,P)$ is called the tolerance limit factor; $\bar{X}$ and s are the sample mean and sample standard deviation of a normally distributed random sample with mean μ and variance σ^2.

For the normal random effects model (1),

$$E(X_{ij}) = \mu \tag{10}$$

$$\sigma_{X_{ij}}^2 = \sigma_x^2 = \sigma_b^2 + \sigma_w^2 \tag{11}$$

$$E(s_B^2) = \sigma_B^2 = J\sigma_b^2 + \sigma_w^2 \tag{12}$$

$$E(s_w^2) = \sigma_w^2. \tag{13}$$

Procedures for obtaining lower tolerance limits for normal random effects models have previously been given by Lemon [1977] and Mee and Owen [1983].

Lemon proposed the tolerance limit

$$\hat{\mu} - k_L s_x, \tag{14}$$

where

$$s_x^2 = s_B^2/J + (1-1/J)s_w^2 \tag{15}$$

and

$$k_L = t_{I-1}(\delta, \gamma) s_B / \sqrt{IJ} s_x; \tag{16}$$

$t_{I-1}(\delta, \gamma)$ is the $100\gamma^{th}$ percentile of a noncentral t distribution with $I-1$ degrees of freedom and noncentrality parameter δ, and

$$\delta = \sqrt{IJ} K_p \frac{\sigma_x}{\sigma_B} = \sqrt{IJ} K_p [(R+1)/(JR+1)]^{1/2}, \tag{17}$$

where $R = \sigma_b^2/\sigma_w^2$. Since the tolerance limit factor k_L depends on R, Lemon proposed replacing R by $\hat{R} = s_b^2/s_w^2$ in (17), where $s_b^2 = (s_B^2 - s_w^2)/J$. Lemon's procedure gives extremely conservative tolerance limits since his estimator of variability is based on s_B^2 which has $I-1$ degrees of freedom.

Mee and Owen [1983] proposed the tolerance limit

$$\hat{\mu} - \frac{t_f(\delta, \gamma)\ s_x}{\sqrt{IJR^*}}, \tag{18}$$

where $t_f(\delta, \gamma)$ is the $100\gamma^{th}$ percentile of the noncentral t distribution with

$$f = \frac{(R^*+1)^2}{(R^*+1/J)^2/(I-1) + (1-1/J)/IJ} \tag{19}$$

degrees of freedom and noncentrality parameter $\delta = \sqrt{IJ} K_p \sigma_x / \sigma_B$,

$$R^* = \max(0, (F_\eta Z - 1)/J); \tag{20}$$

(19) is obtained from the Satterthwaite approximation and F_η is the $100\gamma^{th}$ percentile of the F distribution with $I(J-1)$ and $I-1$ degrees of freedom and $Z = s_B^2/s_w^2$.

The tolerance limit given by (18) is also conservative (but less conservative than (14)), Since R^* is basically an upper confidence limit for R, R is intentionally overestimated. Also η is set to .85, which is approximately correct when $R \to \infty$ but is much too high for low values of R.

This problem and related problems have also been considered by others. Specifically, Bhaumik and Kulkarni [1991] discussed obtaining one-sided tolerance limits for the unbalanced one-way ANOVA random effects model. If the variance ratio R is known, they claimed improvements over the Mee-Owen procedure for large values of R. Vangel [1992] treated one-sided tolerance limits for one-way balanced random effects models using methods based on techniques developed by Welch [1947] and subsequently Trickett and Welch [1954] and Aspin [1948] for dealing with the Behrens-Fisher problem. Bagui, Bhaumik and Parnes [1996] derived tolerance limits for unbalanced m-way random effects ANOVA models based on estimation of the variance components. Also Beckman and Tietjen [1989] generated two-sided tolerance limits for the balanced random effects ANOVA model. They provided extensive tables of tolerance limit factors.

The purpose of the present method is to obtain tolerance limits with actual coverage virtually equal to the nominal coverage γ.

2 A Tolerance Limit for the Normal Random Effects Model

It is customary to describe the balanced normal random effects model by means of the ANOVA table given below.

Table 1. One-Way Analysis of Variance

Source	df	SS	MS	EMS
Between	$I-1$	SS_B	s_B^2	$J\sigma_b^2+\sigma_w^2$
Within	$I(J-1)$	SS_w	s_w^2	σ_w^2
Total	$IJ-1$	SS_t	s_i^2	

Let α, β, γ be arbitrary positive constants and let U and V be independent chi-square random variables with m and n degrees of freedom respectively.

Let $Z = \gamma U/V$ and $W = \alpha U + \beta V$. Also $Z = \frac{\gamma m}{n} F_{m,n}$, where $F_{m,n}$ is a random variable with the F-distribution with m and n degrees of freedom. It is easy to see that the conditional probability density function of W given $Z = z$ is

$$f_{W|Z}(w) = \frac{\left(\frac{\gamma+z}{\alpha z+\beta\gamma}\right)^{(m+n)/2} w^{(m+n)/2-1}}{\Gamma(\frac{m+n}{2})\ 2^{(m+n)/2}} \ \exp\left[\frac{-(\gamma+z)w}{2(\alpha z+\beta\gamma)}\right]. \tag{21}$$

Let

$$T = \frac{(\gamma+Z)W}{\alpha Z+\beta\gamma}. \tag{22}$$

Then T and Z are independent and T is a chi-square random variable with $m+n$ degrees of freedom.

If we now set $U = (I-1)s_B^2/\sigma_B^2$ and $V = I(J-1)s_w^2/\sigma_w^2$, then

$$T = \left[(I-1)\frac{s_B^2}{s_w^2} + I(J-1)\frac{\sigma_B^2}{\sigma_w^2}\right]\frac{s_w^2}{\sigma_B^2} \tag{23}$$

has the chi-square distribution with $(I-1)+I(J-1) = IJ-1$ degrees of freedom.

Let u_p denote the $100(1-P)^{th}$ percentile of the normal population X with mean μ and variance σ_x^2, K_p the $100P^{th}$ percentile of the standard normal distribution, and let $\hat{\mu} - k\sqrt{T}$ be a $100\gamma\%$ tolerance limit for u_p; then $u_p = \mu - K_p\sigma_x$ and

$$P_{\hat{\mu},T}\{\hat{\mu} - k\sqrt{T} \le \mu - K_p\sigma_x\} \ge \gamma. \tag{24}$$

Define $Y = \sqrt{IJ}(\hat{\mu}-\mu)/\sigma_B$ and

$$\delta = \sqrt{IJ}K_p\sigma_x/\sigma_B, \tag{25}$$

then k must satisfy

$$P_{Y,T}\left\{\frac{Y+\delta}{\sqrt{T/(IJ-1)}} \le \sqrt{IJ(IJ-1)}k/\sigma_B\right\} \ge \gamma. \tag{26}$$

Since Y has the standard normal distribution and T has the chi-square distribution with $IJ-1$ degrees of freedom, the random variable $(Y+\delta)/\sqrt{T/(IJ-1)}$ has the noncentral t distribution with $IJ-1$ degrees of freedom and noncentrality parameter δ. Hence, we have

$$k = \frac{t_{IJ-1}(\delta,\gamma)\sigma_B}{\sqrt{IJ(IJ-1)}}, \tag{27}$$

where $t_{IJ-1}(\delta,\gamma)$ denotes the $100\gamma^{th}$ percentile of the noncentral t distribution with $IJ-1$ degrees of freedom and noncentrality parameter δ. Let τ denote the ratio of the expected mean squares (EMS), that is

$$\tau = \frac{\sigma_B^2}{\sigma_w^2} = JR + 1. \tag{28}$$

Set

$$\hat{\mu} - k\sqrt{T} = \hat{\mu} - k_T \cdot s_w, \tag{29}$$

where

$$k_T = \frac{t_{IJ-1}(\delta,\gamma)}{\sqrt{IJ(IJ-1)}}\sqrt{(I-1)\frac{s_B^2}{s_w^2} + I(J-1)\tau}. \tag{30}$$

Therefore, the tolerance limit can be written as

$$\hat{\mu} - \frac{t_{IJ-1}(\delta,\gamma)}{\sqrt{IJ(IJ-1)}}\sqrt{(I-1)\frac{s_B^2}{s_w^2} + I(J-1)\tau} \cdot s_w. \tag{31}$$

Referring to the noncentrality parameter δ in (25), let

$$\frac{\sigma_x}{\sigma_B} = \left[\frac{1}{J}\left(1 + \frac{J-1}{\tau}\right)\right]^{1/2}. \tag{32}$$

The noncentrality parameter δ and the tolerance limit factor k_T depend on the parameter τ, which is unknown. Consequently it is natural to replace the parameter τ by a suitable estimator. Let $\hat{\tau} = F_\eta Z$, where Z denotes the mean square ratio s_B^2/s_w^2, and F_η the $100\eta^{th}$ percentile of an F-distribution with degrees of freedom $\nu_1 = I(J-1)$ and $\nu_2 = I-1$. We next study the relationship between η and (I, J, R).

Substituting the estimator $\hat{\tau}$ into (30), we set

$$\begin{aligned} \hat{k}_T(\eta) &= \frac{t_{IJ-1}(\hat{\delta}, \gamma)}{\sqrt{IJ(IJ-1)}}\sqrt{(I-1)\frac{s_B^2}{s_w^2} + I(J-1)F_\eta Z} \\ &= \frac{t_{IJ-1}(\hat{\delta}, \gamma)}{\sqrt{IJ(IJ-1)}}\sqrt{(I-1) + I(J-1)F_\eta} \cdot \frac{s_B}{s_w} \end{aligned} \quad (33)$$

The properties of the coverage rate,

$$\gamma(\eta) = P_{\hat{\mu},T,\hat{\tau}}\{\hat{\mu} - \hat{k}_T(\eta)s_w \le u_p\}, \quad (34)$$

was studied by simulation for various values of I, J, η. Specifically, the coverage rate was generated for 10,000 simulations for $P = .90, \gamma = .95; \quad I = 4, 8;$ $J = 3, 5, 7, 9, 13$; $\eta = .50, .60, .70, .85$. For the sake of brevity, we provide only a summary of the conclusions.

$\gamma(\eta)$ is less than γ for $\eta = .50, .60, .70$, except for small values of R. $\gamma(\eta)$ is larger than γ for $\eta = .85$, confirming the conservative property of the Mee-Owen procedure. For fixed I and J, $\gamma(\eta)$ decreases as R increases and increases with I for fixed J and R.

A second set of simulations was carried out to ascertain the behavior of η as a function of I, J and R. For $P = .90$, $\gamma = .95$, 10,000 replications were calculated for $I = 4$ and various values of R form 0.00 to 20.00 and $J = 3, 5, 7, 9$. It was noted that η exhibits very regular behavior with a maximum (for the selected values) of .84. Thus regularity will be subsequently exploited.

We now describe the determination of η. Let I and J be fixed. If $R = 0$, $\tau = 1$ and the noncentrality parameter $\delta = \sqrt{IJ}K_p$. Hence the tolerance limit

(31) is given by

$$\hat{\mu} - t_{IJ-1}(\delta, \gamma) s_t / \sqrt{IJ}, \tag{35}$$

where s_t^2 is defined in the ANOVA table (Table 1). That is, if $R = 0$, there is no variation between batches and the model reduces to a normal random sample of size IJ. Then (35) is the tolerance limit obtained by Owen [1963].

However, we have assumed that R is unknown and will still require an estimate of τ. Since Z has the F-distribution with $I-1$ and $I(J-1)$ degrees of freedom, when $R = 0$, we determine η by

$$\eta = P\{Z \leq 1\}. \tag{36}$$

Next, we determine η for large values of R ($R \longrightarrow \infty$). Writing (25) as

$$\delta = \sqrt{IJ} K_p \sqrt{\frac{1}{J}\left(1 + \frac{J-1}{\tau}\right)}, \tag{37}$$

Since $R \longrightarrow \infty$ implies $\tau \longrightarrow \infty$, $\delta \longrightarrow \delta^* = \sqrt{I} K_p$. Then (31) becomes

$$\begin{aligned} &\hat{\mu} - \frac{t_{IJ-1}(\delta^*, \gamma)}{\sqrt{IJ(IJ-1)}} \sqrt{(I-1)\frac{s_B^2}{s_w^2} + I(J-1) F_\eta Z} \cdot s_w \\ &= \hat{\mu} - \frac{t_{IJ-1}(\delta^*, \gamma)}{\sqrt{IJ(IJ-1)}} \sqrt{(I-1) + I(J-1) F_\eta} \cdot s_B \quad . \end{aligned} \tag{38}$$

We can now solve for F_η using (24) and (38), determining

$$P_{Y, s_B}\left\{ \frac{Y + \delta}{s_B/\sigma_B} \leq \frac{t_{IJ-1}(\delta^*, \gamma)}{\sqrt{IJ-1}} \sqrt{(I-1) + I(J-1) F_\eta} \right\} \geq \gamma, \tag{39}$$

where $Y = \sqrt{IJ}(\hat{\mu} - \mu)/\sigma_B$, $\delta = \sqrt{IJ} K_p \sigma_x / \sigma_B$. Y has the standard normal distribution and $(I-1)s_B^2/\sigma_B^2$ has the chi-square distribution with $I-1$ degrees of freedom. Then $(Y+\delta)/(s_B/\sigma_B)$ has the noncentral t distribution with $I-1$ degrees of freedom and noncentrality parameter δ.

Writing

$$\int_{-\infty}^{c} f_{I-1}(t; \delta) dt = \gamma \tag{40}$$

where $f_{I-1}(t;\delta)$ is the probability density function of the noncentral t distribution with $I-1$ degrees of freedom and noncentrality parameter δ. Then F_η can be determined by solving (40) for c.

Consequently we write

$$t_{I-1}(\delta,\gamma) = \frac{t_{IJ-1}(\delta^*,\gamma)}{\sqrt{IJ-1}}\sqrt{(I-1)+I(J-1)F_\eta}, \tag{41}$$

where $t_d(\delta,\gamma)$ denotes the $100\gamma^{th}$ percentile for a noncentral t distribution with d degrees of freedom and noncentrality parameter δ. As $R \longrightarrow \infty$, $\sigma_x/\sigma_B \longrightarrow 1/\sqrt{J}$, we have $\delta \longrightarrow \delta^*$. Therefore

$$F_\eta \longrightarrow F_\eta^* = \frac{IJ-1}{I(J-1)}\left(\frac{t_{I-1}(\delta^*,\gamma)}{t_{IJ-1}(\delta^*,\gamma)}\right)^2 - \frac{I-1}{I(J-1)}. \tag{42}$$

Regarding F_η as the $100\eta^{th}$ percentile of the F distribution with $I(J-1)$ and $(I-1)$ degrees of freedom, η is ready available. We now proceed to determine η for all R, $0 \le R < \infty$. For various I, J values and $P = 0.9, \gamma = 0.95$, η values were simulated for $R = 0, .1, 1, 5, 40$. $R \longrightarrow \infty$ and $R = 0$ were determined analytically. The relationship,

$$\eta = \eta_\infty - (\eta_\infty - \eta_0)\exp(-\theta R^\zeta), \tag{43}$$

where η_∞ and η_0 are the η values when $R = \infty$ and $R = 0$ respectively. (43) satisfies the plots of the simulations with negligible error. The parameters θ and ζ can be determined by a curve fitting algorithm. Some η values for $P = .90, \gamma = .95$ are listed in Table 2. Simulation study shows that linear interpolation in η would give satisfactory results for values of I, J and R being not tabulated.

We now proceed to a verification of the efficiency of this procedure and a comparison with the Mee-Owen [1983] method.

