



Proceedings of the Eighth Annual U.S. Army Conference on Applied Statistics, 30 October – 1 November 2002

Barry A. Bodt, Edward J. Wegman
EDITORS

Hosted by:
Army Research Office

Cosponsored by:
U.S. ARMY RESEARCH LABORATORY
TRADOC ANALYSIS CENTER—WHITE SANDS MISSILE RANGE
WALTER REED ARMY INSTITUTE OF RESEARCH
UNIFORMED SERVICES UNIVERSITY OF THE HEALTH SCIENCES

Cooperating Institutions:
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 Eighth Annual U.S. Army Conference On Applied Statistics, 30 October – 1 November 2002

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

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

Hosted by:
U.S. Army Research Office

Cosponsored by:
U.S. Army Research Laboratory
TRADOC Analysis Center—White Sands Missile Range
Walter Reed Army Institute of Research
Uniformed Services University of the Health Sciences

TABLE OF CONTENTS

EIGHTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

Abstract and Foreword.....	viii
Short Course	
Statistical Data Mining	
Edward J. Wegman, Jeffrey L. Solka.....	1
General Session 1	
Improving Confidence Intervals for Proportions, Differences of Proportions, and Odds Ratios (Abstract)	
Alan Agresti (Keynote).....	2
Special Session 1	
Skew-Elliptical Distributions and Their Applications (Abstract)	
Marc. G. Genton.....	2
Time Series Models for Zero-Inflated Data with Applications to Nuclear Power Plants (Abstract)	
Sujit Ghosh.....	2
A Simple Boolean Model for Assessing Particle Flow from Type II Counter Data (Abstract)	
Jason A. Osborne.....	3
Testing Monotonicity of Regression (Abstract)	
Subhashis Ghosal.....	3
Special Session 2	
Development of Statistical Software (Abstract)	
James Gentle.....	4
A Tour of New Directions in SAS Statistical Software (Abstract)	
Robert Rodriguez.....	4
Using SAS/IM: Workshop for Dynamic Statistical Graphics (Abstract)	
Simon L. Smith.....	4

Contributed Session 1

Efficient Simulation Experimental Designs (Abstract)	
Thomas M. Cioppa, Thomas W. Lucas.....	5

Estimating Probabilities for Simulation (Abstract)	
D. H. Frank.....	5

A Space-Time Model That Combines Simulations and Flight Test Data for Assessing Missile Launches from the F-22 Prototype Aircraft (Abstract)	
Dave Higdon, Mark McNulty, Bruce Lettallier.....	6

Contributed Session 2

Evaluation of the Mental Health Impact of 9/11: Results from the Pentagon Post Disaster Health Assessment Survey (Abstract)	
Nikki Jordon.....	6

Performance Evaluation of Temporal Alerting Algorithms for ESSENCE Syndromic Surveillance Data (Abstract)	
Eugene Elbert, Howard S. Burkom, Kevin Nelson.....	7

Effects of Missing or Incomplete Data on Military Medical Research (Abstract)	
T. E. Powers, Y. Li, L. B. Trofimovich.....	8

Contributed Session 3

A Comparison of the Outcomes of a Course Taught in the Traditional Classroom Setting with the Outcomes of the Same Course Taught in the Distance Learning Format (Developing a Methodology) (Abstract)	
Gene Dutoit.....	9

A Probability Programming Language (Abstract)	
Andrew Glen, Diane Evans, Larry Leemis.....	9

A Lemma Useful in Combinatorics and Some of its Applications (Abstract)	
Bernard Harris.....	10

Clinical Session 1

Multivariate Goodness-of-Fit Testing for M256 Gun Tube Profiles (Abstract)	
David W. Webb, Mark L. Bundy.....	10

Study Design to Assess Autonomous Mobility of the Experimental Unmanned Vehicle (XUV)	
Barry A. Bodt, Ann E. M. Brodeen.....	11

Contributed Session 4

New Binomial and Multinomial Distributions from Graph Theory	
Milton Sobel, Yontha Ath.....	13

Class Cover Digraphs for Latent Class Discovery	
J. L. Solka, C. E. Priebe, D. J. Marchette.....	15

Encoding of Text to Preserve “Meaning”	
Angel R. Martinez, Edward J. Wegman.....	27

General Session 2

A Study of Denial of Service Attacks on the Internet	
David J. Marchette.....	41

Wilks Award Banquet Address

Considerations of Inspection for Homeland Security with Cross Linkages to Quality Control, Game Theory, and Stochastic Simulation (Abstract)	
James R. Thompson.....	61

General Session 3

Stress-Strength Testing: Some Classical Approaches and Some New Formulations and Results (Abstract)	
Francisco J. Samaniego.....	62

Random Disambiguation Paths (Abstract)	
Carey E. Priebe.....	62

Contributed Session 5

Benefits of Non-Destructive Evaluation (NDE) vs. Destructive Testing for Bayesian Reliability Estimation (Abstract)	
Paul Deininger, Shane Reese, Michael Hamada, Robert Krabill.....	64

Munitions Stockpile Reliability Assessment (Abstract)	
Alyson Wilson, Nicholas Hengartner.....	64

Hierarchical Models for Software Testing and Reliability Estimation (Abstract)	
Todd L. Graves, John C. Kern II.....	65

Contributed Session 6

- An Almost Natural Application of Bayesian Statistics in Packaging Quality Control
John A. Wasko, David B. Kim.....67

- Relationship between Toxicity Values for the Healthy Subpopulation and the General Population
Ronald B. Crosier, Douglas R. Sommerville.....76

- Finding the Season Effect and Trend of Attrition: Detect the Attrition Changes in the Early Stage (Abstract)
T. E. Powers, Y. Li.....86

Contributed Session 7

- Determination of the LD50 for Chemical and Biological Threat Agents (Abstract)
Nancy A. Niemuth.....87

- Homogeneity of the Loss Rate and Individual Factor Effect Across MEPS: A Meta-Analysis on Attrition (Abstract)
Y. Li, T. E. Powers.....87

- Relationship between the Dose-Response Curves for Lethality and Severe Effects for Chemical Warfare Nerve Agents
Douglas R. Sommerville.....89

- A Method for Assessing Randomness in the United States Army's Biochemical Testing Program (Abstract)
Kevin P. Romano.....105

Contributed Session 8

- The Analytic Challenges of the Army's Network-Enabled Future Combat Systems (Abstract)
Duane E. Brucker, Paul J. Deason.....106

- Threat Management Using Passive Inference of Network Infrastructure Topology (Abstract)
John Rigsby, Jeff Solka.....107

- A Statistical Methodology for Automatic Target Recognition in Satellite Imagery
John Bart Wilburn.....108

- Finding Clusters (Abstract)
Jon R. Kettenring.....116

Special Session 3

Assessing Uncertainty in Mesoscale Numerical Weather Prediction (Abstract)	
Montserrat Fuentes, Adrian Raftery.....	117
Local Probability Propagation Algorithms for Approximate Inference in Graphical Models (Abstract)	
Martin Wainwright, Tommi Jaakkola, Alan Willsky.....	117
C4ISR and the Future Force (Abstract)	
Monica Farah-Stapleton.....	118

Particle Filtering and Spatial Prediction in the Battlespace (Abstract)	
Noel Cressie, Mark Irwin, John Kornak.....	118

General Session 4

A Microarray Lesson from Dear Old Dad (Design-Analyze-Display) (Abstract)	
Russell Wolfinger.....	119

Contributed Session 9

Statistical Techniques for Breaking Steganography (Abstract)	
R. Chandramouli.....	120

Classifier Optimization via Graph Complexity Measures	
J. L. Solka, D. A. Johannsen.....	121

Statistical Classification Based on Contours (Abstract)	
Mark Fitzgerald, Karen Kafadar.....	136

Contributed Session 10

A Human Dimension Methodology for Assessing Future Combat Systems' C4ISR (Abstract)	
Jock O. Grynivicki, Kragg P. Kysor.....	137

Assessing and Removing Unexpected Collinearity in Designed Experiments (Abstract)	
Trevor A. Craney.....	137

General Session 5

SiZer for Simple, Direct Inference in Exploratory Data Analysis (Abstract)	
Steve Marron.....	138

Author Index	139
------------------------------	-----

EIGHTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

ABSTRACT

The eighth U.S. Army Conference on Applied Statistics was hosted by the United States Army Research Office (ARO), during 30 October – 1 November 2002 on the campus of North Carolina State University. The conference was cosponsored by the U.S. Army Research Laboratory (ARL), the U.S. Army Research Office, 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), and the Uniformed Services University of the Health Sciences (USUHS). Cooperating organizations include Los Alamos National Laboratory (LANL), George Mason University (GMU), the 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 this conference and fifty-one papers were given. This document is a compilation of available papers offered at the conference.

FOREWORD

The eighth U.S. Army Conference on Applied Statistics was hosted by the United States Army Research Office, during 30 October – 1 November 2002 on the campus of North Carolina State University. The conference was cosponsored by the U.S. Army Research Laboratory (ARL), the U.S. Army Research Office (ARO), the United States Military Academy (USMA), the Training and Doctrine Command (TRADOC) Analysis Center-White Sands Missile Range, the Walter Reed Army Institute of Research (WRAIR), and the Uniformed Services University of the Health Sciences (USUHS). Cooperating organizations include Los Alamos National Laboratory (LANL), George Mason University (GMU), the 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 eighth conference was preceded by a two-day short course, “Statistical Data Mining,” given by Edward Wegman of George Mason University and Jeff Solka of the Naval Surface Warfare Center. Robert Launer of ARO opened the conference. Several distinguished speakers spoke during invited general sessions: Alan Agresti (keynote), University of Florida; David Marchette, Naval Surface Warfare Center; Francisco J. Samaniego, University of California, Davis; Carey Priebe, Johns Hopkins University; and Russell Wolfinger, SAS Institute. In addition to outstanding invited speakers, three special sessions were featured at the conference: Statistical Research at North Carolina State University (organized by Leonard Stefanski, North Carolina State University), Statistical Software (organized by Edward Wegman, GMU), and Command Control and Communication (organized by Wendy Martinez, ONR). Thirty-three contributed papers rounded out the program.

An important moment in the conference was the awarding of the Army Wilks Medal to Eugene Dutoit of Troy State University and a recently retired Army civilian from the Dismounted Battle Space Battle Lab at Fort Benning, Georgia. Dr. Dutoit was honored for the years of service in statistical application for the Army.

The Executive Board for the conference recognizes Robert Launer, ARO, for hosting the conference and Edmund Baur, ARL, for maintaining the conference web site, David Webb, ARL, for handling many administrative details, Jock Grynovicki, ARL, for chairing the conference, Edward Wegman, GMU, for assembling the conference proceedings, and Barry Bodt, ARL, for chairing the conference.

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

EIGHT U.S. ARMY CONFERENCE ON APPLIED STATISTICS

SHORT COURSE Statistical Data Mining

Dr. Edward J. Wegman

George Mason University

Dr. Jeffrey L. Solka

Naval Surface Warfare Center



Abstract: Used to extract knowledge hidden from large volumes of raw data, data mining has been applied in recent years to a wide variety of research areas such as medical imaging, high-energy physics, consumer behavior, atmospheric sciences, and security and surveillance. Data mining is an extension of exploratory data analysis and has basically the same goals, the discovery of unknown and unanticipated structure in the data. The chief distinction between the two topics resides in the size and dimensionality of the data sets involved. Data mining in general deals with much more massive data sets for which highly interactive analysis is not fully feasible. In this course we shall discuss the scales of data set sizes and the limits of feasibility for the various data set sizes. We will introduce some visualization tools and indicate how they may be used to accomplish data mining tasks. We shall review some structure finding algorithms including: density estimation and bump hunting; clustering and classification; visual clustering strategies; CART and related methods; time domain time series methods; nonparametric regression including convolution, LOESS and ridges and skeletons methods will be illustrated with application to several data sets. Particular emphasis will be placed on visualization techniques. The course will cover basic techniques used in visual data mining, including parallel coordinates, grand tour, and saturation brushing. These techniques will be illustrated in further discussions on rapid data editing, density estimation, inverse regression, tree-structured decision rules, classification and clustering, structural inference and outlier investigation.

[**Part 1**](#) | [**Part 2**](#) | [**Part 3**](#) | [**Part 4**](#) | [**Part 5**](#) | [**Part 6**](#)

[**Part 7**](#) | [**Part 8**](#) | [**Part 9**](#) | [**Part 10**](#) | [**Part 11**](#)

EIGHTH U.S. ARMY CONFERENCE ON APPLIED STATISTICS

General Session 1

Improving Confidence Intervals for Proportions, Differences of Proportions, and Odds Ratios

Alan Agresti, University of Florida

Abstract: “Exact” small-sample methods for categorical data are exact in term of using probability distributions that do not depend on unknown parameters. However, they are conservative inferentially, having actual error probabilities for tests and confidence intervals that are bounded above by the nominal level. We examine the conservatism for interval estimation and suggest ways of reducing it, illustrating for the binomial proportion, the difference between two proportions, and the odds ratio. We also summarize simple ways of adjusting standard large-sample confidence intervals to improve dramatically their small-sample performance. The standard intervals for proportions and their differences have poor performance, the actual confidence level often being much lower than the nominal level. Simple adjustments based on adding four pseudo observations, half of each type, perform well even for small samples with interval estimation of the proportion and the difference of independent or dependent proportions.

Special Session 1

Skew-elliptical Distributions and Their Applications

Marc G. Genton, North Carolina State University

Abstract: This talk introduces generalized skew-elliptical distributions (GSE), which include the multivariate skew-normal, skew-t, skew-Cauchy, and skew-elliptical distributions as special cases. GSE are weighted elliptical distributions but the distribution of any even function in GSE random vectors does not depend on the weight function. In particular, this holds for quadratic forms in GSE random vectors. This property implies that standard inferential methods might be misleading when applied to time series and spatial processes with GSE distributions. However, the same property is beneficial for inference from non-random samples. Several applications are presented for illustration.

Time Series Models for Zero-Inflated Data with Application to Nuclear Power Plants

Sujit Ghosh, North Carolina State University

Abstract: Statistical methods for analyzing count data with excess zeros are very important in various scientific fields. For instance in a data set presented in Martz et al. (1999), it is observed that the number no ‘scrams’ out of sample of 66 nuclear power plants has increased from 1.5% in 1986 to 33.3% in 1993. The goal of their study was to find if there was any plant specific annual trend in the rate at which unplanned scrams occur. The objective of this talk is to develop, assess, and provide convenient tools for implementing flexible models for time dependent zero-inflated data. A class of

hierarchical models has been found to be useful in modeling time dependence and heterogeneity (e.g. due to several nuclear plants). Some preliminary results and illustrations will be presented based on a data set obtained from the U.S. Nuclear Regulatory Commission (NRC) annual reports.

A Simple Boolean Model for Assessing Particle Flow from Type II Counter Data
Jason A. Osborne, North Carolina State University

Abstract: For aerial application of granular fertilizers and pesticides, knowledge of the amount of material flowing from the aircraft can help guide even distribution over a field. Grift (2001) has developed a system to help measure mass flow in aerial spreader ducts. The system involves an optical sensor that turns off and on as particles flow through the duct, providing clump-length measurements. Clumps occur when multiple particles pass the sensor before a space in between particles is encountered. If particle arrivals constitute a Poisson process, lengths of particles are independent of this arrival process and particles begin passage through the duct immediately upon arrival, the system forms an infinite-server M/G queue. Of interest is mass flow, or the number of particles, which pass through the duct for application to the target region below. Method-of-moment and likelihood-based approaches towards estimation of this quantity from clump-length data are developed.

Testing Monotonicity of Regression
Subhashis Ghosal, North Carolina State University

Abstract: In many situations encountered in practice, the relationship between two variables of interest is expected to follow some order restriction. In economics, the relationship between, for instance, income and expenditure on some specific item such as food or housing is likely to be monotonically increasing, although natural constraints rule out a linear relationship. A similar situation also arises in bio-medical or nutrition studies, where the response variable may be expected to increase with the dose of a drug, until the drug becomes toxic. Another familiar hypothesis is that blood pressure increases with salt consumption. Although estimation under the monotonicity constraint has been well addressed in the literature, the issue of testing the assumption of monotonicity has only begun to receive attention. Bowman, Jones and Gijbels (1998) construct a test based on the idea of the critical bandwidth, while Hall and Heckman (2000) test the hypothesis by testing the positivity of slope for nearest neighbor linear regression. In our approach, we construct test statistics as functionals of a natural U process obtained by locally calculating Kendall's measure of discordance. If monotonicity holds, discordances should be minimum, and so the process should lie below the level 0. Therefore natural test statistics may be obtained by looking at the maximum of the process or the time spent by the process above a certain level. We obtain the asymptotic distribution of the test statistics through linearization of U-statistics, strong approximation methods and extreme value theory for stationary Gaussian processes. Our tests are consistent against the general alternatives and also have reasonably high power. One considerable advantage of our method is that the limiting distribution of the statistics are explicitly obtained, so we

do not have to depend on computationally intensive bootstrap sampling required by the other methods.

The talk will be based on the paper Ghosal, S., Sen, A. and van der Vaart, A. W. (2000). Testing monotonicity of regression. Annals of Statistics, vol 28, No 4, pages 1054--1082.

Special Session 2

Development of Statistical Software **James Gentle, George Mason University**

Abstract: In the past few years, major changes have occurred in how statistical software is developed and distributed. The open source movement has encouraged the participation of a much large set of people in the development. The Internet has allowed for wide and rapid distribution of new software.

A Tour of New Directions in SAS Statistical Software **Robert Rodriguez, SAS Institute**

Abstract: Statistical software in SAS is expanding in a variety of new directions, which are motivated by methodological advances, changes in computing technology, and requests from applied statisticians in many fields. In Version 9, these directions include multiple imputation for missing data, survey data analysis, power and sample size computation, nonparametric modeling, robust regression and outlier detection, and statistical graphics. Heavy-duty analytic procedures are being multithreaded for scalable performance on servers with multiple CPUs, and a number of traditional statistical tools have been enhanced.

Using SAS/IML Workshop for Dynamic Statistical Graphics **Simon L. Smith, SAS Institute**

Abstract: SAS/IML Workshop is a Windows-based programming environment for high-end data analysis that provides the user with the flexibility to combine matrix computing, statistical modeling, and dynamic graphics. Programs are written using an extended version of the Interactive Matrix Language (IML) in SAS. This presentation will demonstrate the power of creating dynamic displays using results from SAS statistical procedures. Examples will show how the user can subset data graphically using local selection, which offers a more powerful version of dynamic brushing for complex multivariate data integrate geographical data with statistical models create graphical displays of statistical results such as fits and outlier diagnostics link diagnostic results back to the data.

Contributed Session 1

Efficient Simulation Experimental Designs

LTC Thomas M. Cioppa, TRADOC Analysis Center; Thomas W. Lucas, Naval Postgraduate School

Abstract: The Department of Defense uses complex high-dimensional simulation models as an important tool in its decision making process. To improve on our ability to efficiently explore larger subspaces of these models, we develop a set of experimental designs for searching over as many as 22 variables in as few as 129 runs. These new designs combine orthogonal Latin hypercubes and uniform designs to create designs having near orthogonality and excellent space-filling properties. Multiple measures are used to assess the quality of candidate designs and to identify the best one. For situations in which more than the minimum number of required runs are available, the designs can be permuted and appended to create additional design points that improve upon the design's orthogonality and space-filling.

The designs are used to explore two surfaces. For a known 11-dimensional stochastic response function containing nonlinear and interaction terms, it is shown that the near orthogonal Latin hypercube is substantially better than the orthogonal Latin hypercube in estimating model coefficients. The other exploration uses the agent-based simulation MANA to analyze 22 variables in a complex military peace enforcement operation. The need for maintaining the initiative and speed of execution during these operations is identified.

Estimating Probabilities for Simulation

D. H. Frank, Indiana University of Pennsylvania

Abstract: In simulating a battle between opponents 1 and 2 we need to estimate the probability of 1 defeating 2, $p_{1,2}$. Usually we have empirical evidence from prior confrontations between 1 and 2 to get a statistical estimate. But we may have data to estimate p_1 , the probability that 1 defeats a randomly chosen opponent and also p_2 . For example in sports we may wish to simulate a bowl game between two teams who have never faced each other but have seasonal records to estimate p_1 and p_2 .

In this paper we examine 3 estimates of $p_{1,2}$ as functions of p_1 and p_2 based on pseudo probability arguments. We examine these 3 methods based on certain desirable properties and also try to test the goodness of these with a chi-square type statistic. We examine cases based on known probabilities and simulation case based on estimates.

It is possible to somewhat reverse the process and use estimates of $p_{1,2}$ to improve estimates of p_1 in cases where this is contact between 1 and 2. This would lead to ranking college football things among other applications

A Space-time Model that Combines Simulations and Flight Test Data for Assessing Missile Launches from the F-22 Prototype Aircraft.

Dave Higdon, Mark McNulty, and Bruce Lettallier; Los Alamos National Laboratory

Abstract: The US Air Force must test new aircraft in order to certify that it performs to pre-set specifications. One component of this certification process is missile separation - ie. can the plane safely launch a missile without the missile knocking off its nose or a piece of wing? Before ever conducting an in-flight missile launch, numerous launch simulations are carried out on a scale model in a wind tunnel. Many factors affect the actual missile trajectory.

Particularly important factors are the altitude, mach number, dynamic pressure, and angle of attack of the aircraft. We develop a hierarchical space-time model that incorporates these wind tunnel simulations, expert judgment, and several actual flight tests to give a predictive model for missile trajectories as a function of the flight factors: altitude, mach number, dynamic pressure, and angle of attack. This is joint work with Mark McNulty and Bruce Letellier of Los Alamos National Lab.

Contributed Session 2

Evaluation of the Mental Health Impact of 9/11: Results from the Pentagon Post Disaster Health Assessment Survey

Nikki Jordon, USACHHPM

Abstract: BACKGROUND: In the aftermath of September 11, 2001, the Pentagon Post Disaster Health Assessment (PPDHA) survey was created to identify healthcare needs/concerns among Pentagon personnel and assure that appropriate care/information was provided. Fundamental in this assessment was the evaluation of the mental health impact due to the attack.

METHODS: The PPDHA incorporated a short screening instrument covering mental health symptom domains, mental health functioning and possible predictive risk factors. High-risk groups for Post Traumatic Stress Disorder (PTSD), depression, panic attacks, generalized anxiety, and alcohol abuse were determined; predictive factors believed to be associated with risk groups were assessed through both univariate analysis and logistic regression; and validation of risk groups was assessed across functional levels using similar statistical methods.

RESULTS: A total of 19,450 Pentagon employees were asked to complete the survey; 4,739 responded, representing approximately 25% of the population. Overall, 1,838 (40%) of respondents met the screening criteria for being at high risk for any of the symptom domains of interest: PTSD (8%), depression (18%), panic (23%), generalized anxiety (27%), and/or alcohol abuse (3%). Mental health risk groups were found to strongly correlate with reduced daily functioning ($OR=14.4$, 95%CI: 11.9-17.4) and use of counseling services ($OR=4.2$, 95%CI: 3.6-5.0). In addition, risk factors known to be

associated with mental health problems following traumatic events were found to be strongly predictive of the high-risk categories identified.

CONCLUSION: These data suggest that the approach used within the survey had validity, and that the short mental health questionnaire could serve as a prototype for the rapid public health assessment of the mental health impact of future traumatic events.

Performance Evaluation of Temporal Alerting Algorithms for ESSENCE Syndromic Surveillance Data

Eugene Elbert, Howard S. Burkom, and Kevin Nelson; Walter Reed Army Institute of Research

Abstract: The U.S. Department of Defense Global Emerging Infections System (DoD-GEIS) has developed the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) to enable outbreak alerting using syndromic surveillance. The ESSENCE system monitors over 100 primary care and emergency clinics in the National Capital Area, and since the terrorist attacks in September 2001, approximately 100,000 per day are collected several times daily from military treatment facilities (MTF) worldwide. These data include active duty forces as well as other TRICARE beneficiaries. Analysts from DoD-GEIS and the Johns Hopkins Applied Physics Laboratory have implemented statistical alerting algorithms enabling prompt notification of anomalous data counts.

Syndromic surveillance facilitates the monitoring of this large volume of data. For each MTF, we classify diagnoses resulting from outpatient and emergency room visits according to seven syndrome groups. Each group is defined by a list of ICD-9 codes. The ESSENCE system increments the count for a syndrome group each time a diagnosis code falls in the corresponding list. For each MTF, we apply temporal alerting algorithms to the data streams corresponding to each syndrome. The utility of these algorithms depends on their detection performance; they must be sensitive but with low false alarm rates. We have developed these algorithms according to the customary behavior of the syndromic data streams. In the ESSENCE II system under development, these algorithms will also be applied to various nontraditional data sources in both military and civilian sectors.

For syndromes of more common conditions and/or for larger MTF populations, the data streams typically have structure characterized by seasonal and weekly features. We have implemented modeling methods to exploit this behavior. To avoid overfitting, we apply tests of trend and serial correlation to determine whether modeling is applicable and, if so, which techniques to apply. Alerting methods based on regression and autoregressive (AR) models include categorical variables for known systematic features such as holiday effects. These methods also feature data cleaning procedures to avoid modeling based on outliers.

For more rare syndromes and/or smaller populations, data streams appear more random, and direct statistical tests are used for alerting. We have compared a variety of methods,

all depending on estimates of the data stream variance. Determining the length of data history required for such tests involves a tradeoff between stability and stationarity.

We have developed methods of performance analysis of these algorithms specific to the surveillance problem. This work has generalized ROC methodology to methods recently adopted in the field of data mining. For this purpose, Signals designed to emulate an outbreak epicurve are added to the authentic data streams for this analysis. This approach permits investigation of how small an outbreak is detectable, how soon alerting can be expected in an outbreak, and other related issues.

Effects of Missing or Incomplete Data on Military Medical Research

TE Powers, Y Li, and LB Trofimovich; Walter Reed Army Institute of Research

Background: Increasingly, military planners and policymakers are incorporating statistical and epidemiological research into decision-making. Statistics on the numbers of applicants for military service, the numbers in training, illness rates and attrition rates are available from several different sources. Like virtually all large-scale data systems, however, the military data sets on which this research is based are often incomplete. An added complication is that the more data that are missing in a database, the greater the need is to address the problem of incomplete cases, yet those are precisely the situations where imputing or filling in values for the missing data points is most questionable due to the small proportion of valid data points relative to the size of the data matrix. The degree to which and manner in which such incompleteness may impact results of military epidemiologic research is examined.

Types of missing data

The most appropriate way to handle missing or incomplete data will depend upon how data points became missing. Little and Rubin (1987) define three unique types of missing data mechanisms.

- A. Missing Completely at Random (MCAR): Cases with complete data are indistinguishable from cases with incomplete data.
- B. Missing at Random (MAR): Cases with incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other variables in the database rather than being due to the specific variable on which the data are missing.
- C. Nonignorable: The pattern of data missingness is non-random and it is not predictable from other variables in the database.

In this presentation, we will study the types of missing data in the military database, and discuss/suggest the statistical methods of handling missing data in the military data. We will give a list of methods for handling missing data appear below. This list is not exhaustive, but it covers some of the more widely recognized approaches to handling databases with incomplete cases.

Contributed Session 3

A Comparison of the Outcomes of a Course Taught in the Traditional Classroom Setting with the Outcomes of the Same Course Taught in the Distance Learning Format (Developing a Methodology)

Gene Dutoit, Troy State University

Abstract: Distance learning is an education method where the students and instructors do not meet in the same classroom setting. Some students prefer these types of classes because they can work at their own pace and time at home without leaving their families to attend classes held in the traditional setting. Troy State University, and several other colleges and universities, are in a partnership with the United States Army in an educational service called eArmyU Access Online and offers degrees through distance learning to soldiers stationed all over the world. The method of instruction is use of the Internet. Distance learning courses are designed to have the same academic standards as the traditional courses and students must meet the same requirements as the traditional form of the course.

This paper presents a case study and methodology for comparing the outcomes of a course taught in the distance-learning format with the same course taught in the traditional mode of in-class instruction. It is a template for further analysis to insure that Army distance learning students receive the same level of instruction as the traditional students. It compares the outcomes for the same course for four consecutive terms and compares the outcomes on parallel forms of examinations. The statistics of item analysis are used to identify differences in outcomes and sources of student difficulties. The results of item analysis and test analysis are used to provide feedback to students within a class and in follow-on classes. A course instructional learning curve is then developed as a tool to be used by the course instructors.

A Probability Programming Language

LTC Andrew Glen, United States Military Academy; Diane Evans, Rose-Hulman University; Larry Leemis, The College of William and Mary

Abstract: A probability programming language (APPL) is presented. Statisticians have traditionally used statistical packages, such as SPSS, Splus and SAS, to analyze large data sets. However, symbolic algebra languages, such as Maple and Mathematica, allowed the development of a “probability package” capable of solving intractable probability problems involving the creation of new, often complicated distributions. The purpose of APPL is to encapsulate algorithms and generalized theorems used in probability into a programming environment with the computer algebra system Maple to provide the applied community with automated probability capabilities. The advantage of such software allows for finding exact distributions in lieu approximations, often producing new distributions with highly desirable properties. Automated functions in APPL include the following operations on random variables and distributions: convolutions, products, transformations, truncations, plots of CDF PDF HF SF and CHF, percentiles, expectations, order statistics, piecewise functions (e.g. the triangular

distribution), minimum distributions, maximum distributions, bootstrap distributions. A demonstration of the language and a short summary of the resultant advances in research as well as teaching Mathematical Statistics is presented. Applications that encompass a wide range of applied topics including goodness-of-fit testing, probabilistic modeling, central limit theorem augmentation, generation of mathematical resources, and estimation are presented. Copies of this free software will be made available.

A Lemma Useful in Combinatorics and Some of its Applications
Bernard Harris, University of Nebraska, Lincoln

Abstract: The purpose of this report is to exhibit a summation formula, which despite the apparent simplicity of its statement, has substantial generality and a large number of applications in enumerative combinatorics.

Clinical Session 1

Multivariate Goodness-of-Fit Testing for M256 Gun Tube Profiles
David W. Webb, Mark L. Bundy; Army Research Laboratory

Abstract: An electronic database of gun barrel centerline measurements for the majority of M256 120mm gun tubes produced at Watervliet Arsenal in the last decade provides a very good estimate of the population of centerline profiles for the entire fleet of M256 gun tubes. In several recent 120mm tank ammunition studies, the selection of tubes based on their profiles has been an important issue, since the centerline profile is known to affect the center of shot impacts. In order to properly evaluate ammunition performance, test designers want assurance that the tanks chosen for their studies have centerline profiles that are representative of the entire fleet.

If centerline profiles were a univariate measurement, then a Kolmogorov-type goodness-of-fit (g.o.f.) test would be an appropriate method for testing if the sample of tanks proposed in a study has the same distribution as the population. However, the true centerline profile for a given tube is a continuous line through 3-dimensional space. In practice, the profile is measured as a discrete collection of ordered pairs. Each ordered pair is the displacement of the profile in the azimuth and elevation plane at one of 23 positions along the barrel. The current method used at the Army Research Laboratory to assess representativeness of the fleet only considers the ordered pairs of displacements from 4 of these positions, and analyzes these values independently using Kolmogorov's univariate g.o.f. test. Rejection of the null hypothesis in any one of the 8 tests is grounds for rejecting that sample of gun tubes as representative of the fleet.

We recognize that our current methodology has its shortcomings, mainly, that the data dependencies existing between neighboring stations are completely lost in our analysis. Of the panel, we wish to ask if a multivariate analog to the Kolmogorov test or some other multivariate g.o.f. test exists that would allow us to determine whether the complete profiles from a sample are representative of the fleet.

*Study Design to Assess Autonomous Mobility of the
Experimental Unmanned Vehicle (XUV)*

Barry A. Bodt and Ann E. M. Brodeen
U.S. Army Research Laboratory

Unmanned ground vehicles (UGVs) will provide scout functions for our forces on the future battlefield. A current study of the Experimental Unmanned Vehicle (XUV) seeks to demonstrate autonomous mobility. The study design attempts to balance military operational and development-related technical concerns, multiple sites, restrictions on randomization, and resource constraints to maximize the information quality and content resulting from testing. The principal design follows a split-split plot scheme, and this answers most questions. However, some additional testing is necessary to address subordinate issues. As in most field trials, trade-offs must be made between the statistical ideal and the practical reality. Considerations run the gambit of pooling, confounding, and nesting, and also the questions of fixed versus random factors and the advisability of using a portion of data in two separate analyses. The panel of experts to which this paper is presented will be asked to respond to these trade-offs with any guidance they have regarding the design or the analysis.

More information from Barry Bodt ...

Enclosed in this e-mail is a briefing on this design that I delivered on 23 October at Fort Indian Town Gap, PA before representatives from NIST, ARL, and General Dynamics. With the caveats mentioned in the briefing, the design as it stands has been fairly well received.

One specific area in which the panel might provide guidance is pooling. Douglas Montgomery's 2002 fifth edition text on Design and Analysis of Experiments suggests on page 536 that what terms given up to pooling might be determined by first testing the significance of the term—perhaps with a high alpha level, say 0.25. I don't have a feel for an appropriate approach, but some pooling will be required because we simply do not have enough degrees of freedom in the denominator for many of the tests. Some rough power computations, convince me I probably should have double the number of replicates I can afford in this test.