Table 2. η values for $P = .90, \gamma = .95$

R	J	I: 2	3	4	5	6	8	15
0	2	.58	.54	.52	.52	.51	.51	.50
	4	.64	.60	.57	.56	.55	.54	.53
	6	.66	.61	.59	.57	.57	.55	.54
	8	.67	.62	.59	.58	.57	.56	.54
	16	.67	.62	.50	.59	.58	.57	.55
	32	.68	.63	.60	.59	.58	.57	.55
.1	2	.63	.63	.62	.61	.60	.59	.55
	4	.63	.63	.63	.61	.60	.59	.56
	6	.67	.63	.63	.63	.61	.60	.57
	8	.68	.65	.63	.63	.62	.61	.61
	16	.74	.68	.67	.66	.64	.64	.63
	32	.78	.71	.70	.71	.66	.66	.63
1	2	.75	.75	.74	.72	.70	.71	.64
	4	.80	.78	.76	.74	.74	.72	.71
	6	.83	.79	.79	.77	.76	.74	.72
	8	.84	.80	.79	.77	.76	.75	.73
	16	.85	.82	.80	.78	.76	.76	.73
	32	.86	.82	.80	.79	.78	.78	.73
5	2	.84	.81	.79	.79	.76	.76	.70
	4	.87	.84	.82	.80	.78	.76	.76
	6	.89	.84	.83	.81	.80	.78	.76
	8	.89	.84	.83	.81	.81	.79	.77
	16	.89	.85	.84	.82	.81	.80	.77
	32	.89	.85	.85	.83	.82	.81	.77
∞	2	.86	.83	.81	.80	.78	.77	.74
	4	.88	.86	.83	.82	.81	.79	.77
	6	.89	.86	.84	.83	.82	.80	.78
	8	.89	.87	.85	.83	.82	.81	.79
	16	.90	.87	.85	.84	.83	.82	.79
	32	.90	.87	.86	.84	.83	.82	.79

Table 3. Simulated Coverage for $P = .90,\ \gamma = .95$
(10,000 observations)

	Mee-Owen		Present	
η	max	min	max	min
0	.990	.965	.952	.948
.1	.988	.961	.952	.948
1	.978	.948	.952	.948
5	.965	.946	.952	.948
40	.965	.948	.954	.948

3 Comparison of Tolerance Limits

Extensive numerical computations were made for the purpose of evaluating the present procedure and providing comparisons of the coverage and the actual tolerance limit with the Mee-Owen [1983] technique. Table 3 and 4 exhibit the results.

The Mee-Owen procedure as noted in the preceding section is most conservative for low values of R, particularly for small values of I and J. The coverage of the present procedure is particularly uniform across the range of values where the sampling error of the simulation is taken into account. Finally, their tolerance limits are also substantially higher for the present procedure. Some sample values are given below.

Note that the Mee-Owen procedure is less conservative for large values of R. The standard errors of the tolerance limits have also been calculated. The present procedure has been shown to have consistently smaller standard errors over most of the range of values and about the same standard error as the Mee-Owen procedure for large R.

Table 4. Tolerance limits for $P = .90, \gamma = .95$

R	I	J	Mee-Owen	Present
0	2	2	−14.11	−4.57
	2	8	−8.33	−2.33
	8	2	−2.14	−2.01
	8	8	−1.67	−1.60
.1	2	2	−15.14	−5.16
	2	8	−11.30	−3.16
	8	2	−2.27	−2.15
	8	8	−1.84	−1.76
1	2	2	−21.38	−10.62
	2	8	−21.40	−11.34
	8	2	−3.26	−3.17
	8	8	−3.00	−2.88
5	2	2	−38.05	−30.43
	2	8	−39.04	−34.47
	8	2	−5.96	−5.93
	8	8	−5.83	−5.77
40	2	2	−97.57	−99.75
	2	8	−99.85	−105.79
	8	2	−15.93	−15.95
	8	8	−16.02	−15.82

4. Numerical Illustration.

To provide a numerical comparison with the results of Lemon (1977) and Mee and Owen (1983), we will apply the present procedure to data analyzed in the above papers. Lemon's data consisted of five batches with six observations per batch. The data and specifications for the tolerance limit are summarized as follows.

$$P = .9,\ \gamma = .95,\ \hat{\mu} = 186,\ s_b = 6.87,\ s_w = 5.86,\ s_x = 9.04,$$

hence $\hat{R} = 1.37$.

Lemon obtained a lower tolerance limit of 156.3. The procedure of Mee and Owen resulted in a lower tolerance limit of 160.0, which is less conservative than Lemon's procedure. The present methodology utilizes the non-central t distribution with non-centrality parameter $\delta = 3.56$ and 29 degrees of freedom, resulting in a lower toleraance limit of 165.0, which is less conservative than the result of Mee and Owen..

REFERENCES

Aspin, A.A. (1948) "An Examination and Further Development of a Formula Arising in the Problem of Comparing Two Mean Values," *Biometrika*, 35, 88-96.

Bagui, S.C., Bhaumik, D.K. and Parnes, M. (1996) "Two-Sided Tolerance Limits for Unbalanced m-Way Random-Effects ANOVA Models," *Journal of Applied Statistical Science*, 3, 135-148.

Beckman, R.J., and Tietjen, G.L. (1989) "Two-Sided Tolerance Limits for Balanced Random-Effects ANOVA Models," *Technometrics,* 31, 185-197.

Bhaumik, D.K. and Kulkarni, P.M. (1991) "One-Sided Tolerance Limits for Unbalanced One-Way ANOVA Random-Effects Model," *Communications in Statistics: Theory and Methods*, 20, 1665-1675.

Lemon, G. H. (1977). "Factors for One-Sided Tolerance Limites for Balanced One-Way-ANOVA Random-Effects Model," *Journal of the American Statistical Association,* 72, 676-680.

Mee, R.W., and Owen,D. B. (1983). "Improved Factors for One-Sided Tolerance Limits for Balanced One-Way-ANOVA Random Model," *Journal of the American Statistical Association,* 78, 901-905.

Trickett, W.H. and Welch, B.L. (1954) "On the Comparison of Two Means: Further Discussion of Iterative Methods for Calculating Tables," *Biometrika*, 41, 361-374.

Vangel, M.G. (1992) "New Methods for One-Sided Tolerance Limits for a One-Way Balanced Random-Effects ANOVA Model," *Technometrics,* 34, 176-185.

Welch, B.L. (1947) "The Generalization of Student's Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

WEDNESDAY, OCTOBER 18

CONTRIBUTED SESSION IV (1530 - 1700)

Statistical Augmentation of a Database for Use in Optical Character Recognition Software Evaluation

Ann E. M. Brodeen, Frederick S. Brundick
U.S. Army Research Laboratory

Malcolm S. Taylor
OAO Corporation

Abstract

In this paper, we consider a statistical approach to augment a limited database of groundtruth documents for use in evaluating optical character recognition (OCR) software. We require groundtruth documents to assign a performance measure to the OCR component of the Forward Area Language Converter (FALCon) system. A modified moving-blocks bootstrap procedure is used to construct surrogate documents for this purpose which prove to serve effectively, and in some regards, indistinguishably, from groundtruth. The proposed method is validated through a rigorous statistical procedure.

Introduction

The Forward Area Language Converter (FALCon) is a portable, field-operated, translation system designed to assist in intelligence collection. It enables an operator with no foreign language training to convert a foreign language document into an approximate English translation for an assessment of military relevance. The principal components of FALCon are an optical scanner, an optical character recognition (OCR) module, and a machine translation (MT) module. In order to assign a performance measure to the FALCon system, measures of effectiveness of the components must be developed and then aggregated into an overall measure. The focus of this paper is limited to evaluation of the OCR module.

A current procedure for determining a quantitative measure of the efficacy of an OCR product is as follows: A selection of carefully prepared source-language documents, called groundtruth, is stored in the computer; hardcopy of the same document set is then scanned into bitmap images; the OCR software partitions a gross bitmap image into homogeneous zones that are processed according to content. For zones that are identified as text, specialized scoring software then compares the OCR output against the corresponding groundtruth to produce accuracy statistics, usually including percentage agreement for both words and characters, and a confusion matrix.*

A central database of groundtruth documents, accepted as a baseline, would enable the evaluation of OCR products to proceed from a common benchmark.

*A confusion matrix displays the number of character insertions, substitutions, and deletions required to reconcile the groundtruth and OCR output files.

Unfortunately, such a database does not exist, making the comparison of OCR software more difficult and any conclusions drawn more tentative. Fundamental questions regarding sample size requirements, and suitable document composition for such a database, remain to be addressed.

Collection of a corpus that is sufficient for evaluation of an OCR product is likely to remain, even in the best of circumstances, a burdensome task. Access to a sufficient number of source-language documents, representative of the document classes of interest, may not be feasible; and, even if obtained, the expensive and time-consuming process of preparing groundtruth remains. To address this problem, we are proposing a statistical approach to corpus generation based on a small set of source-language documents. Coincident with the statistical inquiry, substantial work involving language transliteration must be accomplished.

Time Series Model

Consider the passage of Serbian text shown in Figure 1. Every character—letters, punctuation marks, interword spaces—is represented numerically in the computer. The set of character and numeric equivalents (the mapping) is called a codeset. For a specific language, the codeset representation may not be unique. Russian, for example, has four commonly used 8-bit encodings and some Asian languages even more [1]. A representation of the Serbian text in Figure 1 for a particular codeset assignment is shown in Figure 2.

In Figure 2, the first 80 letters (emboldened in Figure 1) of the Serbian text are portrayed. The vertical dashed lines mark the location of interword spaces, which have been removed, along with most punctuation, to facilitate our methodology. The x-axis indexes the order of occurrence of the characters in the text, and the corresponding codeset values (numeric equivalents suppressed for presentation purposes) are plotted along the y-axis. If the characters are processed sequentially, then we can assign to each character an associated time epoch, and Figure 2 can be considered as a time series representation of the first 80 letters. The scale of measurement for the y-axis is nominal; an alternative codeset, if appropriate, would lead to a different graphic representation with no attendant loss or gain of information.

In attempting to generate a corpus, we would like a core of authentic documents to serve as a basis from which to generate additional pseudodocuments. An analogous situation, arising in the analysis of time series data collected as part of a clinical study, has been described and addressed using the bootstrap [2, 3].

Bootstrap Application

In this section, we present an abridged description of the bootstrap procedure, modified for application to the textual model. Notice the time series has an inherent structure: the time series represents a block of text—it is not a random sequence. Moreover, the words themselves are subject to lexical constraints; hence, the patterns they assume in the codeset representation have meaning. These word patterns are, however, interrupted with great frequency; the interword spaces play the role of interventions in time series modeling. As a consequence, the time series has local structure contributed by the word patterns but little in the way of global structure due to the high frequency of interventions.

"**Нецес остати пусто Невесиње равно,но цес бити оно сто си вазда било:расадник Српства и колевка л**ава!" "Србија мора постати велико радилисте и родилисте!" Ово су само два изватка из скорасњих говоранција Вука Драсковица. Анахроницна, срцепарајуца реторика, примерена политицарима осамнаестог века, представља данас најбољи пример лази-говора или језика-маске. А онај ко на себе стави маску лази-говора, пре или касније, изабраце и лаз као основни политицки принцип. Нико није изрекао толико лази, подвала и лазних доказа о Косову као г. Драсковиц и његова телевизија. Раније смо ту анахроницну реторску маску примали као некакав његов особењацки избор, као сто примамо нецији цудацки стил у одевању. Требало је за његову реторику реци оно сто је одувек и била - да је обицан киц.

Figure 1. Serbian Text

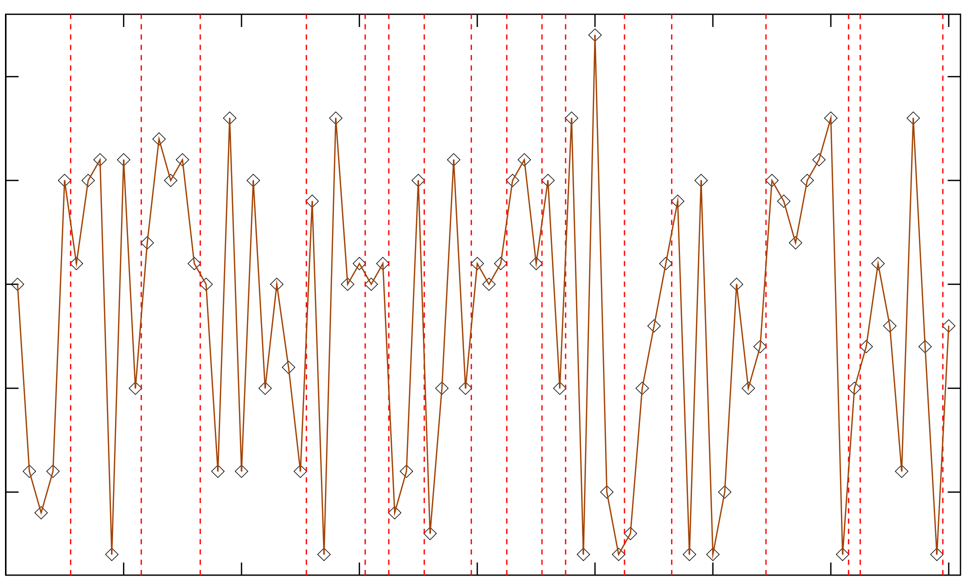


Figure 2. Time Series Representation

Denoting the time series as a sequence of ordered pairs $(x_1, y_1), (x_2, y_2), \ldots,$ (x_n, y_n), we begin the bootstrap procedure by choosing a random location within the time series, say (x_r, y_r). Starting with (x_r, y_r), we copy the subsequence $(x_r, y_r), (x_{r+1}, y_{r+1}), \ldots, (x_{r'}, y_{r'})$ into an array. The length of the subsequence, $r' - r + 1$, is determined by sampling from the distribution of word-lengths found in the authentic document. A second random location, (x_s, y_s), is then determined, and a second subsequence, $(x_s, y_s), (x_{s+1}, y_{s+1}), \ldots, (x_{s'}, y_{s'})$, is copied and appended to the subsequence already in the array. Figure 3 illustrates a situation in which three subsequences have been chosen, two of them overlapping.[†] The overlap does not create a problem since the sampling is done with replacement. This process continues until terminated by a stopping rule. At that point, a bootstrapped time series, the first 80 values of which are shown in Figure 4, has been produced. The shaded regions appearing in Figure 3 are aligned in Figure 4 in order of their occurrence. Inverting the codeset mapping, subject to inherent lexical modeling constraints, yields the bootstrap document shown in Figure 5.

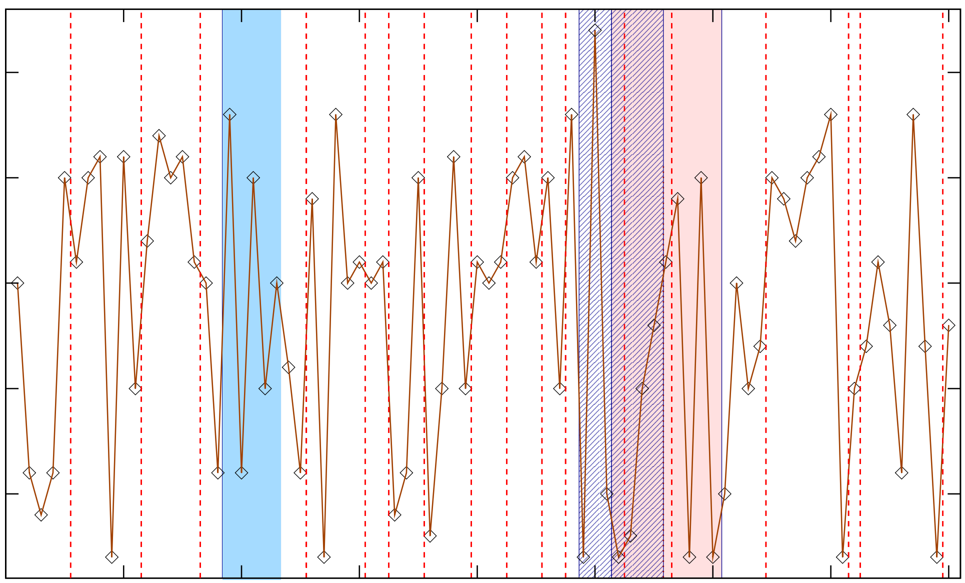


Figure 3. Intermediate Results

Empirical Results

The bootstrap procedure under which the document in Figure 5 was constructed[‡] precludes its being "read" by an individual. Our intent, however, was to produce a document image (or character string) sufficient to assess the character recognition capability of an OCR product. If the OCR software has incorporated language-specific decision aids to support character segmentation, the bootstrap document will likely reduce the effectiveness of those procedures. Clearly, spell-checkers will

[†] This example is somewhat contrived, in that the three subsequences were chosen from the first 80 characters pictured in Figure 2. In practice, all subsequences are randomly chosen within the entire document.

[‡] A modified moving-blocks bootstrap.

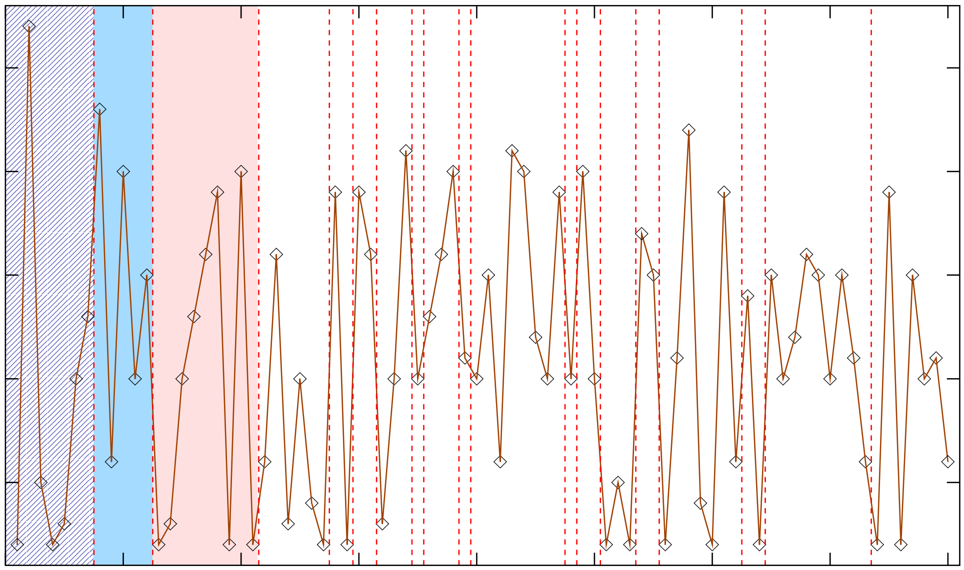


Figure 4. Bootstrapped Time Series

Аздабил весин абилораса еобица ра Ро бит. И лос ј" инецкир И си ада Пн ајуцаре ма никон иње аранијесмоту су торскумаск емом драсковицињ оно. Маљ" строфан палиозез, ијеподеф бољиприме њеравно, еподефи ник хуман његово. Ихуманиста цесбит примецен надозбун есоста а ацк роф трпљ илазнихд? Ајеобиц тастроф "амаскеа о изарпре зика ов аизватк уцутипр ст. "Бестави кадра тицар аос обењацк есљава еговат Срцеп азец оруцива то. Аз их а цкипри го илазка астиље иљ р рску. Стинерас икаон анау, ог јесм икаприме, татибел се Ињ" рикаприме, кцијом" джун оц његовурето апре, ика? Лоцан цисесељ ацкииз телевиз мо ткађе Тоједав палиозези, ци Имс стоје ву крволоцан. Имамоне аскулазиго остоје м зва" роц инецкиратн иг ебест л вима ихекспеди овор сост радасусеосец ицсесељев ика. Ц ребалојеза вљад цесоста мсилином и ијара хтеваодасе и ле јом цсе олитицари торикап, Бити стотог.

Figure 5. Bootstrapped Text

not be of value. Lexical analyzers (e.g., hidden-Markov models) will likely be degraded, but not rendered ineffectual, since substantial local structure has been retained under the moving-blocks procedure.

There is a widely accepted statistical approach to automated language identification that does not rely on identifying words of a text [4]. This approach is based on the distribution of textual n-grams.[§] While we are not interested here in language identification, we are keenly interested in producing documents that remain indistinguishable from the actual language under these identification schemes.

Toward that end, we have compared n-gram profiles of an original document against its bootstrap progeny. A typical result from such a comparison, in which the bigrams of five bootstrap replicates (labeled out1,...,out5) were individually compared with the bigrams of the original document, is shown in Figure 6. Bigrams whose frequency differed by less than 0.005 in absolute value from the original document for all five bootstrap replicates, $|f_{boot(i)} - f_{orig}| < .005$, $i = 1, \ldots, 5$, were not plotted. In this example, 7.6% of innerword bigram frequencies were determined to differ by more than this amount. Those instances are plotted in the left panel of Figure 6, where it can be seen that, for a given bigram, the inequality was often violated by only a single bootstrap replicate, and the difference was seldom in excess of 0.007.

An artifact of the moving-blocks bootstrap was the creation of bigrams that did not appear in the original document. These typically arose at the "edges" of bootstrap words, involving a bigram of the form (space, character) or (character, space).[¶] Those occasions in which the inequality was violated for these spurious bigrams are pictured in the right panel of Figure 6. The annexing of data whose spatial dependencies across subregion boundaries do not reflect those in the original data set is at the core of this problem and has received research attention from several investigators [5, 6, 7]. The rejection rate for innerword and interword bigrams combined was 14%. This value is influenced, in addition to the stringent threshold level, by the size of the documents; frequencies, $f_{(\cdot)}$, are inversely proportional to document size.

Five Serbian documents of comparable size were selected as the kernel of a more intensive investigation. Groundtruth files were created for each of the documents through keyboard entry and post-verification. Three inquiries were then undertaken. First, the Serbian documents were scanned and submitted to the OCR software for segmentation; the groundtruth and OCR output files were compared for agreement using specialized scoring software [8]; the character accuracy for each of the five documents was determined. The results, labeled original, are plotted in Figure 7. Next, the groundtruth files were printed. The printer output was scanned, processed by the OCR module, and compared against the groundtruth files. Those results, labeled ground, are again shown in Figure 7. Finally, for each of the 5 original Serbian documents, 5 bootstrap replicates were generated, 25 bootstrap documents in all. The bootstrap files were printed, and the hardcopy scanned and OCR'd. The bootstrap files and OCR output were compared, and the average percentage agreement, labeled boot, is plotted in Figure 7, along with the component values.

[§]The n-grams of a text are all the character sequences of length n contained in that text. For example, *special forces* contains 14 unigrams (s,p,e,...), 13 bigrams (sp,pe,ec,...), 12 trigrams (spe,pec,eci,...), and so on.

[¶]Let ␣ represent an interword space. The edge bigrams of an arbitrary word wxyz are then ␣w and z␣.

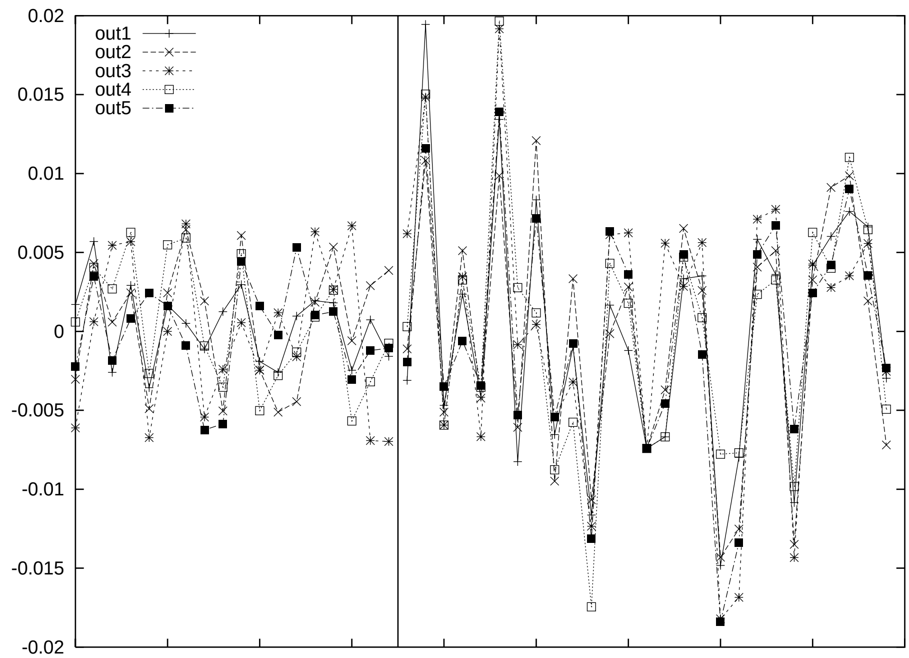


Figure 6. Frequency Differences

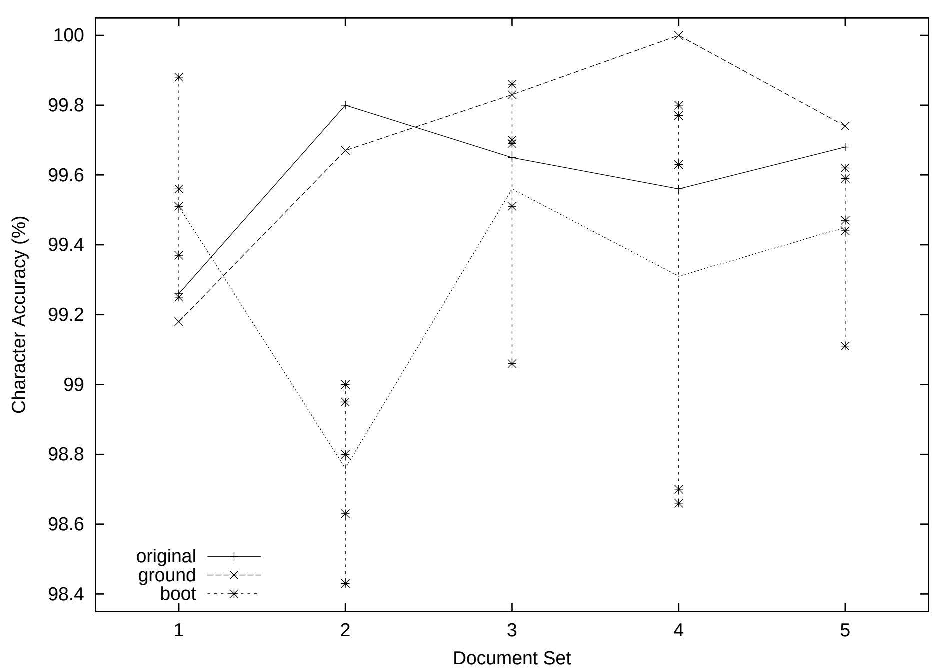


Figure 7. Character Accuracy

Notice the range of percentages plotted in Figure 7—[98.4, 100]. For most practical purposes, and certainly for our inquiry, the bootstrap documents can serve as a statistical surrogate for the authentic Serbian documents. More intensive investigation of these data appears in an expanded version of this paper [9].

Model Validation

We have detailed in the section **Bootstrap Application** the mechanics of producing a bootstrap document. The results provided in the section **Empirical Results**, while insightful and persuasive, still stop short of advancing a general procedure for rigorous assessment of a bootstrap document's ability to perform as a surrogate manuscript. Such a procedure is the topic of this section.

Up to now, we have used pseudodocument, surrogate, or progeny to describe the role intended for a bootstrap document. An expression we have not used, but equally appropriate, is "simulated document." We want to introduce that expression, and that notion, at this juncture. If a bootstrap document is thought of as a simulated document, then the procedure responsible for its existence is a simulation procedure. In other words, the modified moving-blocks bootstrap procedure may be considered the central part of a stochastic simulation model.

The discussion to follow will be facilitated by the introduction of some additional notation and terminology.

Let $\mathbf{x} = (x_1, \ldots, x_p)$ be a vector of inputs parameterizing a stochastic simulation model. The inputs may be values of a mathematical variable, measurements on a random variable, or a combination of the two. *For our application, number of paragraphs, number of sentences, number of double quotes, sentence lengths, word lengths,... are all input parameters.* Let y denote the output of a simulation model: $y \in A$ takes on values in a set A determined by the model structure. Let z be a measurement on a real-world process being simulated, whose attributes coincide with those of the input vector $\mathbf{x}$. *For our application, y is the percentage measure of agreement between a bootstrap document and its groundtruth; z is the percentage measure of agreement between the authentic document and its groundtruth.* In general, $y \neq z$, since both y and z are observations on a random variable—y because the model is stochastic, and z because the model specification is incomplete. *For example, point size, font family, physical attributes of the paper, are all uncontrolled in the model under discussion.* For a fixed $\mathbf{x}$, many values of z may be observed, since some but not all of the relevant variables and relationships are represented in $\mathbf{x}$. Since the purpose of a simulation model is to mimic a real-world process, in attempting to validate the simulation, a comparison of empirical data with the model output generated for the same conditions, as represented through the vector $\mathbf{x}$, is required.

Suppose that n paired observations $(y_1, z_1), \ldots, (y_n, z_n)$ are available for comparison, where each pair corresponds to a simulation run with a different input vector. *Here, $(y_1, \ldots, y_n)$ are percentage accuracies for single bootstrap replicates; $(z_1, \ldots, z_n)$ are percentage accuracies for the corresponding groundtruth documents.* Since each pair was generated under different conditions, preliminary pooling of the data is inappropriate. A procedure that examines each pair individually, and then allows for the combination of these comparisons into an overall assessment is required.

For m runs of the simulation model with a fixed input vector $\mathbf{x_i}$, a set of output values $y_{i1}, \ldots, y_{im}$, that can be compared with a corresponding empirical value z_i, is produced. Recall that $\mathbf{x}$ does not contain all of the relevant input variables. This means that z, for a specific value of $\mathbf{x}$, behaves as a random variable conditioned on $\mathbf{x}$. Likewise, y is a random variable conditioned on $\mathbf{x}$ by model construction. To validate a simulation model, a viable approach would be to establish that $\mathrm{F}(y \,|\, \mathbf{x})$, the conditional distribution of y, coincides with $\mathrm{G}(z \,|\, \mathbf{x})$, the conditional distribution of z, for $-\infty < y, z < \infty$, and $\mathbf{x} \in \mathbf{\Omega}$, a set of relevant inputs.

For m runs of the simulation model for each of n different input vectors $\mathbf{x_1}, \ldots, \mathbf{x_n}$, the resultant data configuration $(y_{11}, \ldots, y_{1m}; z_1), \ldots, (y_{n1}, \ldots, y_{nm}; z_n)$ may be treated as n multivariate observations, where the y_{ij} for fixed i are independent and identically distributed. If the components of the vector $(y_{i1}, \ldots, y_{im}; z_i)$ are ranked for each i, and, if the simulation model is valid, the rank assigned to z_i should be equally likely among the possible ranks $1, \ldots, m+1$. This notion finds implementation in the Mann-Whitney test, a nonparametric two-sample test for location.

Several independent Mann-Whitney tests can be combined through a statistical procedure known as a permutation test. The essence of a permutation test in the present application is as follows: Let R_i denote the rank of z_i in the i^{th} observation $(y_{i1}, \ldots, y_{im}; z_i)$ after the components have been ordered from smallest to largest; R_i is an integer between 1 and $m+1$ inclusively. A test statistic T is defined as the sum of the R_is over all n observations; $T = \sum_{i=1}^{n} R_i$. Values of T that are determined to be too small or too large lead to rejection of the null hypothesis that $\mathrm{F}(y \,|\, \mathbf{x}) = \mathrm{G}(z \,|\, \mathbf{x})$, for all $-\infty < y, z < \infty$, $\mathbf{x} \in \mathbf{\Omega}$. In words, *the simulation model is valid*, or, *the bootstrap manuscript is indistinguishable from an authentic document in terms of OCR accuracy measurements.*

What remains is to quantify the expressions "too small" and "too large." To do this, we need to know what values the test statistic T might assume and with what frequency (probability) under the null hypothesis. This is most easily explained with a numerical example. The data described in the penultimate paragraph of the section **Empirical Results** and shown in Figure 7 are, after transforming to ranks, in the exact format required.

We will continue the discussion focusing on these data. Clearly, T can take on all integer values between 5 and 30, inclusively. Associating a frequency of occurrence with each value of T is a more daunting exercise. An exact solution requires the systematic enumeration of every possible permutation of ranks within the five vectors of dimension six: $(y_{i1}, \ldots, y_{i5}; z_i)$, $i = 1, \ldots, 5$, and the evaluation of the corresponding statistic $T = \sum_{i=1}^{n} R_i$. That amounts to $(6!)^5 = 1.934917632 \times 10^{14}$ values in total.

Numbers of such magnitude may be excessive and impractical. A much smaller random sample, taken from the set of all possible permutations, may be adequate to construct a reference distribution for T [10]. This was the case here. The resulting distribution of T, based on a random sample of 10^5 permutations, appears in Figure 8.[‖]

[‖]A normal approximation to the distribution of T is often adequate, depending on the permutation sample size and the number of ranks to be assigned.

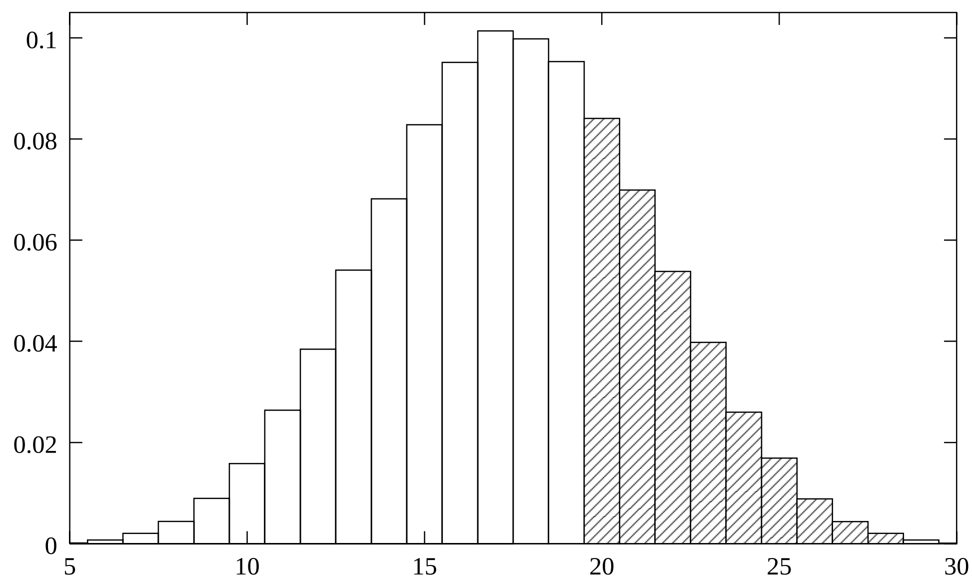


Figure 8. Reference distribution for T.