Another issue that has been brought up is the fact that the XUV/team factor is inseparable in the present design. That was a concession we made. The problem is that if we include XUV and Team as separate factors, to support randomization we may have to shuttle teams back and forth to the two test courses. Logistics cost us time and consequently runs. So this was a trade-off we agreed to. Still, if there was a way to block or in some other way cleverly arrange the design so that Team and XUV could be separated, it would be nice. Any suggestions along that line would be most welcome.

Ultimately, a second site will be tested. The way this is set up now, I would be driven to a split-split plot design. However, I need to be careful on the analysis. For example, the test course—even with specified difficulty level—is really nested within site. I hope not, but it is even possible that technical operators will be, at least, somewhat different the second time around. Alerting me to any landmines in the analysis would be very useful.

Another issue is the borrowing of runs from Tech-T1 in the principal experiment (12 to be exact) for use in another comparison involving soldiers and night conditions. That's not ideal either and if weather conditions change for the Tech-T1 runs from the principal experiment to Excursion 2, I am only going to have to settle for an unbalanced design. Randomization restrictions are not well accounted for either. Any thoughts on this are also welcome.

Currently, manned HMMWV runs occur on a separate day than the principal experiment involving the XUVs. Given the manned HMMWV's are the baseline, this is not ideal either. There is some thought about lengthening the work day to accommodate some manned runs in the morning of say, days 8 and 9, and perhaps in the evening on days 2-7. What we play against is fatigue on the part of the safety crew and the test administrators. That could also influence the end test result.

The responses in this study are geared toward autonomous mobility. The robot is supposed to carry out a mission, traveling to certain GPS designated points in three mission distance configurations. The real issue is, generally, how much help did it need to get there? The way it is addressed is in terms of the number of operator interventions necessary, the percentage of time the robot is truly autonomous, the number of emergency stops (e-stop) that safety invokes to protect the equipment, operator workload to keep track of the robot, and the percent of mission distance completed. It is possible the robot will get stuck or e-stopped and will not be able to continue.

There are other questions that you could focus on and I certainly don't expect you to address all of these in the time we will have. Moreover, you may choose one of the other questions I allude to in my briefing or another that has completely escaped me.

My final comment is to let you know this is a real test and the program is a very big deal in the Army robotics community. Never before have we tried to collect data that addresses military-operational considerations and never before has a test of this size been attempted. And, although the schedule is tight, testing does not begin until the 2nd of December, although a shake-out test at Fort Indian Town Gap will occur in mid November. NIST has been contracted to administer the test, with us involved, and so provides test directors, terrain assessment, and some personnel.

Contributed Session 4

New Binomial and Multinomial Distributions From Graph Theory

Milton Sobel, University of California, Santa Barbara; Yontha Ath, California State University, Dominguez Hills

Abstract: We'd like to apply statistical methodology to graph theory. Although we are not experts in the various applications of statistical methodology, we think that by bringing graph theory into the statistician's province we are opening up new areas of research.

On a non-directed graph the edges can be traversed in either direction. If the graph is simple there are no loops and only one edge between any pair of nodes. If the graph is regular then each node has the same number of edges (same, r) emanating from it. A step is the traversal of an edge from one node to another. In a random walk on a graph the various edges emanating from any node have probabilities adding to one; for the present we assume these are equal, so that for a regular graph of degree r the common probability is $\frac{1}{r}$ for going along any one edge.

At this point several combinatorial questions arise: In analogy with the usual binomial distribution where one can use a fixed sample size rule or on with random sample size, we first separate our graph random walk problems into those with a fixed number of steps and those with a random number of steps. Suppose we consider first the family of complete graphs K_n with a total of n nodes ($n = 1, 2, 3, \dots$). Let SP denote the starting node and NSP any other node. For SP (resp. NSP).

I. Fixed Number of Steps

1. What is the generating function (GF) for the number of visits to a specified node in a random walk on K_n with 5 steps. What are the results (mean, variance, etc) obtained from the above GF? What are the corresponding results for 10 steps? Can these be summarized by writing a general formula for any n ? Is there an analogous problem for traversing every edge of the graph (every traversal being in either direction)?

II. Random Number of Steps

2. What is the GF for visiting every node (resp., traversing every edge) in the graph (if the starting point is not regarded as a bona-fide visit) (NG is the non-gratis version; we use G for the gratis version)? A new node is simply one that hasn't been visited before.
3. What is the GF for visiting j new nodes in a graph?
4. What is the multinomial analogue for the joint distribution of visiting x_1 i times and x_2 j times if neither node is the SP?

5. What is the GF for the number of new nodes visited before returning to the SP (without counting the SP)?
6. What is the GF for the number of steps needed to visit a specific node (NSP) r times?

Besides the family of complete graphs K_n , we also consider (for some of the same goals as above) a number of additional families, which we list below.

1. The Complete Family (K_n)
2. Bipartite Family (CBP) (analog of the Binomial)
3. Wheel Graphs (W_n) with center and radial spokes
4. Circular Graphs (a closed regular graph of degree 2)
5. Circulant Graphs (Dividing a clock equally with diameters)
6. Familiar Geometric Shapes.

A variety of results are obtained for each of these families. Each problem calls for an expectation as well as a variance. For family #6 there are a few scattered results in the literature; we apologize for those that were not cited.

Class Cover Digraphs for Latent Class Discovery

J. L. Solka

C. E. Priebe

D. J. Marchette

Code B10

Mathematical Sciences Department

B10

NSWCDD

JHU

NSWCDD

Dahlgren, VA 22448

Baltimore, MD 21218

Dahlgren, VA 22448

Abstract

This paper examines the robustness of a new graph theoretic method of latent class discovery. This new method allows for the discovery of new latent classes within sets of observations residing in a high dimensional space. The robustness of the methodology is studied using a single gene expression data set.

Keywords

latent class discovery, gene expression data, statistical pattern recognition, graph theory, clustering

Introduction

One is often presented with a set of observations from various labeled classes for construction of a classifier. For the discussions within a classifier can be thought of as a mapping from our set of observations, residing in \mathcal{R}^n to a set of class labels $1, 2, \dots, j$. There are numerous ways that one may construct such a mapping. Some of the ways are based on estimating the unknown distributions of the observations that make up each of the individual classes while other methods merely require one to be able to estimate the discriminant boundary that separates the sets of observations.

Other times one is presented with a set of unlabeled observations and one is interested in identifying structural clusters within the group of observations. In this case one is not provided with any sort of class labels and one is interested in clustering the observations based on some other criteria. As can be imagined there are numerous criteria that one might use in order to cluster the observations or evaluate a clustering after the clustering has been obtained.

A more interesting case, which is the focus of the discussions within, is the case where we are provided labeled observations, for which we wish to create a classifier, but we are also interested in the identification of clusters or latent classes that reside within the labeled observations. In particular we are interested in the identification of said clusters based on the relationship of the observations to the discriminant boundary.

This is the focus of our discussions within. We examine the robustness of a graph theoretic method of latent class discovery. This method is essentially based on a clustering of observations obtained using a graph theoretic surrogate for their relationship to the discriminant boundary.

The paper is laid out as follows. First we provide the reader with some background material on the latent class discovery problem and our previous endeavors in this area. This section will also detail some of the interesting questions, such as the robustness of the discovered latent class structure to the derived classifier. In the next or results section, we will illustrate the application of this methodology to a gene expression dataset. In this case the identified latent classes correspond to previously identified disease types. We will also present some preliminary results in this section that attempt to characterize the robustness of the discovered latent class structure to the particular classifier that was employed. Finally we wrap up the paper with some discussions that illuminate our planned future activities. As with many of the papers that we write, we hope that the reader will be inspired to go forth and investigate some of the new strategies that might be sparked by the discussions outlined below.

Background

Given a set of possibly hyperdimensional observations we are interested in identifying any latent classes, from a discriminant analysis standpoint, that reside within the set of class labeled observations. The first step in the process is the construction of a graph theoretic classifier. This classifier construction methodology is predicated on the existence of a metric or pseudometric that measures the distances between the observations. Our previous work has illustrated how the choice of this distance measure may markedly effect the performance of the classifier but that is not the focus of our discussions within [Priebe et al., 2000]. The obtained models are then subjected to the latent class discovery process. The obtained latent classes are then analyzed using multidimensional scaling or nonlinear dimensionality reduction methods. Figure 1 provides a flowchart of our overall research strategy.

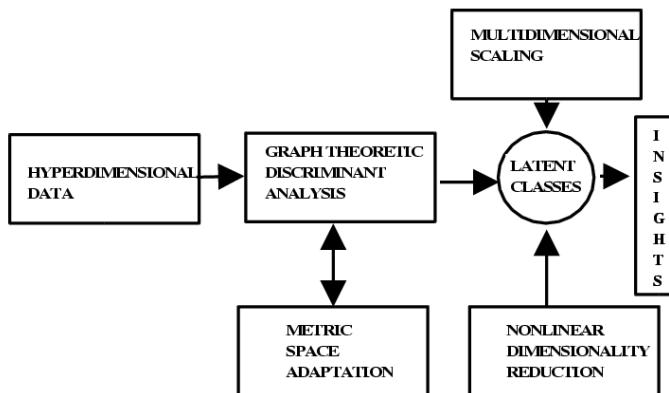


Figure 1: A flowchart detailing our latent class discovery research strategy.

One of the classification methods that can be employed in the latent class discov-

ery process is the class cover catch digraph (CCCD) method of Priebe [Priebe et al., 2003a]. The construction of this classifier begins by placing a ball about each of one of the class's observations. One then obtains a subset of the balls that cover the observations through a greedy based algorithm. If one considers each of the observations as a vertex in a graph and then produces a directed graph by drawing a directed edge from vertex a to vertex b if the ball centered on vertex a covers vertex b then the construction of this reduced cover can be cast as the solution of a dominating set. The reader is referred to [Chartrand and Lesniak, 1996] for a full discussion of domination in graphs.

Given this reduced set of balls that cover one of the class's observations the latent class discovery process proceeds as follows. We cluster the balls based on their radii. The exact manner in which the number of ball clusters is determined is discussed later. Once the balls are clustered then the discovered latent classes consist of those observations residing within a particular cluster of balls. In Figure 2 we illustrate the latent class discovery process using a toy problem. We have colored each ball based on the latent class that it belongs to and hence there are 4 latent classes. The observations are colored according to their original two classes and the CCCD solution that was used during the latent class discovery process originally covered the red observations.

We note that the proposed latent class discovery process is clustering observations based on their relationship to the discriminant boundary. This clustering could of course be accomplished using any sort of classifier that produces a discriminant boundary that allows one to measure distances to the discriminant boundary and subsequently cluster the observations based on this distance. Figure 3 illustrates this idea using a quadratic classifier.

Results

We now illustrate our proposed latent class discovery process on an example gene expression data set. We have chosen to use an ALL/AML data set first proposed by Golub [Golub and Slonin, 1999]. This particular study investigated the ability to distinguish between two forms of leukemia, ALL and AML, and measure the responses of roughly 7000 genes on 72 patients using an Affymetrix microarray system. Figure 4 illustrates our strategy for the application of our latent class discovery process to the gene expression data.

We will now discuss the determination of the number of clusters during the latent class discovery process. We first define the estimated error rate of our CCCD-based classifier as follows. For each candidate number of clusters $k = 1, \dots, \hat{\gamma}$ an empirical risk (resubstitution error rate estimate) \hat{L}_k is calculated as

$$\begin{aligned} \hat{L}_k &:= (1/(n+m)) \left(\sum_{i=1}^n I\{x_i \notin \cup_{j=1, \dots, k} \cup_{v \in \hat{S}_j} B(v, \min_{w \in \hat{S}_j} r_w)\} \right. \\ &\quad \left. + \sum_{i=1}^m I\{y_i \in \cup_{j=1, \dots, k} \cup_{v \in \hat{S}_j} B(v, \min_{w \in \hat{S}_j} r_w)\} \right) \end{aligned}$$

We proceed by defining the "scale dimension" \hat{d}^* to be the cluster map dimension that minimizes a dimensionality-penalized empirical risk; $\hat{d}_\delta^* := \min\{\arg \min_k \hat{L}_k + \delta \cdot k\}$ for some penalty coefficient $\delta \in [0, 1]$. This scale dimension determines the number of clusters to be used during the latent class discovery process.

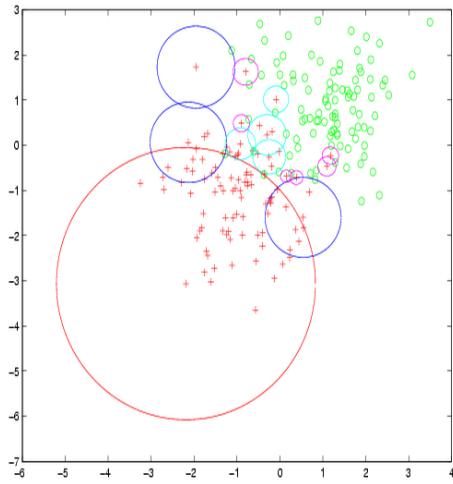


Figure 2: CCCD-based latent class discovery. Ball color indicates ball cluster membership. The CCCD solution was created based on the red observations.

Figure 5 presents performance as a function of scale space dimension for the gene expression data. A visual inspection of the plot indicates a scale space dimension, indicated by the abscissa of the elbow of the plot, of 5.

The latent classes discovered in the gene expression data correspond to the well known B-cell and T-cell ALL subtypes. These latent classes were first discovered using a custom in-house developed visualization framework known as the Interactive Hyperspectral Exploratory Data Analysis Tool (IHEDAT). It turns out that the measured distance from the ALL B-cell to the AML observations differs from the measured distance from ALL T-cell observations to the AML observations. In fact the ALL T-cell observations are in general more distant from the AML observations than the ALL B-cell observations. This distance is apparent in Figure 6 where we plot ALL and AML observations in the MDS projection space. The reader is referred to [Priebe et al., 2003b] for a more detailed description of how the original latent class discovery was performed. The reader is referred to [Solka et al., 2002] for an

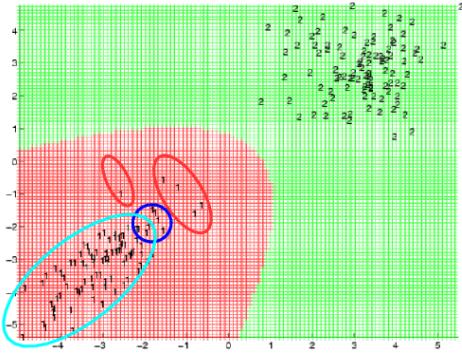


Figure 3: Quadratic classifier based latent class discovery. Circles indicate clusters of observations based on the distance to the quadratic discriminant boundary.

in-depth treatment of IHEDAT.

After the initial success of our latent class discovery methodology, we began to wonder how robust the procedure was to other possible CCCD coverings. In order to answer this question one of us, DJM, developed a method to enumerate all possible coverings of the ALL observations. Some of the these possible solutions correspond to greedy solutions while others do not. There are 180 21 node, ball, solutions. Sixteen of the nodes remain fixed across the solutions. There are 14 greedy solutions. In Figure 7 we plot scale dimension as a function of the various solutions. The red symbols indicate the locations of the greedy solutions. The green symbol indicates the location of the previous solution used to discover the T/B subclass.

In Figure 8 we present a histogram of the scale dimension across all solutions. We note that the value of 5 obtained in our original analysis was a little high as compared with the perceived mean/mode of this distribution. It is an open question as to how the choice of the scale dimension would effect the latent class discovery process.

In Figure 9 we present the dominating sets for each vertex. The triangles at the top of the plot indicate the 16 vertices that appear as part of all 180 solutions. For each of the other vertices we plot the number of times that the vertex appears in one of the covers. We have also used color to indicate whether the ball that corresponds to this vertex is centered on an ALL B-cell, blue, or ALL T-cell, red, observation. We note that only one T-cell vertex is in the set of 16 that does not change.

We may also analyze the variety of coverings through a characterization of the graphs that make up the coverings. In Figure 10 we present the unique induced subgraphs for the 5 changing vertices of the 180 dominating sets (top) and the unique induced subgraphs in the 5 changing vertices of the 14 greedy dominating sets (bottom). We note that the greedy based graphs are a proper subset of those graphs obtained based on the 5 changing vertices of the 180 dominating sets.

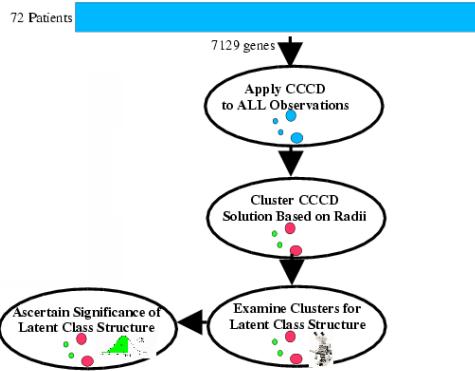


Figure 4: Latent class discovery strategy for the Golub gene expression data.

The analysis, with regard to the robustness of the procedure, presented so far in the paper has been interesting but not particularly compelling or even necessarily relevant to ascertaining whether the latent class structure identified during the original analysis would be identified if one were to use a different covering. The original greedy solution contained 3 clusters of balls that only contained ALL B-cell observations and 1 cluster of balls that contained 8/9 of the ALL T-cell observations. One way to evaluate the other coverings would be through the use of a figure of merit that captured the information contained within this first solution. We have chosen to use a figure of merit consisting of the percentage of B points that are in pure B clusters and the highest percentage of T points in any one cluster. In Figure 11 we present the calculated figures of merit for each solution. We note that all of the greedy folutions contain eight ninths of the T points in one cluster. We also note that .4 or more of the B points are in pure B clusters. It is a little hard to discern the exact nature of the solutions, based on ther plot, due to overplotting but it is reasonable to infer that one may have been able to idnetify the B-cell T-cell distinction utilizing any of a number of the other greedy solutions.

Conclusions

We have discussed a new method of latent class discovery that is appropriate for application to hyperdimensional data sets. This method allows one to discover previously unidentified classes within a class based on the relationship of the observations to the discriminant boundary. We have illustrated the application of this methodology to one gene expression data set. We have also presented some preliminary results that attempt to quantify the robustness of this method with regards to the dominating set that was used to facilitate the latent class discovery process. We have performed some rudimentary exploratory data analysis on the enumerated dominating sets. We have presented a strategy that we have developed to study the robustness of the latent class discovery process to choice of dominating set. This

'Scale dimension' for Golub gene expression data: $d^*=5$

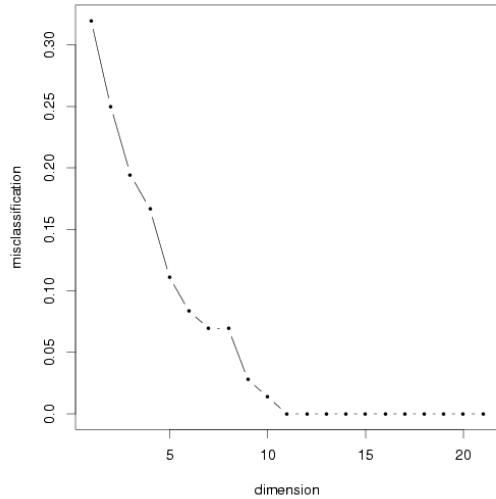


Figure 5: Cross validated performance as a function of scale dimension for the ALL Golub gene expression data with respect to the AML data.

problem is a very difficult one and will continue to be the subject of our ongoing research efforts.

Acknowledgments

The work of the first author, JLS, was sponsored by the NSWCDD ILIR Program. The work of the second author, CEP, was partially supported by the Office of Naval Research Grant N00014-01-1-0011 and the Defense Advanced Research Projects Agency Grant F49620-01-1-0395, and the work of the third author, DJM, was sponsored by the NSWCDD ILIR Program.

References

- [Chartrand and Lesniak, 1996] Chartrand, G. and Lesniak, L. (1996). *Graphs and Digraphs*. Chapman and Hall/CRC.
- [Golub and Slonin, 1999] Golub, T. R. and Slonin, D. K. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.
- [Priebe et al., 2000] Priebe, C. E., J., M. D., and Solka, J. L. (2000). On the selection of distance for a high dimensional classification problem. *ASA Proceedings of the Sections on Statistica and Statistical Graphics*, (58–63).
- [Priebe et al., 2003a] Priebe, C. E., Marchette, D. J., DeVinney, J., and Scolinsky, D. (2003a). Classification using catch cover digraphs. *to appear Journal of Classification*.

Weighted MDS using 7129-dim data

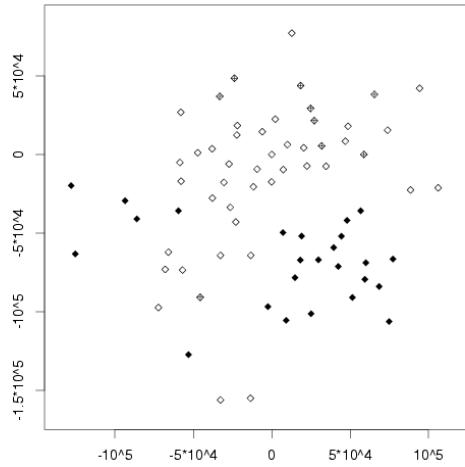


Figure 6: ALL, T and B cell, and AML observations in the MDS derived space. T-cell observations are cross hatched diamonds, B-cell observations are open diamonds, and AML observations are filled diamonds.

[Priebe et al., 2003b] Priebe, C. E., Solka, J. L., Marchette, D. J., and Clark, B. T. (2003+b). Class cover catch digraphs for latent class discovery in gene expression monitoring by dna microarrays. *to appear the Special Issue, of Computational Statistics and Data Analysis on Statistical Visualization.*

[Solka et al., 2002] Solka, J. L., Priebe, C. E., and Clark, B. T. (2002). A visualization framework for the analysis of hyperdimensional data. *International Journal of Image and Graphics*, 2(1):145–161.

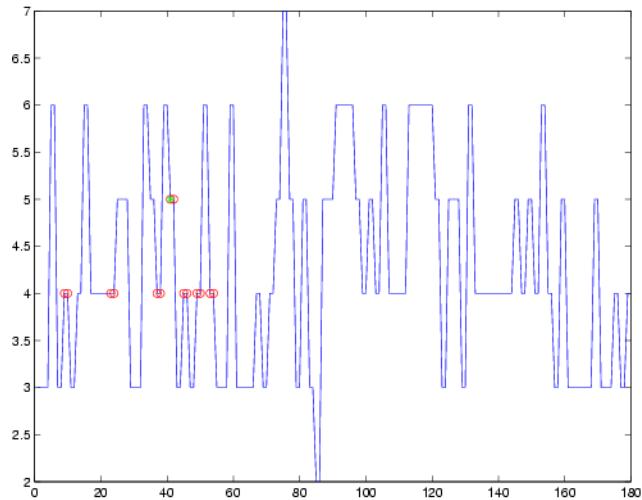


Figure 7: Scale dimension plotted as a function of the various Golub ALL dominating sets with respect to the AML observations. Red symbols indicate greedy solutions. The green symbol indicates our original solution used to make the T/B cell discovery.

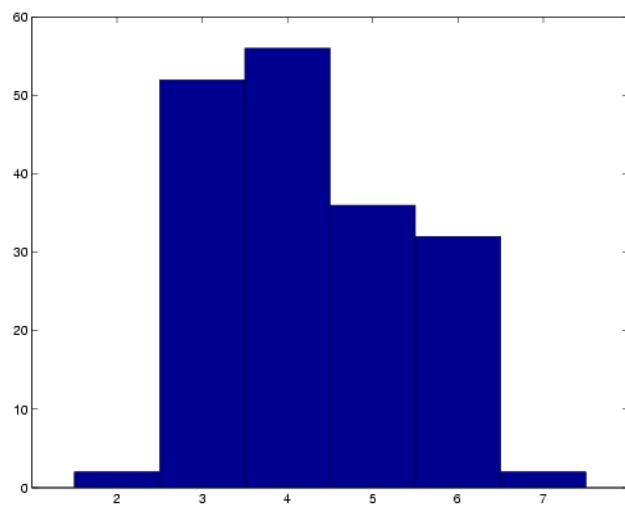


Figure 8: Histogram of the scale dimensions obtained with all 180 Golub ALL dominating sets.

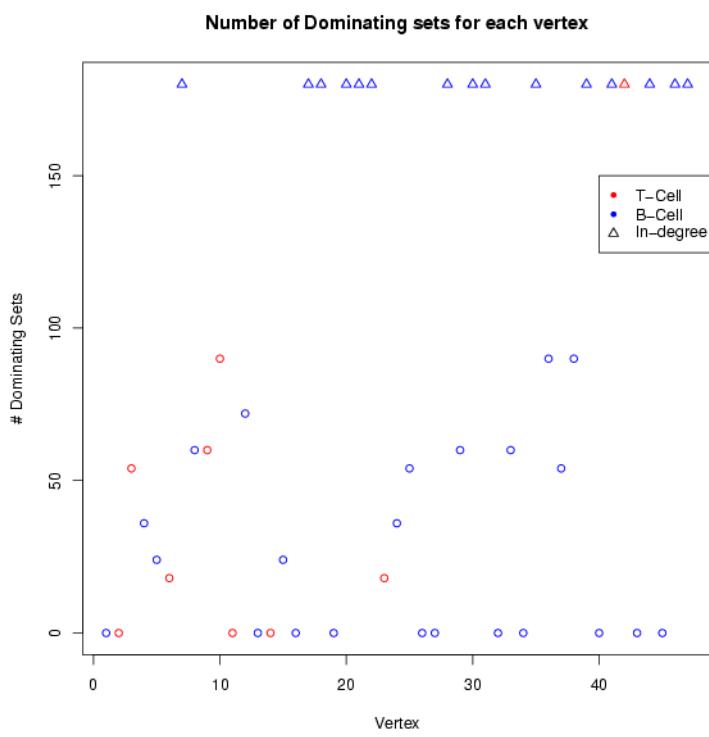


Figure 9: Dominating sets for each vertex in the Golub ALL data. Triangles represent the 16 vertices that are members of all 180 solutions. Color indicates T-cell or B-cell membership.

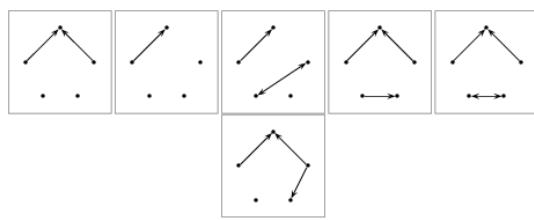
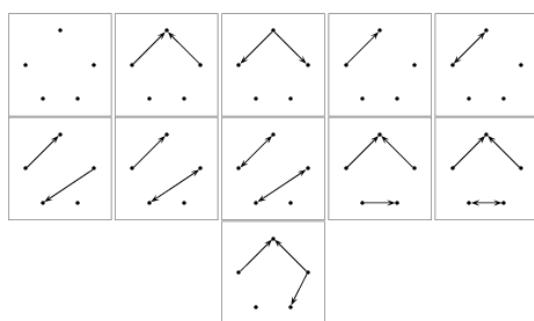


Figure 10: Unique induced subgraphs in the 5 changing vertices of the 180 dominating sets for the Golub ALL data with respect to the Golub AML data (top) and the unique induced subgraphs in the 5 changing vertices of the 14 greedy dominating sets for the Golub ALL data with respect to the Golub AML data (bottom).

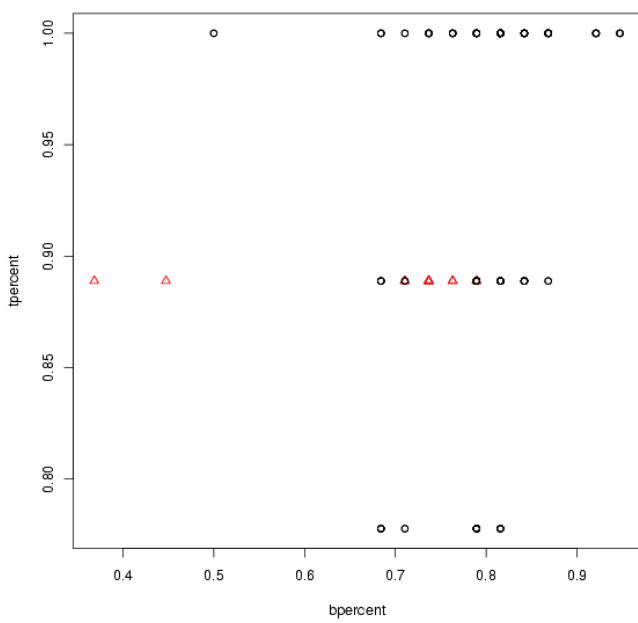


Figure 11: Proportion of B points that are in pure B clusters vs. highest proportion of T points in any one cluster for the 180 dominating sets of the Golub ALL data with respect to the Golub AML data. Red triangles indicate greedy solutions.

Encoding of Text to Preserve “Meaning”

Angel R. Martinez

George Mason University¹

Edward J. Wegman

George Mason University

Abstract:

A novel way to encode text streams is provided. The encoding allows for the application of computational methods in the determination of semantic similarity between text units. Supervised and unsupervised learning methods are used on a subset of the TDT Pilot Corpus to determine the effectiveness of the encoding in classification and clustering processes.

Key Words: bigram proximity matrix, trigram proximity matrix, classification, clustering, semantics, dimensionality reduction, similarity measures

1. The Problem

A large percent of information available to command and control systems exists in the form of text. Fast and effective methods to classify and group automatically related information for further processing are desirable. A critical step to the application of efficient computational methods is the encoding of the text stream. This paper introduces a novel method for encoding the text stream, with highly desirable computational properties. Tests show that the encoding preserves enough semantic distinctiveness to allow for very high rates of correct classification, where classification is based on measures of ‘semantic’ similarity computed between encoded text units.

In Section 2 we will introduce the bigram proximity matrix (BPM) and the trigram proximity matrix (TPM), the two structures resulting from the text stream encoding. The corpus used as a testbed, as well as the similarity measures used, are presented in Section 3. Section 4 discusses the experiments conducted to determine the encoding capacity to preserve distinctive features of the text stream content. The ‘shape of meaning’ via parallel coordinates is discussed in Section 5, and Section 6 offers conclusions and possible future work.

2. The Bigram Proximity Matrix (BPM) and the Trigram Proximity Matrix (TPM)

The BPM and the TPM are matrix structures used to encode each text unit, i.e., paragraph, section, chapter, book, etc. A simple example using a sentence will make the encoding process clear. The BPM for the sentence or text stream,

“The wise young man sought his father in the crowd.”

is shown in Table 1. We see that the matrix element located in the third row (*his*) and the fifth column (*father*) has a value of one. This means that the pair of words *his father* occurs once in this unit of text. It

1. Email: martinezar@nswc.navy.mil (Angel Martinez) and ewegman@galaxy.gmu.edu (Edward Wegman)

should be noted that in most cases, depending on the size of the lexicon and the size of the text stream, the BPM will be very sparse.

Table 1. Example of Bigram Proximity Matrix^a

	.	crowd	his	in	father	man	sought	the	wise	young
.										
crowd	1									
his					1					
in								1		
father				1						
man							1			
sought			1							
the		1							1	
wise										1
young						1				

a. Zeros in empty boxes are removed for clarity.

By preserving the ordering of words of the discourse stream, the BPM captures a substantial amount of information about meaning. Also, by obtaining the individual counts of word co-occurrences, the BPM captures the ‘intensity’ of the discourse’s theme. Both features make the BPM a suitable tool for capturing meaning and performing computations to identify semantic similarities among units of discourse (e.g., paragraphs, documents).

The TPM captures the occurrence of consecutive triplets of words by constructing a cube with the lexicon on three axes. A trigram is the point of intersection of the row, column and page in the cube, as illustrated in Figure 1. The figure expands the same sentence given above. The trigram “*sought his father*,” is the point (*sought, his, father*), that is, the array element in the 7th row, 3rd column, and 5th page. As can be seen, the resulting N^3 array structure, where N is the size of the lexicon is very sparse. The TPM is a trivial extension of the BPM. Preliminary testing seems to indicate that for some applications (e.g., change of topic determination) and for larger sizes of text units, the TPM performs better than the BPM.

Notice that the BPM and TPM are arrays whose rows, columns, and pages (in the case of the TPM) are indexed by the lexicon of the text unit. We chose alphabetical ordering of the lexicon; however, this is not essential. In the pre-processing of the text, all punctuation marks, except the ending period, were deleted. The end period was considered a word and placed at the head of the lexicon.