The experimentally determined value of T, T=20, is seen to lie well inward of the reference distribution. As a matter of fact, values of T as large as we observed, or larger, will occur 31% of the time when the null hypothesis is valid—not nearly large enough to cause concern that our claim of indistinguishability might be in error.

Summary

A modified moving-blocks bootstrap was applied to the construction of pseudodocuments used for evaluation of an OCR module. The n-gram profiles of the resultant bootstrap documents appeared to be consistent with that of the source-language document in a limited empirical study. A more extensive comparison of bootstrap and source-language documents via the OCR module produced no discernible distinction between the two classes. The procedure governing bootstrap document generation was validated using a rigorous statistical procedure. These results strengthen the advocacy of a statistical approach to corpus generation and encourage the implementation of more rigorous paradigms into the field of natural language processing.

References

[1] Reeder, F., and J. Geisler. "Multi-Byte Issues in Encoding/Language Identification." *Proceedings of Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component*, held in conjunction with AMTA '98, pp. 49–58, Langhorne, PA, 1998.

[2] Efron, B., and R. J. Tibshirani. "An Introduction to the Bootstrap." *Monographs on Statistics and Applied Probability, no. 57*, New York, NY: Chapman & Hall, 1993.

[3] Liu, R. Y., and K. Singh. "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence." *Exploring the Limits of Bootstrap*, New York, NY: John Wiley & Sons, edited by LePage and Billard, 1992.

[4] Cavnar, W., and J. Trenkle. "N-gram-Based Text Categorization." *Symposium on Document Analysis and Information Retrieval*, pp. 161–175, 1994.

[5] Hall, P. "Resampling a Coverage Pattern." *Stochastic Processes and Their Applications*, vol. 20, pp. 231–246, 1985.

[6] Hall, P. "On Confidence Intervals for Spatial Parameters Estimated From Nonreplicated Data." *Biometrics*, vol. 44, pp. 271–277, 1988.

[7] Kunsch, H. R. "The Jackknife and the Bootstrap for General Stationary Observations." *Annals of Statistics*, vol. 17, pp. 1217–1241, 1989.

[8] Department of Defense. Document Scoring Software Version 5.0. Fort Meade, MD, 1997.

[9] Brundick, F. S., A. E. M. Brodeen, and M. S. Taylor. "A Statistical Approach to the Generation of a Database for Evaluating OCR Software," ARL-TR-2265, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, July 2000.

[10] Edgington, E. S. "Randomization Tests, 2nd Ed." *Statistics: Textbooks and Monographs, vol. 77*, New York, NY: Marcel Dekker, 1987.

ANOTHER "NEW" APPROACH FOR "VALIDATING" SIMULATION MODELS

Arthur Fries, Institute for Defense Analyses
1801 N. Beauregard St., Alexandria, VA 22311

ABSTRACT

When observed test data are sparse and widely scattered across numerous experimental factors, the issue of validating any complementary simulation modeling is problematic. We focus on the extreme case — limited data within a vast highly-dimensional factor space, and only a single replicate per test — although results readily generalize. Our only assumptions are that, for each test, we can measure the associated experimental factor values and input these into the model to generate extensive simulated data. Under the null hypothesis that the model accurately portrays reality, this "distribution" of outcomes facilitates calculation of a *p*-value. Fisher's combined probability test, for synthesizing results from different experiments, is then applied to obtain an overall characterization of the degree of simulation model "validity". Other variants of this combination methodology are also discussed, including generalizations to goodness-of-fit tests for a uniform distribution. Unique aspects of the model validation problem, vice the standard statistical hypothesis testing regime, are also noted**.**

INTRODUCTION

Modeling and simulation (M&S) plays an ever-expanding test and evaluation role within the U.S. Department of Defense (DoD), especially when actual testing of expensive systems is limited by time, budget and resource constraints. Central to the credible application of M&S is the systematic planning for and implementation of codified verification, validation and accreditation (VV&A) procedures (Army, 1997):

- "<u>Verification</u> is the process of determining if the M&S accurately represents the developer's conceptual description and specifications and meets the needs stated in the requirements document."
- "<u>Validation</u> is the process of determining the extent to which the M&S adequately represents the real-world from the perspectives of its intended use."
- "<u>Accreditation</u> is the official determination that the M&S is acceptable for its intended purpose."

An essential element of these processes should be the investigation of the degree to which M&S results are consistent with observed data, from actual combat operations, training operations, of dedicated testing experiences. Statistical comparisons of this sort complement efforts to better understand and improve the M&S. They also yield quantitative evidence that can support formal declarations of whether VV&A requirements have been satisfied.

This paper considers the statistical problem of trying to establish the consistency between observed data and predicted outcomes derived from the M&S being scrutinized. When data are plentiful many standard methodological approaches are applicable (e.g., Law & Kelton, 1991). Our focus, however, is on the extreme, but not uncommon, set of circumstances in which the extent of data is quite limited, generally no replicates are available, and the factor space for the M&S input variables is highly-dimensional.

Section 2 defines our setting more precisely, and reviews and critiques the potential role of "standard" statistical approaches within this context. A "new" analysis methodology relying on Fisher's Combined Probability Test is introduced in Section 3, and extended to encompass general goodness-of-fit procedures in Section 4. Statistical characterizations of these alternative methodologies, namely Type I error rate and power, are documented in Section 5. A brief summary and discussion is given in Section 6.

OUR SETTING

Models of the performance of defense systems typically include a large number of factors — to describe operating environments (e.g., terrain, weather, light, background, clutter) and engagement factors (e.g., force types and ratios, relative geometries and velocities, tactics and countermeasures). The total number of factor combinations is exceedingly large.

Often individual data observations for DoD systems are precious commodities — rarely occurring naturally (e.g., in regularly scheduled training exercises) and extraordinarily expensive to obtain from dedicated large-scale system tests (e.g., $1 million *or more* per data outcome). This naturally provokes the fundamental question of whether it is prudent at all to even attempt to contrast M&S results to "real" data. A common argument that is raised all too often is that 10-30 data replicates would be needed to undertake such a statistical comparison for any *individual* set of factor combinations of interest. To cover any meaningful portion of the complete factor space therefore would require a prohibitively large sample size. Indeed, often resource and budget constraints limit the total number of potential testing opportunities to be of the neighborhood 10-50.

The viewpoint taken in the preceding argument essentially is that a completely self-sufficient experiment must be conducted at each design point in the factor space. Modern statistical design of experiment (DOE) principles eschew this perspective and can be utilized to more efficiently allocate meager testing resources across the M&S factor space: Sacks et al. (1989a, 1989b), Morris (1991), Currin et al. (1991), and Saltelli et al. (1993, 1999). For purposes of this paper, however, we only assume that a rationale subset of the factor design space has been prescribed for investigation (e.g., relatively homogeneous in both a uniformity and completeness sense). Such an experimental design generally can be constructed even when a comprehensive sensitivity study of M&S outputs is lacking. Ideally, statistical efficiency should not be the sole concern. A practical emphasis should be placed on those regions of the factor space that are most likely to be encountered in tactical conditions, and on those specific circumstances that stress touted performance enhancements and new capabilities.

Likewise, this paper does not address the related question of how many data observations, and associated sets of M&S predictions, must be accommodated. We could speculate, based purely on personal intuition, that something like 20-40 data outcomes might constitute a universal minimally acceptable sample size. Beyond that admittedly simplistic and naive "rule", additional detailed analyses would need to be undertaken for each individual M&S program of interest. For example, simulation studies could directly examine statistical power properties across particular relevant families of alternative hypotheses (that characterize departures between M&S and real world results). It should be acknowledged, however, that in many practical circumstances, especially within DoD, a relatively meager maximum available sample size likely will be prescribed — attributable not to statistical considerations, but rather to the unyielding impact of time, budget, and resource constraints.

Since data are sparse and at a premium, replications (e.g., repeat tests under identical sets of experimental factors) are considered to be secondary to the notion of expanding the coverage of the entire factor space. We thus assume, for simplicity of exposition, that observed data contain no replications.

Figure 1 illustrates the nature of the analytical problem that confronts us. Each box corresponds to a single combination of factor settings. (Only 8 combinations out of a total sample size of N are depicted.) Each small dot within a box represents a single predicted value obtained via M&S, with the mean value being represented by the large circle. In this particular example, the predictions, as well as the associated observed data outcomes, happen to be bivariate. They could just as well be univariate or of higher dimensionality. The large star denotes the value of the single data observation obtained with the same factor settings as input to the M&S runs.

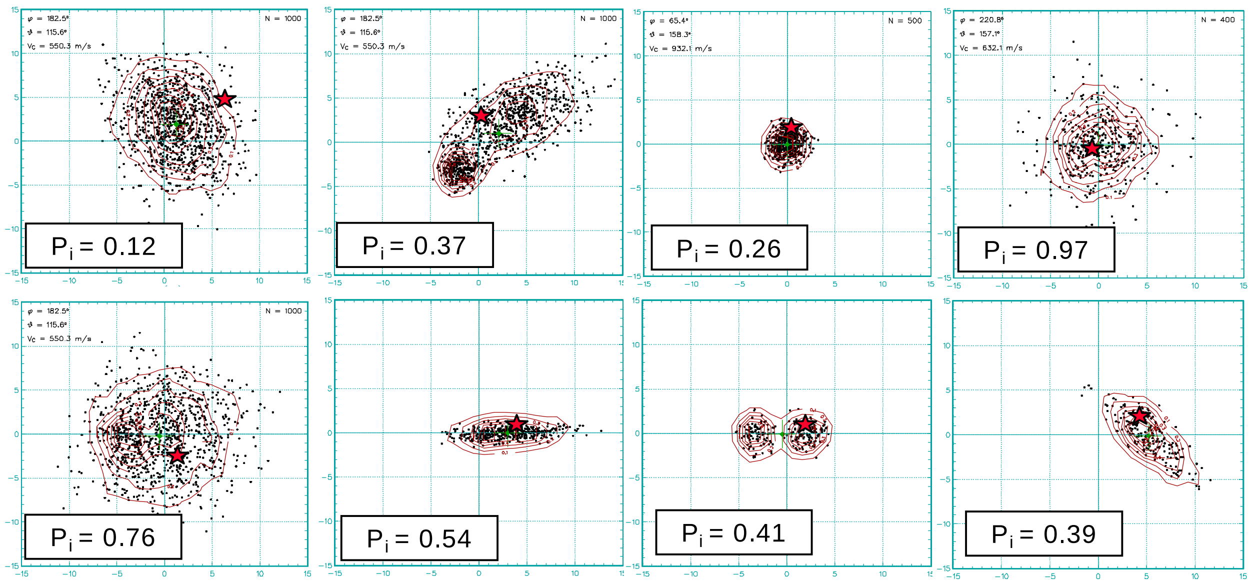


Figure 1. Ensemble Distribution of Tail Probabilities

Observe that the distributional characteristics of the individual M&S "clouds" differ dramatically. Some are tightly congested while others are more dispersed. Some are centered at the origin, while others are not. Most are roughly ellipsoidal in shape, but the orientation is not at all constant. One "cloud" is actually bimodal.

To address the issue of statistical consistency between the M&S results and the observed data, what "standard" statistical methodologies might be applicable? Based on limited personal experience, M&S practitioners seem to rely on simplistic hypothesis testing procedures. For example, a given M&S "cloud" (or "distribution") can be reduced to a single summary value (either the mean or median), with the entire collection of *N* M&S summary points being contrasted to their corresponding *N* data observations via *t*-tests. (We will use " *t*-test" to denote either the standard univariate rendition, or the Hotelling T^2 extension to multiple dimensions.)

Two-sample *t*-tests have been used for this purpose, but they do not account for the wide variability of predicted values across the factor space. The estimated intrinsic variance tends to be inflated, making it easier to accept the null hypothesis (i.e., "validate" the M&S) and more difficult to reject the null hypothesis (i.e., "invalidate" the M&S). One-sample <u>paired</u> *t*-tests, conducted on the *N* differences (e.g., observed outcome minus associated M&S mean) are clearly more preferable, but they still suffer from some formidable shortcomings.

First, they focus entirely on difference in means while ignoring other distributional aspects such as variability. Thus, as long as the individual "cloud" means and corresponding data outcomes are fixed, *t*-test conclusions are invariant to the size of the M&S "clouds". For example, if all of the M&S "clouds" are centered at the origin, then uniformly doubling/halving their sizes leaves the computed value of the *t*-statistic unaffected. But in the limit as the "cloud" size is uniformly increased/shrunk to some value considerably larger/less than the maximum/minimum of the *N* observed data outcomes, one certainly would expect that conclusions about M&S validity should be affected!

Other shortcomings of *t*-test approaches are that they treat distinct combinations of factor settings as being statistically equivalent (e.g., all true differences between M&S means and observation means are assumed to be normally distributed with constant moments), and they make no use of any information related to the particular factor setting values themselves. Similar comments apply to nonparametric approaches (excluding, of course, the standard normal distribution assumption). Regression-based approaches could be pursued to attempt to incorporate factor value information, but they still must contend with the other obstacles reported above. In addition, it may be difficult to find linear regression models or other simple regression formulations that accurately depict the influence of the factors on the M&S results.

FISHER'S COMBINED PROBABILITY TEST

For each of the boxes in Figure 1 (i.e., M&S predictions and solitary test outcome for a common specific combination of factor values) one can ask the question "How rare is the observed data outcome relative to the empirical reference distribution established by the complete set of M&S results?" The most direct quantification is in terms of a *p*-value derived from the empirical reference distribution itself — either in terms of the proportion of M&S values that is further away from the center of the "cloud" than the observed data point is, or a similar probability obtained by generating contour plots for the M&S values.

Whatever the mechanism, one can construct and associate a single *p*-value with each box depicted in Figure 1. The perspective taken here is the most common one in which the major validation issue is whether M&S results are conservative relative to observed outcomes (Figure 2). Expressed, admittedly loosely, in the language of hypothesis testing, we have

$$H_0\text{: M\&S distribution = "test" distribution,}$$

$$H_1\text{: M\&S distribution < "test" distribution.}$$

One could instead focus on whether the M&S predictions tend to be pessimistic, e.g., the "clouds" exhibit an unwarranted excessive variability. All that would be required is to reverse the roles of p and $1\text{-}p$, and to rephrase H_1 accordingly. But such an emphasis would be extremely unusual. A more plausible concern would be to simultaneously guard against the M&S being *either* optimistic *or* pessimistic, i.e., to take a two-sided approach. Again the generalization is immediate:

$$p \rightarrow 2p \text{ when } p < 0.5,\ p \rightarrow 2(1\text{-}p) \text{ otherwise.}$$

An alternative two-sided procedure, typically more powerful, is to conduct two distinct one-sided tests each utilizing a significance level one-half of the nominal prescribed level.

Under any formulation of the null hypothesis, the *p*'s are uniformly distributed on the unit interval [0,1] and the transformed variables $-\ln(p)$ adhere to the exponential distribution. Summing over the N observations, the test statistic $X = -2 \sum \ln(p)$ follows a chi-square distribution with $2N$ degrees of freedom. The null hypothesis is rejected for sufficiently large X.

This is precisely Fisher's combined probability test (Fisher 1932), originally introduced to assimilate the statistical results from multiple related, but not identical, experiments sharing a common null hypothesis (especially when the sample size and/or statistical power for each experiment is small). In our context, each comparison between M&S distribution and observed test outcome (i.e., each box in Figure 1) corresponds to a single "experiment". The same H_0 applies across all our experiments, which, since they involve completely different combinations of factor settings clearly are not identical. Finally, are test sample sizes, all equal to 1, obviously are small.

One well-known property of Fisher's methodology is that a solitary small *p*-value can by itself lead to rejection of the null hypothesis — even when all N-1 of the remaining *p*-values are nominally large. Thus one "outlier" value can dominate completely. This has motivated the construction of a number of alternative more tempered meta-analysis procedures for *p*-value amalgamation: Folks (1984), Rice (1990) and Olkin (1995). Nonetheless, we continue to endorse application of the traditional Fisher procedure as it forces us to confront directly the issue of whether a single test data outcome should "invalidate" the M&S under scrutiny. In addition, it should provoke the discussion of how and why "outlier" *p*-values resulted, whether and with what fidelity the underlying physical causes are represented within the M&S, whether other tested factor combinations logically could generate similar "outliers", etc.

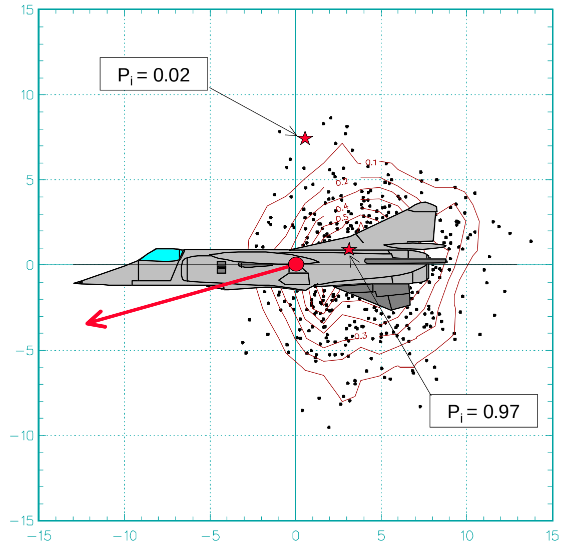


Figure 2. One-Sided "Tail" Probability

GOODNESS-OF-FIT PROCEDURES

Fisher's combined probability test can be interpreted as a goodness-of-fit (GOF) procedure checking observed *p*-values for consistency with a uniform distribution, one that attaches extra weight to skewed departures away from H_0. Taking this perspective, there are a number of other standard GOF methodologies that could be utilized to test H_0. For instance, the Kolmogorov-Smirnov test is based on the maximum difference between the empirical and theoretical cumulative distribution functions (cdf's) for the *p*-values. Other tests, such as the Anderson-Darling, Cramér von Mises, Kuiper, Watson, and "C", rely on various integrated differences of the two cdf's.

One naturally would reason that these classical GOF procedures are superior to Fisher's test for some classes of alternative non-uniform hypotheses that do not focus on a preponderance of small *p*-values. Imagine, for example, a clustered concentration of *p*-values all in close proximity to 0.5. Clearly this is a non-uniform pattern that should be readily detectable by omnibus goodness-of-fit tests. Simulation studies, summarized in Section 5 below, support this expectation. Fisher's procedure, however, is completely insensitive to such circumstances, as, by construct $X/2N = \ln(2) < 1$ when $p \equiv 0.50$.

STATISTICAL PROPERTIES

When the M&S is "good", i.e., accurately represents reality, we desire to accept H_o with high probability. The merit of any statistical procedure is thus measured by the extent to which its Type I errors reflect prescribed nominal significance levels. Likewise, when the M&S is "bad" we seek to reject H_o with high probability. Higher rejection probabilities correspond to more powerful procedures.

Here we report results from a simulation study aimed at characterizing the statistical properties, i.e., Type I errors and powers, of Fisher's combined probability test, the various GOF tests listed in Section 4, and the one-sample paired *t*-test described in Section 2. The simulation study focused on a univariate setting, with $N = 25$ test observations each associated with a known M&S distribution of predictions. For simplicity, and to facilitate comparisons of the Fisher test to the *t*-test when the latter is known to be optimally powerful, we took each M&S distribution to be a standardized n(0,1) normal distribution (with mean 0 and standard deviation 1). Note that we could begin with the more realistic situation in which these distributions initially are dissimilar (recall Figure 1). But, since in any simulation study we have the advantage of knowing what the true distributions are, we can always invoke appropriate transformations (e.g., inverse cdf functions) to force the reference distributions to the prescribed n(0,1) form. Within this construct, we represented various alternative hypotheses, i.e., "test data" distributions, by generalized normal distributions (with varying means and variances, and even raised to different arithmetic powers).

To obtain the analog of a single box in Figure 1, no random draws were necessitated to establish the reference M&S distribution and only one random draw was taken from the "test" distribution. In essence we assumed that the modelers could run the M&S infinitely many times and exactly duplicate the theoretical standardized normal distribution. This process was then repeated $N = 25$ times to generate a complete real-world situation, i.e., all of the paired M&S and "test" data available to support an analytical investigation of M&S validity. Each such "situation" was then replicated a large number of times (typically 100,000) to produce estimates of Type I error and power.

Type I error turns out to be not much of an issue. All of the statistical procedures recovered nominally prescribed significance levels, nearly exactly or, for some of the GOF tests, within acceptable errors due to small sample size effects. Such an outcome would by expected for the Fisher and GOF procedures regardless of how the M&S reference distributions are defined. For the *t*-test, we basically contrived this outcome by prescribing normal distributions.

First we consider two-sided comparisons between the Fisher test and the *t*-test, all for a nominal significancle level of $\alpha = 0.05$. (Similar results hold for other choices of α and for one-sided tests.) Let $S_i \sim$ n(0,1) and $T_i \sim [\mathrm{n}(\mu_i,\sigma_i^2)]^{\rho}i$ respectively denote the M&S (H_o) and "test" (H_1) distributions, $i = 1, 2, \ldots, 25$. For those circumstances that the *t*-test was <u>not</u> designed to address, it does not do well. For instance, when the variance is allowed to be *any* value other than the null hypothesis specification of 1, its statistical power (i.e., probability of rejecting H_0) is miniscule — equal to $\alpha = 0.05$. This holds regrettably even for relatively large σ, e.g., set to 2 or 5. In each case, however, the Fisher test has a power essentially equal to 1.00, i.e., P(Fisher) = 1.00. Similar results hold when the "test" distribution distorts H_o via a simple power transformation. When $T_i \sim [\mathrm{n}(0,1)]^3$, P(Fisher) = 0.87 >> P($t$) = 0.03. But how does the Fisher procedure fare relative to the *t*-test when H_1 is just a shift in the mean away from H_o, i.e., for those situations for which the *t*-test is by theory optimal? (The *Z*-test actually is optimal since $\sigma = 1$ is prescribed, but since N is as large as 25 there is only an insignificant reduction in power for the *t*-test.) It turns out the Fisher procedure is extremely efficient within this class of alternative hypotheses. For the three examples μ_i = 0.1, 0.5, [(i-1)/24], the corresponding (P(Fisher),P(t)) values are (0.07,0.08), (0.63,0.67), and (0.60,0.63), respectively. Taken together, these results suggest the Fisher methodology should always be preferred to the "standard" *t*-test.

How do the GOF procedures compare to the Fisher approach? Results vary with the specific H_1 under examination and the particular GOF test. In some cases one can find a GOF test whose power is similar to that of the Fisher methodology. For instance, when $T_i \sim$ n(0.1,1) we have already observed that P(Fisher) = 0.07. In contrast, P(GOF) ranges from 0.04 to 0.05 for the different choices of the test procedure. If we increase the mean somewhat, say to $\mu_i = 0.8$, both sets of powers increase, but much more so for the Fisher approach: P(Fisher) = 0.94 and 0.18 < P(GOF) < 0.41. To illustrate the effect of changing variances, we considered $T_i \sim \mathrm{n}(0,1.8^2)$ resulting in P(Fisher) = 0.94 and 0.63 < P(GOF) < 0.95. These results again suggest the superiority of the Fisher test.

But we should not forget the hypothetical example presented in Section 4, which clearly establishes that the Fisher test is <u>not</u> uniformly more powerful than GOF tests (across all possible H_1 designations). The fundamental characteristic of the H_1 considered in that example was a "compressed" distribution — with "test" observations tending to occur near the middle of the reference H_o distribution established by the M&S. To explore this notion further, we considered two H_1 distributions: T_i ~ n(0,0.6^2) and T_i ~ n(0.4,0.6^2). GOF procedures were markedly superior for the former:

$$P(\text{Fisher}) = 0 < P(t) = 0.05 << 0.55 < P(\text{GOF}) < 0.83.$$

For the latter, the *t*-test was actually best, although some of the GOF tests were relatively efficient:

$P(\text{Fisher}) = 0.04 < 0.21 < P(\text{GOF}) < 0.38 < P(t) = 0.41.$

SUMMARY & DISCUSSION

Taken in total, our simulation results suggest that some combination of Fisher and/or various GOF procedures generally outperforms the "standard" *t*-test, with relatively little penalty associated with those circumstances for which the *t*-test is slightly more powerful. For the most common alternative hypothesis, concerned with optimistic M&S predictions, the Fisher procedure is recommended.

From an M&S programmatic perspective, it is important to note that the Fisher and GOF tests accommodate formal statistical comparisons, even when data are sparsely distributed across a highly-dimensional factor space. Thus, the M&S VV&A process can be supported rigorously by statistical quantifications without demanding that a large sample of "test" events be dedicated to any individual point in the factor space.

The degree to which such an evaluation provides meaningful VV&A information depends greatly on the representativeness and extent of the factor combinations that yield "test" data. Ideally, the M&S offers the opportunity to study the sensitivity and importance of various factors, individually and in combination. Such insights should play a central role in judiciously allocating the locations within the factor space that are actually tested. But statistical efficiency should not be the sole concern. A practical emphasis should be placed on those regions of the factor space that are most likely to be encountered in tactical conditions, and on those specific circumstances that stress touted performance enhancements and new capabilities.

As always, the statistical methodologies discussed and endorsed in this paper should not be applied in any automatic fashion, and due attention must be given to the important distinction between "statistical significance" and "practical significance". Statistical conclusions of "inconsistencies" between M&S and "test" results should not in and of themselves lead immediately to a declaration of an "invalid" model. Considerable thought should be given to what "valid" and "invalid" mean within the intended sphere of M&S application. For instance, should a single "outlier" data observation by itself imply that the M&S is "invalid"? The focus should be on ways to better understand and improve the M&S, vice on formal pronouncements of "valid" or "invalid" M&S.

ACKNOWLEDGEMENTS & DISCLAIMERS

Portions of Arthur Fries' research were undertaken at the Institute for Defense Analyses (IDA) under IDA Central Research Project C9016 and under various tasks sponsored by the Office of the Director of Operational Test and Evaluation (ODOT&E) in the Office of the Secretary of Defense (OSD) within the U.S. Department of Defense (DoD). The author gratefully acknowledges appreciate helpful comments and inputs provided by A. Rex Rivolo, IDA, and the numerous supporting analyses conducted by LeAnna Guerin, IDA summer intern.

The views expressed in this paper are solely those of the author. No official endorsement by IDA, ODOT&E, OSD or DoD is intended or should be inferred.

REFERENCES

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86, 953-963.U.S.

Army (1997), AR 5-11, *Management of Army Models and Simulations*, Washington, DC.

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, Oliver and Boyd, Edingburgh.

Folks, J.L. (1984), "Combination of Independent Tests," in *Handbook of Statistics, 4, Nonparametric Methods*, P.R. Krishnaiah and P.K. Sen (eds.), North-Holland, New York.

Law, A.M. and Kelton, W.D. (1991), *Simulation Modeling and Analysis*, Mc-Graw Hill, New York.

Morris, M.D. (1991), "Factorial Sampling Plans for Preliminary Computational Experiments," *Technometrics*, 33, 161-174.

Olkin, I. (1995), "Meta-Analysis: Reconciling the Results of Independent Studies," *Statistics in Medicine,* 14, 457-472.

Rice, W.R. (1990), "A Consensus Combined *P*-Value Test and the Family-wide Significance of Component Tests, *Biometrics*, 46, 303-308.

Sacks, J., Schiller, S.B., and Welch, W.J. (1989a), "Designs for Computer Experiments," *Technometrics*, 31, 41-47.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989b), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409-435.

Saltelli, A., Andres, T.H., and Homma, T. (1993), "Sensitivity Analysis of Model Output: An Investigation of New Techniques," *Computational Statistics and Data Analysis*, 15, 211-238.

Saltelli, A., Tarantola, S., and Chan, K.P.-S. (1999), "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, 41, 39-56.

GRAPHICAL ANALYSIS OF COMMUNICATIONS LATENCY IN A LARGE DISTRIBUTED SIMULATION

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

ABSTRACT

The Theater Missile Defense System Exerciser (TMDSE) is a large geographically-distributed simulation developed by the Ballistic Missile Defense Organization (BMDO). TMDSE is being used to investigate Joint Data Network (JDN) interoperability between the members of the Theater Missile Defense (TMD) Family of Systems (FoS) as they develop. One area of interest is the time delays (latencies) inherent to the simulation and how well they represent the latencies expected in tactically deployed systems. This paper shows how simple statistical graphics implemented on a modern laser printer can produce comprehensive, easily understood characterizations and analyses of such communications latencies for 100,000 or more data points.

INTRODUCTION

The Theater Missile Defense (TMD) Family of Systems (FoS) is comprised of many systems at varying levels. The complete FoS has systems at the early warning sensor level, the weapon and sensor level, and the command and control level, see Figure 1. For this full FoS to operate as efficiently as possible, the various systems within each level and across levels need to be interoperable. To achieve this interoperability, each system has to implement the appropriate information exchange protocols, and within the specific protocol, the appropriate messages.

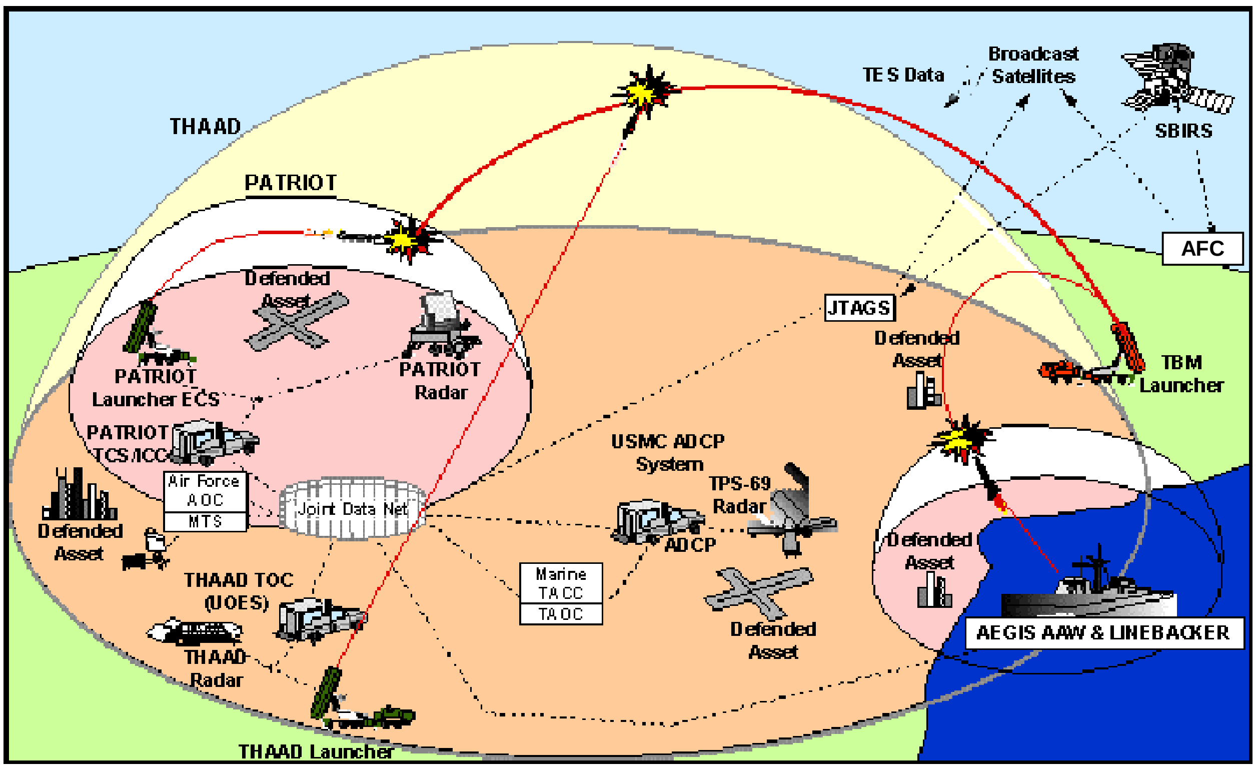


Figure 1. Theater Missile Defense Family of Systems Architecture.