3. The Test Suite and Similarity Measures

Documents from the Topic Detection and Tracking (TDT) Pilot Corpus (Linguistic Data Consortium, Philadelphia, PA) were used as the textual testbed. The TDT corpus is comprised of close to 16,000 newscasts collected from July 1, 1994 to June 30, 1995 from the Reuters newswire service and CNN broadcast news transcripts. A set of 25 events are defined in the TDT. Each of the 16,000 newscasts is flagged with one of three possible flags: *Yes*, *No*, or *Brief*. The flags are used to indicate that a newscast discusses one of the 25

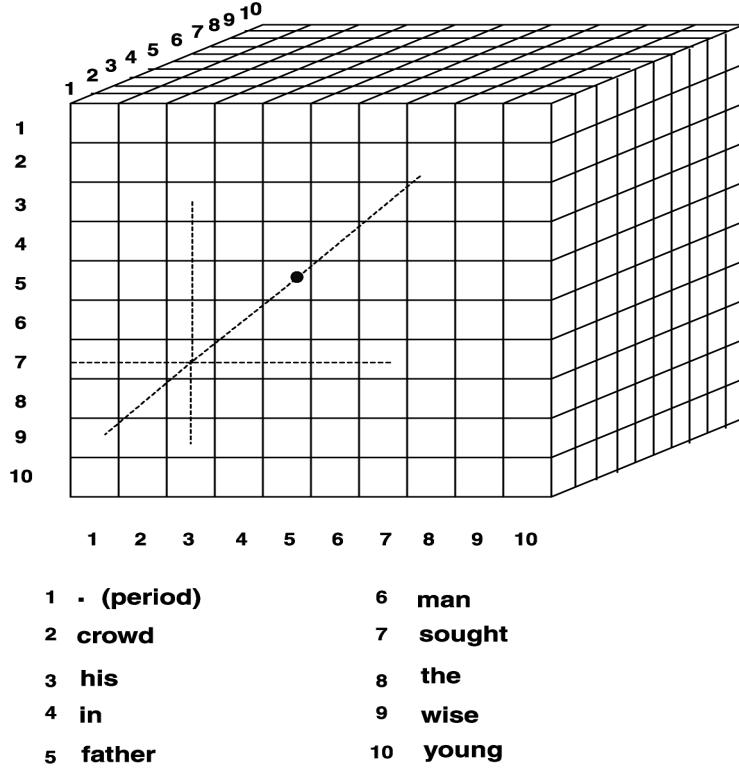


Figure 1 Here we see an example of a trigram proximity matrix (TPM). Note that the indexes on each side of the cube correspond to a word in the lexicon. The element in the i th row, j th column and k th page indicates the number of times that sequence of three words appears in the text unit.

events, or it does not, or it does so only briefly. In order to meet computational requirements, a subset of the TDT corpus was used in this work. A total of 503 newscasts were chosen from the 16,000 available. These news stories comprise 16 of the 25 events discussed in the TDT. See Table 2 for a list of topics. The 503 documents chosen contain only the *Yes* or *No* flags. This choice stems from the need to demonstrate that the BPM and TPM capture enough meaning to make a correct or incorrect topic classification choice.

The 503 stories selected produced a lexicon of 11,103 unique words. Conflated forms are counted as different words. Using this lexicon and the structure already described, a BPM and TPM were created for each of the 503 stories. The assertion is that each of the structures captures enough of the meaning in the newscast to serve as a classification feature. We test this assertion using various methods in exploratory data analysis and computational statistics.

Table 3 lists the measures of semantic similarity used in this study. For definitions of these measures, see [Martinez, 2002]. It should be noted that some of these are distances and some are similarities; however, for ease of exposition, we will refer to them all as measures of semantic similarity. All similarity measures were first converted to distances so they could be used with the various methods. Also, some of these measures are binary, in which case the frequencies of word pairs and triples are changed to a 0 or a 1. Similarly, the proximity matrices are converted to distributions to use the probabilistic measures.

Two variants of the lexicons are considered. In one variant, common high-frequency words have been removed from the lexicon and the documents. In another variation, we stemmed the words as well as removed the common high-frequency words from the documents. Part of this research examines how this

affects the discriminating power of the proximity matrices. Many NLP applications [Kimbrell, 1988], [Salton, Buckley and Smith, 1990], [Frakes and Baeza-Yates, 1992], [Berry and Browne, 1999] use a shorter version of the lexicon by excluding words often used in the language. These words, usually called *stop words* or *noise words*, are said to have low informational content and thus, in the name of computational efficiency, are deleted. Not all agree with this approach [Witten, Moffat and Bell, 1994].

Taking the denoising idea one step further, we stemmed the words in the denoised text. The idea is to convert all conflated forms of the words to their stem or root to increase the frequency of key words and thus enhance the discriminatory factors of the features. Stemming is routinely applied in the area of information retrieval (IR). In the IR application, stemming is used to enhance the performance of the IR system, as well as to reduce the total number of unique words and save on computational resources. A popular stemmer is the Porter stemmer [Baeza-Yates and Neto, 1999], [Porter, 1980]. The Porter stemmer is simple; however, its performance is comparable with older established stemmers. The Porter stemmer works on the suffix of words. These are stripped according to several parsing rules and replaced with one of a list of endings or the null ending. On most occasions, these simple replacements work well, as is the case of the words: *protecting, protected, protects, protection*, which are conflated forms of ‘*protect*.’ However, in other cases, it does not work as well. For example, a word like *probate* will be stemmed to *probe*, which carries a totally different meaning. The same thing applies for *relativity*, which is conflated to *relate*. In information retrieval applications, these anomalies of the stemmers do not seem to affect their usefulness. The issues are not so clear in our application.

Table 2. List of 16 Topics

Topic Number	Topic Description	Number of Documents Used	Topic Number	Topic Description	Number of Documents Used
4	Cessna on the White house	14	15	Kobe, Japan Quake	50
5	Clinic Murders (Salvi)	38	16	Lost in Iraq	30
6	Comet into Jupiter	45	17	NYC Subway bombing	24
8	Death of Kim Il Sung	35	18	Oklahoma City bombing	76
9	DNA in OJ trial	29	21	Serbians down F-16	16
11	Hall's copter in N. Korea	75	22	Serbs violate Bihać	19
12	Humble, TX, flooding	16	24	US Air 427 crash	16
13	Justice-to-be Breyer	8	25	WTC bombing trial	12

Table 3. Measures of Semantic Similarity

1. Matching Coefficient	8. Sokal - Sneath
2. Sokal - Michner	9. Gower - Legendre 1
3. Dice Coefficient	10. Gower - Legendre 2
4. Jaccard Coefficient	11. Normalized Correlation Coefficient
5. Cosine - Ochiai	12. L_1 - Probabilistic Measure
6. Russell - Rao	13. IRad - Probabilistic Measure
7. Roger - Tanimoto	

Table 4. Lexicon Sizes

Type of Lexicon	Size of Lexicon
Full Lexicon	11,103
Denoised Lexicon	10,997
Stemmed Lexicon (also denoised)	7,146

To determine if there is any deleterious effect of using the stemmer and denoising with the BPMs and TPMs, the same experiments are conducted on all three versions: full, denoised, and stemmed documents. Table 4 summarizes the size of the lexicons in these three cases.

4. The Tests

In this section, we present the results of applying supervised (k NN classification) and unsupervised learning (model-based clustering) to the collection of 503 news stories. In supervised learning experiments, the class membership of each observation is known, and we are interested in how well we can classify the topics given the BPM and TPM features. In the case of unsupervised learning, experiments start with very little information, essentially the observations. No knowledge of the number of classes nor the observations' membership to these are available.

4.1 k NN Classification

We first apply supervised learning approaches to the BPM and TPM features to determine whether these features allow us to classify documents according to their meaning. If we can accurately classify using the proximity matrix as a feature, then this is an indication that these features preserve meaning. In k NN classification, the decision rule is to assign \mathbf{x} to the class that has the greater number of members amongst the k nearest neighbors [Web, 1999], [Cover and Hart, 1967]. Due to the high-dimensional nature of these features, we cannot apply classical supervised learning methods such as linear or quadratic classifiers. However, the k NN classifier is well-suited for our BPM and TPM features, since all we need are the pairwise distances between all BPMs and TPMs.

The k NN classification method was applied using the following parameters:

- Thirteen measures of semantic similarity (see Table 3)
- k values: $k = 1, 3, 5, 7$, and 10
- Proximity matrix: BPM and TPM
- Three text conditions: full, denoised and stemmed

The dotplot shown in Figure 2 shows the results for the rate of correct classification (CCR) using denoised text and is representative of the other experimental results. Of the thirteen measures of semantic similarity, 10 resulted in high CCRs. Three of them resulted in very low CCRs. These are the Sokal-Michener, Roger-Tanimoto, and Gower-Legendre 1. Summary conclusions from the experiments are:

1. The BPM and TPM seem to contain sufficient semantic information to allow for almost perfect classification results.
2. Binary similarity measures based solely on word pairs and triples that are common to two documents worked better with denoised text
3. Probabilistic measures of semantic similarity performed the best with denoised and stemmed text.
4. The Dice, Jaccard, Sokal-Sneath, Gower-Legendre 2, L_1 norm and IRad measures were the best performers.

4.2 Unsupervised Learning

The method chosen for our unsupervised learning experiments is called model-based clustering [Banfield and Raftery, 1993], [Fraley and Raftery, 1998]. This method is based on finite mixtures [Everitt and Hand, 1981] where the output model is a weighted sum of c multivariate normals. See Martinez [2002] for more information on model-based clustering for this application. One of the benefits of using model-based clustering instead of some other method is that it includes a mechanism for determining the number of groups in the data set. Thus, we hoped to see results with approximately 16 clusters (or topics).

As a way to compare supervised and unsupervised learning methods, similar experiment variables were used with model-based clustering. The following variable combinations were used:

- Thirteen measures of semantic similarity (see Table 3)
- Two values of k nearest neighbors for the Isomap dimensionality reduction: $k = 7, 10$
- Proximity matrix: BPM and TPM
- Three text conditions: full, denoised and stemmed
- One ‘best’ dimension value from Isomap

In order to use model-based clustering, the dimensionality of our observations (i.e., the BPMs and TPMs) had to be drastically reduced from $11,103^2$ (in the case of the full lexicon) to 2, 3, 4, 5, and 6 dimensions. This reduction was effected through the Isometric Figure Mapping (Isomap), a nonlinear dimensionality reduction method [Tenenbaum, deSilva and Langford, 2000]. Isomap is essentially an extension of multi-dimensional scaling (MDS) methods, where geodesic distances between k nearest neighbors are used as inputs to MDS.

The assessment of the results was done via a visualization aid we developed called ReClus. ReClus takes the output from the model-based clustering procedure and draws one large rectangle. This rectangle is subdivided into n smaller rectangles, where n is the number of clusters chosen according to the model-based clustering procedure. The area of each smaller rectangle is proportional to the number of cases in the cluster. Inside each rectangle, and for each case assigned to that cluster, the class number is printed, or optionally, the case number is printed. Each number is color-coded to denote the degree of certainty that the particular case belongs to the cluster. A threshold is set to print in black bold type when the certainty is 0.8 or above. ReClus, thus, provides a quick visual way to examine the results from model-based clustering. Although, judging between two results entails a degree of subjectivity, this is a problem only where

results are close. Additionally, ReClus provides information to guide the examination of confounding factors in the clustering process. An example of a ReClus plot is given in Figure 3.

We now offer some specific observations on the results, keeping in mind that it is difficult to assess the goodness of clusters. Of the 312 experiments, thirteen showed the correct number of clusters, sixteen. Not surprisingly, however, none of these - and for that matter, none of the 312 - showed sixteen correct (i.e., ‘pure’) clusters. In each of the thirteen results, two rectangles contained the same class cases (topic number 6 was split into two groups). We note that the same situation for topic 6 arose in those results containing 15 and 17 clusters. Usually, more than half of the rectangles suffered from some degree of contamination. If we consider a good result as one with the highest number of ‘pure’ rectangles, followed by a high number of only lightly ‘contaminated’ ones, and the fewest number of jumbled rectangles, then the following are the best results:

- Ochiai measure, full text, dimensionality 6, BPM and $k = 7$
- Jaccard measure, stemmed text, dimensionality 6, BPM, and $k = 7$

The above categorization of the best results is naive. It assumes that a mix of 2 or more classes in a rectangle is an undesirable result. However, in the case of our test bed, a mix could point to a justifiable confusion. For example, in several of the ‘best’ results classes 8 and 11 are usually mixed; however, both sets of documents are about North Korea. Also, topics 18 and 17 are sometimes mixed: both sets of documents deal with bombing, the Oklahoma City bombing and the NY subway bombing. The same happens a few times with classes 21 and 22: both report on two different aspects of the Serbian conflict.

The intriguing case mentioned above, where class 6 had two pure rectangles containing class 6 cases, raises the issue of latent classes or sub-groups within the topics. A reading of the documents involved does show two different foci. The main subject of the set is the crash of fragments of the comet Shoemaker-Levy onto the surface of Jupiter. One group in the set emphasizes background information about the comet as well as the fact that the space shuttle is in orbit ready to observe what is yet to take place. The second group’s focus is predominantly on the event already taking place and observations of the phenomenon.

5. The ‘Shape of Meaning’ and Parallel Coordinates

Examination of the model-based clustering results using ReClus seem to show two aspects of the semantic content of the text units: (1) the possibility of latent topics, as was the case with topic 6, and (2) the detection of similarity between topics, as in the cases of topics 8 and 11, 17 and 18, and 21 and 22. Visual detection of similarity between topics can also be seen using parallel coordinates [Wegman, 1990]. A matrix of parallel coordinate plots was created by placing together in matrix form a parallel coordinate plot for each topic. See Figure 4 at the end of the paper. By looking at the overall shape formed by the lines and the points where these touch the five axes (5 dimensions), we are able to detect patterns. These patterns seem to be manifestations of semantic content of the clusters. Notice the following:

- The parallel coordinates for topics 8 and 11 show exact patterns for a good number of their lines. This corroborates the confusion detected in the model-based clustering results via the ReClus display. The possible common theme repeated is North Korea and US relations.
- The parallel coordinate plots for topics 17 and 18 show a group of lines with the exact pattern in both. This corroborates the confusion detected in the model-based clustering results via ReClus. A possible common theme that is repeated is bombing and its immediate effects.
- The parallel coordinates for topics 21 and 22 show a small group of lines with a common pattern. This pattern may represent a common core of the two topics about the Serbian conflict.

- Topic 6 showed invariably in two clusters in the ReClus figures. Notice the pattern from the lines of the parallel coordinates for topic 6. On the second axis from top to bottom, one notices a separation of lines. This indicates two different groups, separable at the dimension represented by that axis. These groups may represent the two sub-themes found in reading the newscasts of topic 6.

Parallel coordinates were a crucial help in making sense of our model-based clustering results. On occasion, a permutation tour [Wegman, 1990] of a single topic's parallel coordinates was necessary. For example, the parallel coordinate plot for topic 8 shows a simple structure. But, the ReClus view shows topic 8 linked with topic 11 in three different clusters. Is the model-based clustering result wrong? When a permutation tour of topic 8 was performed, about three sub-groups became evident. This seems to match a reading of the 35 newscasts from topic 8 as the next paragraph explains.

As mentioned above, classes 8 and 11 appeared mixed in the experiments. Topic 8 and topic 11 both deal with North Korea, one regarding the death of Kim Il Sung and the other the crash of the American helicopter in North Korean territory. Most of the time there are three rectangles containing cases from 8, of which two are mixed with 11 and one rectangle (almost purely 11) was only very slightly mixed with 8. As is the case with class 6, this may imply the existence of latent classes in groups 8 and 11. A quick reading of the newscasts for topic 8 seems to show three major themes discussed over the background of Kim Il Sung's death and the probable succession of his son Kim Jong-il. The three latent topics are: (1) US and North Korea relations; (2) North Korea and South Korea relations; and (3) North Korea's nuclear plants.

The visualization tool ReClus made the examination of the results from the model-based clustering experiment possible and fruitful. The BPMs and TPMs capture sufficient meaning to produce satisfactory results with this unsupervised learning method. For best results, the Ochiai measure of semantic similarity should be used in the Isomap dimensionality reduction method, and the dimensionality can be reduced to five or six dimensions. Full and denoised text did well with the Ochiai measure. It seems that latent classes are detected by the BPMs and TPMs, as made manifest by the results discussed above.

6. Conclusions and Future Work

We introduced in this paper two transformations of the text stream amenable to computational methods, called the bigram proximity matrix and the trigram proximity matrix. The usefulness of the BPM and TPM depends on how much semantic information they preserve. In order to determine the adequacy of these encodings to preserve semantic information, supervised learning using k NN classification and unsupervised learning, using model-based clustering were applied. Variables in the experiments consisted of combinations of the following:

- Thirteen semantic similarity measures
- Three text conditions (full, denoised and stemmed)
- Various values of k (k NN).

Supervised learning experiments were conducted on the full dimensionality of the feature space (see Table 4 for the lexicon sizes). Dimensionality was reduced to a lower number (2 - 6) using the nonlinear dimensionality reduction procedure called Isometric Figure Mapping (Isomap). With dimensionality reduced, unsupervised learning experiments were conducted using model-based clustering.

Results from supervised learning experiments showed that correct classification ratios in the range of 0.95 - 0.99 were common for many of the semantic similarity measures used. This indicates that the BPM and TPM capture sufficient semantic information for the discrimination of semantically dissimilar

text units. Results from the unsupervised learning experiments showed that the BPM and TPM capture sufficient semantic information to group thematically related documents and seems to detect latent sub-themes. In conclusion, we can state that the text stream transformations do capture enough semantic information to allow for the semantic discrimination of text units.

Several obvious possibilities for future work are:

- To create efficient algorithms for lexicon expansion and subsequent recomputation of BPMs and TPMs.
- To apply BPMs and TPMs to the problem of change of topic determination.
- To explore the capability of BPMs and TPMs in combination with model-based clustering, parallel coordinates and ReClus in the detection and identification of sub-topics.
- To examine the effect of the following: size of documents, type of documents (scientific article, news story, patent description, etc.), and number of text units in a topic.

References:

- Baeza-Yates, Ricardo and Berthier Ribero-Neto, 1999. *Modern Information Retrieval*, ACM Press, New York, NY.
- Banfield, J. D. and A. E. Raftery, 1993. ‘Model-based Gaussian and non-Gaussian clustering,’ *Biometrics*, 49, pp. 803 - 821.
- Berry, Michael W., and Murray Browne.1999. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia, PA.
- Charniak, Eugene. 1996. *Statistical Language Learning*, The MIT Press, Cambridge, MA.
- Cover, T. M. and P. E. Hart, 1967. ‘Nearest neighbor pattern classification,’ *IEEE Transactions on Information Theory*, 13, pp. 21 - 27.
- Everitt, B. S. and D. J. Hand, 1981. *Finite Mixture Distributions*, Chapman and Hall, London, UK.
- Frakes, W. B. and Ricardo Baeza-Yates, 1992. *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, New Jersey.
- Fraley, C. and A. E. Raftery, 1998. ‘How many clusters? Which clustering method? - Answers via model-based cluster analysis,’ *The Computer Journal*, 41, pp. 578 - 588.
- Kimbrell, Roy E., 1988. ‘Searching for text? Send an N-Gram!’, *Byte*, May, pp. 297 - 312.
- Landauer, Thomas K., Darrell Laham, and Peter Foltz. 1998. Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns and S. A. Solla (Eds.). *Advances in Neural Information Processing Systems*, 10, pp. 45 - 51. Cambridge. MIT Press.
- Martinez, Angel R., 2002. *A Statistical Framework for the Representation of Semantics*, Ph.D. Dissertation, George Mason University.
- Nettleton, Dan and T. Bannerjee, 2001. ‘Testing the equality of distributions of random vectors with categorical components,’ *Computational Statistics and Data Analysis*, 37, pp. 195 - 208.

- Porter, M. F., 1980. 'An algorithm for suffix stripping,' *Program*, 14, pp. 130 - 137.
- Révézs, Gyorgy. 1983. *Introduction to Formal Languages*, McGraw-Hill Book Company, New York, NY.
- Salton, Gerard, Chris Buckley and Maria Smith, 1990. 'On the application of syntactic methodologies,' *Automatic Text Analysis, Information Processing & Management*, 26, pp. 73 - 92.
- Tenenbaum, Joshua B., Vin deSilva and John C, Langford, 2000. 'A global geometric framework for non-linear dimensionality reduction,' *Science*, 290, pp. 2319 - 2323.
- Webb, Andrew, 1999. *Statistical Pattern Recognition*, Oxford University Press, Oxford, UK.
- Wegman, E. J., 1990. 'Hyperdimensional data analysis using parallel coordinates,' *Journal of the American Statistical Association*, 85, pp. 664 - 675.
- Witten, I. H., A. Moffat and T. C. Bell, 1994. *Managing Gigabytes: Compressing and Indexing Documents and Images*, van Nostrand Reinhold, New York, NY.

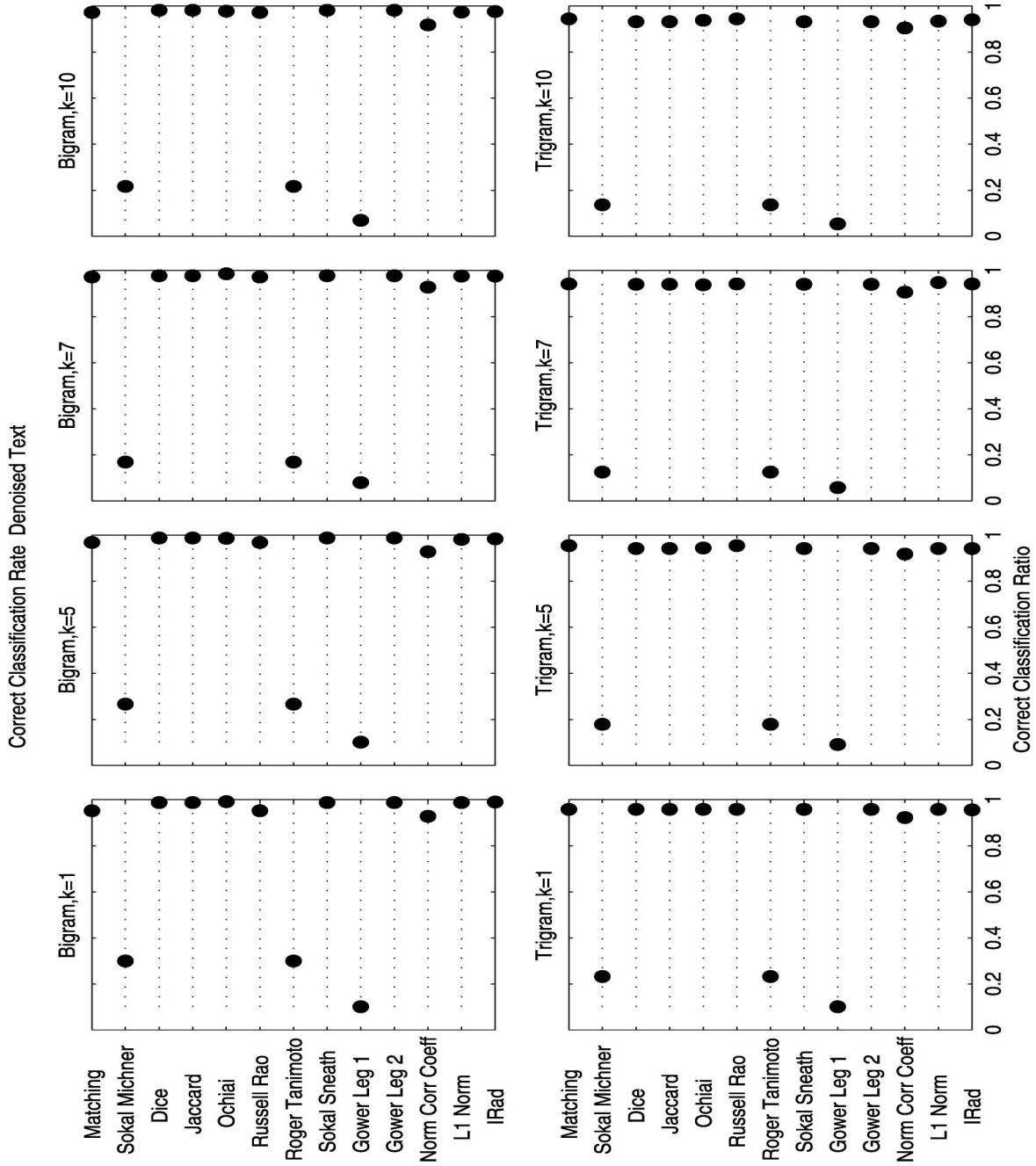


Figure 2 This shows the results of applying the k NN classification method to the problem of correctly classifying the newscasts according to their topic. We see that most of the measures of semantic similarity perform well.

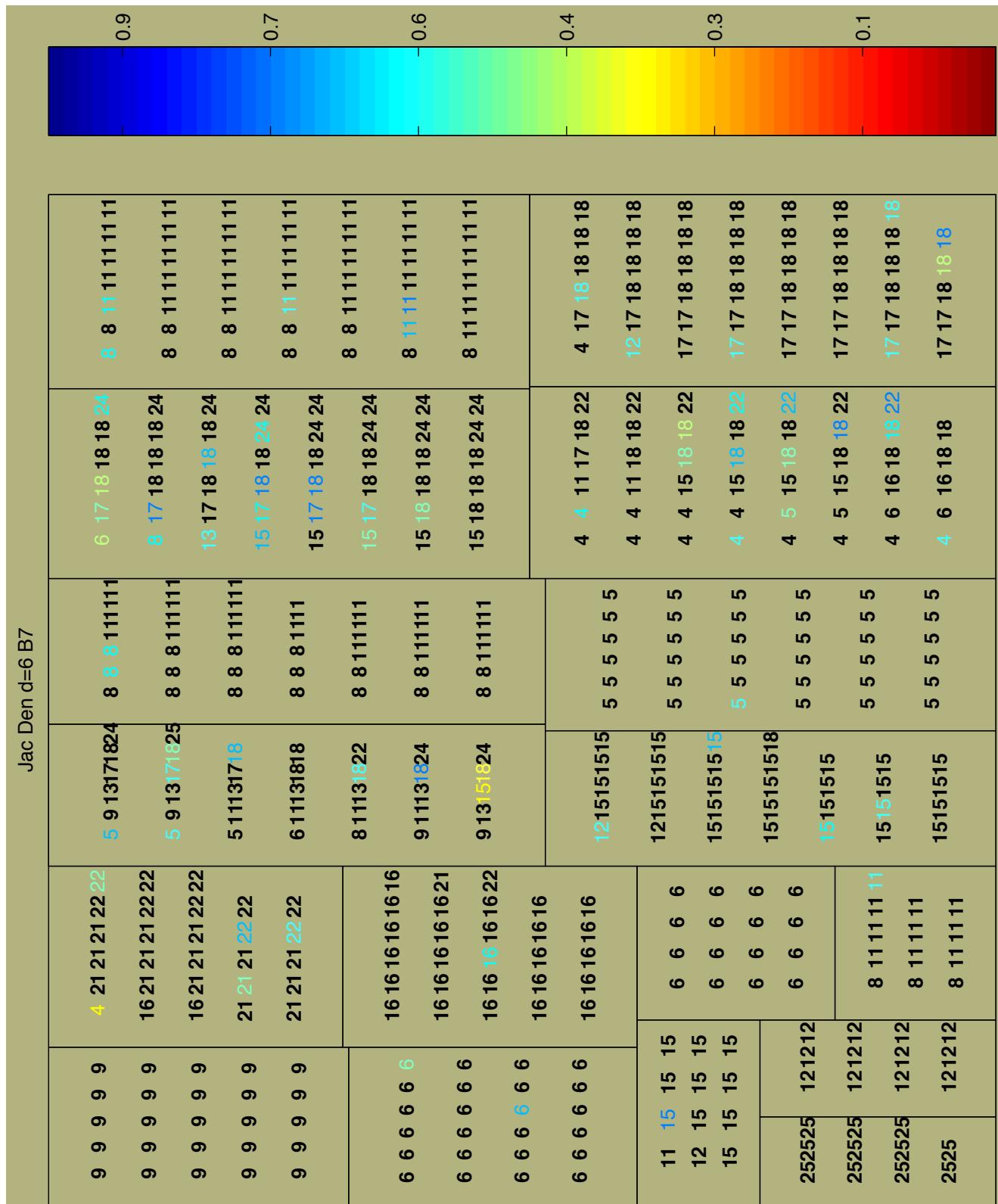


Figure 3. ReClus layout showing the results from the model-based clustering where the Jaccard measure is used with denoised text

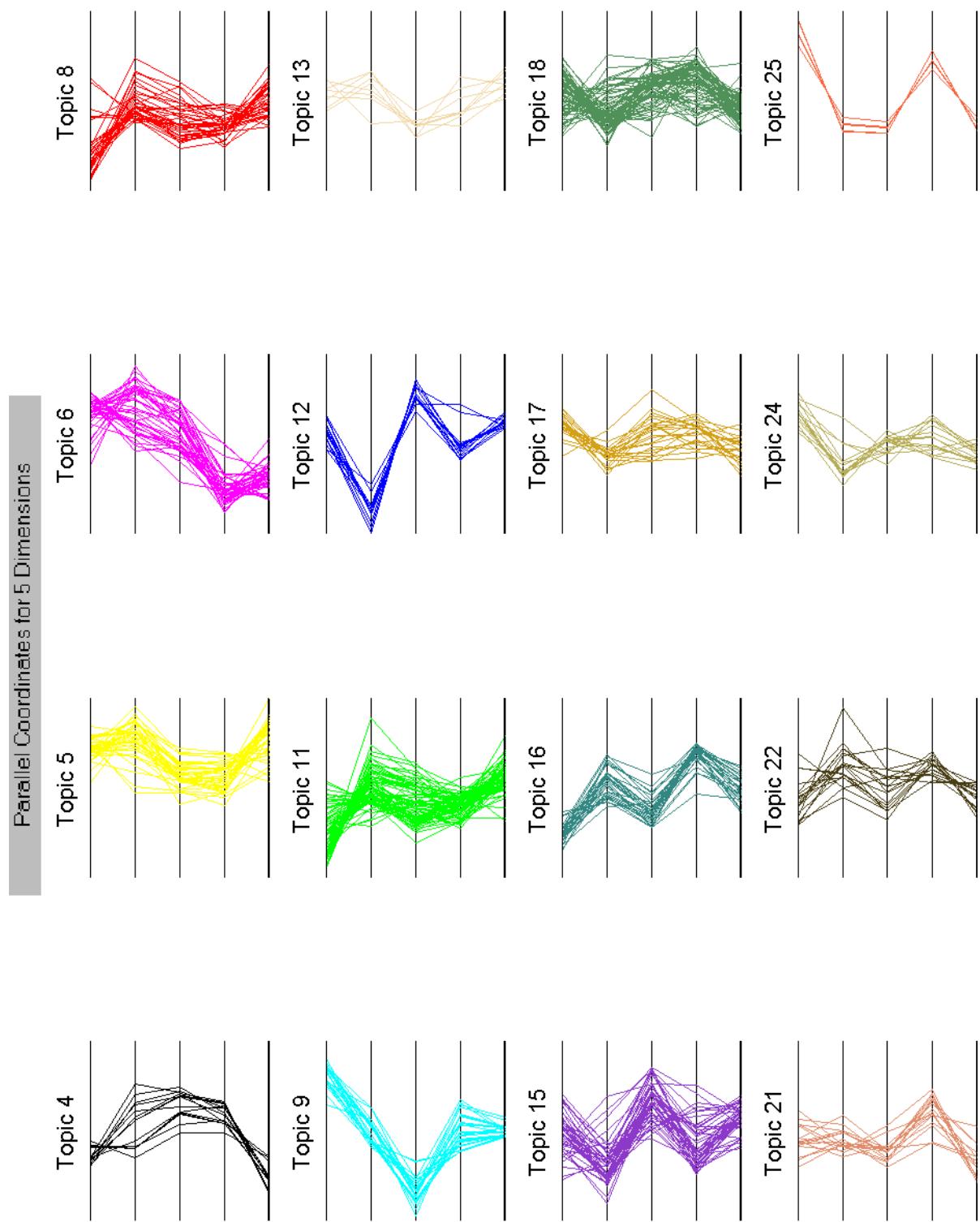


Figure 4. Parallel coordinates plot matrix for the 16 topics.

General Session 2

A study of denial of service attacks on the Internet

David J. Marchette

Abstract

The public Internet is a critical component of the information infrastructure supporting finance, commerce, and civil and national defense. Denial of service attacks on major Internet sites, both the direct effect on the attacked sites and the indirect collateral effects on the Internet as a whole do considerable financial damage on a regular basis. Denial of service attacks could be a part of a concerted attack on the flow of information. This coupled with a physical attack of some kind poses a substantial threat.

Monitoring these attacks in a timely manner is problematic because the institutions under attack often have good financial and security incentives not to share that information, and getting the information in a timely (cyber-scale) manner is difficult without a (computerized) automatic notification process. Remote detection using backscatter allows the detection of attacks in a completely passive manner without any cooperation from the primary target computer(s). This paper discusses some of the mathematical and statistical aspects of backscatter analysis, and illustrates some interesting practical issues in the analysis.

1 Introduction

Suppose there is a coordinated denial of service attack (using one of a selection of freely available tools) on the public banking access sites of the 10 largest US banks (or the largest stock trading sites). Financial institutions are reluctant to share information, so it might take a while (hours or days) to sort out the size and the scope of the attack, or to even find out that the attack took place. A method of determining the scope of the attacks without relying on self-reporting is clearly needed.

The basic idea of most denial of service attacks is to flood a computer with bogus requests, or otherwise cause it to devote resources to the attack at the expense of the legitimate users of the system. A classic in this genre is the SYN flood. The attacker sends SYN packets requesting a connection, but never completes the handshake. One way to do this is to set the source IP address to a nonexistent address (this process of changing the source address is called “spoofing” the address). For each SYN packet, the victim computer allocates a session and waits a certain amount of time before “timing out” and releasing the session. If enough of these “bogus” SYN packets are sent, all the available sessions are devoted to processing the attack, and no legitimate users can connect to the machine.