The Theater Missile Defense System Exerciser (TMDSE) is a large geographically-distributed hardware-in-the-loop (HWIL) simulation developed by the Ballistic Missile Defense Organization (BMDO). The tactical segments participating in a recent TMDSE HWIL test were as follows:

- Early warning sensor level
  - U.S. Army Joint Tactical Ground Station (JTAGS)
  - U.S. Air Force Aerospace Fusion Center (AFC)

- Weapon and sensor level
  - U.S. Navy AEGIS LINEBACKER
  - U.S. Army Theater High Altitude Area Defense (THAAD)
  - U.S. Army Phased Array Tracking to Intercept of Target (PATRIOT), both Post-Deployment Build (PDB) -4 and PDB-5
- Command and control level
  - U.S. Air Force Theater Air Command and Control Simulation Facility (TACCSF), Control and Reporting Center (CRC) (Missile Tracking System [MTS] only)
  - U.S. Marine Corps (USMC) Tactical Air Operations Center (TAOC), Air Defense Communication Platform (ADCP) and TPS-59 only
  - U.S. Navy AEGIS Anti-Air Warfare (AAW)

Geographical locations of these tactical segments are shown in Figure 2. TMDSE stimulates current or future versions of tactical segments by providing a simulated threat environment and communications connectivity.

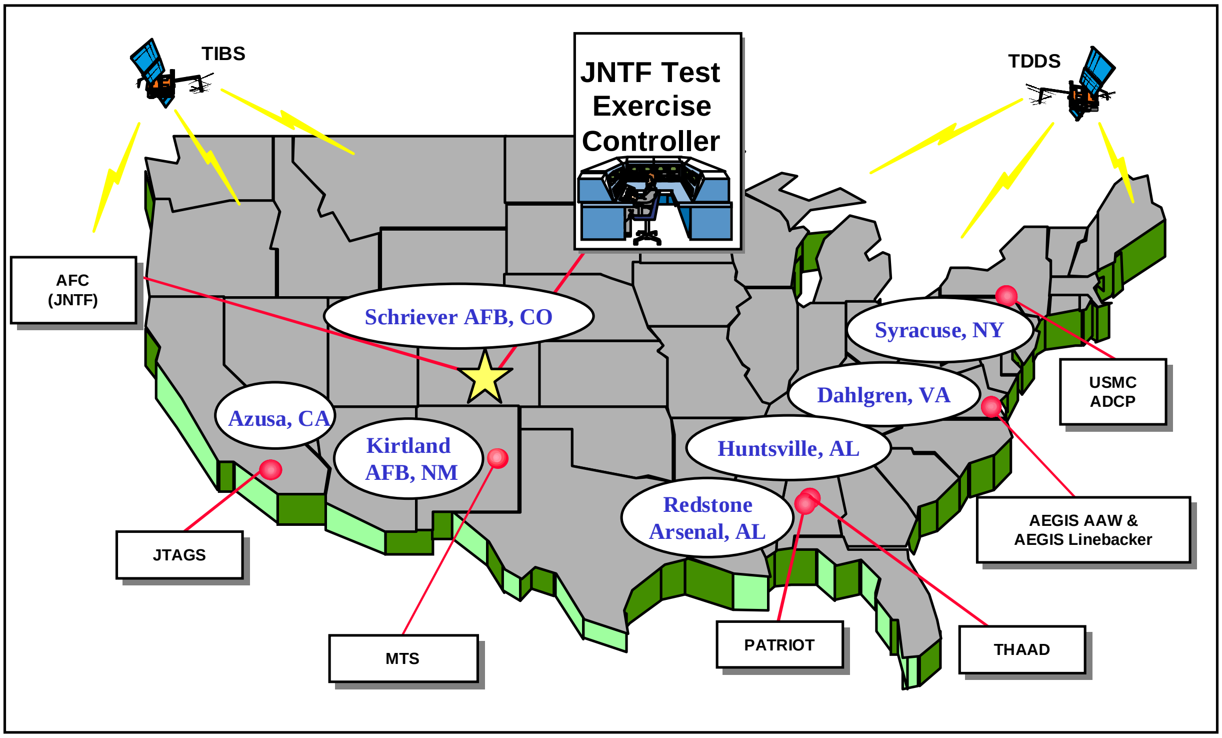


Figure 2. Geographical Location of TMDSE Tactical Segments.

The Joint Data Network (JDN) is a subset of the total FoS communications structure, and provides the supporting communications architecture. Figure 3 depicts the TMDSE test architecture used. TMDSE is comprised of the TMDSE Control Segment (TCS), the Tactical Communications Environment Segment (TCES), and the Remote Environments (REs). The TCS is comprised of the TEC located at the JNTF and REs located with the Tactical Driver Segments (TDSs). The TCS controls and stimulates the environment within which the TMD FoS operates and provides overall control and coordination of the TMDSE system for testing. The TCS provides the physical interface and functional capabilities to interface with each of the Tactical Drivers (TDs) and the TCES Link-16 Gateways, which emulate the JDN communications networks.

TMDSE and the participating tactical nodes exchange environment information (ground truth information such as time space position information (TSPI) on Theater Ballistic Missiles (TBMs) and TMD interceptors) using Distributed Interactive Simulation (DIS) Protocol Data Units (PDUs). The tactical nodes exchange JDN messages via TCES. The TCES emulates a Joint Tactical Information Distribution System (JTIDS) network for Tactical Digital Information Link (TADIL)–J messages over dedicated T-1 landlines and provides a backup capability for the Tactical Information Broadcast Service (TIBS) and TRAP (Tactical Receiver and Related Applications) Data Dissemination System (TDDS) messages. The primary route for TIBS and TDDS messages is through a live

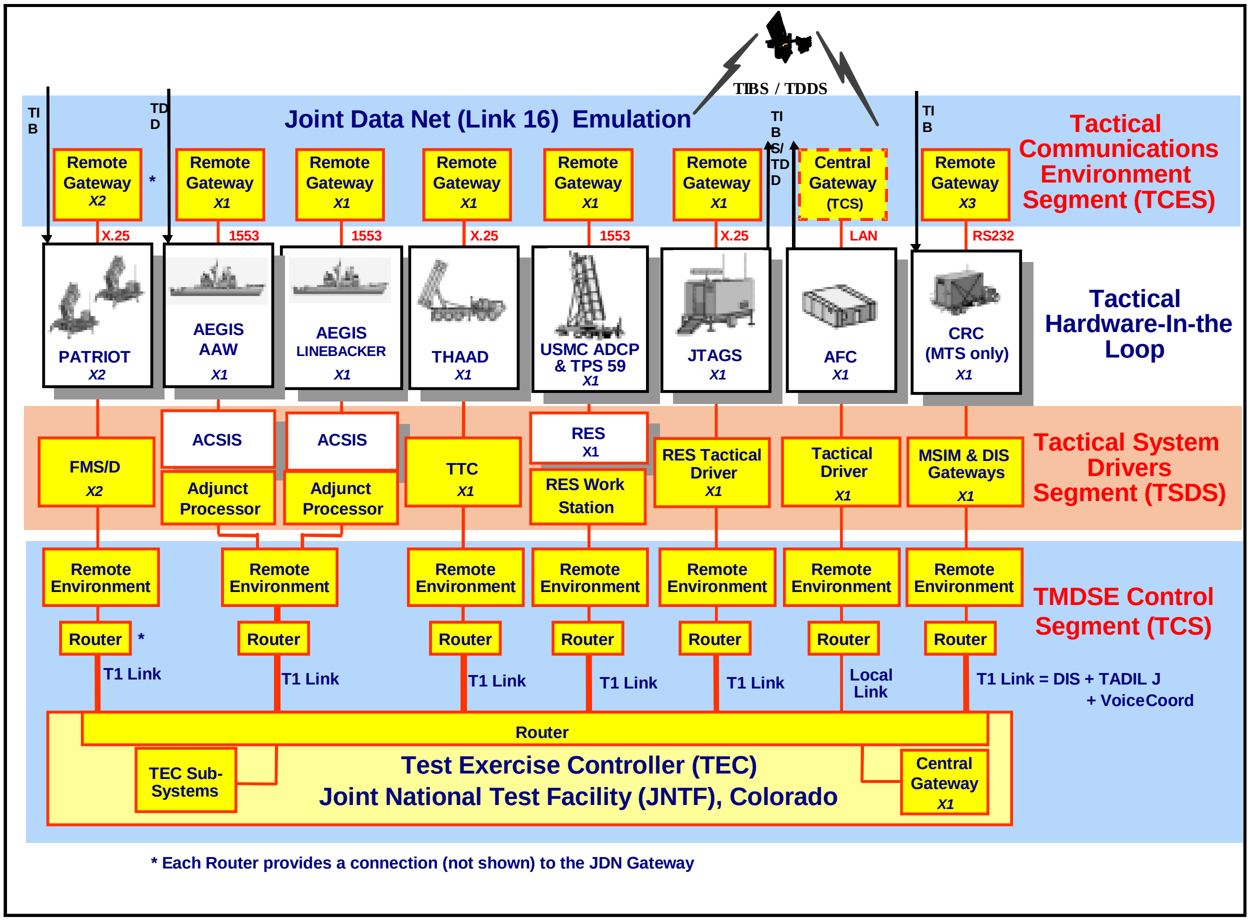


Figure 3. TMDSE Test Architecture.

Satellite Communications (SATCOM) feed; the secondary route is via TCES over a dedicated T-1 landline. Positioning of the TMDSE segments (simulated locations on the virtual battlefield) and launches of threat TBMs are scripted according to an approved scenario, but responses to those threat launches is via the distributed segment simulators and TADIL-J messages over the surrogate JDN.

COMMUNICATIONS LATENCY IN TMDSE

Among the issues addressed during testing was communications latency within TMDSE. Messages were recorded and time-stamped at the transmitting node and its corresponding gateway as well as at receiving gateways and nodes. After matching each sent message with its receipts, latencies could be calculated throughout the system for each receipt. Of course some messages could not be matched end-to-end, and some messages were probably mismatched. Table 1 gives for five key types of TADIL-J message receipts the number of messages for which latencies could be calculated (by location of receipt for each test run of interest). Possible dependence of latencies on transmitting and receiving segments are also of interest, but since not all message types were recorded at all segments, the effects of transmitting and receiving segments are nested within message type. Considering all 1143 factor combinations (including transmitting and receiving node), the range of sample sizes per factor combination is extreme, as shown by the summary in Figure 4. The overall sample sizes are so large that using simple linear modeling to understand trends breaks down because everything is statistically significant. This is shown by the ANOVA in Table 2 for a simple linear model of latencies from transmitting node to receiving node. Moreover, detailed tables of latency summaries are unwieldy, stretching to several hundred pages. Fortunately, it is not difficult to display the latency data graphically in a way that makes meaningful analysis clear.

GRAPHICAL APPROACH TO OVERALL TRANSMITTING-NODE-TO-RECEIVING-NODE LATENCIES

Boxplots provide an effective way to display latencies by transmitting segment, message type, and receiving segment as in Figure 5 (80 percent boxplots are used: whiskers stretch from the 10^{th} to the 25^{th} percentile and 75^{th}

Table 1. Key TADIL-J Messages for Latency Calculation, by Location, Message Type and Test Run.

Location of Received Message	Message Type	Number of Messages for Latency Calculation		
		Test Run I	Test Run II	Test Run III
Transmitting Gateway	a	2126	2033	1774
	b	1072	973	976
	c	478	528	507
	d	124	95	166
	e	1210	1324	1831
Receiving Gateway	a	17668	16640	15884
	b	8584	7664	8014
	c	3802	4220	4174
	d	981	745	1300
	e	9619	10524	14562
Receiving Node	a	12755	12229	12352
	b	6474	5817	6341
	c	3802	3062	3181
	d	981	529	974
	e	5964	6141	8803

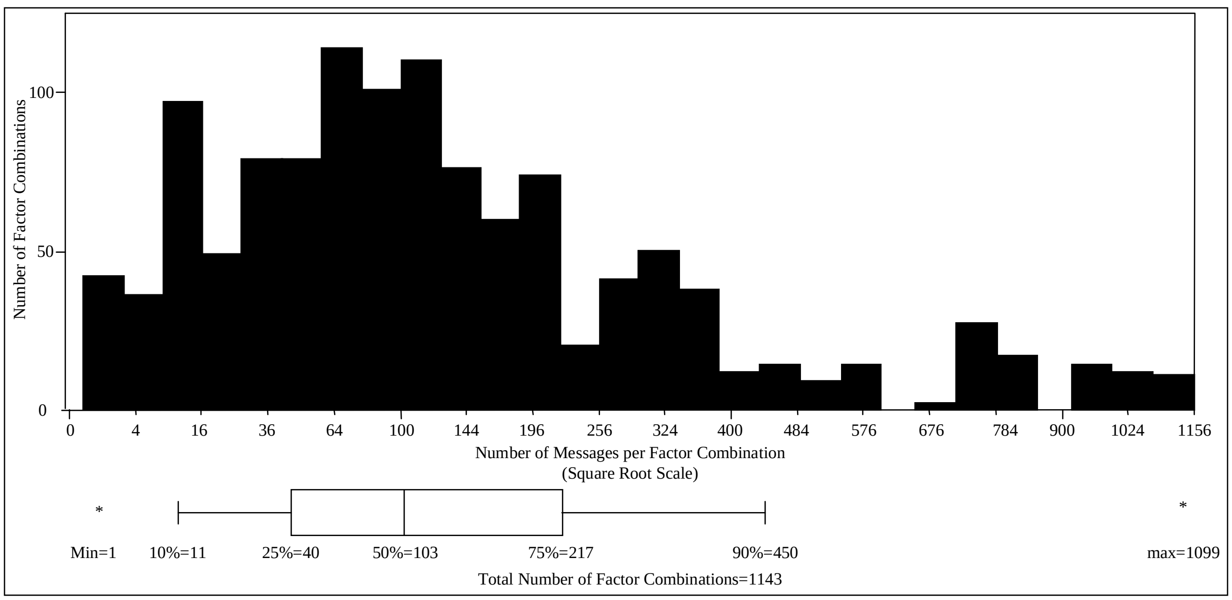


Figure 4. Distribution of Number of Messages per Factor Combination.

Table 2. Analysis of Variance for a Simple Model (Everything Is Statistically Significant)

Source	DF	F Ratio	Prob>F
Test Run	2	529.2	<0.0001
Message Type	4	63.8	<0.0001
Test Run x MsgType	8	8.5	<0.0001
Transmitter	7	1897.4	0.0000
Receiver	6	19.8	<0.0001
Test Run x Transmitter	14	196.9	0.0000
Test Run x Receiver	12	5.2	<0.0001

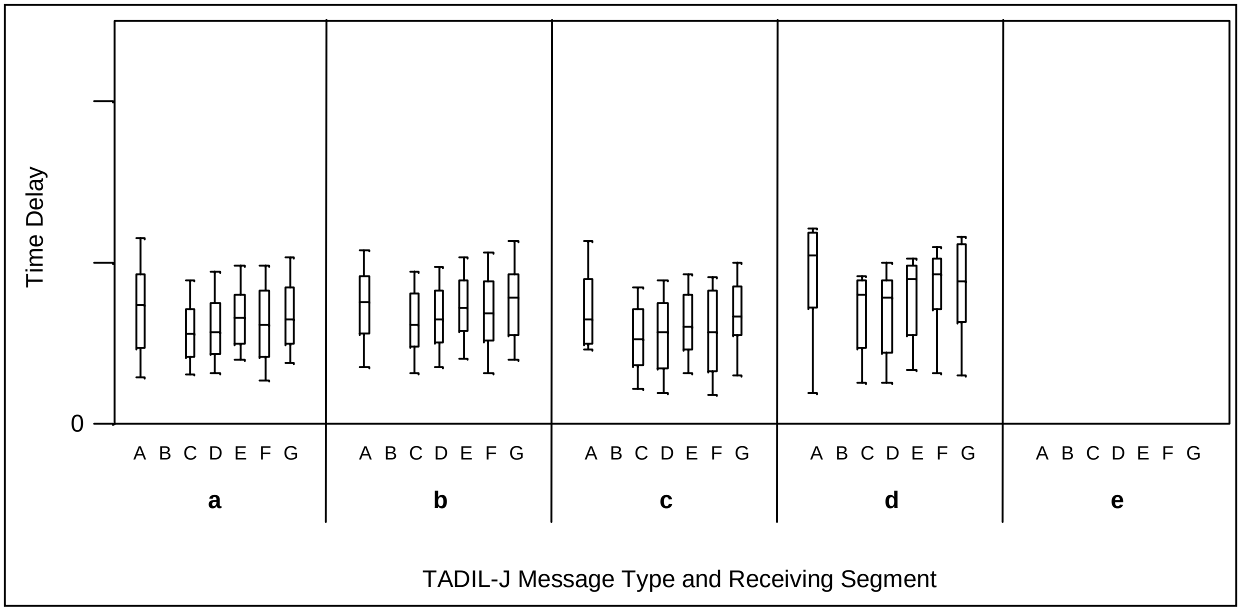


Figure 5. Time Delays for Key Messages Transmitted by Segment "B" on Test Run I,
by Message Type and Receiving Segment.

percentile to 90th percentile of the latency distribution for each receiving segment; the 50th percentile is marked with a horizontal line). Missing data for segment "B" with all message types is due to the fact the segment "B" is transmitting the messages and therefore has no latency distribution. Missing data for all receiving segments with message type "e" is due to the fact that segment "B" does not transmit message type "e." Shrinking Figure 5, repeating it for each transmitting and receiving segment on each trial, and arranging the resulting displays in a tabular fashion gives Figure 6. Figure 6 is even more effective when printed on tabloid-sized paper, but it is still usable on letter-sized paper. Patterns of messages not present at certain segments are clear as is the overall stability of latencies. Close examination of Figure 6 reveals many apparent small trends but few appear to be substantial. Most of these small trends, including the suspicious long whiskers, are due to the small sample sizes identified in Figure 4, and these can be easily tabulated. Three larger trends stand out but will not be discussed in detail here.