A related attack is to send packets that are out of sequence, or errors, forcing the victim computer to spend time handling the errors. For example, if a SYN/ACK packet is sent without having received an initiating SYN packet, the destination computer generates and sends an RST (reset) packet. If the attacker can arrange to have millions of SYN/ACK packets sent, the victim computer will spend all its resources handling these errors, thus denying service to legitimate users. One way to arrange this, is through a distributed denial of service tool, such as trinoo or TFN2k. These tools compromise a set of computers, dispersed across the IP address space, then use these “intermediate victims” to send thousands of packets to the intended victim. Each packet is crafted to have a random (spoofed) source IP address, so the attacking machines cannot be identified. See [Mar01], [Che01] and [INNM01] for descriptions of some distributed denial of service attacks.

The result of such an attack is a number of reset (or other) packets appearing at random sites around the Internet, with no obvious session or initiating packets to explain them. See Figure 1. This is used by [MVS01] to estimate the number of denial of service attacks during three one week periods, by counting how many unsolicited packets are seen addressed to one of the 2^{24} possible IP addresses they monitored.

2 Analysis

Following [MVS01], we can compute some of the probabilities of detection needed to analyze backscatter packets. Assume the spoofed IP addresses are generated randomly, uniformly on all 2^{32} addresses, and independently. Assume there are m packets sent in an attack on a given victim. If we monitor all packets to n IP addresses, then it is easy to see that the probability of detecting an attack is:

$$P[\text{detect attack}] = 1 - \left(1 - \frac{n}{2^{32}}\right)^m. \quad (1)$$

From this, one obtains the result that the expected number of backscatter packets we detect is

$$\frac{nm}{2^{32}}. \quad (2)$$

We would like to determine how many packets were originally sent. This will give an estimate for the severity of the attack, and might allow us to infer whether the attack was likely to have been mounted by multiple attackers, for example through a distributed denial of service tool. To do this, note that the probability of seeing exactly j packets, under our independence assumption, is

$$P[j \text{ packets}] = \binom{m}{j} \left(\frac{n}{2^{32}}\right)^j \left(1 - \frac{n}{2^{32}}\right)^{m-j}. \quad (3)$$

The maximum likelihood estimate for m , using Equation 3, is

$$\hat{m} = \left\lfloor \frac{j2^{32}}{n} \right\rfloor. \quad (4)$$

Thus, if we see j packets, we can use Equation 4 to estimate the size of the attack.

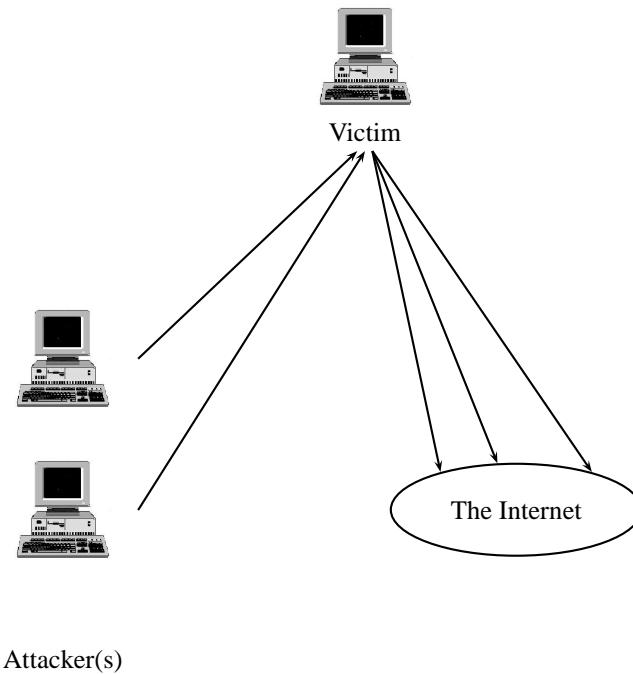


Figure 1: Backscatter from a denial of service attack. Packets are sent to the victim from one or more attackers. These packets have spoofed IP addresses, which cause the victim’s response to be sent to random addresses within the Internet.

Note that if the attacker chooses from a subset of all possible IP addresses, say of size N , then we must replace 2^{32} in Equations 1–4 with N . These equations, then, under the assumptions of uniformity and independence, allow us to estimate the original size of the attack from the packets we see at our network, assuming we know the original size of the pool of IP addresses from which they are selected.

There is also the question of determining the number of attacks. [MVS01] do this by defining an attack as a series of packets with a maximum inter-packet gap less than a fixed value. The idea being that if there is a long enough gap between packets, then it is reasonable to assume that these correspond to different attacks. They then count the number of attacks they detect, and report over 12,000 attacks in the three week period they investigated.

This assumes that all attacks generate packets to the network monitored. We will assume that all attack packets generate backscatter (until the machine ceases to function), ignoring issues such as filtering firewalls or other kinds of mechanisms that may block either the attack or the backscatter packets. If our monitored network (n IP addresses) is sufficiently small, and a sufficiently small number of packets are sent in the

attack, there is a reasonable probability that we will receive no packets.

Several calculations are possible to determine whether the assumptions are valid. For example, [MVS01] suggest using the Anderson-Darling test of uniformity to test that the IP addresses are in fact uniformly distributed. We will discuss this further below. This of course assumes we know how many IP addresses are in the complete pool. A perusal of the attack code available on the Internet shows that the tools often allow the user to choose which octets of the IP address to randomize, thus reducing the pool. Assume the pool contains N addresses, and we monitor n addresses, and that the attack packets are sent every t seconds. Then if we knew how long a gap one should expect to see between detected packets, one could use this to estimate N , and thus be able to use the equations above for the estimate of the size of the attack. This would also be important for the determination of a definition of attack, as one would want the gap to be many standard deviations larger than this expected delay. The calculation is straightforward. The expected number of attack packets between two detected packets (assuming independence) is:

$$\begin{aligned} \sum_{s=1}^N \left(1 - \frac{n}{N}\right)^{s-1} \frac{n}{N} s &= \frac{(1 - (n+1)(1 - \frac{n}{N})^N)N}{n} \\ &\approx \frac{N(1 - e^{-N})}{n} \\ &\approx \frac{N}{n} \end{aligned}$$

The variance of the number of packets between two detected packets is

$$\begin{aligned} &\sum_{s=1}^N \left(1 - \frac{n}{N}\right)^{s-1} \frac{n}{N} s^2 - \left(\sum_{s=1}^N \left(1 - \frac{n}{N}\right)^{s-1} \frac{n}{N} s\right)^2 \\ &= \frac{N(N-n-N(1+n)^2(1-\frac{n}{N})^{2N}-n(1-\frac{n}{N})^N(nN-1))}{n^2} \\ &\approx \frac{N(N-n)}{n^2}. \end{aligned}$$

So, from this we see that we expect a gap of around $\frac{tN}{n}$ seconds between packets from a given victim. For example, if an attacker sends 100 packets per second, and one monitors 2^{24} IP addresses, one expects to see a new packet about every 2.5 seconds, and a spread of three standard deviations gives a 10 second gap. This rate of attack is quite low ([MVS01] claim intensities of as large as 600,000 packets per second), but this is only for illustration's sake (even at this rate, a SYN flood can be quite effective). Similar calculations can easily be done for other values of n , N , and t .

All these calculations have been predicated on the attacker choosing randomly from 2^{32} possible IP addresses. Many attack tools choose from a subset of these, such as only selecting octets from the range 1–254. This can be easily incorporated in the above analysis, by replacing the 2^{32} by the appropriate number.

Table 1: Data sets used in the SYN/ACK study.

Data Set Name	Duration	# days	# packets
April	April 4 – April 17	14	10,449
May	May 9 – May 17	9	23,264
June	June 1 – June 15	15	27,845
July	July 1 – July 15	15	59,666
Sept	Sept 1 – Sept 17	17	210,774
Oct	Sept 19 – Oct 15	26	1,253,714
Dec	Oct 28 – Dec 12	66	5,421,893
Jan	Jan 1 – Jan 31	31	665,392
Total		193	6,672,997

3 Experimental Results

To determine the extent that the assumptions of the theory are met, we consider a data set taken from a network of 2^{16} IP addresses. The data consists of unsolicited SYN/ACK packets received during two periods: April 4, 2001 – Jul 16, 2001 and September 1, 2001 – Jan 31, 2002. During these periods there were times when the sensor was down, for a total of 210 hours. The full data set consisted of 5,842 hours. We refer to the network on which the data was collected as the “protected network” throughout this discussion.

Missing data brings up one of the practical issues in a study of this kind. The protected network is a working network with a moderate load, and so there is the problem of determining which packets were solicited and which were not. This is exacerbated if there are packets that were not captured by the sensor, either because it was unable to handle the load or because the sensor was down. With SYN/ACK packets, we need to know if the SYN packet was sent. If it was, and the sensor failed to capture it, we will notice further packets (ACKs, PUSHs, etc), and can therefore determine that the SYN/ACK is a part of a legitimate session, and therefore not backscatter. For this reason, we focus on SYN/ACK packets in this section.

3.1 The Data

In order to avoid the gaps in our data collection, we broke the data into eight subsets, as depicted in Table 1. These are named according to the last month in which data was collected for that subset. As will be seen, this split was not perfect, as there were still a few gaps within the larger subsets.

We further restrict our investigation to web server (port 80) traffic. Thus we are considering only unsolicited SYN/ACK packets to our network from port 80. Figures 2 and 3 depict the data for the eight data sets. In these, the x-axis corresponds to time (in hours) from the start of the data set, and the y-axis corresponds to the victim (source) IP address. The IP address is always a 32-bit number with the highest octet in the highest bits. One dot is plotted for every packet (there is considerable overplotting in these pictures, but they serve to illustrate the data).

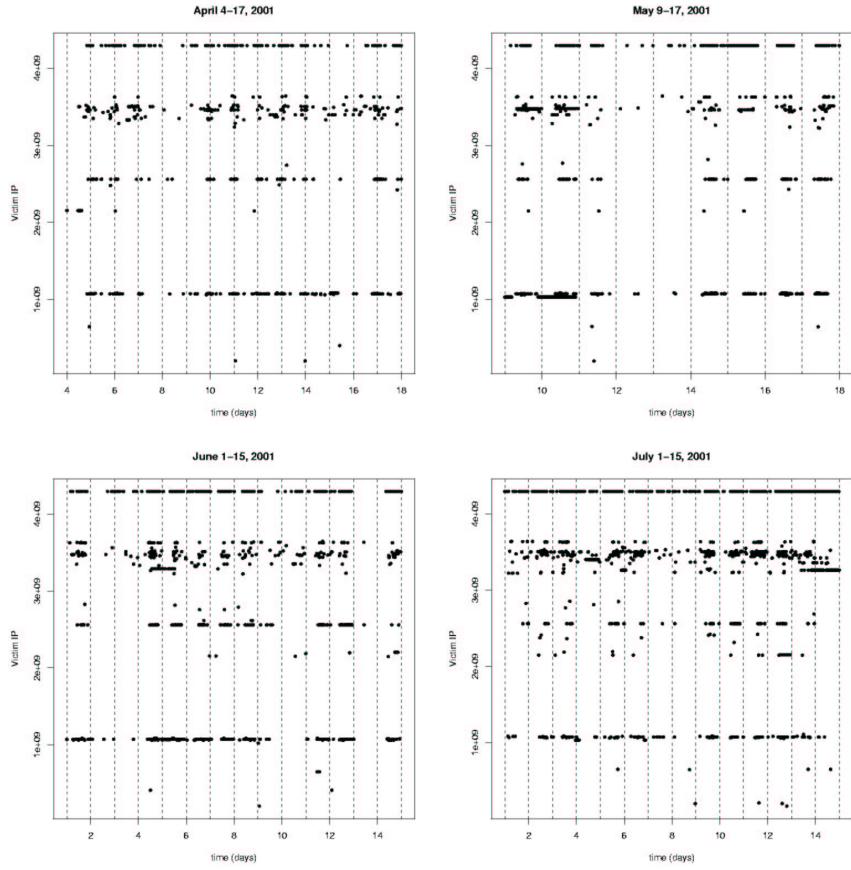


Figure 2: The attacks for the first four data sets. The x-axis is time, the y-axis is a 32-bit number corresponding to victim IP address. A dot is placed for each packet. Days are indicated by dotted lines.

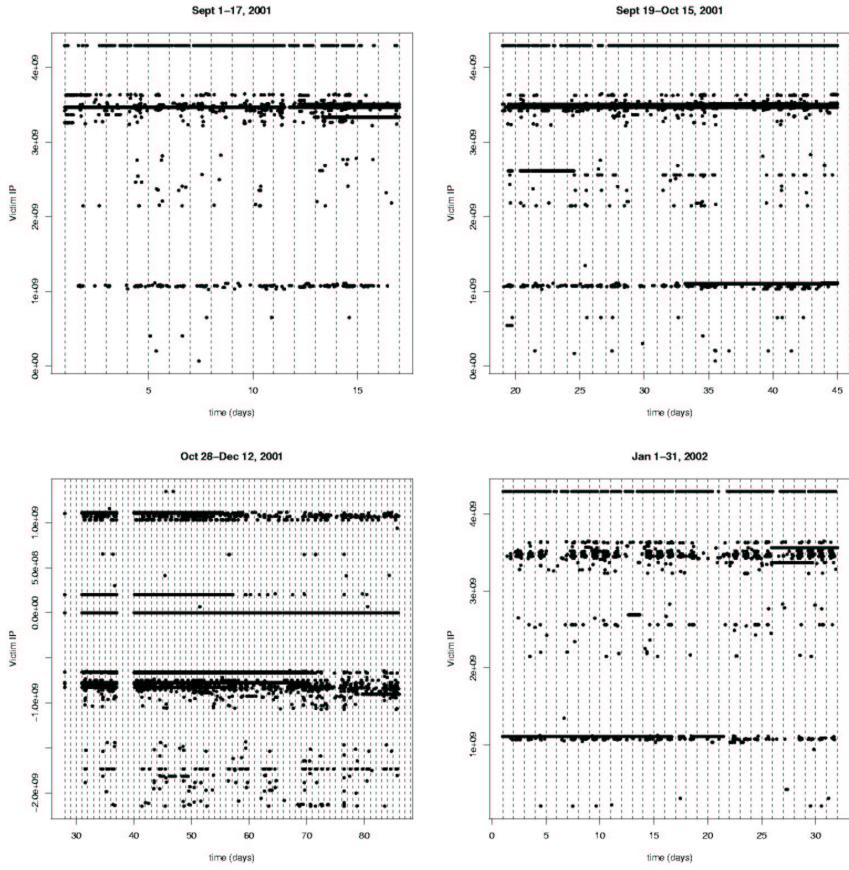


Figure 3: The attacks for the second four data sets. The x-axis is time, the y-axis is a 32-bit number corresponding to victim IP address. A dot is placed for each packet. Days are indicated by dotted lines.

Table 2: Number of attacks in each data set.

Data Set	T = 5 minutes	T = 1 hour
April	1,510	1,231
May	3,072	1,585
June	2,901	2,248
July	1,727	1,220
Sept	3,493	1,520
Sept/Oct	5,216	1,847
Oct/Dec	48,050	3,990
Jan	3,804	3,070

As can be seen in these Figures, there are a number of obvious attacks, as well as some very long-lived attacks. At this resolution it is impossible to count the attacks, and so we need to define exactly what we mean by an attack. For our purposes, we define an attack to be a sequence of packets from a single victim such that no gap between packets exceeds a fixed value (T). The results for two values for this threshold are presented in Table 2. If we restrict our definition to those attacks for which we received more than ten packets, we have the results reported in Table 3.

Table 3: Number of attacks in each data set for which there were more than 10 packets.

Data Set	T = 5 minutes	T = 1 hour
April	54	42
May	62	60
June	97	80
July	149	107
Sept	375	192
Sept/Oct	1,324	177
Oct/Dec	6,551	414
Jan	263	206

Some care is needed in counting the packets in an attack. Figure 4 depicts the packets from one victim. The destination (spoofed) IP addresses are on the y-axis, and time is on the x-axis. Note the characteristic “streaking” in this Figure. This is a result of resent packets. When the victim does not receive an answer to its SYN/ACK, it waits a small amount of time and then assumes the packet was lost in transit and resends the packet. It repeats this several times, each time increasing the wait period. This results in the “streaks” in the Figure, and in an over-estimate of the number of attack packets, if this is not taken into account. We define a resent packet to be one which agrees with a previous packet in the source and destination IPs and ports, and the acknowledgment number, and which is received within 1 minute of the first such packet. The numbers in Table 3 are computed using this definition, and so resends are not counted in the definition of an attack.

Note that the resends can also be used to help determine whether the packets are backscatter from a denial of service attack, or are a scan of the protected network. One

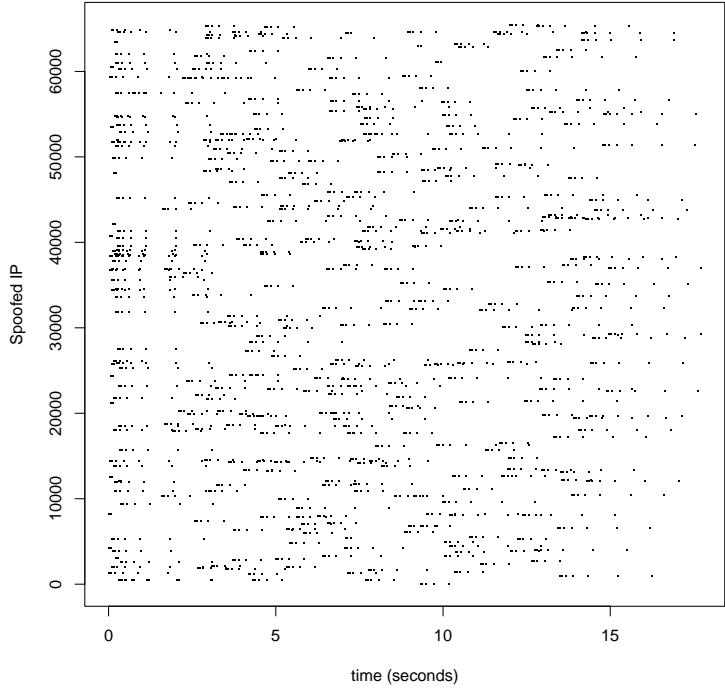


Figure 4: 2,160 packets from a single victim computer.

expects to see resends in backscatter. Scan tools that send a single packet per host/port will not show this pattern, while those that send multiple packets will typically not increase the time between packets, nor will they tend to have as large a time between packets as one sees with resent packets.

3.2 Attack Statistics

Figures 5 and 6 show histograms for the (log base 2 of the) number of packets detected in the attack, after removing resends, for the different data sets, for the two values of T. Our estimate of the number of packets in the original attack (assuming we believe that the attacker is selecting spoofed IP addresses uniformly, independently, from all 2^{32} possible IP addresses) can be obtained by multiplying the x-axis values by 16.

One observation is that the densities are surprisingly similar across all the data sets. The histograms appear to support a hypothesis of roughly three modes to the density, indicating (perhaps) the existence of three different types of attacks.

It is likely that many of the packets in the bin at 0 (corresponding to a single packet detected in the attack) represent errors in the process of selecting “unsolicited” packets

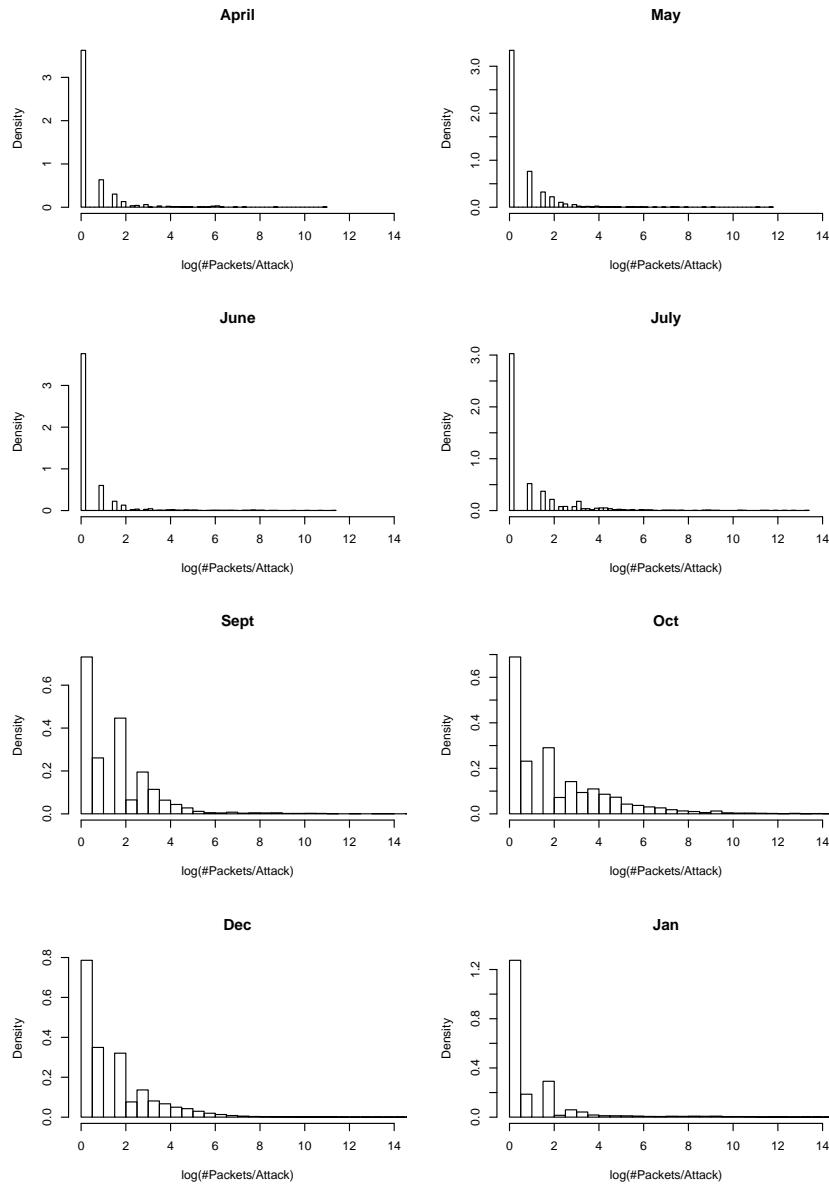


Figure 5: Histogram of the log (base 2) of the number of packets per attack. These counts are computed after the resends have been removed, as described in the text. T=5 minutes.

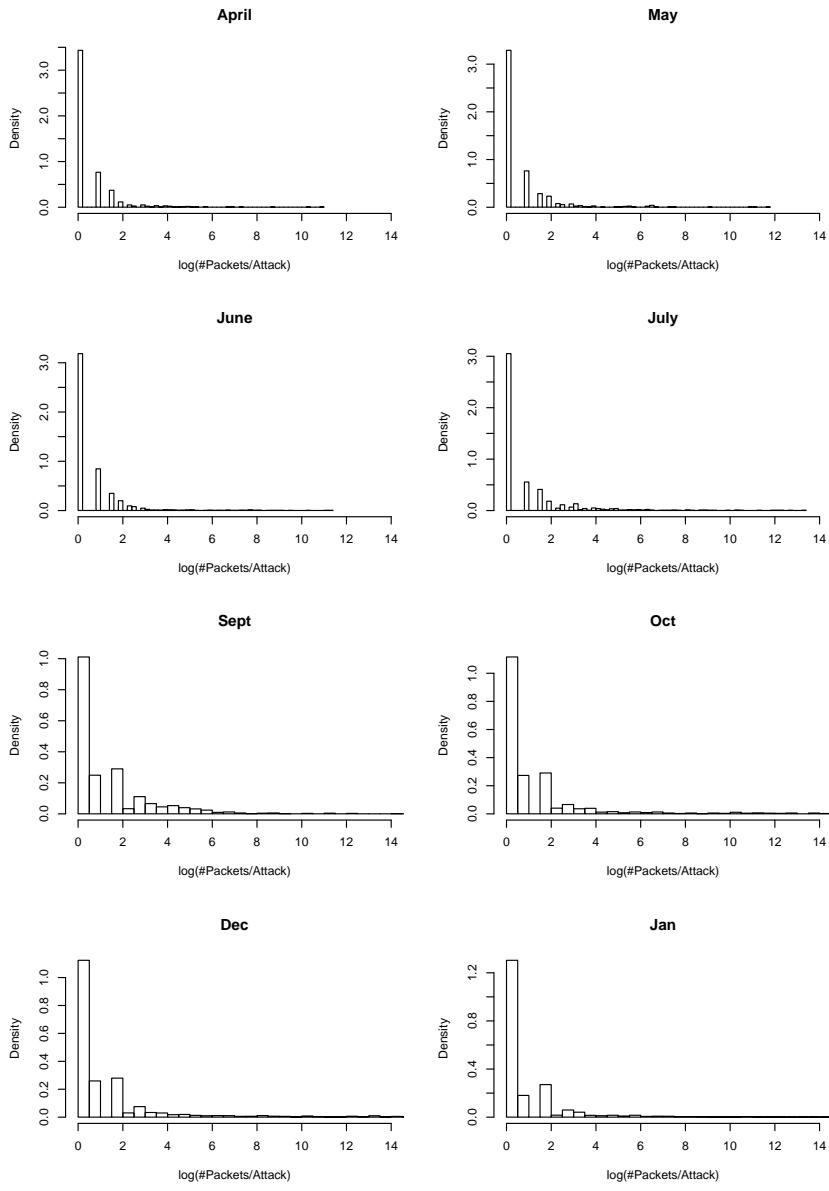


Figure 6: Histogram of the log (base 2) of the number of packets per attack. These counts are computed after the resends have been removed, as described in the text. $T = 1$ hour.

(for example, dropped SYN packets at the sensor), or are the results of other attacks (such as scans) or errors unrelated to cyber attacks.

Another explanation is that these are low packet rate attacks. Since a SYN flood need only fill the connection table of the victim, and keep it filled, an attack lasting only a few hours need not send more than 2^{16} packets (our estimate of the number of packets in an attack in which we observe 1 packet). Thus, it seems reasonable to suggest that for attacks against single servers (that cannot load-balance using a server farm, for example), attacks of this magnitude might be effective, and popular, accounting for the large number of such “attacks” detected.

We now turn to the question of whether the attack is random, that is, whether the spoofed IP addresses have been (uniformly) randomly selected from all 2^{32} possible IP addresses. Some pictures will be informative. While looking at pictures is subjective, and cannot detect subtle deviations from randomness, it can be very effective in detecting unexpected structure. (Note: in all the analysis which follows we use $T = 5$ minutes.) Figures 7 and 8 depict the packets from two victims. In these plots each packet is plotted as a dot, with x value corresponding to time and y value corresponding to the spoofed destination IP address. This is computed from the IP $x.x.y.z$ as $256y + z$. Figure 7 seems to pass a “looks random” test, while Figure 8 shows definite non-random structure. This manifests itself in two ways. First, it is obvious that the intensity of the the attack is not constant throughout the attack. Second, there is a diagonal structure detectable in the packets, showing a high degree of correlation. This attack does not satisfy our assumptions of independent random selection of spoofed IP addresses.

Figure 9 depicts attacks against two victims consisting of 9,674 and 22,716 packets. These show quite different structure, indicating several different attack tools were used. The top figure shows an attack with linear structure, overlapping an attack that looks to the eye to be fairly random. The bottom figure shows an attack with quite complicated dependence structure, with both a linear component, and some measure of clearly deterministic structure. This latter kind of attack was not observed in the data prior to the October data set.

Because of the systematic nature of the IP address selection in the bottom plot of Figure 9, the data passes a goodness-of-fit test (the Kolmogorov-Smirnov test) with flying colors. This test assumes (and does not test for) independence, and so is invalid for these data.

The above observations indicates that the blind use of goodness-of-fit tests will be of little use for these data. The changing intensity, and structure in many of the attacks make any assessment by a goodness-of-fit test problematic at best. Thus, each attack must be assessed individually, testing the different intensity regions separately. Further, it is vital that tests for dependence be used, in addition to distributional tests.

The number of large attacks (attacks with more than 1000 packets) seems to be increasing in these data. In April the average was approximately one such attack every two weeks, while by December the rate was approximately two per day. This may be a short time phenomenon (the rate does appear to have dropped to about 1 per day by January), or it may be a result of the increasing availability of attack tools or new attack paradigms. Further data is needed to assess this trend.

It might seem natural to assume that the attacks with linear structure are actually

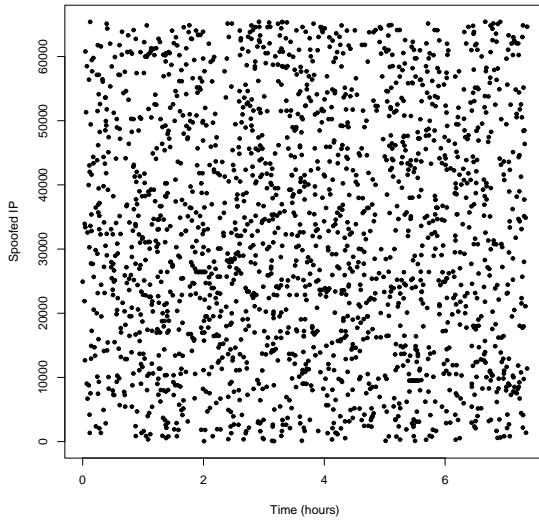


Figure 7: 1,997 packets from a single victim computer in April. The x-axis corresponds to the time of arrival of the packet, the y-axis corresponds to the last two octets of the spoofed destination IP address.

scans of the protected network, rather than backscatter from denial of service attacks. A perusal of the data shows that some of the attacks exhibiting linear structure do not have resent packets associated with them, lending credence to this hypothesis. Of the 69,773 attacks in the $T = 5$ minute data, 60,488 contained resent packets. Also, of the 248 attacks consisting of more than 1000 packets (after eliminating resends), 209 of them had resends associated with them. An alternative explanation would be that these victims have been configured to not send retries, but to rather drop the connection if an ACK packet is not received promptly. There is a technique, referred to as “SYN cookies”, in which the victim encodes state information in the SYN/ACK packet, and thus does not resend packets. See

<http://cr.yp.to/syncookies.html>.

The case against the hypothesis that these attacks represent scans of the protected network rests on three observations: first, it is unusual to scan a network from port 80, although one could certainly do this, provided one had the permission necessary to use this port; second, the linear structure does not manifest itself as a sequential pass through the IPs in the domain, but rather, on a small scale, has an apparent random component to it; third the existence of apparent “resend” packets argues against any of the known scan tools. Thus, regardless of the actual nature of the attack, the linear structure still remains to be explained.

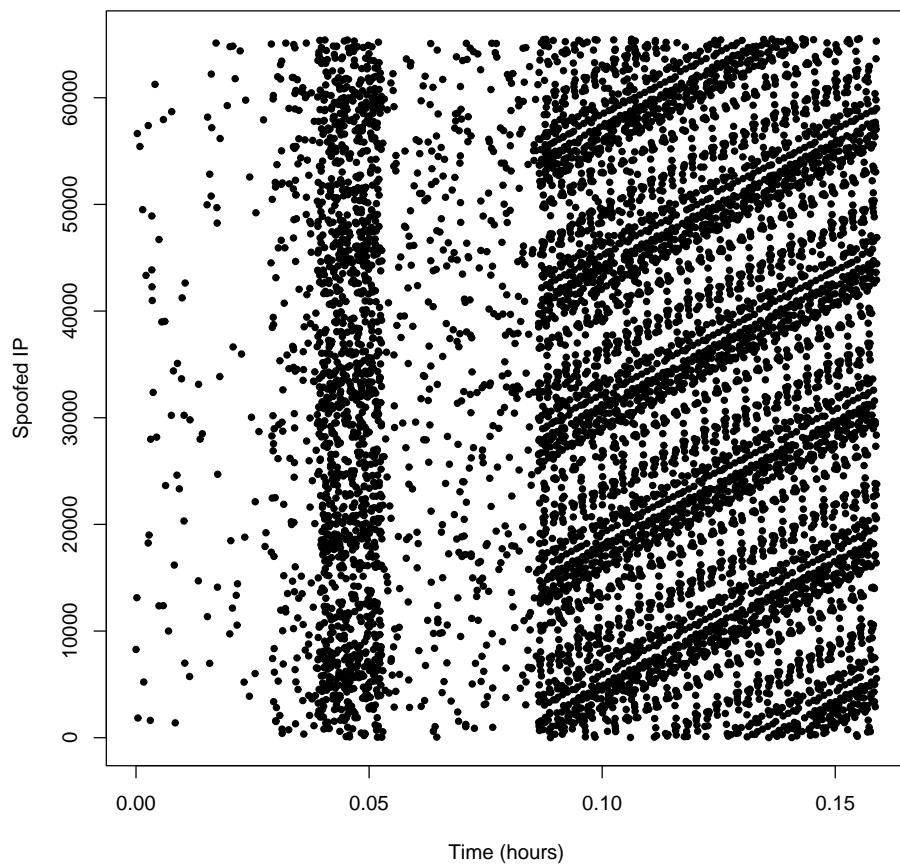


Figure 8: 7,137 packets from a single victim computer in October. The x-axis corresponds to the time of arrival of the packet, the y-axis corresponds to the last two octets of the spoofed destination IP address.

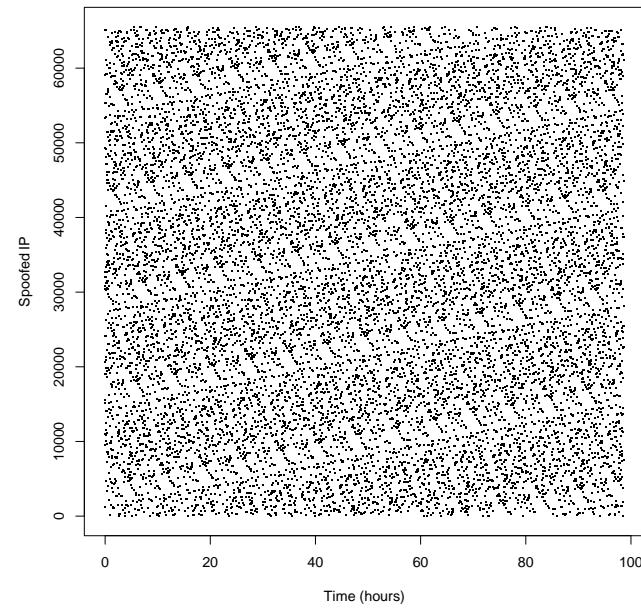
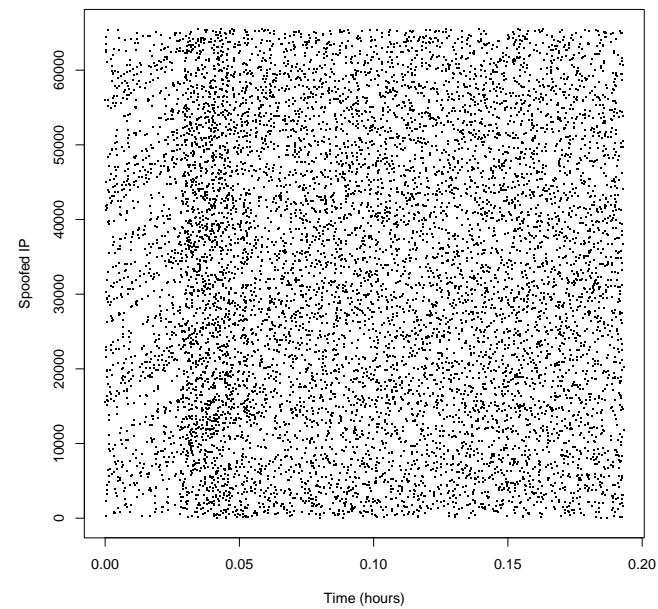


Figure 9: Attacks on two victims, showing nonrandom structure. The top figure represents 9,674 packets collected in July, while the bottom represents 22,716 collected in November.

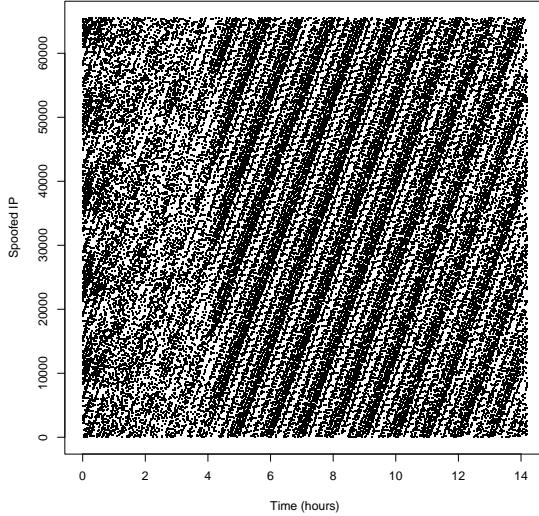


Figure 10: An attack on a United Kingdom Internet Service Provider.

Figure 10 is interesting, in that the same pattern is replicated over six different victims, corresponding to addresses xxx.xxx.xxx.3-8. This is an Internet service provider in the UK, which is obviously using a server farm to load balance. Our hypothesis is that this is a distributed denial of service attack, where different attackers received different IP addresses when the victim IP was resolved through a DNS lookup.

This raises the question of whether the linear pattern that we are seeing is an artifact of the attack or the response of the victim. Perhaps it is the load balancing that is inserting the linear structure into the attack. Perhaps the non-random IPs are a result of the timing of responses from the victim, rather than an error in the attack tool's random number generator.

As can be seen in the Figure, the character of the attack changes approximately four hours into the attack. Is this a change in the packets sent in the attack, or a change in the strategy of the victim(s)? This change occurs approximately simultaneously for all six victims, indicating that in either case the change is coordinated.

Victim action seems unlikely to be the cause, partly from the standpoint that there seems to be little value in it from the point of view of the victim, and partly from further observation of other attacks. A closer look at Figure 9 (top) reveals that there is an overlap between structured and non-structured attack patterns within the same victim. This is hard to reconcile with the hypothesis that victim response is responsible for the pattern. Thus, we believe that the pattern is a result of the activity of the attacker.

As can be seen in Figure 11, these data are highly correlated, which is hardly surprising given the pictures. One can use this information to build a model of the generating process.

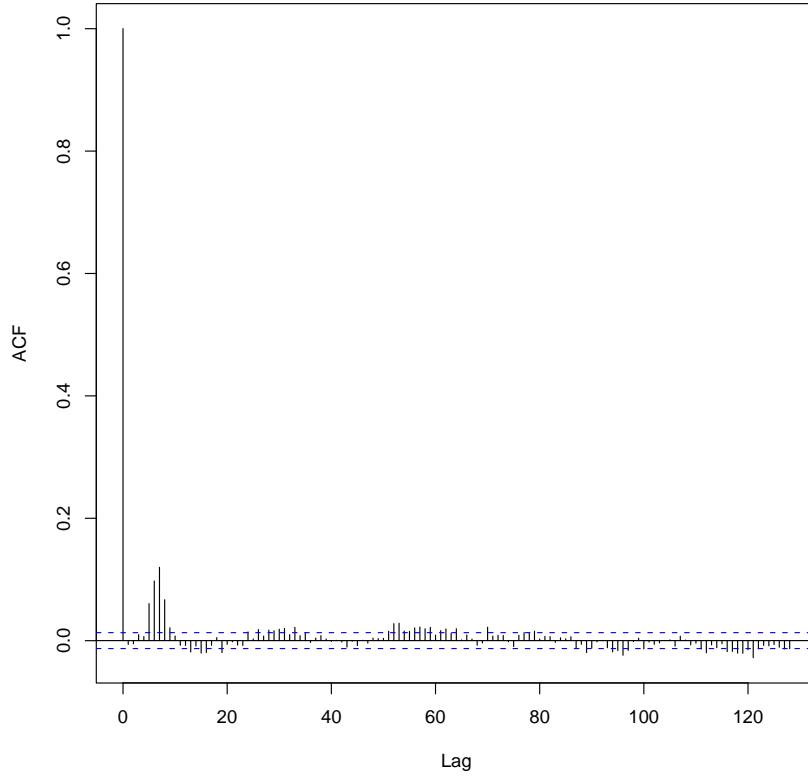


Figure 11: Autocorrelation function for the data in the lower plot of Figure 9, showing statistically significant autocorrelation.

Figures 12 and 13 provide a view of the number of attacks ongoing as a function of time. Attacks are defined as packets from individual victims, with no gap between packets of more than five minutes.

For the first four data sets, we see that while attacks occur throughout the time periods considered, there are rarely more than a few attacks at any time, and attacks typically last less than a day. There is some activity between May 10 and May 11, when there were 9 simultaneous attacks. Otherwise, the attack level is quite low.

The last four data sets show considerable activity. The ramp up in attack levels starts in mid September, and continues through to late November. At the height of the attacks there were over 30 victims under attack, and the attacks lasted for a month.

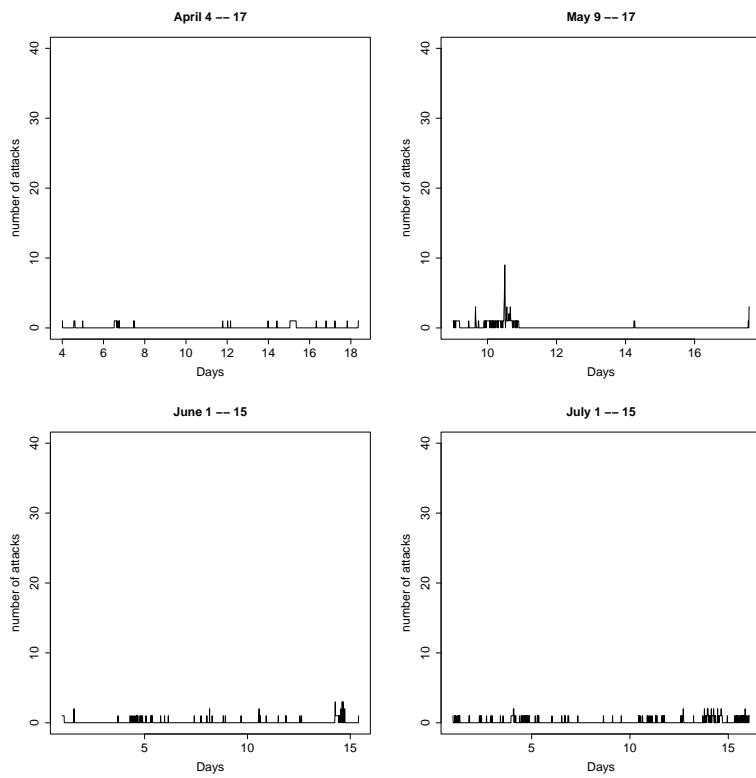


Figure 12: Number of attacks detected as a function of time.

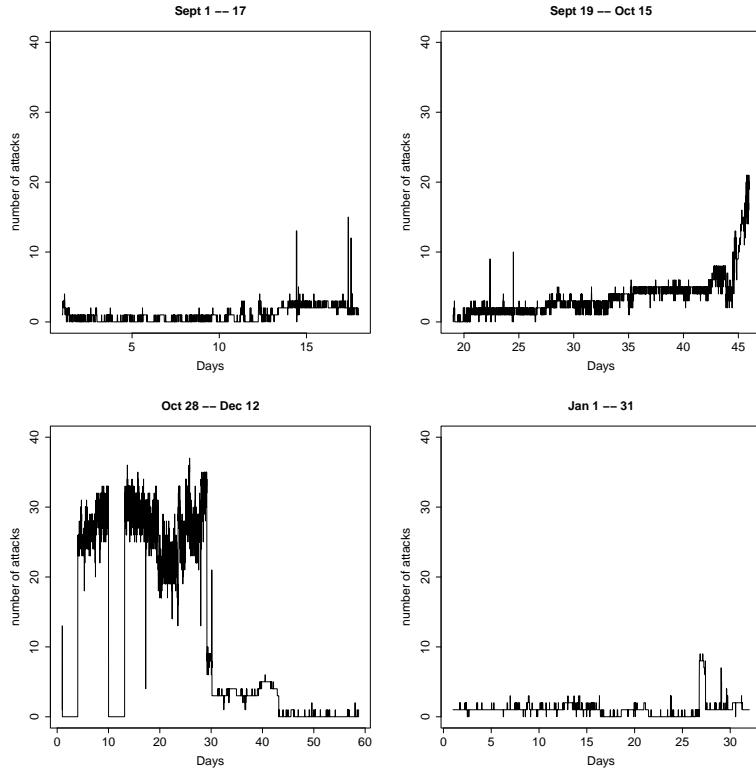


Figure 13: Number of attacks detected as a function of time. The gaps in the October plot are due to sensor drop-out.

4 Conclusions

The problem of measuring the number of denial of service attacks on the Internet is a difficult one, since many organizations are hesitant to report these attacks. Even when they do report them, they are often after the fact, and of little value for warning other potential victims of the threat. By utilizing the backscatter packets from certain classes of attacks, we have demonstrated that one can track these attacks in real-time, and we have shown that the attack level on the Internet can be quite high for extended periods of time.

Further work is needed in modeling these attacks, determining the algorithms used for generating packets, and thus providing some ability to classify the attacks. Knowing something about the attack can provide useful information for potential victims to use in defending against the attacks. Also, by monitoring trends in the attacks we can, potentially, identify when new classes of attacks are created, or when a new massive attack is underway.

The problem of determining the impact of the attack on the victim is a difficult one, which we have not addressed here. The victim machine could go down, in which case the backscatter packets would cease, but this may be indistinguishable from a cessation of the attack. It would be of value to determine whether subtle changes in the backscatter packets can be used as indications of the effect of the attack on the victim.

There are some methods available to defend against denial of service attacks, but these are not perfect and have difficulty with large distributed attacks. It would be valuable to incorporate that defense strategy into our analysis so that we could determine whether the victim is defending against the attack, and measure the effectiveness of the defense. With that said, the mere fact that we can track these attacks in real time without the cooperation of the victims and without adding to the load on the network is a powerful and useful tool. Clearly there are plenty of opportunities for statisticians to aid in the analysis of these data.

References

- [Che01] Eric Y. Chen. AEGIS: An active-network-powered defense mechanism against DDoS attacks. In Ian W. Marshall, Scott Nettles, and Naoki Wakamiya, editors, *Active Networks: IFIP-TC6 Third International Working Conference*, pages 1–15. Springer, 2001. Lecture Notes in Computer Science 2207.
- [Mar01] David J. Marchette. *Computer Intrusion Detection and Network Monitoring: A Statistical Viewpoint*. Springer, New York, 2001.
- [MVS01] David Moore, Geoffrey M. Voelker, and Stefan Savage. Inferring Internet denial-of-service activity. Available on the web at www.usenix.org/publications/library/proceedings/sec01/moore.html, 2001. USENIX Security '01.
- [NNM01] Stephen Northcutt, Judy Novak, and Donald McLacan. *Network Intrusion Detection. An Analyst's Handbook*. New Riders, Indianapolis, IN, 2001.

Wilks Award Banquet Address

Considerations of Inspection for Homeland Security with Cross Linkages to Quality Control, Game Theory, and Stochastic Simulation

James R. Thompson, Noah Harding Professor of Statistics, Rice University

Abstract: It is proposed to develop models for inspection (both of people and containers) utilizing insights from quality control. In Acceptance-Rejection Quality Control, we balance costs of sampling with those of passing bad items. In the Deming Paradigm of Statistical Process Control, we carry out sampling for the purposes of system improvement as opposed to lot validation. Homeland Security issues embody potential for both philosophies and we will be attempting both lot validation and system improvement. The classical Acceptance-Rejection paradigm, in the Homeland Security situation, can be shown readily to lead to inspection of all airline passengers and all baggage. It is extremely costly and leads to a situation where ticket prices soar and/or the state heavily subsidizes security.

In the United States, current allocations of funds for inspection require different strategies than 100% inspection. This leads to the use of covariate information concerning the inspected population so that the inspection should involve stratification. It also leads to strategies whereby the input stream of customers with poor risk profiles might be modified by political and other strategies.

In classical quality control, the system inspected is not sentient. In Homeland Security, the terrorist commanders are intelligent agents who will attempt to use information about inspection protocols to lessen the probability of discovery, both of terrorists and their baggage. Thus, we need to develop mixed strategies (in the sense of von Neuman-Morgenstern) hybrids for inspection strategies. Rather than seeking to deal with models simple enough for closed form solution, it is proposed to go rather toward realistic models whose analysis requires stochastic simulation .

Although the immediate charge here is for Homeland Security, it should be noted that quality control-game theoretic simulation models may also be created when planning inspections for weapons of mass destruction and/or their development in another country.

References

Thompson, James R. and Koronacki, Jacek (2001). *Statistical Process Control: the Deming Paradigm and Beyond*. Boca-Raton, Florida: Chapman & Hall. 431 pages.

Thompson, James R. (2000). *Simulation: A Modelerís Approach*. New York: John Wiley & Sons, 297 pages.

General Session 3

Stress-Strength Testing: Some Classical Approaches and Some New Formulations and Results

Francisco J. Samaniego, University of California, Davis

Abstract: If Y is a random variable representing the breaking strength of a given material and X is a random variable measuring the stress placed on that material, then the probability that the material will survive the stress to which it is subjected is simply $P = P(X < Y)$. In most applications involving stress-strength testing, the variables X and Y are modeled as independent. The probability P arises in a broad array of applications, and has been studied extensively. Some of the highlights of the literature in this area will be reviewed, including Birnbaum's (1956) nonparametric confidence bounds for P and subsequent improvements, parametric inference regarding P pioneered by Mazumdar (1970) and Church and Harris (1970), work on general parametric families based on large sample theory (see Johnson (1988) and Enis and Geisser's (1971) Bayesian treatment of the estimation of P . After surveying these classical results, our focus will shift from estimating P to the problem of estimating the stress and strength distributions themselves from available data. Two particular formulations of stress-strength testing will be discussed. Arcones, Kvam and Samaniego (2002) defined the notion of stochastic precedence as follows: X stochastically precedes Y if $P(X < Y) > 1/2$. In most industrial and military applications, test subjects can, more often than not, withstand the stress to which they will be subjected. This leads to a constrained nonparametric estimation problem whose solution will be described. The second formulation to be discussed involves the analysis of autopsy data, for example, data on welded rebar from a collapsed bridge. Estimation of the distributions of stress and strength is shown to be feasible and efficacious in selected parametric settings.

Random Disambiguation Paths

Carey E. Priebe, Johns Hopkins University

Abstract: We wish to navigate from a specified source to a specified destination through a spatial configuration of detections and associated potential risk regions. Along with each detection comes a mark indicating the probability that entering the associated risk region incurs non-zero risk. In accordance with application we may, upon approaching a detection, disambiguate the associated risk (that is, determine conclusively if that risk is indeed non-zero) at cost c to the overall traversal time.

A random disambiguation path is a path-valued random variable whose various values represent different paths taken depending on the results of disambiguations; the actual path depends on the (unobserved at the outset) actual risks. Our goal is to determine the random disambiguation path achieving the minimum expected zero-risk traversal time.

An illustrative application for random disambiguation paths is mine countermeasures path planning -- navigating through a field of detections, each of which may or may not be an actual mine, but each of which is marked by the detector with a probability that the

detection is indeed a mine. A sensor is available which allows us, when close enough, to determine conclusively whether or not the detection is truly a mine.

Contributed Session 5

Benefits of Non-Destructive Evaluation (NDE) vs. Destructive Testing for Bayesian Reliability Estimation

Paul Deininger, Los Alamos National Laboratory; Shane Reese, Los Alamos National Laboratory and Brigham Young University; Michael Hamada, Los Alamos National Laboratory; Robert Krabill, Los Alamos National Laboratory

Abstract: Can non-destructive testing (NDE) of weapon parts or assemblies produce enough data to effectively obviate the need for destructive evaluation (DE), even in cases where the NDE data is less accurate and precise than the DE data? The data from both NDE and DE are used in either the normal frequency or Bayesian estimation of the reliability of parts or assemblies. It is well known that NDE examination is considerably less monetarily expensive than DE examination, even when the loss of the part or assembly from DE is discounted. For a given set of data desired, one can perform several to many NDE examinations for the price of a single DE. The technical problem of how many NDE tests with inaccurate or imprecise measurements are equal in value to a single DE test is explored by using a statistical framework for determining the possible advantages of NDE data over DE data. Two cases are considered: pass/fail NDE data and continuous measurement NDE data. Examples are shown using NDE and DE radiography data from a weapon component where both the misclassification probability and the number of NDE trials needed are calculated from the data rather than assumed. The analysis shows that given these values of misclassification probabilities, a small number of NDE trials can produce an equal or larger amount of usable information compared to a single DE, information that is suitable for estimating the weapon or weapon component reliability. The results imply that by completing several replications of each type of NDE test on the weapon parts, one may cost-effectively estimate the weapon reliability without destruction of the weapon or component. This conclusion, however, depends on the values of the misclassification probabilities, the relative costs of NDE vs DE, and the assumption that NDE can obtain the same type of desired data as DE without qualitative losses of any kind.

Munitions Stockpile Reliability Assessment

Alyson Wilson, Los Alamos National Laboratory; Nicholas Hengartner, Los Alamos National Laboratory

Abstract: Both the DoD and DOE maintain stockpiles of munitions. These systems are faced with issues of reliability and performance, which are often made more complex due to diverse storage conditions and aging effects. Information about the systems comes from a variety of sources, including component testing, engineering judgment, similar system data, and full-system testing. However, there may not be data collected about every component or subsystem. We will present a Bayesian hierarchical approach to estimating full system and component reliability for munitions stockpiles that integrates a variety of information sources and treats the problem of combining data and priors from various sources in a consistent fashion.

Hierarchical Models for Software Testing and Reliability Estimation

Todd L. Graves, Los Alamos National Laboratory; John C. Kern II, Duquesne University

Abstract: It is generally impossible to test software exhaustively. Instead, one must select examples of inputs to the software, run them, and ascertain what their success or failure imply about untested inputs. Bayesian hierarchical modeling is an excellent methodology for inference based on experimentation in related but nonidentical conditions. We will discuss two different hierarchical modeling approaches, one due to Wooff, Goldstein, and Coolen (2002), and one our own. We will also discuss implementation of analyses based on these two modeling strategies in software.

Contributed Session 6

An Almost Natural Application of Bayesian Statistics in Packaging Quality Control

John A. Wasko

David B. Kim

Department of Mathematical Sciences

United States Military Academy

West Point, NY 10996

October, 2002

Abstract

Lower cost homogeneous items are advertised and sold by quantity in a single package across all business sectors, from food service to medical to manufacturing. Quality control measures come at a price that is somewhat exacerbated by operating in a low cost, high quantity market. It is therefore of interest to find a cost-effective way of ascertaining the number of items in a package.

Straightforward application of the central limit theorem and the Bayes' theorem allows us to find the distribution of the number of items in a package conditioned on the weight of the package, provided a prior distribution of the number of items in the package is known. Initial and updated prior distribution estimates can be formed via manufacturer's reliability data and sampling data. This information combined with a standard quality control measure of package weight provides an almost natural Bayesian framework. We will show that well-established Bayesian techniques and testing can be employed to provide a useful tool which addresses this problem.

1 Introduction

In today's highly competitive business environment, the mass packaging of homogeneous items is a common practice whether it is done by hand or by an ultra precise computer controlled machine. Both the manufacturer of the items and the customer are interested in ensuring that the proper number of items are contained in each package. From the manufacturer's point of view, she needs to implement the necessary quality control measures at the lowest possible cost. The customer may be interested in economical and efficient ways of determining if each package contains at least the contracted amount.

Most of us have seen a jar full of pennies—presumably in thousands—and wondered how many pennies were in it. One sensible solution in a situation like this would be to weigh the jar and try to estimate how many pennies are in the jar based on the weight. We propose that one way of tackling the quality assurance problem described in the previous paragraph is simply by weighing the package and incorporate the information from the distribution of the weights of the items. Even as both frequentist and Bayesian analyses may be employed in the situation, the availability of the prior information and other factors make the situation almost ideal for the Bayesian framework to be used in.

In this report, we concentrate on the situation where a customer is interested in ways of accepting or rejecting a package based on the measurement of its weight. In the next section, both frequentist and Bayesian models are constructed for the problem. In the subsequent section, some numerical results are presented.

2 Modeling

2.1 Motivation and Development of a Bayesian Model

Let N be the number of items; X_i the weight of the i th item; $W_K = \sum_{i=1}^K X_i$. If a single package is given and its weight is measured, then we may treat N as a parameter, and we have one observation $W_N = w_n$. A corresponding frequentist approach as well as a Bayesian approach conditioned on the observed weight can be employed in the situation.

In many practical applications, n should be at least in hundreds, if not in thousands. Whether or not X_i is normally distributed, W_n should be normally distributed by Central Limit Theorem. That is,

$$W_K | K = n \sim N(n\mu, n\sigma^2) \quad (1)$$

where $\mu = E(X_1)$ and $\sigma^2 = \text{Var}(X_1)$. To facilitate the analytical treatment, we will assume from this point on that $X_1 \sim N(\mu, \sigma^2)$ and that both μ and σ^2 are known. (We also assume that the measurement is that of net weight: the tare weight has been subtracted from the raw measurement.)

Clearly, $\frac{W_K}{\mu}$ is an unbiased estimator of n . This is also the solution many will arrive at in the problem of counting the number of pennies in a jar. At this point, let us ask the following question: what are the sources of randomness in the situation? First, there is the item to item difference in weight, which leads us to model the weight of individual item as a random variable X_i . The packaging mechanism is another source of variation. So N is a proper random variable with a probability structure, and we need to model it as such. For the situation we are concerned with in this report where only the weight of one package is measured and used to estimate the number of items in it, we then have $N = n$ for a given box and wish to estimate n , and the probability distribution of N may be used as a prior distribution in a Bayesian framework. If a measurement

yields the weight w_n , we can write the probability mass function of n conditioned on the measured weight w_n as follows:

$$p(n|w_n) \propto f(w_n|n)\pi(n)$$

That is,

$$p(n|w_n) = \frac{f(w_n|n)\pi(n)}{\sum_j f(w_n|j)\pi(j)} \quad (2)$$

Since n , the number of items, is discrete, the marginal of w_n is obtained by summation. Note also that the index n in w_n is not to be changed from a term to another term in the summation since w_n denotes the observed weight of the package. Treating the conditional *pmf* in Eq.(2) as the posterior in a typical Bayesian framework, we may employ any of widely used Bayesian inferential tools that would suit our purpose.

We should also make one remark here concerning any conjugate prior. The key feature of our model is in Eq.(1). The “parameter” n appears both in the mean and the variance of the normal distribution, causing n to appear both in and out of the exponential such that the well known result that the conjugate relation between normal distributions cannot be applied here even if we were to allow n to be continuous for the sake of technical convenience.

2.2 Bayesian Hypothesis Testing Using Generalized 1-0 Loss

The case we are interested in, the case where the customer is going to either accept or reject the shipment based on the weight of the box, is clearly the one where we can apply the hypothesis testing method. The customer has an obvious interest in

$$H_0 : n \geq n_0$$

$$H_1 : n < n_0$$

where n_0 is the contracted amount. One could easily construct the hypotheses of interest for the manufacturer as well, and the following can be just as easily adapted.

In a Bayesian framework, it makes perfect sense to ask what the probability a hypothesis is correct is, since we have, in our model, probability distribution for the parameter(s) which the hypotheses are describing. Given the posterior distribution found using Eq.(2), we can compute the posterior probability that each hypothesis is correct. That is,

$$P(H_0|w_n) = \sum_{n \geq n_0} p(n|w_n) \quad (3)$$

$$P(H_1|w_n) = \sum_{n < n_0} p(n|w_n). \quad (4)$$

Choosing the hypothesis with the higher posterior probability corresponds to the Bayes' rule using 0-1 loss. In many situations, however, the customer may have different loss for different types of error. We can incorporate that information by using a generalized 0-1 loss function. The generalized 0-1 loss function allows us to incorporate differing loss based on the type of error (I or II) whereas the usual 0-1 loss assigns the equal loss on both types of error. Let C_I be the cost

the customer will incur when she rejects a package with the acceptable number of items, $n \geq n_0$ —that is when she makes the type I error—and let C_2 be the cost when she accepts the package when it does not contain the acceptable number of items, $n < n_0$, corresponding to the type II error. It is a well known result that rejecting the null hypothesis when the posterior probability that it is correct, $P(H_0|w_n)$, is less than $\frac{C_{II}}{C_I+C_{II}}$ (or alternatively $P(H_1|w_n) > \frac{C_I}{C_I+C_{II}}$) corresponds to a Bayes rule, having the optimality properties that go with being a Bayes rule (see Casella and Berger (1990) or Berger (1993)).

3 Numerical “Sensitivity” Analysis

Using the decision rule described in the last section, to reject the shipment when $P(H_1|w_n) > \frac{C_I}{C_I+C_{II}}$, we know it will be a “good” decision in theoretical terms knowing that it is a Bayes rule. In practice, μ and σ^2 will also vary from one setting to another, and a user of the method may be interested in how the method performs under such different settings. We have of course performed the usual posterior sensitivity analysis to see if prior misspecification affects the inference in significant manner. The numerical investigation of how the method performs under these different settings was done using Microsoft Excel. We tabulated the values of $P(H_1|w_n)$ for various settings. The loss function we used corresponded to the case where $C_I = 9C_{II}$ so that the value we are comparing the posterior probability to is $\frac{C_I}{C_I+C_{II}} = .9$. The values of $P(H_1|w_n)$ in bold red then means that the shipment is rejected.

Suppose now $n_0 = 200$, $\mu = 3$. We used a triangular prior on 190 and 240. We varied the mode and σ^2 . Table 1 shows the posterior probabilities and the decisions based on them when the observed weight was 605. The robustness in prior specification is demonstrated by the proposed method giving mostly consistent decision in the same rows. In Table 2, we looked at the posterior

	Mode of the prior									
0.9696	193	198	204	209	215	221	226	232	237	
1.00%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
2.00%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
4.00%	0.9904	0.9904	0.9920	0.9920	0.9920	0.9920	0.9920	0.9920	0.9920	
6.00%	0.9286	0.9286	0.9408	0.9408	0.9408	0.9408	0.9408	0.9408	0.9408	
8.00%	0.8756	0.8756	0.8966	0.8968	0.8968	0.8968	0.8968	0.8968	0.8968	
10.00%	0.8529	0.8529	0.8774	0.8781	0.8781	0.8781	0.8781	0.8781	0.8781	
12.00%	0.8485	0.8484	0.8731	0.8744	0.8744	0.8744	0.8744	0.8744	0.8744	
14.00%	0.8524	0.8520	0.8755	0.8777	0.8777	0.8777	0.8777	0.8777	0.8777	
16.00%	0.8597	0.8589	0.8807	0.8838	0.8838	0.8838	0.8838	0.8838	0.8838	
18.00%	0.8679	0.8666	0.8868	0.8907	0.8908	0.8908	0.8908	0.8908	0.8908	
20.00%	0.8761	0.8742	0.8929	0.8976	0.8977	0.8977	0.8977	0.8977	0.8977	
22.00%	0.8839	0.8812	0.8986	0.9040	0.9042	0.9042	0.9042	0.9042	0.9042	
24.00%	0.8910	0.8877	0.9039	0.9099	0.9102	0.9102	0.9102	0.9102	0.9102	
26.00%	0.8974	0.8934	0.9087	0.9152	0.9157	0.9157	0.9157	0.9157	0.9157	
28.00%	0.9031	0.8986	0.9130	0.9199	0.9207	0.9207	0.9207	0.9207	0.9207	
30.00%	0.9082	0.9031	0.9169	0.9242	0.9252	0.9252	0.9252	0.9252	0.9252	
32.00%	0.9128	0.9072	0.9204	0.9280	0.9292	0.9293	0.9293	0.9293	0.9293	
34.00%	0.9168	0.9109	0.9235	0.9315	0.9329	0.9330	0.9330	0.9330	0.9330	
36.00%	0.9204	0.9142	0.9264	0.9346	0.9363	0.9364	0.9364	0.9364	0.9364	
38.00%	0.9236	0.9171	0.9290	0.9373	0.9393	0.9395	0.9395	0.9395	0.9395	
40.00%	0.9265	0.9198	0.9314	0.9398	0.9421	0.9424	0.9424	0.9424	0.9424	

Table 1

robustness for different values of μ , and similar robustness is demonstrated.

		Mode of the prior									
		0.97	193	198	204	209	215	221	226	232	237
% diff in w from the np	3.00%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	2.00%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	.100%	0.9902	0.9899	0.9873	0.9873	0.9873	0.9873	0.9873	0.9873	0.9873	0.9873
	0.50%	0.7975	0.7973	0.7712	0.7712	0.7712	0.7712	0.7712	0.7712	0.7712	0.7712
	0.25%	0.5659	0.5659	0.5392	0.5392	0.5392	0.5392	0.5392	0.5392	0.5392	0.5392
	0.00%	0.4358	0.4358	0.4360	0.4360	0.4360	0.4360	0.4360	0.4360	0.4360	0.4360
	0.25%	0.5550	0.5550	0.5802	0.5802	0.5802	0.5802	0.5802	0.5802	0.5802	0.5802
	0.50%	0.7872	0.7872	0.8102	0.8102	0.8102	0.8102	0.8102	0.8102	0.8102	0.8102
	1.00%	0.9891	0.9891	0.9913	0.9913	0.9913	0.9913	0.9913	0.9913	0.9913	0.9913
	2.00%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	3.00%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2

4 Conclusion

We have proposed a method of assuring or testing the number of items in a package only by its weight measurement and the knowledge of the distribution of the weights of the individual items. The robustness of the Bayes' rule obtained using a generalized 0-1 loss function was demonstrated.

The case we examined in this report is where the testing is done on a box by box basis. We are currently working on an extension of the method which can be phrased in the following question: Suppose we are given the weight measurements of m packages. These packages are supposedly packed with the same mechanism/procedure so that these may be modeled as iid random variables.

From these m measurements of weights, can we find out about the marginal distribution of n , the amount the packing mechanism puts in each box? This could be used to test the reliability of the packing mechanism quickly (without counting items one by one) and also to better specify the prior that is to be used in the case we looked at in this report.