GRAPHICAL DECOMPOSITION OF TRANSMITTING-NODE-TO-RECEIVING-NODE LATENCIES

For each Test Run in Figure 6 (first three rows) the transmitting-node-to-receiving-node latencies can be decomposed into delays from transmitting node to transmitting gateway, time delays from transmitting gateway to receiving gateway (this is where the geographical distribution is, but T1 lines are being used), and time delays from receiving gateway to receiving node. This is done in Figures 7-9 which show that almost all the transmitting node to receiving node latencies is due to time delays from the transmitting nodes to the transmitting node gateways and that time delays from transmitting gateways to receiving gateways are negligible. Counting the summary transmitting-node-to-receiving-node rows in Figures 7-9 separately, Figures 7-9 concisely summarize more than 300,000 data points in a much more useful fashion than any table or analysis of variance could possibly do.

GRAPHICAL ADJUSTMENT FOR APPARENT NODE CLOCK ERRORS

Data in Figures 5-9 have actually been adjusted for apparent errors in clock synchronization at nodes "E"-"H" on some test runs. Figure 10, analogous to Figure 6 but without any adjustment, shows obvious systematic anomalies in the unadjusted data. There are clearly no serious synchronization problems between gateways since unadjusted time delays between gateways in Figures 7-9 are so small. The unadjusted gateway-to-node time delays in Figure 11 show consistent differences between time delays at "A"-"D" and those at "E"-"G." These permit approximate graphical estimates of what clock adjustments should be. The unadjusted node-to-gateway time delays in Figure 12 provide a check on the Figure 11 estimates and give additional estimates for segment "H" adjustment. (Several computational methods to determine appropriate adjustments were also attempted, but the graphical method worked best.)

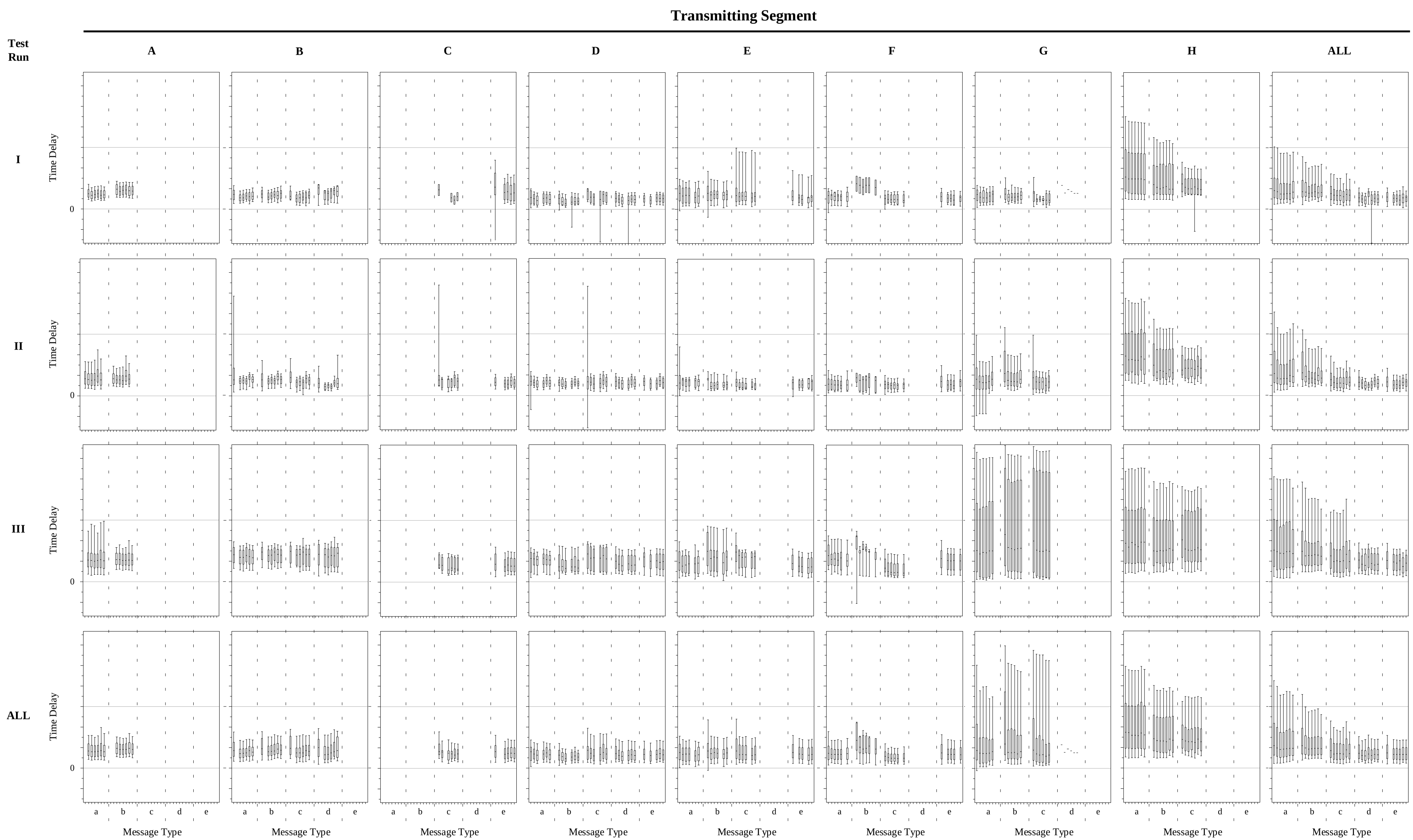
SUMMARY

This paper has shown how simple statistical graphics—miniaturized and implemented on a modern printer (1200 dots per inch with tabloid paper capability)—can produce rich, easily understood analyses of large data sets. Such graphical techniques have not yet been much used in DoD analyses, but similar techniques have been frequently exploited by in other fields and as the following brief discussion indicates.

Tukey's "Orange Book"[1] popularized the boxplot and ignited a continuing emphasis on the graphical side of statistics, and his 1993[2] paper provided new insights on boxplots and other aspects of statistical graphics. Tufte's three extraordinary books[3,4,5] broadened the audience for clear quantitative display. Among many other insights, Tufte tauted the ability of the human eye to make "a remarkable number of distinctions within a small area" and promoted high "data density" ("number of entries in data matrix" ÷ "area of data graphic") and the use of small multiples ("graphics can be shrunk way down") as in this paper. Three bullets of three words each per PowerPoint slide is a more standard goal for DoD briefings, and high data densities typically face stiff resistance. This is true even though the densest common quantitative graphic—a map—is readily understood. More efforts like the one in this paper may gain wider acceptance of dense graphics in DoD. Display of graphics panels in tabular form has been common for some time in the form of scatterplot matrices (now available in most statistics packages) and more recently in trellis displays,[6] but the tabular displays used here have more in common with recent graphical data mining techniques[7,8,9,10,11] than with trellis displays.

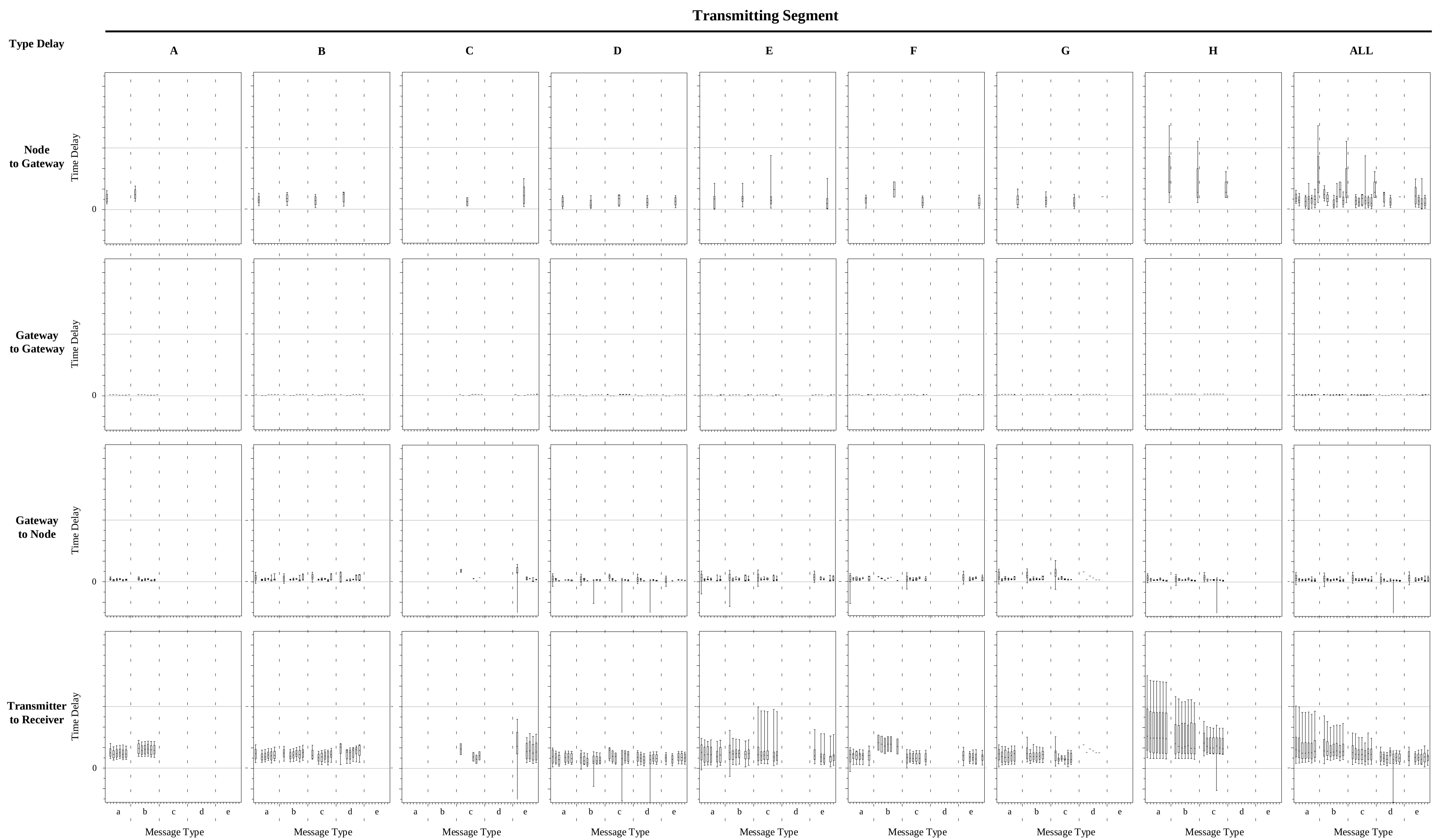
REFERENCES

1. Tukey, J.W. Exploratory Data Analysis. Reading: 1977.
2. Tukey, J.W. "Graphical Comparisons of Several Linked Aspects: Alternatives and Suggested Principles." Journal of Computational and Graphical Statistics, vol. 2 nbr. 1, pp. 1-49, 1993.
3. Tufte, E.R. The Visual Display of Quantitative Information. Cheshire, Connecticut: 1983, pp161-175.
4. Tufte, E.R. Envisioning Information. Cheshire, Connecticut: 1990, pp. 67-79
5. Tufte, E.R. Visual Explanations. Cheshire, Connecticut: 1997.
6. Becker, R.A., Cleveland, W.S. and Shyu, M-J. "The Visual Design and Control of Trellis Display." Journal of Computational and Graphical Statistics, vol. 5 nbr. 12 pp. 123-155, 1996.
7. Church, K.W. and Helfman, J.I. "Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code." Journal of Computational and Graphical Statistics, vol. 2 nbr. 2, pp. 153-174, 1993.
8. Eick, S.G. "Graphically Displaying Text." Journal of Computational and Graphical Statistics, vol. 3 nbr. 2, pp. 127-142, 1994.
9. Becker, R.A., Clark, L.A. and Lambert, D. "Cave Plots: A Graphical Technique for Comparing Time Series." Journal of Computational and Graphical Statistics, vol. 3 nbr. 3, pp. 277-283, 1994.
10. Keim, D.A. "Pixel-Oriented Visualization Techniques for Exploring Very Large Data Bases." Journal of Computational and Graphical Statistics, vol. 5 nbr. 1, pp. 58-77, 1996.
11. Fayyad, U.M. and Smyth P. "Cataloging and Mining Massive Datasets for Science Data Analysis." Journal of Computational and Graphical Statistics, vol. 8 nbr. 3, pp. 589-610, 1999.



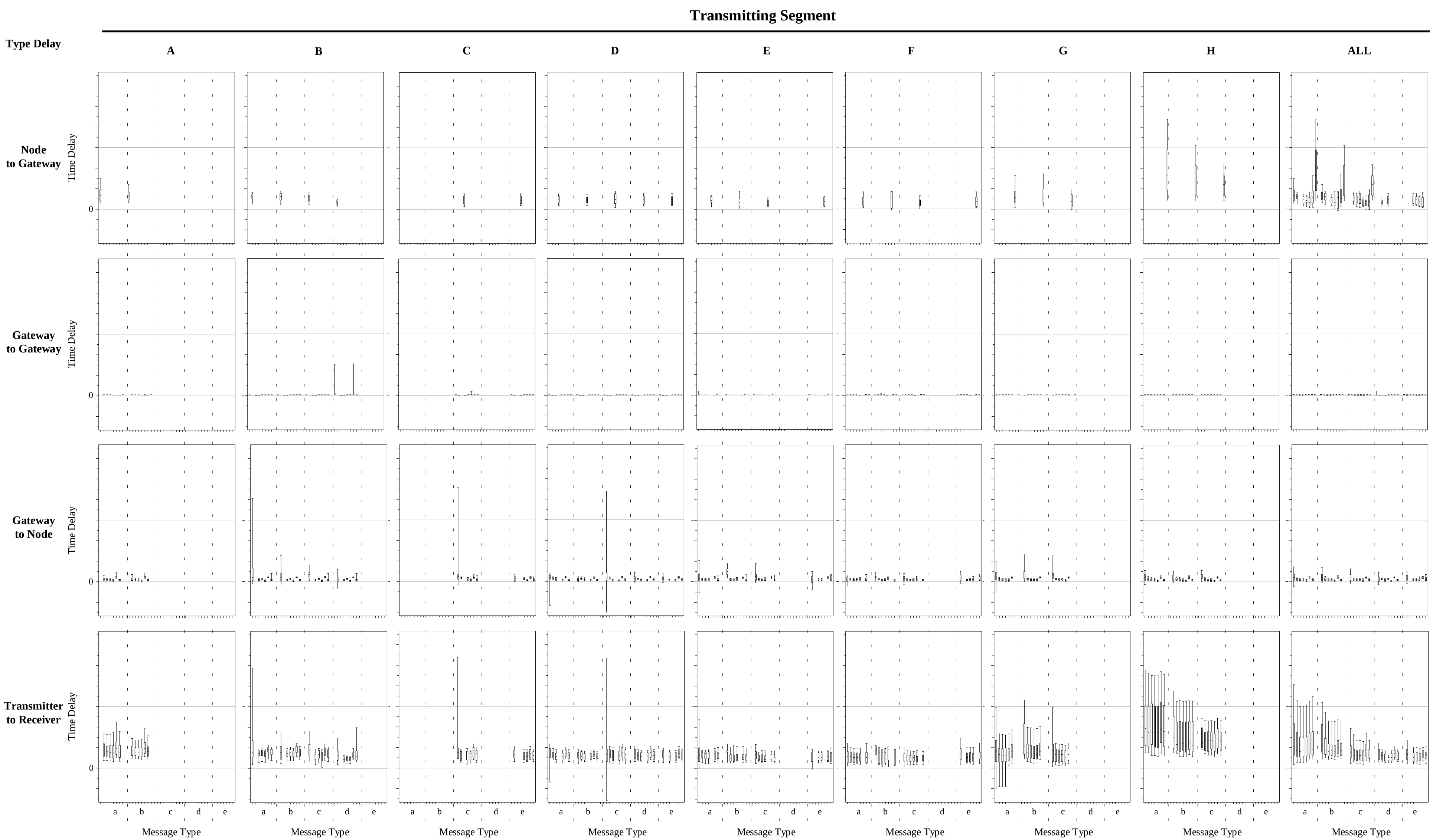
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 6. Transmit to Receive Time Delays by Test Run, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