References

- Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. New York: Springer-Verlag. Corrected reprint of the second (1985) edition.
- Casella, G. and R. L. Berger (1990). *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

RELATIONSHIP BETWEEN TOXICITY VALUES FOR THE HEALTHY SUBPOPULATION AND THE GENERAL POPULATION

Ronald B. Crosier and Douglas R. Sommerville, PE
U.S. Army Edgewood Chemical Biological Center
5183 Blackhawk Road
Aberdeen Proving Ground, MD 21010-5424

douglas.sommerville@sbccom.apgea.army.mil
ronald.crosier@sbccom.apgea.army.mil

ABSTRACT

The present chemical warfare agent toxicity estimates are not suitable for use with the general population (GP) because they are framed for male soldiers. A method was created to convert the median effective dose and probit (or Bliss) slope to estimates applicable to the GP. It was assumed that individual susceptibilities have a log-normal distribution. Two mathematical models were developed to describe a healthy or sensitive subpopulation (SP). In the tail model, the SP consists of all individuals having susceptibilities within a tail of the GP distribution. In the bell model, the SP has a lognormal distribution. The median and the probit slope of an SP were determined as a function of the SP size. The two models gave similar results. Historical military demographics were used to estimate the size of the healthy SP from which military personnel are drawn. Uncertainty factors were obtained from the tail and bell models. Uncertainty factors from both models were consistent with the results of two previous studies that quantified differences between populations. Based on our analysis, revisions are required in the intraspecies uncertainty factors used in establishing proposed acute exposure guideline levels for threshold lethality due to inhalation of nerve agents.

The complete documentation for this presentation is available from the following published technical report:

Crosier, Ronald B. and Sommerville, Douglas R., *Relationship Between Toxicity Values for the Military Population and Toxicity Values for the General Population*, ECBC-TR-224. U.S. Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, MD, March 2002. AD-A400 214. (40 pages).

The technical report has been approved for public release, distribution is unlimited. Registered users should request copies from the Defense Technical Information Center; unregistered users should direct such requests to the National Technical Information Center.

The following are the individual slides for the presentation. The authors wish to thank Ms. Robyn Lee of Robyn B Lee and Associates LLC for presenting this paper (on short notice) in the place of Mr. Sommerville at the Eighth US Army Conference on Applied Statistics, North Carolina State University, Raleigh, NC, 31 October 2002.



Relationship Between Toxicity Values for the Healthy Subpopulation and the General Population

**Eighth US Army Conference on
Applied Statistics
Raleigh, NC
31 October 2002**

Robyn Lee
Contractor for US Army MRICD
Presenting on behalf of
Ronald B. Crosier
Douglas R. Sommerville, PE

U.S. Army Soldier and Biological Chemical Command
Edgewood Chemical Biological Center
5183 Blackhawk Road, ATTN: AMSSB-RRT-IM, Bldg. E5951
Aberdeen Proving Ground, Maryland, USA 21010-5424

Email: Douglas.Sommerville@sbccom.apgea.army.mil
Phone: (410) 436-4253
FAX: (410) 436-2742

Comparison of Populations via Mathematical Modeling

Edgewood Chemical Biological Center

- Goal: To develop a mathematical model to describe differences in agent toxicity between a healthy subpopulation (SP) and the general population (GP)
 - Parameter value conversion between populations—median dose/dosage values and probit slopes
 - No known work previously done on this subject
- Only one model parameter: SP Size
- Key assumptions
 - Individual susceptibilities for the GP have a normal distribution (bell-shaped curve) of Log (Effective Dose) or Log (ED) values
 - SPs (either healthy or sensitive) are represented by one of two models: Bell or Tail
- Disclaimer: The content of this poster is not to be construed as an official Department of the Army position unless so designated by other authorizing documents

1

Application to Decision Support Methods

Edgewood Chemical Biological Center

- Casualty estimations
 - Current CW agent toxicity values (LCT_{50} or ECT_{50} and probit slope) for military subpopulation are not appropriate for use in estimating casualties for the general population exposed to CW agent attacks or incidents
 - Using military toxicity values for the general population will result in the underestimation of civilian casualties
- Method offers a simple means to arrive at reasonable approximation of civilian toxicity values based on an extrapolation using mathematical/statistical modeling from known military values
 - Algorithm for toxicity value conversion can be easily programmed into transport & dispersion models

2

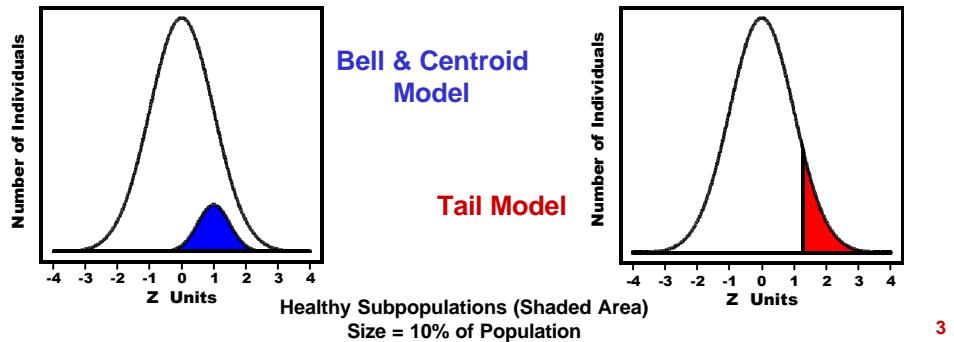
Models Used to Compare Population Differences

Edgewood Chemical Biological Center

$$Z = m_{GP} [\log(ED) - \log(ED_{50})]$$

m_{GP} = Probit slope of GP

ED_{50} = Effective Dose 50% for GP



Defining a Subpopulation

Edgewood Chemical Biological Center

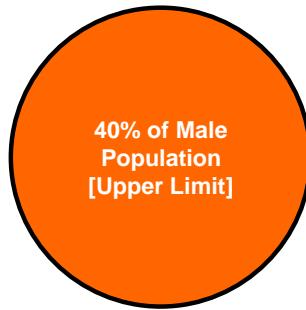
A Subpopulation can be defined in a variety of ways

- Healthy Subpopulations
 - Military
 - Workplace
- Sensitive Subpopulations
 - Infants
 - Elderly
 - People with chronic medical conditions
- Other Subpopulations
 - Gender
- Mathematical modeling can account for gender differences
 - Separately apply either Bell or Tail Model to each gender
- Use of demographics to estimate SP size
 - Existing chemical warfare (CW) agent toxicity values developed for military SP
 - Workplace SP used for industrial chemicals

4

Demographics of U.S. WWII Military Subpopulation

Edgewood Chemical Biological Center



Peak Military Strength (1945)

- Many age eligible males medically unfit for service
- Some healthy males exempted from service

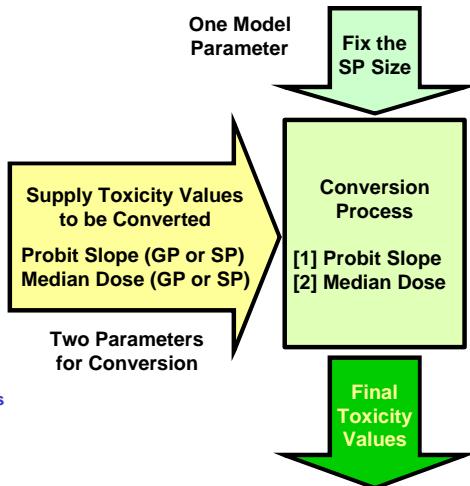
Age Eligible Males (18 to 45 years old)

5

SP Size Role in Models and Conversion Process

Edgewood Chemical Biological Center

- Tail Model**
 - Selection of SP size fixes mathematical relationship between SP and GP
 - Example: If Size = 10%, then ED_{50} of a sensitive SP is located at ED_{05} of the GP
- Bell and Centroid Models**
 - Selection of SP size does not determine SP mean and standard deviation
 - SP bell curve must remain underneath GP bell curve
 - Range of feasible values exists for SP mean and standard deviation
 - Bell Model—Maximum difference in means of SP and GP
 - Centroid Model—Located at centroid of feasible range

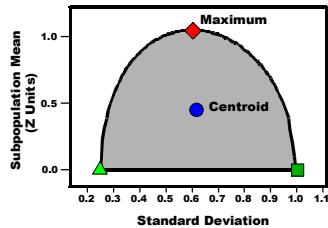


6

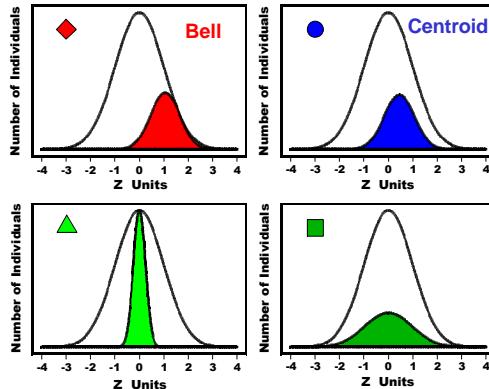
Feasible Region of Mean & Standard Deviation Values

Edgewood Chemical Biological Center

Region of Feasible Values for Healthy Subpopulation Size of 25%



Subpopulation Distributions (Shaded Areas) as Function of Location Within Region



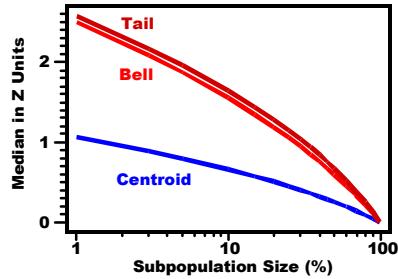
Feasible value pairs do not always produce realistic distributions (see ▲)

7

Subpopulation Model Statistics

Edgewood Chemical Biological Center

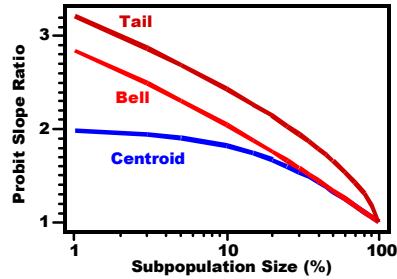
Median of a Healthy Subpopulation



For SP Size
of 25%

Median (Tail) = 1.15
Median (Bell) = 1.06

Probit Slope Ratio
(For Healthy and Sensitive Subpopulations)



Probit Slope Ratio (PSR) = Slope (SP)/Slope (GP)
 $= m_{SP} / m_{GP}$

For SP Size
of 25%

PSR (Tail) = 2.03
PSR (Bell) = 1.66

8

Calculation of Effective Dose Ratio

Edgewood Chemical Biological Center

$$EDR = \frac{ED_B}{ED_A} = \text{antilog} \left(\frac{(Z_B - Z_A)}{m_{GP}} \right)$$

EDR = Effective Dose Ratio

ED_A and ED_B = Effective doses for Populations A and B

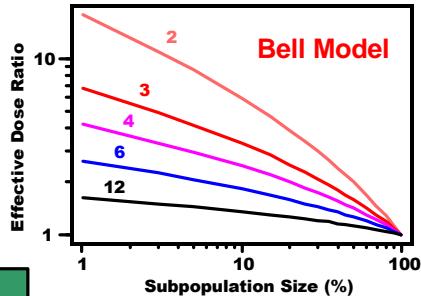
Z_A, Z_B = Distance (in Z units) of ED_A and ED_B from ED_{50} of GP

m_{GP} = Probit slope of GP

Only Three
Values Needed

Subpopulation Size
Probit Slope (m_{SP} or m_{GP})
Median Dose (for either SP or GP)

EDR Dependency on Probit Slope
Ratio of Medians for m_{GP} : 2, 3, 4, 6 and 12



9

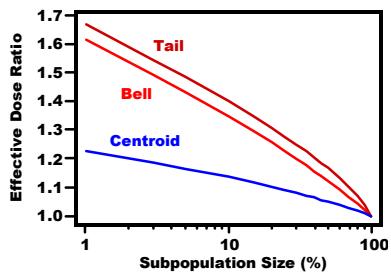
Comparison of EDRs from Different Models

Edgewood Chemical Biological Center

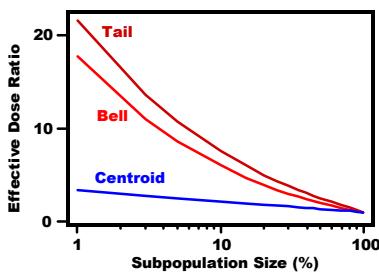
Relative Magnitude of Model EDRs
at a Fixed Probit Slope and SP Size

Tail > Bell > Centroid

EDRs from Tail, Bell and Centroid Models
Ratio of Medians for $m_{GP} = 12$

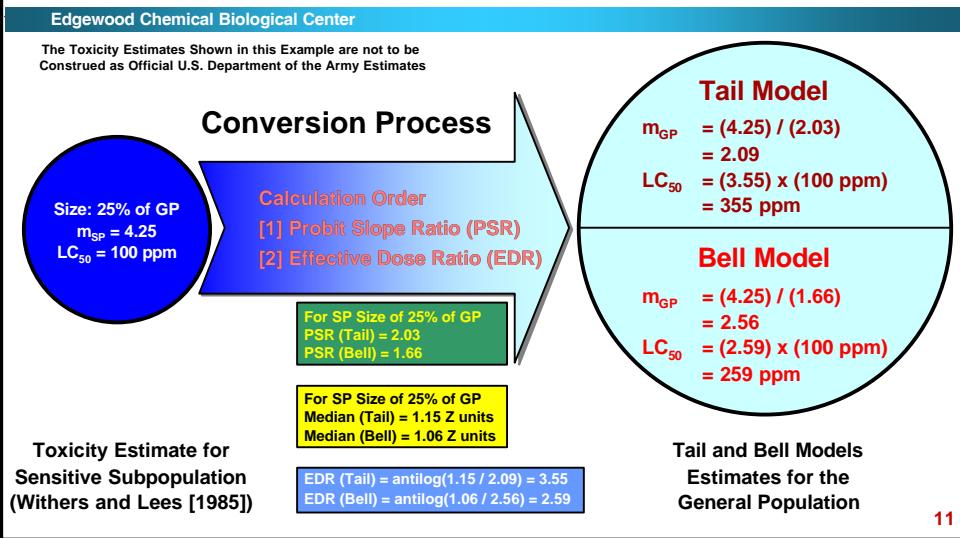


EDRs from Tail, Bell and Centroid Models
Ratio of Medians for $m_{GP} = 2$



10

Sample Calculation—Chlorine Toxicity



CW Agent Acute Exposure Guideline Levels (AEGLs)

Edgewood Chemical Biological Center

- Environmental Protection Agency (EPA) AEGLs—protection of health of sensitive individuals
 - AEGL-1: Threshold notable discomfort
 - AEGL-2: Threshold serious effects
 - AEGL-3: Threshold lethality
- CW agent AEGLs based on most toxic route: inhalation (IH)
- Proposed CW agent AEGLs (posted on EPA website)
 - G-type and VX Nerve Agents (Oct 2000)
 - Sulfur Mustard (HD) (January 2000)
 - Phosgene (CG) (August 2000)
 - Chlorine (October 1997)
- AEGL development involves use of Uncertainty Factors (UF) to account for various sources of uncertainty
 - UF values are usually 1, 3 or 10
 - Examples of UF applications in AEGLs:
 - Healthy to sensitive human (Intraspecies)
 - Laboratory animal to human
 - Incomplete to complete database
- Intraspecies UFs
 - Needed to account for response variability in the human population
 - Used to convert from a healthy human SP to a GP basis for threshold effects
 - Essentially ECT₀₁ (healthy SP) to ECT₀₁ (GP)

12

Comparison of Intraspecies UFs for CW Agent AEGLs

Edgewood Chemical Biological Center

- Tail and Bell Models can be used to calculate intraspecies UFs
 - UFs based on EDR of LCT₀₁ (healthy SP) to LCT₀₁ (GP)
 - Military probit slope values from Grotte and Yang (2001)
 - Probit slopes for CG and Chlorine estimated from review of existing experimental data
 - Models provide mathematical basis for setting intraspecies UF values
- EPA AEGL-3 intraspecies UFs shown for comparison
 - Assignment of values more qualitative in nature

Agent	Route	Military Probit Slope	m_{GP}		Uncertainty Factors (Between 1 st Percentiles)			EPA AEGL
			Tail	Bell	Tail	Bell	EPA AEGL	
G	IH	12.0	5.9	7.2	3.2	1.9	10	
G	PC	5.0	2.5	3.0	16.7	4.6		
HD	IH	6.0	3.0	3.6	10.4	3.6	3	
HD	PC	7.0	3.4	4.2	7.5	3.0		
VX	IH	6.0	3.0	3.6	10.4	3.6	10	
VX	PC	6.0	3.0	3.6	10.4	3.6		
CG	IH		6.7	8.3	2.8	1.7	3	
Chlorine	IH			5.9	3.2	2.2	3	

IH — Inhalation

PC — Percutaneous

13

Conclusions from Comparison of Intraspecies UFs

Edgewood Chemical Biological Center

UF Comparison Summary

Poor

G-Agents:
AEGL (10) >> Tail (3) & Bell (2)

Caution

VX:
AEGL (10) >> Bell (4)
AEGL (10) = Tail (10)
CG & Chlorine:
AEGL (3 & 3) > Bell (1.7 & 2.2)
AEGL (3 & 3) = Tail (2.8 & 3.2)

Excellent

HD:
AEGL (3) » Bell (4)
AEGL (3) << Tail (10)

- Both models are conservative
 - Tail Model the most conservative
 - Sets an absolute upper limit on UF value
 - Bell Model gives more realistic SP distribution shape
 - Important for comparing the 1st percentiles of two distributions
- Suggested course of action on current CW agent AEGL intraspecies UF values
 - G-Agent should be strongly reconsidered
 - VX, CG and Chlorine should be reassessed
 - Strong mathematical support for HD—no change need be considered
 - Any changes should be kept in context of ALL other assumptions made in developing AEGLs for a particular agent

14

Summary

Edgewood Chemical Biological Center

- New method developed for converting toxicity
 - Based on the mathematical modeling of a SP and its relationship to the GP
 - Conversion from SP to GP basis
 - Addresses a critical parameter gap (GP CW agent toxicity estimates)
- Method needs only three values:
 - Model parameter: SP size
 - Two toxicity values for conversion
 - Probit slope for either SP or GP
 - Median dose for either SP or GP
- Both healthy and sensitive SPs can be modeled with either of two models: Tail or Bell/Centroid
- Historical military demographics reviewed for modeling military SP
- Intraspecies UF for EPA CW Agent AEGL-3s investigated with method
 - Method provides mathematical basis for calculation of intraspecies UF values
 - Strong argument exists for current G-agent UF being too high
 - Current VX UF value is questionably high

15

Additional Information

Edgewood Chemical Biological Center

- Work documented in U.S. Army technical report
 - Crosier, RB and Sommerville, DR, *Relationship Between Toxicity Values for the Military Population and Toxicity Values for the General Population*, ECBC-TR-224. Edgewood CB Center, Aberdeen Proving Ground, MD, March 2002. UNCLASSIFIED/UNLIMITED. AD # A400214.
- Work funded by U.S. Department of Energy, National Security Administration, Chemical and Biological National Security Program
 - Technical point of contact: John E. Brockmann, Sandia National Laboratory
- Authors' acknowledgment
 - Dr. Sharon A. Reutter, Edgewood CB Center, for her technical advice and assistance

16

Finding The Season Effect And Trend Of Attrition: Detect The Attrition Changes In The Early Stage

T.E. Powers, Walter Reed Army Institute of Research; Y. Li, Walter Reed Army Institute of Research

Abstract: Background: Early attrition is an expensive problem for the United States military, and attrition reduction targets are frequently discussed as a sensible means of cost savings. Progress toward such goals is often measured over relatively short time spans. Specifically, time series examination of attrition by time of enlistment is a standard tool for assessing progress. The aim of this paper is to see if there is a season effect or there is a sudden increase in attrition

Subjects and Methods: All first-time enlistees beginning active duty military service during January, 1995 - December, 2000 will be categorized according to the month and year of beginning duty. Attrition percentages during the first 3 months of service will be determined within each month/year group for the 1995-1999 period, and regressed against several factors seen in previous studies to be predictive of attrition, including service branch distribution, gender distribution, race distribution etc. Then homogeneity of the attrition rates by the months of enrollment will be tested, and time trends examined. A second level regression model will be developed to predict the attrition rate for future months. A measure of the agreement between the predicted attrition rate and actual attrition rate of the following month will be used to detect any sudden in actual attrition from expected levels.

Results: Raw 3-month attrition percentages were seen to fluctuate considerably according to the month and year of beginning duty, ranging from about 5% to over 25% within the study period. Strong seasonal patterns seen in the raw attrition percentages were still present in the standardized percentages, although not as pronounced, indicating that some of the apparent seasonal pattern was related to changing demographic profiles of recruits over the course of a year. Time series modeling of the adjusted results yielded harmonious fit, and precise predictions of the CY 2000 attrition rates. Modeling of the unadjusted results was less precise, although also reasonably accurate.

Conclusions: The apparent seasonal trend in attrition is found to be more than a purely seasonal phenomenon - it is in part related to seasonal variation in the demographic profile of incoming recruits. Any changes to that pattern, such as might be introduced by changes in delayed entry procedures, would not be noticed by a seasonal time series analysis. A prior regression to distill the effects of demographic factors from the attrition data makes for a more harmonious and robust predictive model.

Contributed Session 7

Determination of the LD₅₀ for Chemical and Biological Threat Agents

Nancy A. Niemuth, Battelle

Abstract: The LD₅₀, defined as the dose of a substance expected to kill 50 percent of the test population within a given time, is a commonly used measure of toxicity. Three studies will be presented to illustrate and compare approaches to experimental design and LD₅₀ estimation using examples of chemical and biological threat agents tested in our laboratory. A modified up-and-down design was used to establish the LD₅₀ for GD in treated and untreated animals in an assessment of treatment efficacy. In developing a model of passive protection of human botulinum immunoglobulin against botulinum toxin, stagewise adaptive dose allocation was used to determine the LD₅₀ in the test species, while a fixed dose design was used to monitor toxin potency over time. Fixed dose designs generally require more animals, but less time and laboratory resources, than the stagewise and up-and-down strategies. For each study, the LD₅₀ was calculated using four statistical methods, including probit dose-response models, logistic regression, and point estimates obtained using the Reed-Meunch and Spearman-Karber methods. These methods generally give consistent point estimates, but varying information about the dose-response curve.

Homogeneity of the Loss Rate and Individual Factor Effect Across MEPS: A Meta-Analysis on Attrition

Y. Li, Walter Reed Army Institute of Research; T.E. Powers, Walter Reed Army Institute of Research

Abstract: Introduction: A major aim of the AMSARA project is to develop predictive models for military attrition based on information that can be reliably detected at the time of applicant screening. Previous AMSARA studies have found significant contributions to early attrition several such variables, including: medical disqualification at MEPS, gender, age, race, academic measures, number of dependents, body mass index, etc. Separately, it has been found that attrition rates vary according to the MEPS through which individuals are processed, with the highest attrition rate more than double that of the lowest one. A natural question then is whether the effects of medical and demographic factors found previously are homogeneous across all MEPS. A hierarchical model is used to make this determination, and results are used to develop a more accurate overall attrition model.

Methods: An initial attrition model examines for homogeneity of effects of individual variables mentioned above, as well as the attrition rate across the MEPS. Then a hierarchical model is used to study the effects of these variables according to their overall distributions at the various MEPS. For example, the model will help assess whether the effect of being male is the same across MEPS with vastly different percentages of male applicants. As another example, the effects of medical disqualification will be examined in relation to the percentage of individuals disqualified at each MEPS. Finally, all MEPS

variables that show significance and interactions in the hierarchical model are controlled for in the attrition model.

Results: Several of the demographic factors considered are found to differ according to the MEPS through which individuals were processed. Interestingly, while the medical factors were stronger predictors of attrition, none was found to differ by MEPS, perhaps reflecting the fact that medical waiver decisions are handled centrally. The variation of the effect of the heterogeneous variables were predicted using the improved model, and the chi-square test shows the model is significantly improved.

RELATIONSHIP BETWEEN THE DOSE-RESPONSE CURVES FOR LETHALITY AND SEVERE EFFECTS FOR CHEMICAL WARFARE NERVE AGENTS

DOUGLAS R. SOMMERVILLE

U.S. Army Edgewood Chemical Biological Center

5183 Blackhawk Road, ATTN: AMSSB-RRT-IM (Bldg. E5951)

Aberdeen Proving Ground, MD 21010-5424

douglas.sommerville@us.army.mil

In recent years, the U.S. Environmental Protection Agency has developed a categorical (or ordinal) logistic regression approach for regressing ordered categories of toxic responses (*e.g.*, no effects, non-adverse effects, other effects of increasing severity, *etc.*) on one or more factors (*e.g.* dose, exposure time, type of agent, *etc.*) due to a chemical agent exposure. The advantage of such an approach is that two types of dose-response curves (severity of effect and percent of individuals versus dose) are fitted simultaneously. For chemical warfare (CW) agent toxicity, ordinal logistic regression provides a means to statistically demonstrate and quantify the steepness of both types of dose-response curves for acute exposures to organophosphate-type CW agents (or nerve agents). Experimental animal data from three acute inhalation studies were reviewed and analyzed separately using a probit link-function: (1) monkeys exposed to GA (tabun), GB (sarin) or GF (cyclosarin); (2) rats exposed to GB; and (3) rats exposed to GB or GF. For each study, both vapor concentration and exposure time were varied. Clinical signs and mortality were recorded for all three studies. From these signs three categorical responses were defined: death, severe effects and less than severe effects. An animal was categorized as having severe effects if it exhibited at least one of the following signs (yet did not die within 24 hours post-exposure): convulsions, gasping, collapse or prostration. The regression analysis indicated that for all three studies slightly more than one standard deviation (1.0 to 1.4) separated an effective concentration (EC_{XX}) for severe effects from a lethal concentration (LC_{XX}) for XX% affected. This method provides a better way of estimating threshold lethality, because for the data analyzed, threshold lethality (approximately a LC_{01} or LC_{05}) is the equivalent to about an EC_{16} (for severe effects). The 16% effect level can be estimated with greater confidence from experimental quantal data via probit analysis than the 1% effect level. Thus, questionable extrapolation of the dose-mortality curve down to the 1% level can be avoided by using the dose-severe effect curve in its place.

INTRODUCTION

Human toxicity estimates for chemical warfare (CW) agents are required to properly evaluate agent-related health hazards under a variety of situations: military deployment operations; handling, storage and destruction of CW agents; and emergency response procedures. Modeling and simulation (M&S) plays an important role towards this end. For use in such models, it is required that toxicity is expressed as a function of exposure parameters (dosage and exposure duration). Knowledge is also needed of the dose-response (DR) curves for the population at risk: severity of effect (DR-S) and percent of affected individuals (DR-P) as a function of the dose.

To address these needs, data for CW organophosphate (or nerve) agents have been traditionally analyzed via the probit analysis (or binary logistic regression)¹⁻³ of the quantal (or binary) data taken for a particular toxicological endpoint (*e.g.*.. alive or dead, presence or absence of miosis, *etc.*) as a function of one or more factors (dosage, vapor concentration, exposure duration, *etc.*). It has been standard practice to define the resulting mortality-response relationships in terms of a linear time-integrated concentration (*i.e.*, vapor concentration (C) multiplied by the exposure time (T), or CT for short—a dosage).⁴ Two important parameters are produced by probit analysis that characterize the DR-P curve for any particular toxicant: median dosages (either median effective (ECT_{50}) or lethal (LCT_{50}))) and the probit slope. Both Mioduszewski *et al.*^{4,6} and Anthony *et al.*^{7,8} employed this method. However, on its own probit analysis can only

characterize the DR-P curve for a particular endpoint. Knowledge about the other dose-response curve, DR-S, for nerve agents is also very important, especially since it is known to be very steep.⁹

However, defining the DR-S curve requires additional measures. The simplest approach has been to compare the reported literature values of ECT₅₀'s* and probit slopes calculated via probit analysis for a range of endpoints.^{10,11} However, the accuracy of calculated ECT₅₀ ratios for different endpoints is reduced when the values in the ratio come from separate studies (*i.e.* comparing the ECT₅₀ (miosis) from Study A to the ECT₅₀ (convulsions) from Study B).

A better approach is to investigate multiple endpoints in the same study, as was done by Cresthull, *et al.*¹² who reported both severe effects and lethality as a function of C and T. A probit analysis was performed separately on each endpoint. The estimated ECT₅₀'s (incapacitation) and LCT₅₀'s were compared to estimate the steepness of the DR-S curve. Unfortunately, regressing toxicological responses using a binary format implicitly assumes that the responses are independent of each other, which is not the case here. The result is that important information is ignored which could better characterize the steepness of the DR curve.

One solution to this problem is to employ categorical (or ordinal) logistic regression.¹³ The U.S. Environmental Protection Agency (EPA) is currently developing this method for its own applications (such as supporting the Benchmark Dose (BMD) model).¹⁴⁻¹⁷ Instead of the binary response used in probit analysis, categorical logistic regression uses ordered categories of toxic responses (*e.g.*, no effects, non-adverse effects, other effects of increasing severity, *etc.*), which are regressed as a function of one or more factors (*e.g.* dose, exposure time, type of agent, *etc.*). The advantage of this approach is that the two types of DR curves (DR-P and DR-S) are fitted simultaneously. Thus, for CW nerve agents, ordinal logistic regression provides a means to statistically demonstrate and better quantify the steepness of both types of curves for acute inhalation (IH) exposures. Data from three CW nerve agent studies^{5-8,12} were reviewed and re-analyzed using ordinal regression. The purpose of the analysis was to determine the relationship between the DR-P curves for lethality and severe effects resulting from IH exposures to G-type nerve agents. Potential risk assessment applications¹⁸ of this type of knowledge were then explored.

TOXICITY STUDIES REVIEWED

Overview. The three studies reviewed for this work were Cresthull, *et al.* (1957),¹² Mioduszewski, *et al.* (2001 and 2002),⁴⁻⁶ and Anthony, *et al.* (2002),⁸ all of which were conducted at what is now the Edgewood Chemical Biological Center (ECBC). A brief summary of the studies is presented in Table 1. The Toxicology Team, ECBC, maintains raw data and other materials associated with these studies.

In all three studies, the animals were exposed (whole body) in dynamic airflow inhalation chambers.¹⁹ For Cresthull, *et al.*, the agent vapor concentrations were allowed to reach equilibrium at the target value for the run; after which, the animals were quickly introduced into (and removed from) the chamber *via* a sliding animal carriage. The exposure duration was, thus, the time between introduction and removal.

* The definition of an effective dosage (ECT_{XX}) includes aspects from both types of dose-response curves. An ECT_{XX} for Effect A is the dosage needed to produce either Effect A or an effect of greater severity (from the same route of exposure) in XX% of the subjects exposed to that dosage. Thus, cumulative measures are found for both the effect severity curve (an effect of equal or greater severity) and the percent affected (XX%) curve.

In the other two studies, the animals were placed into the chamber prior to the introduction of agent vapor. Then, the chamber was quickly brought to equilibrium at the target vapor concentration. The concentration was kept constant once equilibrium had been reached. The concentration-time profile generated was described by MacFarland (1987).¹³ His definition of exposure duration was the one used in these studies--the interval from the start of the flow of agent into the chamber to the time-point when the agent flow is stopped. Following exposure, the chamber was purged with air for 10 minutes, and the animals were then removed from the chamber. None of the animals were restrained during an exposure run. Both Cresthull *et al.* (1957) and Mioduszewski, *et al.* (2001 and 2002) have been previously used in the development of acute exposure guideline level values for CW nerve agents.^{10,11}

Table 1
Summary of CW Nerve Agent Studies Reviewed

Name of Study	Cresthull, <i>et al.</i> (1957)	Mioduszewski, <i>et al.</i> (2001 and 2002)	Anthony, <i>et al.</i> (2002)
Year(s) Conducted	1953 to 1954	1998 to 2000 in two separate phases	2001 to 2002
Agent(s) Investigated	GA, GB and GF	GB	GF and GB
Species Used	Rhesus Monkey	Sprague-Dawley Rat	Sprague-Dawley Rat
Total Number of Subjects	152	700	500
Gender	Mostly Female	Equal number of Males and Females	Males (240) Females (260)
Breakdown by Agent of Number of Subjects	GA (56), GB (52), GF (44)	All GB	GF (320), GB (180)
Number of Subjects per Exposure Group	4	10 or 20	5, 10 or 20
Number of Runs	36	43	38
Vapor Concentrations (mg/m ³)	GA: 18.1 to 81 GB: 6.6 to 29.1 GF: 7.3 to 59	GB: 2.0 to 54.4	GB: 3.5 to 35.9 GF: 2.0 to 41.9
Exposure Times (minutes)	2 and 10	Phase I: 10, 30, 90, 240 Phase II: 5, 60, 360	10, 60 and 240
Primary Endpoint(s) of Interest	Incapacitation and Lethality (1 day)	Lethality (1 and 14 days)	Lethality (1 and 14 days)

Definition of Severe Effects and Lethality. The ordered ternary responses defined for the present work (lethality (L), severe effects (S) and less than severe effects (M)) were defined from the clinical signs and mortality, which were recorded in the three studies. Mortality within 24 hours post exposure was counted as a lethal effect. An animal was categorized as having severe effects if it exhibited convulsions, gasping, collapsed or prostration (yet did not die within 24 hours post exposure). Mortality occurring between one and 14 days post exposure was treated as a severe effect response. In Cresthull, *et al.*, incapacitation was defined as collapse or convulsions.