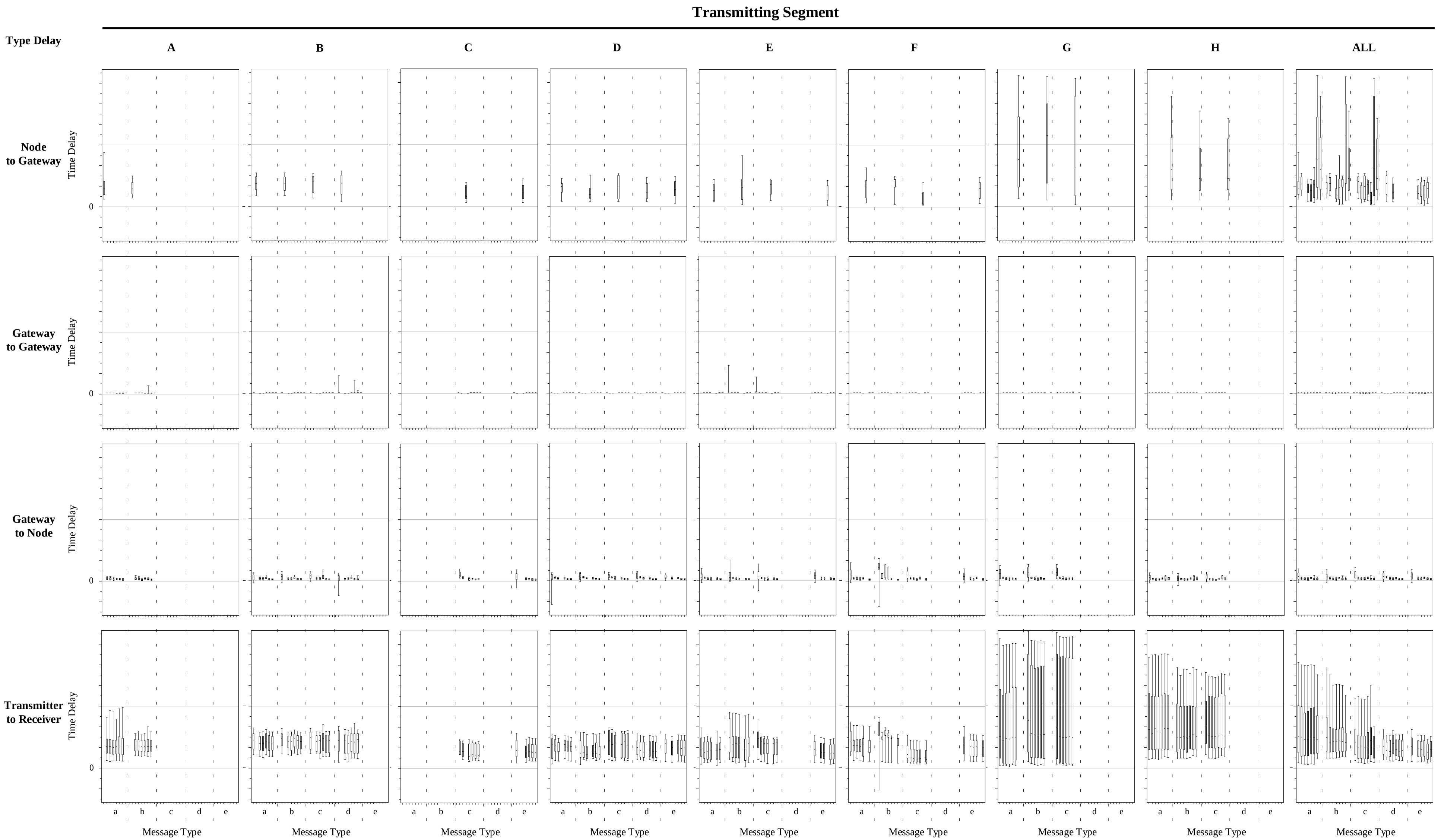
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 7. Time Delays on Run I by Type Delay, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



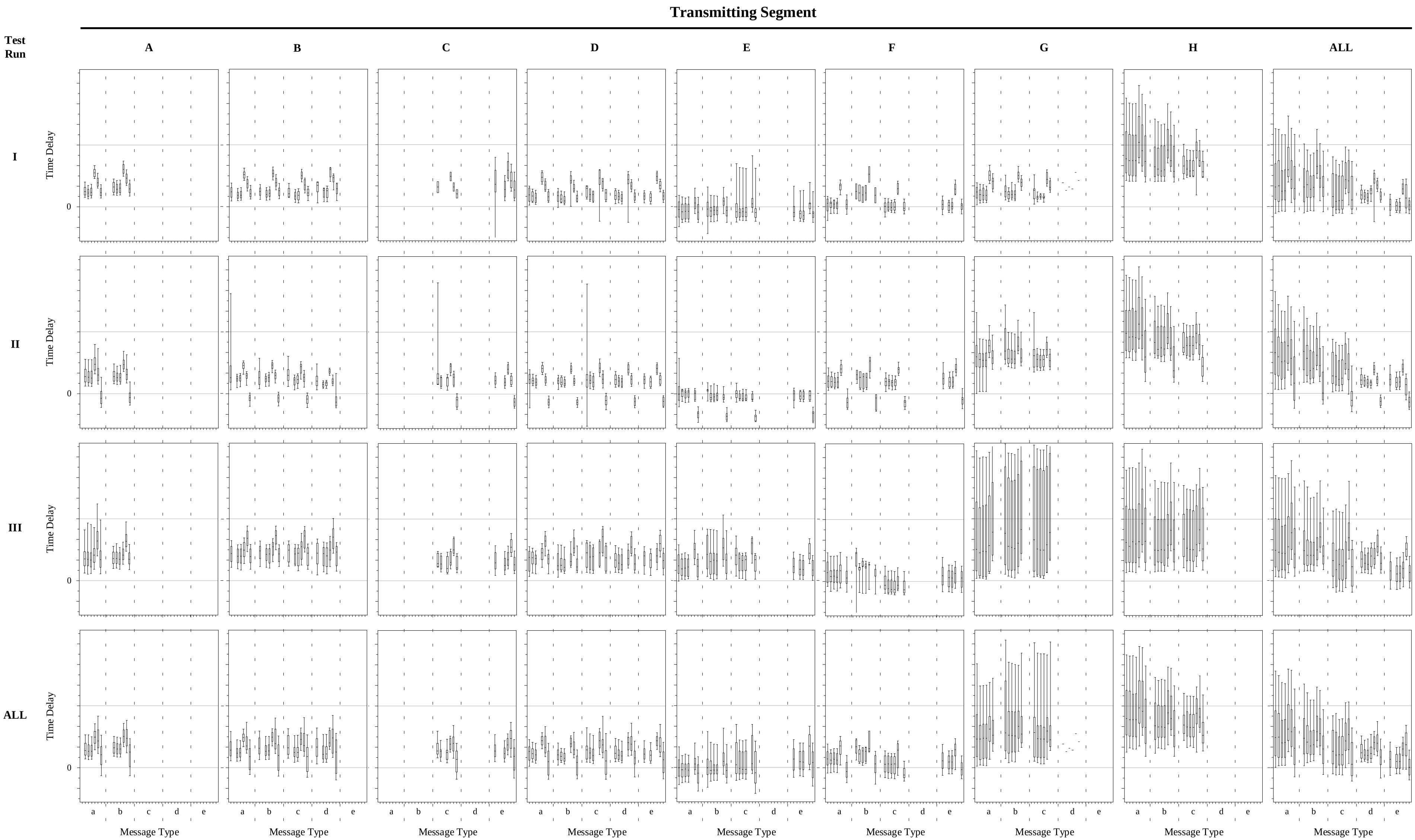
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 8. Time Delays on Run II by Type Delay, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



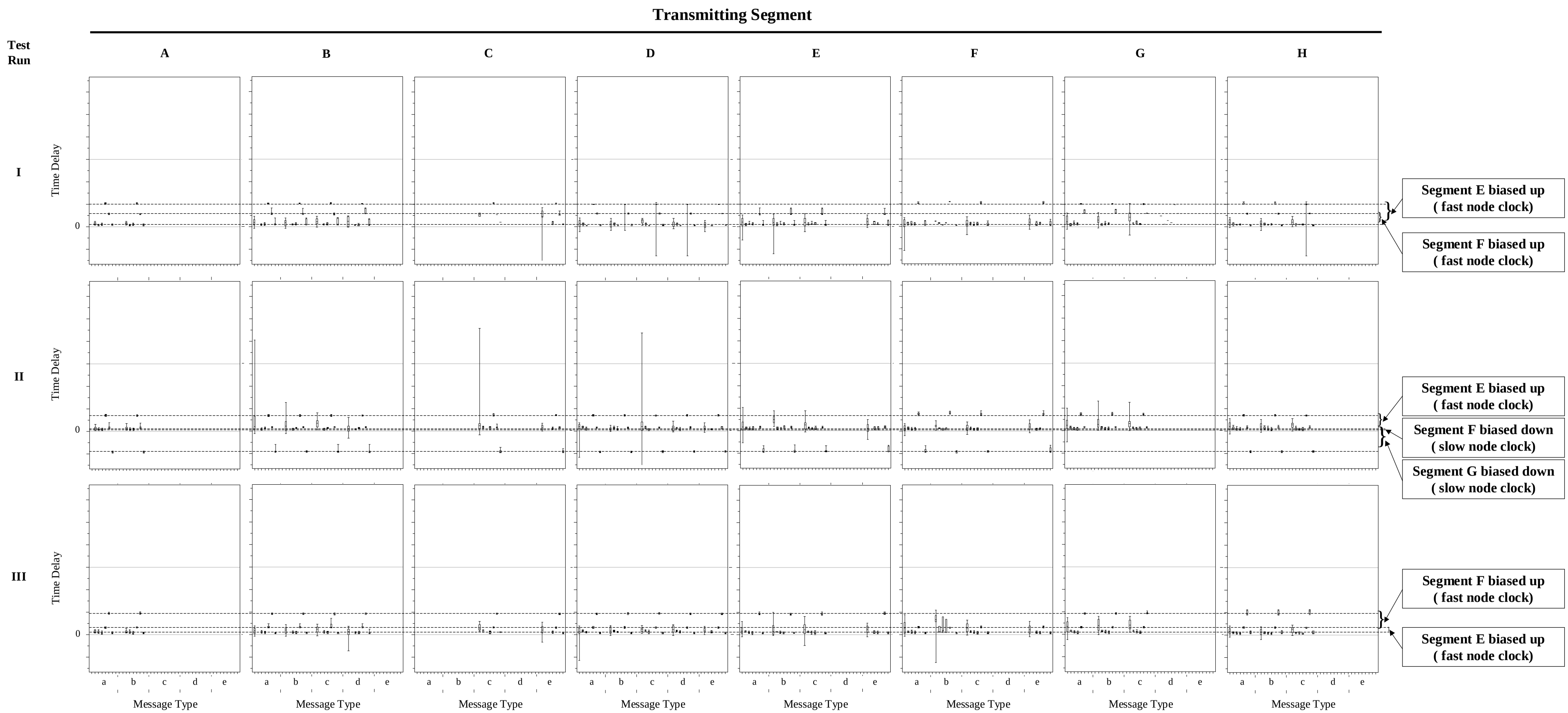
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 9. Time Delays on Run III by Type Delay, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Adjusted for Apparent Node Clock Errors.



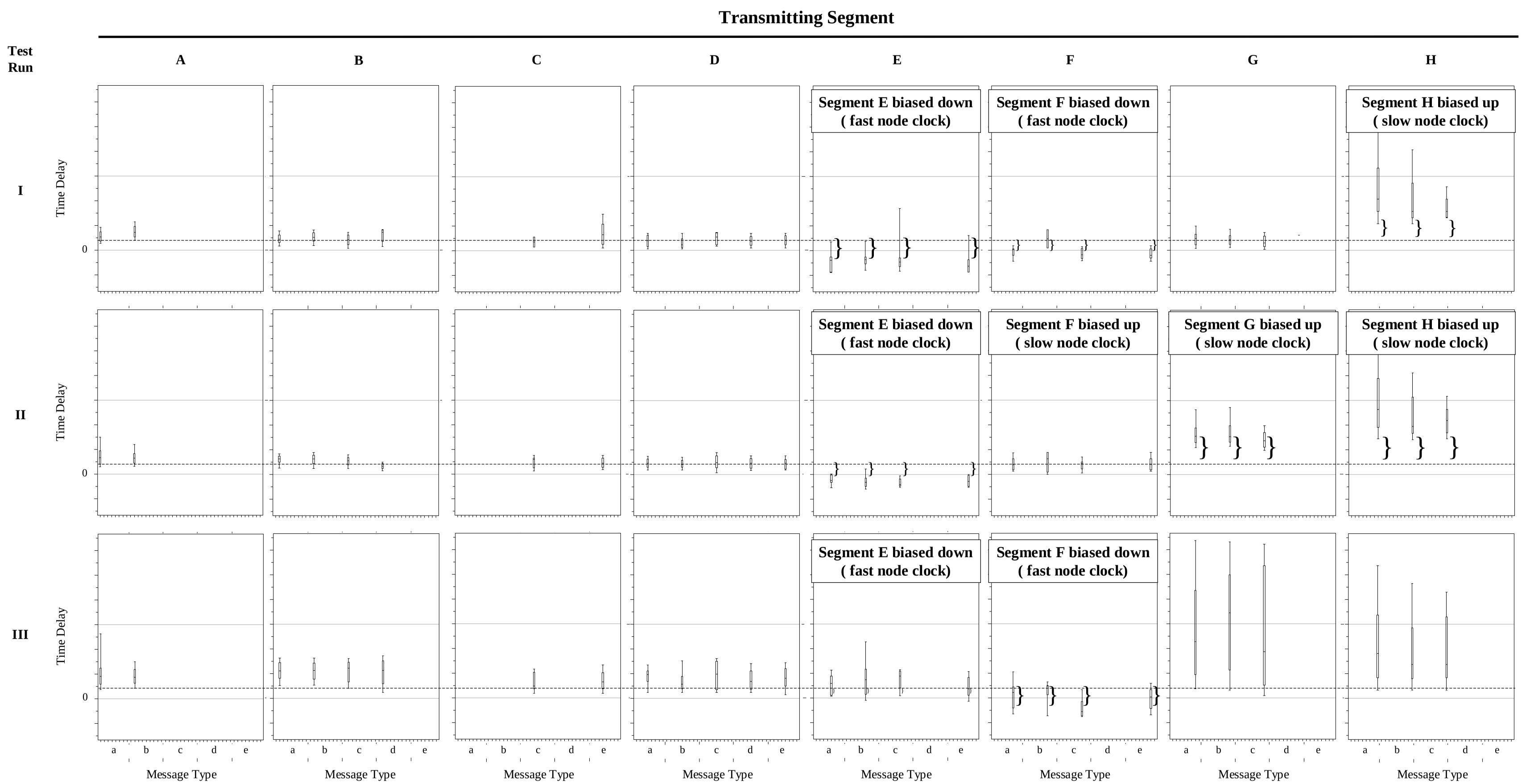
Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 10. Transmit to Receive Time Delays by Test Run, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Not Adjusted for Apparent Node Clock Errors.



Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.
Within message type, left to right receiving segment order is: A, B, C, D, E, F, G.
Delays corresponding to Segment H receipts are not displayed, and the sending node for each cell has no receiving latency distribution except node to gateway.

Figure 11. Gateway to Node Time Delays by Test Run, Type, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Not Adjusted for Apparent Node Clock Errors.



Within message type, the "whiskers" on the boxplots extend from the 10th to the 25th percentile and the 75th to the 90th percentile of the latency distribution for each receiving segment; the 50th percentile (median) is marked with a horizontal line.

Figure 12. Node to Gateway Time Delays by Test Run, Type, Transmitting Segment, Message Type and Receiving Segment (80% Boxplots). Times Not Adjusted for Apparent Node Clock Errors.