Experimental Quantal Data. The experimental quantal data from Cresthull, *et al.*, Mioduszewski, *et al.*, and Anthony, *et al.* are presented (using a discrete format) in Tables 2 to 4. For example, in Table 2 for T = 10 minutes and C = 18.1 mg/m³ GA, there are two animals with no effects severe or above, one animal having at least severe effects

and no mortality, and one animal that died or in shorthand—[2, 1, 1].^{*} In the discrete format, the sum of the values in a table row equals the total number of individuals exposed using a particular set of test parameters (agent, C, T and gender), which for the present example equals four (or 2+1+1). The original notebooks were also reviewed to gather additional information on the categorical response distributions.

Table 2
Monkey G-Type IH Quantal Data from Cresthull, et al.

Agent	T (min)	C (mg/m ³)	CT (mg-min/m ³)	Number per Category		
				M	S	L
GA	2	33.3	67	4	0	0
		50.8	102	2	2	0
		54.3	109	0	4	0
		62.0	124	0	1	3
		62.3	125	0	3	1
		65.0	130	0	2	2
		68.5	137	2	2	0
		71.0	142	0	0	4
	10	81.0	162	1	0	3
		18.1	181	2	1	1
GB	2	18.8	188	0	2	2
		21.7	217	0	0	4
		23.0	230	1	0	3
		24.6	246	0	0	4
		8.8	18	4	0	0
		13.7	27	1	0	3
		16.4	33	2	1	1
		17.0	34	4	0	0
	10	18.6	37	1	0	3
		19.7	39	2	1	1
		23.5	47	1	1	2
		29.1	58	0	1	3
GF	2	6.6	66	1	2	1
		8.1	81	0	1	3
		8.2	82	0	2	2
		10.0	100	0	0	4
		31.0	62	2	1	1
		42.0	84	1	1	2
	10	44.0	88	0	1	3
		48.0	96	0	0	4
		59.0	118	0	0	4
		7.3	73	3	1	0
		10.0	100	2	1	1
		12.2	122	0	2	2

Note: Shaded row was not used in the final analysis after having been identified as a statistical outlier in the initial analysis.

* This is in contrast to a common toxicology convention of displaying quantal data in a cumulative format, where the number of animals having an effect of equal or greater severity are included in an effect category. In which case, the above example [2, 1, 1] would be written instead as [4, 2, 1].

Table 3
Rat GB IH Quantal Data from Mioduszewski, et al.

T (min)	C (mg/m ³)	CT (mg-min/m ³)	Male			Female				
			Number per Category			C (mg/m ³)	CT (mg-min/m ³)	Number per Category		
			M	S	L	M	S	L		
5	36.3	182	1	7	2	25.6	128	5	3	2
	44.0	220	3	4	3	28.2	141	9	1	0
	48.1	241	0	6	4	31.5	158	1	3	6
	51.4	257	1	2	7	36.3	182	1	3	6
	54.4	272	0	3	7	44.0	220	0	1	9
10	15.3	153	9	1	0	9.6	96	10	0	0
	18.7	187	7	2	1	12.0	120	9	1	0
	21.8	218	0	6	4	15.3	153	1	7	2
	27.1	271	0	2	8	18.7	187	2	4	4
	34.3	343	0	0	10	21.8	218	0	1	9
30	6.0	180	10	0	0	6.0	180	6	4	0
	7.4	222	8	2	0	7.4	222	0	9	1
	9.0	270	1	6	3	8.5	255	0	5	5
	10.3	309	0	0	10	9.0	270	0	3	7
	12.1	363	0	0	10	12.1	363	0	1	9
60	6.0	360	2	5	3	5.9	354	1	7	2
	6.4	384	3	5	2	6.0	360	0	4	6
	7.0	420	1	8	1	6.4	384	1	6	3
	7.6	456	1	3	6	7.0	420	0	5	5
	8.1	486	3	1	6	7.6	456	0	0	10
90	4.0	360	6	4	0	4.0	360	0	8	2
	4.1	369	9	1	0	4.1	369	4	5	1
	4.5	405	1	6	3	4.5	405	0	4	6
	4.9	441	0	6	4	4.9	441	2	1	7
	5.5	495	0	2	8	5.5	495	0	0	10
240	2.1	504	10	0	0	2.1	504	10	0	0
	2.7	648	9	0	1	2.7	648	0	6	4
	3.3	792	7	2	1	3.3	792	0	4	6
	4.2	1008	0	6	4	4.2	1008	0	2	8
	4.4	1056	0	4	6	4.4	1056	0	0	10
360	2.3	828	5	3	2	2.3	828	5	4	1
	2.7	972	2	6	2	2.4	864	0	10	0
	2.8	1008	2	7	1	2.7	972	2	4	4
	3.0	1080	0	2	8	2.8	1008	0	5	5
	3.5	1260	0	1	9	3.0	1080	0	1	9

Table 4
Rat GB and GF IH Quantal Data from Anthony, et al.

Agent	Gender	T (min)	C (mg/m ³)	CT (mg-min/m ³)	Number per Category			
					M	S	L	
GF	Female	10	17.2	172	9	1	0	
			21.5	215	7	3	0	
			23.3	233	10	0	0	
			23.9	239	2	3	5	
			25.2	252	2	2	6	
			26.9	269	1	3	6	
			31.1	311	0	0	10	
	Male		17.2	172	10	0	0	
			21.5	215	10	0	0	
			31.1	311	1	8	1	
			34.4	344	5	3	2	
			41.9	419	0	1	9	
	60	4.9	294	6	2	2		
		5.7	342	4	4	2		
GB	Female	240	5.9	354	0	1	9	
			6.4	384	0	0	10	
			7.2	432	0	0	10	
			4.9	294	10	0	0	
			5.7	342	5	4	1	
			6.4	384	0	6	4	
			7.2	432	1	2	7	
	Male		7.8	468	0	0	10	
	10	2.0	480	7	2	1		
		2.0	480	1	8	1		
		2.2	528	0	3	7		
		2.5	600	0	2	8		
		3.3	792	0	0	10		
		2.0	480	3	4	3		
		Male		2.0	480	4	6	0
				2.2	528	0	8	2
				2.5	600	1	3	6
				3.3	792	0	1	9
GB	Female	60	18.0	180	0	10	0	
			21.6	216	4	5	1	
			22.7	227	0	8	2	
			23.8	238	0	3	7	
			24.8	248	0	3	7	
			26.6	216	0	0	10	
			22.7	227	8	2	0	
	Male		26.7	267	1	8	1	
			28.7	287	0	6	4	
			32.8	328	0	5	5	
			35.9	287	0	2	8	
			5.6	336	0	4	1	
			6.1	366	0	4	6	
			6.6	396	0	0	5	
GB	Male	240	6.6	396	1	4	0	
			7.0	420	1	5	4	
			7.5	450	0	1	4	
			3.5	840	0	5	5	
			4.3	1032	8	2	0	
	Male		5.6	1344	0	3	7	

STATISTICAL THEORY

Probit analysis was the method used by Cresthull, *et al.*,¹² Mioduszewski, *et al.*,⁴⁻⁶ and Anthony, *et al*^{7,8} for the analysis of their data. A brief review of probit analysis is presented herein, followed by a review of its extension for use with ordered categorical responses with three or more levels (or ordinal logistic regression), whose application towards CB nerve agents is the subject of this work.

To perform either a binary or ordinal logistic regression, a link-function is used to connect the random and systematic components of the regression model.¹³ This is accomplished by transforming the probability of an effect or response to a linear scale. Several probability distributions are commonly used for this transformation: probit, logit and complementary log-log.^{2,13,22} Historically, CB nerve agent toxicology has used a probit link-function, which is implicit in the use of probit analysis. For ease of comparison, ordinal regression with a probit link-function is used in this work. Thus, the following discussions implicitly assume the use of a probit link-function.

Probit Analysis. For each individual, there is a dose or dosage* that is just sufficient to produce a specified biological response. These just-sufficient dosages are called effective dosages to distinguish them from administered dosages. The distribution of effective dosages for a homogeneous population is usually lognormal.^{1,5,20,21}

Although statisticians typically describe the lognormal distribution of effective dosages by the mean and variance of log(effective dosage), toxicologists usually describe the distribution by the median effective dosage, ECT₅₀, and the probit (or Bliss) slope, *m*:

$$(1) \quad ECT_{50} = \text{antilog}(\eta)$$

$$(2) \quad m = 1 / \sigma$$

$$(3) \quad Z = \frac{\{\log(CT) - \log(ECT_{50})\}}{s}$$

where η is the median of log(effective dosage), σ^2 is the variance of the distribution, and Z is the standard normal random variable. The ECT₅₀ is used in a cumulative fashion by toxicologists: 50% of the exposed individuals will exhibit a specified biological response of equal or greater severity for the same exposure route. Effective dosages for response levels other than 50% can be calculated using Eqn. (3) with known values for ECT₅₀ and *m*, and using the *Z* value corresponding to the cumulative probability of interest (*e.g.* *Z* equals 0 for a 50% response). Toxicologists traditionally use base 10 logarithms to calculate the probit (Bliss) slope.^{1,3,5,21} This convention is used herein.

Although the normal distribution is continuous, quantal (binary) data are used to estimate the distribution parameters (ECT₅₀ and *m*).¹ Probit analysis and maximum likelihood estimation (MLE) are used to estimate these parameters from data.^{1,22} The following equation is fitted via probit analysis/MLE for vapor toxicity studies:¹

$$(4) \quad Y_N = (Y_P - 5) = k_0 + k_C \log C + k_T \log T + k_i (\text{other factors})$$

where Y_N is a normit, Y_P is a probit, and the *k*'s are fitted coefficients. The constants k_C and k_T are the probit slopes for concentration and time, respectively. Often, experiments are conducted with exposure time held constant, which reduces Eqn. (4) to the traditional probit equation.¹ Thus, the probit slope for a vapor exposure usually refers

* The terms dose and dosage are often used interchangeably, but they do have different definitions. Dose is the total amount of a substance that is administered, while dosage is an amount administered relative to some other quantity (*e.g.*, body mass, body surface, and/or time).²⁰ For inhalation exposures, dosage is the term used.²⁰

to the slope on vapor concentration ($m = k_C$) instead of the slope on time. The greater the probit slope, the smaller the variance is in the distribution of individual susceptibilities.

When fitting Eqn. (4), all variability in the data will contribute to the estimate for m , be it from variance due to individual susceptibilities, batch effects, experimental error, etc. Probit analysis performed on a compilation of data from many sources will not produce an accurate measure of variance among individuals due to the heterogeneity introduced by differences among the studies (e.g., experiment procedures, type of animals used, etc.).²³ The effect of such heterogeneity will be to reduce the probit slope. Also, as was previously noted, probit analysis on its own can only characterize the DR-P curve for a particular endpoint.

Ordinal Logistic Regression. Conceptually, ordinal regression simply involves the division of ordered multi-level categorical responses into a series of cumulative binary responses.²³ In the case of ternary data, with ordered discrete response levels of low {0}, medium {1} and high {2}, the following ordered binary combinations are produced: {0} vs. {1 and 2}; and {0 and 1} vs. {2}. Thus, one way to express the model is to apply Eqn. (4) to each binary combination:¹

$$(5) \quad Y_N \{0|1,2\} = k_{\{0|1,2\}} + k_C \log C + k_T \log T + k_i (\text{other factors})$$

$$(6) \quad Y_N \{0,1|2\} = k_{\{0,1|2\}} + k_C \log C + k_T \log T + k_i (\text{other factors})$$

where $Y_N \{0|1,2\}$ and $Y_N \{0,1|2\}$ are the normits for the binary responses of {0} vs. {1 and 2}; and {0 and 1} vs. {2}, respectively. The constants, $k_{\{0|1,2\}}$ and $k_{\{0,1|2\}}$, are the intercepts for the normits of the cumulative probabilities of an effect exceeding in severity the low {0} and medium {1} responses, respectively.²³

When using Eqns. (5) and (6), it is implicitly assumed that the values of the individual various probit slopes (i.e. k_C , k_T , k_i , etc.) are constant (e.g. k_C , (in Eqn. (5)) equals k_C (in Eqn. (6))). Otherwise there would be conditions where Eqns. (5) and (6) would intersect, a probabilistic impossibility for ordered responses.¹ As with probit analysis, MLE is used to provide fits for Eqns. (5) and (6).^{13,22} An iterative-reweighted least squares algorithm is used to obtain maximum likelihood parameter estimates.^{22,23}

Dose-Percent Response Curves (Severe and Lethality). For the present study, Eqns. (5) and (6) are used to solve for ECT₅₀ (severe) and LCT₅₀, respectively. To calculate the ECT₅₀ / LCT₅₀ ratio, Eqns. (5) and (6) can rearranged to produce:

$$(7) \quad \log_{10} \frac{\hat{e} \text{ECT}_{50}}{\hat{e} \text{LCT}_{50}} = ? / k_C$$

$$(8) \quad \kappa = \left[k_{\{0,1|2\}} - k_{\{0|1,2\}} \right] = [k_{\text{severe}} - k_{\text{lethal}}]$$

where κ is the distance in normits between the percent affected levels of the severe and lethality DR curves. For instance, when κ equals one, the ECT₅₀ equals a LCT₁₆ (since the 50 and 16% cumulative effect levels from a standard normal distribution are separated by one standard deviation), or if κ equals two, then ECT₈₄ equals a LCT₁₆.

Confidence limits on estimates for both $\{\kappa / k_C\}$ and κ can be calculated. The standard error of a ratio, (a/b) , is given by Barry (1978),²⁵ which is based upon the propagation of error formula for a ratio:

$$(9) \quad \text{std err of } \left(\frac{a}{b} \right) = \left(\frac{a}{b} \right) \sqrt{\left(\frac{\text{var}(a)}{a^2} \right) + \left(\frac{\text{var}(b)}{b^2} \right) - (2) \left(\frac{\text{cov}(a,b)}{ab} \right)}$$

where $\text{var}(a)$, $\text{var}(b)$, and $\text{cov}(a,b)$ are the variance of the quantities, a and b , and their covariance, respectively. The 95% confidence limits for the ratio will equal $(a/b) \pm (2)(\text{std err})$. The following relations from Mood, *et al.* (1974)²⁶ were also used to get the necessary information for determining the limits for both $\{\kappa / k_C\}$ and κ :

$$(10) \quad \text{var}(a \pm b) = \text{var}(a) + \text{var}(b) \pm (2)\text{cov}(a, b)$$

$$(11) \quad \text{cov}(a \pm b, c) = \text{cov}(a, c) \pm \text{cov}(b, c)$$

where $\text{cov}(a \pm b, c)$ is the covariance of the quantity, $(a \pm b)$, with a third quantity, c .

DATA ANALYSIS

An ordinal logistic regression program (a component of MINTAB® Version 13) was used to perform the calculations. The three datasets (in Tables 2 to 4) were analyzed separately. The ternary data consisted of the number of subjects having less than severe effects (M), severe effects (S), and lethality (L), as previously defined.

Only one continuous predictor, $\log C$, was used in the present analysis. The other available continuous predictor, T , was treated as a categorical factor instead, since the emphasis was on the estimating the relationship between severe and lethal DR-P curves. Complications were avoided by not trying to directly model the non-linear dependence of toxicity on $\log T$. Both Mioduszewski, *et al.*^{4,6} and Anthony, *et al.*^{7,8} have found that $\log(LCT_{50})$ versus $\log T$ was non-linear for G-agent IH toxicity.

In addition to $\log C$, full factorial designs were used in each of the three studies to investigate the effect of two or more of the following factors: agent type, exposure duration (T) and gender. Cresthull, *et al.* investigated agent type (3 levels) and exposure duration (2 levels), for a total of 6 groupings. Mioduszewski, *et al.* studied gender (2 levels) and exposure duration (7 levels), for a total of 14 groupings. Anthony, *et al.* explored all three predictors, using a total of 12 groupings [agent type (2 levels), gender (2 levels), and exposure duration (3 levels)].

For the present analysis, the following model was used in the ordinal regression programs (modifications of Eqns. (5) and (6)):

$$(12) \quad Y_N \{\text{severe}\} = k_{\text{severe}} + k_C \log C + \sum_i^N k_i G_i$$

$$(13) \quad Y_N \{\text{lethal}\} = k_{\text{lethal}} + k_C \log C + \sum_i^N k_i G_i$$

where G_i equals one when modeling the i -th group (from the total number (N) of groups from the full factorial) of a dataset and zero for all other groups, and the k_i 's are fitted coefficients. This approach produces only one value each for the probit slope (k_C), $\{\kappa / k_C\}$ and κ for the whole dataset, as well as individual ECT_{50} and LCT_{50} values for each group. By dividing a dataset into smaller independent subsets (for separate analyzes using MINTAB®), it is possible to obtain multiple values for k_C , $\{\kappa / k_C\}$ and κ as a function of the various factors within a dataset. However, it was found for each parameter that the individual subset values were not significantly different (statistically) from other subset values within the larger dataset. Thus, it was assumed that k_C , $\{\kappa / k_C\}$ and κ were constant in value for the whole dataset.

In addition to calculating values for k_C , $\{\kappa / k_C\}$ and κ for each dataset, Eqns. (9) to (11) were used (in conjunction with the variance-covariance matrix of the model fit returned by MINTAB[®]) to estimate the errors associated with these values. Also, error estimates for individual group ECT₅₀ and LCT₅₀ values were made in the same fashion.

RESULTS

The results of the data analysis are presented in Tables 5 to 11. Tables 5 to 8 contain the estimates for individual group ECT₅₀ and LCT₅₀ values, while Tables 9 to 11 present the estimated probit slope (k_C), $\{\kappa / k_C\}$ and κ values for each dataset. When available, values previously reported by the researchers are shown for comparison.

Table 5
Monkey G-Type IH ECT₅₀ (Severe) Values from Ordinal Logistic Regression and Cresthull, et al.

		Estimates from Ordinal Logistic Regression		Cresthull, et al (1957) (24 hours Post-Exposure)	
Agent	T (min)	ECT ₅₀ (Severe) (mg-min/m ³)	95% Fiducial Limits	ECT ₅₀ (Severe) (mg-min/m ³)	95% Fiducial Limits
GA	2	102	90 to 115	102	none reported
GB		36	31 to 40	30	none reported
GF		58	49 to 70	62	none reported
GA	10	145	121 to 173	<180	none reported
GB		56	46 to 67	<66	none reported
GF		96	82 to 112	100	none reported

Table 6
Monkey G-Type IH LCT₅₀ Values from Ordinal Logistic Regression and Cresthull, et al.

		Estimates from Ordinal Logistic Regression		Cresthull, et al (1957) (24 hours Post-Exposure)	
Agent	T (min)	LCT ₅₀ (mg-min/m ³)	95% Fiducial Limits	LCT ₅₀ (mg-min/m ³)	95% Fiducial Limits
GA	2	131	118 to 146	135	123 to 152
GB		46	40 to 53	42	29 to 60
GF		76	65 to 88	75	63 to 87
GA	10	187	161 to 217	187	164 to 221
GB		72	61 to 85	74	62 to 87
GF		124	108 to 143	130	112 to 151

Table 7
Rat GB IH ECT₅₀ (Severe) and LCT₅₀ Values from Ordinal Logistic Regression and Mioduszewski, et al (2001) (24 hours Post-Exposure)

		Estimates Derived from Ordinal Logistic Regression				Mioduszewski, et al (2001) (24 hours Post-Exposure)	
Gender	T (min)	ECT ₅₀ (Severe) (mg-min/m ³)	95% Fiducial Limits	LCT ₅₀ (mg-min/m ³)	95% Fiducial Limits	LCT ₅₀ (mg-min/m ³)	95% Fiducial Limits
Female	5	136	128 to 145	173	163 to 184	166	151 to 186
Male		184	173 to 196	234	220 to 248	240	211 to 287
Female	10	144	134 to 183	183	171 to 196	184	167 to 205
Male		185	173 to 198	235	220 to 252	231	211 to 255
Female	30	196	183 to 209	249	233 to 265	263	241 to 292
Male		225	211 to 240	286	268 to 305	undefined	undefined
Female	60	300	281 to 320	381	360 to 404	387	357 to 417
Male		354	334 to 375	450	425 to 476	459	412 to 472
Female	90	319	300 to 340	406	383 to 430	404	385 to 426
Male		366	346 to 388	466	440 to 493	448	427 to 482
Female	240	589	547 to 633	748	697 to 803	741	654 to 825
Male		801	749 to 857	1018	952 to 1090	1040	917 to 1466
Female	360	780	735 to 827	991	938 to 1048	987	946 to 1039
Male		830	781 to 882	1055	996 to 1117	1048	973 to 1150

Mioduszewski, et al.

Table 8
Rat GB and GF IH ECT₅₀ (Severe) and LCT₅₀ Values from Ordinal Logistic Regression and Anthony, et al.

		Estimates Derived from Ordinal Logistic Regression				Anthony, et al (2001) (24 hours Post-Exposure)		
Agent	Gender	T (min)	ECT ₅₀ (Severe) (mg-min/m ³)	95% Fiducial Limits	LCT ₅₀ (mg-min/m ³)	95% Fiducial Limits	LCT ₅₀ (mg-min/m ³)	95% Fiducial Limits
GF	Female	10	222	213 to 231	267	256 to 278	253	244 to 266
	Male		305	288 to 324	367	347 to 389	371	344 to 405
GB	Female	60	187	179 to 197	226	216 to 235	235	228 to 243
	Male		253	236 to 271	304	283 to 326	316	297 to 348
GF	Female	240	286	271 to 302	344	328 to 361	334	317 to 349
	Male		335	319 to 352	403	384 to 423	396	376 to 416
GB	Female	240	288	266 to 311	346	322 to 372	355	332 to 376
	Male		359	335 to 384	432	405 to 461	433	409 to 464
GF	Female	240	447	425 to 471	539	513 to 565	533	506 to 566
	Male		470	448 to 494	566	540 to 594	595	550 to 677
GB	Female	240	686	623 to 757	826	753 to 907	840	766 to 922
	Male		1090	1016 to 1169	1312	1222 to 1408	1296	1152 to 1486

Table 9
*Probit Slope(Concentration) Estimates for G-Type Nerve Agents IH Exposures from
 Ordinal Logistic Regression and Original Researchers.*

Dataset	Estimates from Ordinal Logistic Regression		Median and Range Reported by Original Researchers (24 hour post-exposure)	
	Probit Slope (k_C)	95% Conf. Limits	Probit Slope (k_C)	Range of Values
Cresthull, et al (1957)	9.1	6.4 to 11.9	11.0	6.6 to 15.4
Mioduszewski, et al. (2001)	13.9	12.3 to 15.5	13.2	8 to 24.4
Anthony, et al. (2002)	18.0	15.4 to 20.5	23.5	13.3 to 31.2

Note: For shaded blocks above, Cresthull, et al. arrived at essentially one probit slope value for their entire dataset, along with an estimate for the standard error. Thus, instead of a range of values, the 95% confidence limits calculated from their standard error are shown in the table.

Table 10
*Estimates for Distance (\mathbf{k}) Between Severe (S) and Lethality (L) Dose-Response Curves
 for G-Type Nerve Agents IH Exposures from Ordinal Logistic Regression*

Dataset	Species	Estimates from Ordinal Logistic Regression		
		S to L Distance (\mathbf{k}) (normits)	Variance (S to L Dist)	95% Conf. Limits
Cresthull, et al (1957)	Monkey	1.02	0.0225	0.72 to 1.32
Mioduszewski, et al. (2001)	Rat	1.44	0.0069	1.28 to 1.61
Anthony, et al. (2002)	Rat	1.44	0.0100	1.24 to 1.65

Table 11
ECT₅₀/LCT₅₀ Ratio Estimates for G-Type Nerve Agents IH Exposures from Ordinal Logistic Regression and Original Researchers

Dataset	Estimates from Ordinal Logistic Regression		Median and Range Reported by Original Researchers (24 hour post-exposure)	
	(ECT ₅₀ /LCT ₅₀) $10^a(k/k_C)$	95% Conf. Limits	(ECT ₅₀ /LCT ₅₀) $10^a(k/k_C)$	Range of Values
Cresthull, et al (1957)	0.77	0.70 to 0.85	0.80	0.71 to 0.96
Mioduszewski, et al. (2001)	0.79	0.76 to 0.81		
Anthony, et al. (2002)	0.83	0.81 to 0.85		

DISCUSSION

Group ECT₅₀ and LCT₅₀ Estimates. The estimates for median effective dosages for severe effects and lethality from ordinal logistic regression are in agreement with those reported by the original researchers for the datasets that were reviewed (see Tables 5 to 8). The means of the absolute percent differences (see Eqn. (14) below) were

found to equal 4.9, 2.1 and 2.6%, for the datasets from Cresthull, *et al.* (1957), Mioduszewski, *et al.* (2001) and Anthony, *et al.* (2002), respectively.

$$(14) \quad \text{abs \% diff} = (100) \left| \frac{\text{XCT}_{50} \text{ (original)} - \text{XCT}_{50} \text{ (ordinal)}}{\text{XCT}_{50} \text{ (original)}} \right|$$

In the cases of Mioduszewski, *et al.* (2001) and Anthony, *et al.* (2002), values for ECT₅₀ (severe) were not reported, thus the ECT₅₀ (severe) values in Tables 7 and 8 from the ordinal regression analysis are the first such reported values for these datasets.

Probit Slopes (k_C). For each dataset, the probit slope (concentration) estimates from the ordinal regression are in agreement with those reported by the original researchers (see Table 9). These results confirm previous findings on the steepness of the DR-P curves for G-type nerve agents.^{9,10} For the ordinal regression k_C values, the differences between the k_C values from the three datasets are statistically significant. The larger k_C values (less individual variability) from the two rat studies (Mioduszewski, *et al.* and Anthony, *et al.*) (*vs.* the monkey study) is probably due to the genetically defined laboratory rats as compared to the monkeys used by Cresthull *et al.* However, other reasons for differences between the rat and monkey studies (batch effects, experimental error, *etc.*) cannot be entirely ruled out. Within the two studies investigating two or more agents (Cresthull, *et al.* and Anthony, *et al.*), the difference in probit slopes between the agent subsets are not statistically different; so, it is unlikely that the changes in probit slopes are due to differences between the agents.

Distance (Normit) Between Severe and Lethality Dose-Response Curves.

The distance (κ) (see Eqn. (8)) is found to range from 1.02 to 1.44 normits for the three datasets reviewed (see Table 10). The average of κ values equals 1.30. Values for κ from these datasets were not previously reported.

Using $\kappa = 1.30$ for G-type nerve agent IH exposures, it is found that an ECT₁₆ (severe) approximately equals the LCT₀₁. Going both further up and down the dose-percent response curves, other equivalencies can be calculated (see Table 12). The steepness of the DR-S curve is readily demonstrated by the fact that the dosage causing incapacitation (or greater effect) in 84% of exposed individuals will also kill about half (45.4%) of those within the incapacitated (or greater) group. Furthermore, trying to use a G-type nerve agent to achieve complete incapacitation with minimal fatalities among a target group is an impossibility, since there will be an 85% lethality rate among the 99 out of 100 incapacitated subjects at an ECT₉₉ (severe).

Table 12
Comparison of Equivalent ECT_{XX} and LCT_{YY} Levels for G-type Nerve Agent IH Exposures

Y _N Severe (normits)	Y _n Lethal (normits)	XX% Severe (or greater)	YY% Lethal	Ratio YY% to XX%
-2.00	-3.3	2.3	0.0	2.1
-1.00	-2.3	15.9	1.1	6.8
0.00	-1.3	50.0	9.7	19.4
1.00	-0.3	84.1	38.2	45.4
2.00	0.7	97.7	75.8	77.6
2.31	1.01	99.0	84.4	85.3

Based on the estimated variances of the individual κ values, there is a significant difference (with 99% confidence) between the monkey κ value of Cresthull, *et al.* and the two rat κ values of Mioduszewski, *et al.* and Anthony, *et al.* This suggests that the existence of a species effect on κ values for G-type agent IH toxicity, particularly since the two separate rat studies produced identical κ values. However, additional work is needed before any definitive conclusions can be reached.

In addition to using ordinal logistic regression to estimate κ from quantal data sets, it is also possible to use Eqn. (7) to estimate κ from historical studies where no raw quantal data is provided. All that is needed are estimates for ECT_{50}/LCT_{50} and k_C , and it is not a requirement that the parameter estimates be taken from the same study.

Ratio of ECT_{50} and LCT_{50} Values. The ECT_{50}/LCT_{50} ratio is found to range from 0.77 to 0.83 for the three datasets reviewed (see Table 11 and Eqn. (7)). Based on the estimated 95% confidence limits of the individual ratio values, there is no significant difference between the values from the three datasets. The average of the ratio values equals 0.80. Only Cresthull, *et al.* reported an estimate for the ratio, 0.80, which is in agreement with the ordinal regression ratio value of 0.77 for this dataset.

Steepness of Dose-Response Curves. The ECT_{50}/LCT_{50} ratio represents a comparison between the steepness of the two DR curves (DR-P and DR-S) (see Eqn. (7)). There is no statistically significant species effect on ECT_{50}/LCT_{50} (as mentioned previously). However, there is a species effect on both κ (smaller for the monkey than for the rat) (see Table 10) and k_C (smaller for the monkey than for the rat) (see Table 9). Thus, there is no change in ECT_{50}/LCT_{50} values, since changes in both κ and k_C have roughly the same dependence on species. In practical terms, this means that the monkeys in Cresthull, *et al.*, had more individual variability (lower k_C value), but a steeper DR-S curve (lower κ value), than the rats in Mioduszewski, *et al.* and Anthony, *et al.*

Defining Threshold Lethality. Historically, defining the threshold lethality for a nerve agent has been a difficult task.¹⁸ The operational community needs threshold lethality estimates for purposes of modeling, exposure criteria, risk assessment, *etc.* Level 3 of the Acute Exposure Guideline Levels (AEGL-3) is an example of a threshold lethality exposure estimate.²⁷ In practical terms, a threshold lethality dosage is commonly defined as the dosage that will cause mortality in about 1% of the exposed individuals (a LCT_{01}).^{10,11,18} Unfortunately, probit analysis is not suitable for accurate extrapolation from the 50% down to the 1% effect level.¹ Extrapolations beyond the 16% to 84% range are not recommended, as demonstrated by the widening fiducial limits in the example probit analysis plot shown in Figure 1. The shape of the probit plots (both fit and fiducial limits) is typical of what is expected:¹ large random errors are involved in estimating the two key values needed for the extrapolation, the LCT_{50} and k_C .

The use of ordinal logistic regression provides a better approach to the problem of defining threshold lethality. For G-type agent IH exposures, the results from Table 12 demonstrate that an ECT_{16} (severe) is equivalent to an LCT_{01} . Thus, instead of the questionable extrapolation from the median lethal dosage down to the 1%, the more statistically defensible extrapolation from the median effective (severe) dosage down to the 16% level can be performed instead. Thus, the concerns of the toxicologist about the limitations of probit analysis in estimating threshold lethality are satisfactorily addressed.

CONCLUSIONS

Estimation of the relationship between the DR curves for lethality and severe effects has been accomplished for inhalation exposures to G-type nerve agents via the use of ordinal logistic regression on data from three previously conducted animal studies.

Knowledge of the mathematical relationship between the two curves provides a better means to define threshold lethality dosage by using the dose-severe effect curve in its place. The use of ordinal logistic regression is statistically and toxicologically defensible for this application, thereby addressing concerns with the known limitations of probit analysis (the previously used approach).

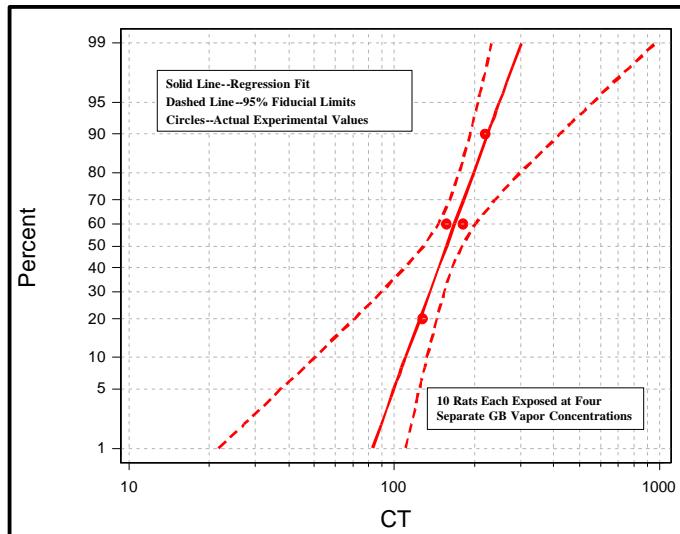


Figure 1
Percent Cumulative Probability as a Function of Dosage for Female Sprague-Dawley—Five Minute GB IH Exposure from Mioduszewski, et al. (2001)

For inhalation exposures to G-type agents, it was found that an ECT₁₆ (severe) is equivalent to a LCT₀₁ (a distance of 1.27 standard deviations). At the 16% level for severe effects, it is not improbable that an occasional death will occur among any small group of untreated victims with severe effects (convulsions, etc.)—exactly what is meant by threshold lethality. By defining threshold lethality using a sub-lethal endpoint, a safe and conservative approach is achieved, with a higher degree of statistical confidence.

Acknowledgements. I wish to thank Ms. Robyn B. Lee of Robyn B Lee and Associates LLC for presenting this paper (on short notice) in my place at the Eighth US Army Conference on Applied Statistics, North Carolina State University, Raleigh, NC, 31 October 2002. In addition, the technical assistance of Mr. Ronald B. Crosier, Dr. Sharon A. Reutter, Dr. Robert J. Mioduszewski and Mr. J. Steven Anthony (US Army Edgewood Chemical Biological Center) is greatly appreciated.

REFERENCES

- [1] Finney, DJ, **Probit Analysis**. Third Edition, Cambridge University Press, Cambridge, 1971.
- [2] Hosmer, DW and Lemeshow, S, **Applied Logistic Regression**. John Wiley & Sons, New York, 1989.
- [3] Bliss, CI, *The Determinations of the Dosage-Mortality Curve from Small Numbers*, **Quarterly Journal of Pharmacology**. 2: 192-216, 1938.
- [4] Mioduszewski, RJ, Manthei, JH, Way, RA, Burnett, DC, Gaviola, BP, Muse, WT Jr., Sommerville, DR, Crosier, RB, and Thomson, SA, *Interaction of Exposure Concentration and Duration in Determining Acute Toxic Effects of Sarin Vapor in Rats*, **Toxicological Sciences**. 66: 176-184, 2002.
- [5] Mioduszewski, RJ, Manthei, JH, Way, RA, Burnett, DC, Gaviola, BP, Muse, WT Jr., Anthony, JS, Durst, HD, Sommerville, DR, Crosier, RB, Thomson, SA, and Crouse, CL, *ECBC Low Level Operational Toxicity Program: Phase I—Inhalation Toxicity of Sarin Vapor in Rats as a Function of Exposure Concentration and Duration*, **ECBC-TR-183**. US Army ECBC, APG, MD, August 2001. AD# A394372.

- [6] Mioduszewski, RJ, Manthei, JH, Way, RA, Burnett, DC, Gaviola, BP, Muse, WT Jr., Thomson, SA, Sommerville, DR, Crosier, RB, Scotto, JA, and McCaskey, DA, *Low Level Sarin Vapor Exposure in Rats: Effect of Exposure Concentration and Duration on Pupil Size*, ECBC-TR-235. US Army ECBC, APG, MD, May 2002. AD# A402869.
- [7] Anthony, SJ, Haley, MV, Manthei, JH, Way, RA, Burnett, DC, Gaviola, BP, Sommerville, DR, Crosier, RB, Mioduszewski, RJ, Jakubowski, EM, Montgomery, JL, and Thomson, SA., *Inhalation Toxicity of GF Vapor in Rats as a Function of Exposure Concentration and Duration and Its Potency Comparison to GB*, presented at **Biosciences 2002 Medical Defense Review**, US Army MRICD, Hunt Valley, MD, 2-7 June 2002.
- [8] Anthony, SJ, Haley, MV, Manthei, JH, Way, RA, Burnett, DC, Gaviola, BP, Sommerville, DR, Crosier, RB, Mioduszewski, RJ, Thomson, SA, Crouse, CL, and Matson, KL., *Inhalation Toxicity of GF Vapor in Rats as a Function of Exposure Concentration and Duration and Its Potency Comparison to GB*, ECBC-TR-XXX. US Army ECBC, APG, MD, In publication.
- [9] Reutter, S, *Hazards of Chemical Weapons Release during War: New Perspectives*, **Environmental Health Perspectives**. 107(12): 985-990, December 1999.
- [10] US EPA, Office of Pollution Prevention and Toxics, **NERVE Agents GA, GB, GD, GF (CAS Reg. No. 77-81-6, 107-44-8, 96-64-0, and 329-99-7): Proposed Acute Exposure Guideline Levels (AEGLs)**. October 2000. <http://www.epagov/docs/fedrgstr/EPA-Tox/2001/May/Day-02/4940.pdf> (September 2001).
- [11] US EPA, Office of Pollution Prevention and Toxics, **NERVE AGENT VX (CAS Reg. No. 50782-69-9): Proposed Acute Exposure Guideline Levels (AEGLs)**. October 2000. <http://www.epagov/docs/fedrgstr/EPA-Tox/2001/May/Day-02/4945.pdf> (September 2001).
- [12] Cresthull, P, et al., *Inhalation Effects (Incapacitation and Mortality) for Monkeys Exposed to GA, GB, and GF Vapors*, CWLR 2179. US Army Chemical Warfare Laboratories, Army Chemical Center, Edgewood Arsenal, MD, 16 September 1957. AD# 145581.
- [13] Agresti, A, **Categorical Data Analysis**. John Wiley & Sons, New York, 1990.
- [14] Allen, BC, Hertzberg, RC, Strickland, JA, and Teuschler, LK, *Categorical Regression for Dose-Response Modeling of Toxicity Data and It's Application to RfD/C Development*, **Department of Defense Workshop**. Sponsored by the US EPA, National Center for Environmental Assessment, Cincinnati, OH. Workshop held 27 April 1998, Wright-Patterson Air Force Base, Dayton, OH.
- [15] US EPA, *The Use of the Benchmark Dose Approach in Health Risk Assessment*, **EPA/630/R-94/007**. US EPA, Office of Research and Development, Washington, DC, 1995.
- [16] Guth, DJ, Carroll, RJ, Simpson, DG, and Zhou, H, *Categorical Regression Analysis of Acute Exposure to Tetrachloroethylene*, **Risk Analysis**. 17(3): 321-332, 1997.
- [17] Dourson, ML, Hertzberg, RC, Hartung, R, and Blackburn, K, *Novel Methods for the Estimation of Acceptable Daily Intake*, **Toxicology and Industrial Health**. 1(4): 23-41, 1985.
- [18] Sommerville, DR, *A Novel Idea on How to Define Threshold Lethality for Nerve Agent Inhalation Toxicity*, **Third Annual Decontamination Commodity Area Conference**. Salt Lake City, UT, sponsored by the Joint Service Material Group's Commodity Area Manager for Decontamination, 23-26 May 2000.
- [19] MacFarland, HN, *Designs and Operational Characteristics of Inhalation Exposure Equipment*, in Salem, H, ed. **Inhalation Toxicology**, Marcel Dekker, New York. pp. 93-120, 1987.
- [20] Salem, H, *Principles of Inhalation Toxicology*, in Salem, H, ed. **Inhalation Toxicology**. Marcel Dekker, Inc., New York, pp. 1-33, 1987.
- [21] Crosier, RB, and DR Sommerville, *Relationship Between Toxicity Values for the Military Population and Toxicity Values for the General Population*, ECBC-TR-224. US Army ECBC, APG, MD, UNCLASSIFIED, March 2002. AD# A400214.
- [22] MINITABTM **Statistical Software**, Release 13.32. MINITAB Inc., 3081 Enterprise Drive, State College, PA 16801-3008, website: [/www.minitab.com/](http://www.minitab.com/). 2002.
- [23] Franks, AP, Harper, PJ, and Bilo, M, *The Relationship Between Risk of Death and Risk of Dangerous Dose for Toxic Substances*, **Journal of Hazardous Materials**. 51: 11-34, 1996.
- [24] McCullagh, P, and Nelder, J, **Generalized Linear Models**, Chapman and Hall, NY, 1992.
- [25] Barry, BA, **Errors in Practical Measurement Science, Engineering and Technology**. John Wiley & Sons, Inc., NY, 1978.
- [26] Mood, AM, Graybill, FA, and Boes, DC, **Introduction to the Theory of Statistics**. Third Edition, McGraw-Hill, NY, 1974.
- [27] Crossgrove, RE, ed., **Standing Operating Procedures for Developing Acute Exposure Guideline Levels for Hazardous Chemicals**. National Research Council, Committee on Toxicology, Subcommittee on Acute Exposure Guideline Levels (Krewski, D, Chair), National Academy Press, Washington, DC, 2001.

A Method for Assessing Randomness in the United States Army's Biochemical Testing Program

CPT Kevin P. Romano, United States Military Academy

Abstract: Each year the United States Army spends millions of dollars combating its number one threat to soldier readiness. Some assert the futility of combating it, acknowledging its pervasive presence. Those with lives closely tied to the soldier at risk are deeply affected also, straining the Army's resources in their struggle to support their service member. What is this threat to soldier readiness, safety and the well being of Army families? Substance abuse.

In the early 1980's the Department of Defense realized that a more aggressive program was needed to curb the amount of drug and alcohol abuse among the armed services. At that time DoD instituted mandatory biochemical testing in the Armed Forces. The Army chose to implement what has been termed "Smart Testing." The foundation of Smart Testing is based upon randomness- randomness of collection date and individual. Randomness of collection date is the one aspect of Smart Testing that is hard to address. There is currently no automated method that assigns random collection dates. Nor is there a metric for Army officials to assess the randomness of unit collection dates.

This paper presents a user-friendly method, the Testing Order of Merit statistic, to qualitatively analyze the randomness of biochemical testing within the Army. The method presented expands upon theory originally presented by Claude E. Shannon of Bell Labs in the 1940s. Additionally, historical data is analyzed and presented using the Testing Order of Merit statistic as a metric.

Contributed Session 8

The Analytic Challenges of the Army's Network-enabled Future Combat Systems

**LTC Duane E. Brucker, U.S. Army TRADOC Analysis Center - White Sands
Missile Range; Paul J. Deason, U.S. Army TRADOC Analysis Center - White Sands
Missile Range**

“...The science and technology insights and breakthroughs are being discovered today in labs, workshops and simulations centers all across the country. We’re looking for capabilities that will gird a capabilities-based force for the full spectrum of missions we will face in the 21st century”

Army Chief of Staff Eric K. Shinseki

Abstract: The danger of the Army not transforming into a force that can project real sustainable combat power anywhere in the world is the Army becoming irrelevant to national security. The vision GEN Shinseki has for the future of the Army is encompassed in the transformation to a future Objective Force. The primary instrument for Transformation is the yet-to-be-developed Future Combat System. The envisioned FCS will be networked to allow real-time situational awareness and require less logistics and maintenance than current combat systems. It will also be able to operate more effectively in joint operations. This Objective Force will be more deployable than current heavy divisions, yet have more lethal firepower than today’s light and heavy divisions. “We must be able to project power anywhere in the world – not just in the easily accessible areas with multiple air and sea ports of debarkation, but in the most remote, landlocked and infrastructure-poor areas as well,” Shinseki said. “That goal was critical as we crafted the Army Vision over two years ago. Our current operations in Central Asia reinforce the need for Objective Force capabilities as we balance this global war against the asymmetries of international terrorism with the regional Threats that demand our attention and a need for conventional warfighting process.”

The Army may be required to deal with a wide range of potential operations from stability and security operations (SASO) through small-scale contingencies (SSC) and regional conflicts to major theaters of war (MTW) in the next decade and beyond. The globalization of societies, the urbanization of populations, and the advances in technology, information technology in particular, combine to produce unique challenges to the superiority of the Army in any operation that it may undertake. A comprehensive description of the FCS operational environment may be found in “The Future Threat 2015”

Principal assumptions in analyses are that the models and simulations, and their data and algorithms represent the FCS, Stryker Brigade Combat Team, and legacy forces sufficiently accurately and robustly for meaningful analyses and comparisons to be accomplished. This is especially vital in representing the electronics, sensors, communications and fusion systems that transform data to information to intelligence to knowledge to decision following the guidance of the mission and the tactics, techniques, and procedures necessary to enable the force.

The network centric and information focus of the FCS-equipped Objective Force requires the combat simulation models be able to represent the gathering, processing, dissemination, interrelation and interaction, and effect of information to a degree that heretofore has not been necessary. Or possible.

Technical solutions projected for FCS do not exist. Data will be derived from models and not from test results. These solutions will be expected from the well-designed experiments and analytic rigor from the entire body of experimental statisticians throughout the United States. This is the challenge to this body. The challenge for the future to be a part of the necessary transformation of the Army to an effective, efficient, and relevant arm of decision for the national security of the United States and its allies in the 21st Century.

Threat Management Using Passive Inference of Network Infrastructure Topology

John Rigsby, Naval Surface Warfare Center; Jeff Solka, Naval Surface Warfare Center

Abstract: Understanding how network infrastructure changes with time is essential to protecting an organization's network. Multiple methods for discovering network topology using different areas of graph theory and concepts of social network relationships will be discussed. Passively detecting changes in network topology and presenting this to network engineers and analysts will increase an organization's threat management capabilities to counter malicious network activities; the application of change point detection to social networks will be the backbone of this approach. This research project is in the early stages of development and will be presented as such.

A Statistical Methodology for Automatic Target Recognition in Satellite Imagery

John Bart Wilburn
Recognition Research
Tucson, Arizona
Wilburn@dakotacom.net

Abstract: A methodology is presented for Automatic Target Recognition (ATR) of missiles in satellite imagery of a cloud-covered earth. The method is based on the results of a two dimensional local maximum filter applied to a sequence of simulated images capturing the boost phase of a missile, and is an algorithmic estimate of the probability of detection on a trajectory, i.e., a “track”. Track association is determined by the significance level for rejecting hypotheses that detection events in sequential frames of imagery are not associated with missile targets or satellite motion.

Introduction:

The problem addressed here is a calculation of statistical confidence in the detection and tracking of missile targets acquired by a satellite sensor in the boost phase of missile deployment. The specific task of the system is to detect a small, relatively bright, moving target against a scene cluttered by clouds. The criteria for success are a high probability of detection and a low probability of a false alarm, and computational efficiency. Further, assumptions for this system must be minimal as follows:

1. The elements of the scene are either background features, e.g., clouds and terrestrial lights, or missile targets.
2. Target radiance is additive.
3. The target moves along a line of predictable shifts between frames of imagery,
4. The target object is small, generally much less than the resolution of the optical system, thus it is represented in the image of the point-spread function (psf) of the optical system, adequately sampled by the focal plane array (FPA), with the peak expected most of the time to be in the instantaneous-field-of-view (IFOV) of one pixel.

The method of detection and tracking proposed is based on the notion that detection is an event, i.e., the occurrence of an isolated and distinct pixel. To satisfy this requirement, the satellite imagery must be filtered in a manner that results in distinct and isolated pixels, and the filter that satisfies this requirement is the local maximum filter¹. As will be shown, this filter also satisfies the requirements of a high probability of detection and a low probability of false alarm, is computationally efficient, and it is well behaved in the context of cloud-cluttered imagery

The Local Maximum filter:

The local maximum filter is a form of ranked-order (RO) filters developed from theory described in prior work¹⁻³ on RO filters. The application of the local maximum filter to target recognition in a cluttered image is a form of feature extraction derived from a fundamental departure of local RO filters from classical approaches to image filtering. The local maximum filter, and its RO family relative the local median filter, are a window type of filter and function according to the satisfaction of a reflexive relationship between the location and relative values of the data in the sample space of the filter window. Morphological considerations of the filter, then, apply to the structure of the data with respect to the reflexive relationship defining the filter, and the morphology of a particular object, or class of objects, of interest are represented in the filter coincidentally as properties of an object that satisfy the reflexive relationship of the filter. A

presupposition of an extensional property, such as the shape of a particular object, or class of objects, evident in some context, is not represented in the filter.

The approach to target recognition described here is not to focus on suppressing the clutter to reveal the target, but rather to focus on finding the target embedded in the clutter. A canonical description of this approach is looking for properties shared by any significant part of all targets, but not shared by other features in the image. This is the mode of object recognition employed by human observers⁴, indeed by all animals. This principle, and our application of it, and may be illustrated by the analogy of a looking for a baseball in a pile of leaves. A baseball has the properties of sphericity, grayness and it has a seam. All baseballs share these properties; wherever you see a baseball under any conditions you see it, it has these properties. If we can see as much as any quarter of any baseball in any situation of the clutter, we see these properties. We expect nothing else in the pile of leaves to have these properties, and if we see anything else with these properties and claim it to be a baseball, then we commit a type-II error - a false alarm. Conversely, if we see some part of a baseball and reject it, then we commit a type-I error.

Filters can be constructed in a variety of geometries (Figure 1) that include adjacent pixels on linear intersecting arms, denoted by the “0”, with a single common pixel at the intersection, or center position indicated by “⊗”, that is also the output port of the filter. The orthogonal filter prohibits adjacent maxima on the horizontal and vertical axes while the diagonal filter prohibits adjacent maxima on the diagonal axes. The hexagonal filter prohibits adjacent maxima on all but the horizontal axis while the octagonal filter prohibits adjacent maxima on all axes. The latter is well suited to detection and track of a missile launch in any direction.

Orthogonal (17)	Diagonal (17)	Hexagonal (25)	Octogonal (33)
$\begin{bmatrix} 0 & & & \\ 0 & & & \\ 0 & & & \\ 0 & & & \\ 0 & 0 & 0 & 0 \otimes 0 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & & \\ 0 & & & \\ 0 & & & \\ 0 & & & \end{bmatrix}$	$\begin{bmatrix} 0 & & & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & 0 & 0 & \otimes \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & & & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & & & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & 0 & 0 & \otimes \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

Figure 1. Filter Geometry (Number of Elements)

The local maximum filter functions according to a rank-order relationship of the data sampled by the filter constrained by a *predicate*. A predicate has semantic content as it is either true or false, and the filter is constrained by it to be defined if the predicate is true, or undefined if it is false. The predicate is a description of the data in terms of conditions. If the data do not satisfy the conditions, then the predicate is false and the filter returns a null result. If the data satisfy the conditions, then the predicate is true and the filter returns the result of the filter function at the center position of the window.

The predicate conditions of the local maximum filter involve the value and position of an individual datum with respect to the value and position of all data within the scope of the window, and they apply to each of the 1-d arms of the filter simultaneously for each increment of the filter in the field of data. The predicate conditions are: (a) the datum sampled by the center position of the filter, “⊗”, is simultaneously the maximum value of the data of every axis of the filter, and (b) the satisfaction of a threshold established by the SNR statistics of the image. This threshold defines the lower limit of a reasonably expected target in terms of its SNR determined by its brightness and the standard deviation of the image brightness. If the data sampled by the filter satisfy these conditions, the predicate is true and the filter returns the datum sampled by the center position of the window. If the predicate is not true, then the filter returns a zero. This functioning of the octagonal filter results in isolated and distinct pixels as

intended, but of particular interest is the additional realization that the output image of all local maximum filters is a subset of the input image, thus radiometric information is preserved.

The effectiveness of the octagonal local maximum filter in isolating local maxima is demonstrated (Figure 2) for a simulated cloud scene (102×102 pixels) in a solar band from the Synthetic Scene Generation Model (SSGM). The imagery is presented as a bit map on an 8-bit scale, thus all pixel values are on a scale of 0 to 255. The figure compares the pixel imagery and histogram/exceedance statistics before and after application of the filter. The comparisons indicate a significant reduction in the number of pixels at all brightness levels: the total number of pixels decreases from greater than 10^4 to less than 200. The filtered image consists of a pattern of isolated pixels as intended, and the pixel levels of the original, or input, image are preserved in the output, or filtered, image, thus radiometric information is preserved. As can be seen, many regions of a natural image above some threshold do not have local maxima in the sense defined here.

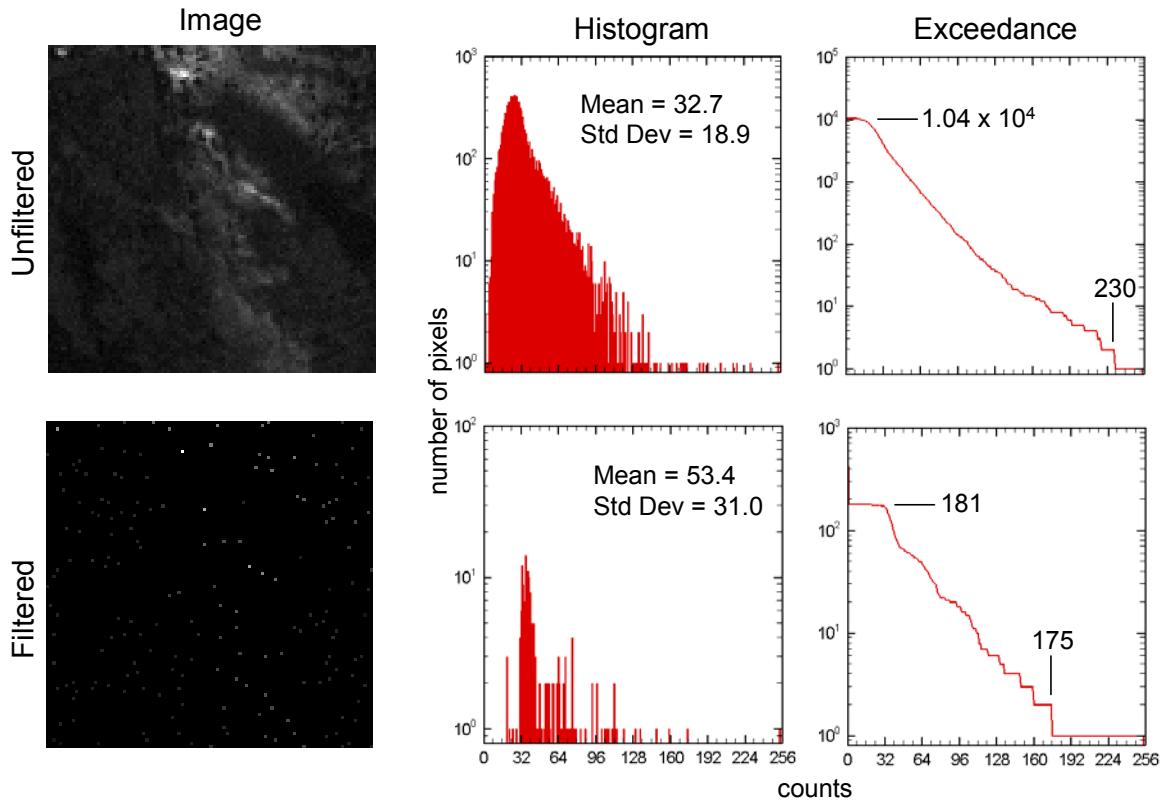


Figure 2. Filter Performance for SSGM Scene

The performance of the filter is specific to a class of image clutter represented by figure 2. The image in figure 2 is one frame of a sequence of 26 frames simulating the continuous nadir view at one frame per second of a cloud cluttered earth by satellite sensor in low earth orbit. This set of imagery is reasonably stationary in the strict sense: the image mean of the set $\langle I \rangle = 31.1$ with, $\sigma_I = 2.72$, and mean standard deviation: $\langle \sigma_I \rangle = 19.4$ and $\sigma_o = 1.1$. The statistics of the image in figure 2 are: $\langle I \rangle = 33.7$; $\sigma_I = 18.9$, $skew = 2.58$, and $Kurtosis = 15.42$. The filter performance is described by two parameters: the probability of detection, p_d , and the probability of false alarm, p_{fa} . These probabilities are measured in two kinds of Monte Carlo tests of 300 trials of the filter against the raw image shown in figure 2 for randomly inserted targets of different brightness levels. The p_d is the empirical probability in the Monte

Carlo test that the actual target is one of the output pixels of the filtered image, and p_{fa} is the empirical probability of pixels appearing in the output image that are not the target.

The first kind of Monte Carlo test, shown in figure 3, is of a fixed composite of target plus background randomly inserted in place of a background pixel. The second kind of Monte Carlo test, shown in figure 4, is of a fixed target brightness added to a randomly selected background pixel. These two Monte Carlo tests satisfy different purposes. The first kind is a measure of the filter to detect fixed pixel brightness levels in the test image. This first kind of Monte Carlo test measures the probability of the filter to detect an unknown target parameterized by an average brightness of the pixel containing the target measured over all images having an intensity distribution as shown in figure 2. The second kind of Monte Carlo test is of a simulated target of known, absolute brightness to satisfy a given probability of detection in imagery having an intensity distribution as shown in figure 2. As will be seen, the average signal-to-noise of the target plus background pixels, $\langle \text{SNR}_t \rangle$, measured in the second test, and the $\langle \text{SNR}_t \rangle$ inferred from $\langle I_t \rangle$ of target pixels, and $\langle I \rangle$ and σ_I of the total image in the first test are remarkably close.

The first kind of test is parameterized by the signal-noise (SNR_i) of the target *pixel* for the i^{th} level of brightness: T_i .

$$\text{SNR}_i = \frac{T_i}{\sigma_I}, \text{ where } \sigma_I \text{ is the standard deviation of the image brightness.} \quad (1)$$

The filtered image is subject to a threshold test by setting to zero all pixel output of the filter satisfying the integer calculation of $p_n < \text{SNR}_i * \sigma_I$. The purpose of the test was to describe the behavior of the filter parametrically in terms of p_d and p_{fa} as a function of a variable SNR_i in the space of 0 to 7.5. The p_d and p_{fa} results are shown in figure 3.

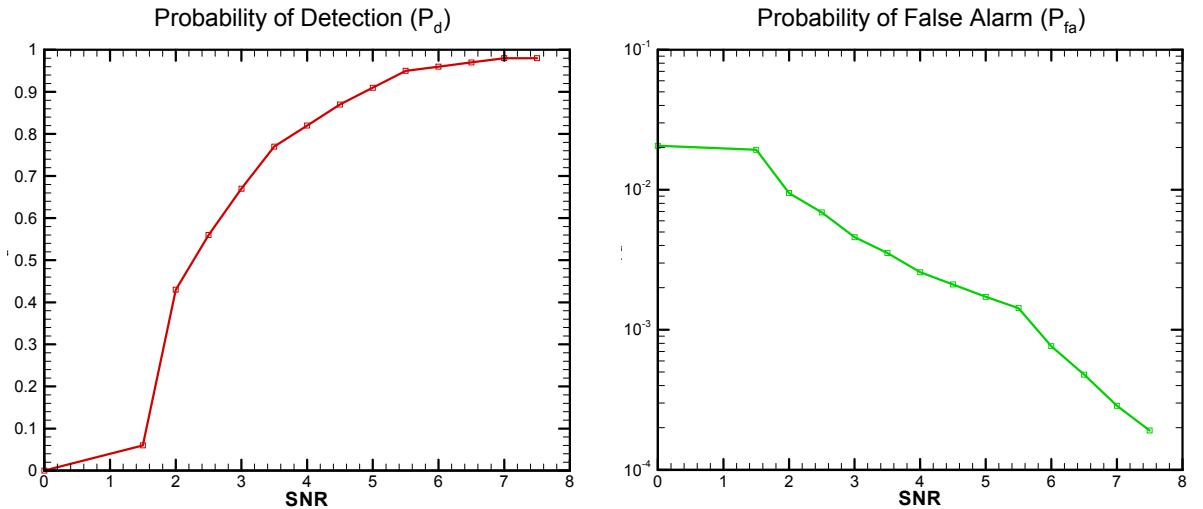


Figure 3. Composite Target Pixel Detection and False Alarm from Monte Carlo Statistics

The results of the Monte Carlo Testing show that the filter is well behaved and that the region of a reasonable target is one having a $\text{SNR} \geq 2.5$ satisfying a probability of detection of $p_d > 0.5$ and a probability of false alarm of $p_{fa} = 6.4 \times 10^{-3}$.

The second kind of Monte Carlo test follows from the first kind with a criterion for detection of $p_d > 0.5$. In this test, we fix the p_d and search for the minimum absolute target brightness satisfying this criterion as a function of the threshold of the filter in terms of the SNR. The results are curves of absolute target brightness as a function of SNR, or p_{fa} by figure 3, parameterized by p_d . The average (target + background) pixel signal-to-noise, $\langle \text{SNR}_t \rangle$ satisfying p_d is computed over all measures of SNR_t in the Monte Carlo tests of the image.

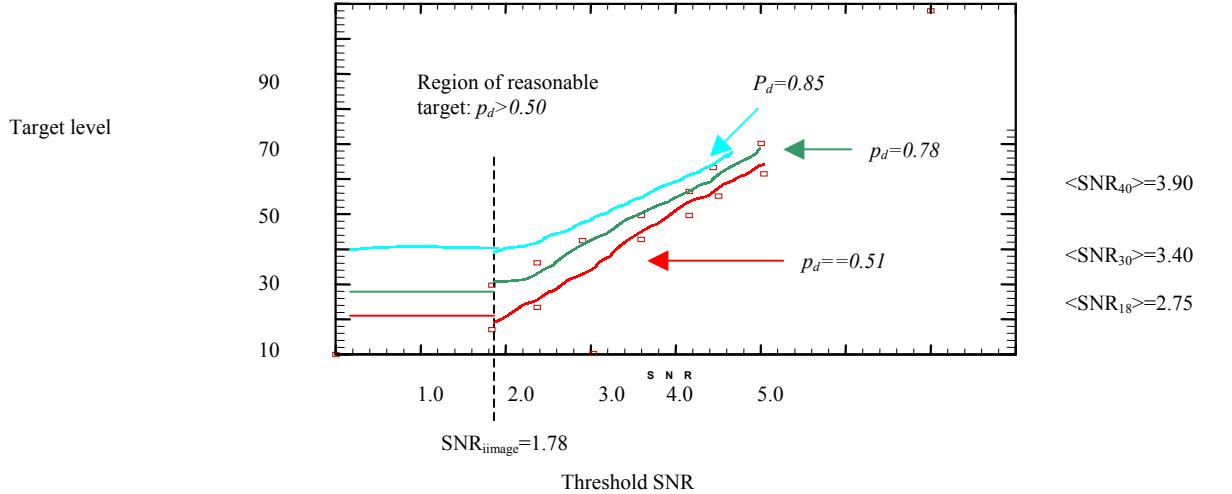


Figure 4. Target Level Satisfying a p_d from Monte Carlo Statistics

The distribution of the probability of detection of a given target level in this test follows the distribution of the image intensity (figure 2) according to conventional signal detection theory in terms of the skewed statistics of the image data. Thus it follows that the target level for $p_d=0.50$ is constant for thresholds less than or equal to the SNR of the image, as that threshold level corresponds to the image mean value, and increasing for thresholds greater than the image mean. By the same reasoning, we expect that target levels for $p_d>0.5$ is constant to thresholds somewhat greater than the SNR of the image, and increasing thereafter, and we see that this is indeed the case. The results show that the minimum target level for $p_d=0.51$ is *target level*=18 having a $\langle \text{SNR}_t \rangle=2.75$ at a threshold of $\text{SNR}=1.78$, or by figure 3, a $p_{fa}=1.58 \times 10^{-2}$, and increases to *target level*= 26 ($\langle \text{SNR}_t \rangle=3.18$) at a threshold of $\text{SNR}= 2.5$ ($p_{fa}=6.4 \times 10^{-3}$). The results for higher probabilities of detection follow a similar pattern: The minimum target level at $p_d=0.78$ is *target level*=30 ($\langle \text{SNR}_t \rangle=3.40$) at a threshold of $\text{SNR}= 2.0$ ($p_{fa}=1.2 \times 10^{-2}$), and for $p_d=0.85$, the minimum target level is a *target level*=40 ($\langle \text{SNR}_t \rangle=3.90$) at a threshold of $\text{SNR}= 2.5$ ($p_{fa}=6.4 \times 10^{-3}$). These results compare favorably with the inferences drawn from figure 3, and the expected results from signal detection theory.

Target Recognition:

The schema for target recognition combines the notions of detecting local maxima and persistent linear motion. Persistent linear motion is a high probability of detection on a line of predictable shifts between frames, and it is referred to as a *track*. Determining a high probability of detection on a track is *track association*. We may combine these notions with our assumptions of the imagery and conclude that any track consistent with background motion, which we know a priori, is not like a target, and any track inconsistent with background motion is like a target. In this way, track association inconsistent with satellite motion, i.e., inconsistent with background motion, constitutes target recognition.

We may illustrate the schema for recognition of missile targets in boost phase by application of the Octagonal filter to a dynamic sequence of 25 SSGM images simulating sensor acquisition at the rate of 1 frame per second of a missile target. The missile simulated in this set of imagery is a theater missile and has a variable target pixel brightness with $\langle \text{SNR}_t \rangle = 3.8$ and a $\sigma_t = 1.7$. The results of the Monte Carlo test of the filter shown in figure 3 indicate that we may expect this target to have a $p_d \geq 0.8$, thus it is a target we may reasonably expect to detect and it is a typical target by definition. Recognition of the target is determined by the measure of confidence we may have in associating detection of this target with a specific track as a reasonable and typical target.

The detections of the target and a background feature suggest a hypotheses of a target track $r_t = x - 4y$, i.e., a shift of $\Delta x = 1$ and $\Delta y = -4$ between frames, and a satellite motion track of $r_s = -2x - 3y$. Detections by the local maximum filter satisfying the hypotheses of r_t and r_s at a threshold of $\text{SNR} = 2.5$ are shown in figure 5 as a composite of frames plotted on an arbitrary frame of the input set of imagery for reference. The detection events are indicated by a bright cross reflecting a tolerance in detection of $+/- 1$ pixel to allow for non-integer multiples of pixel movement between frames. Failures of the filter to detect an object satisfying this hypothesis are shown by a null result.

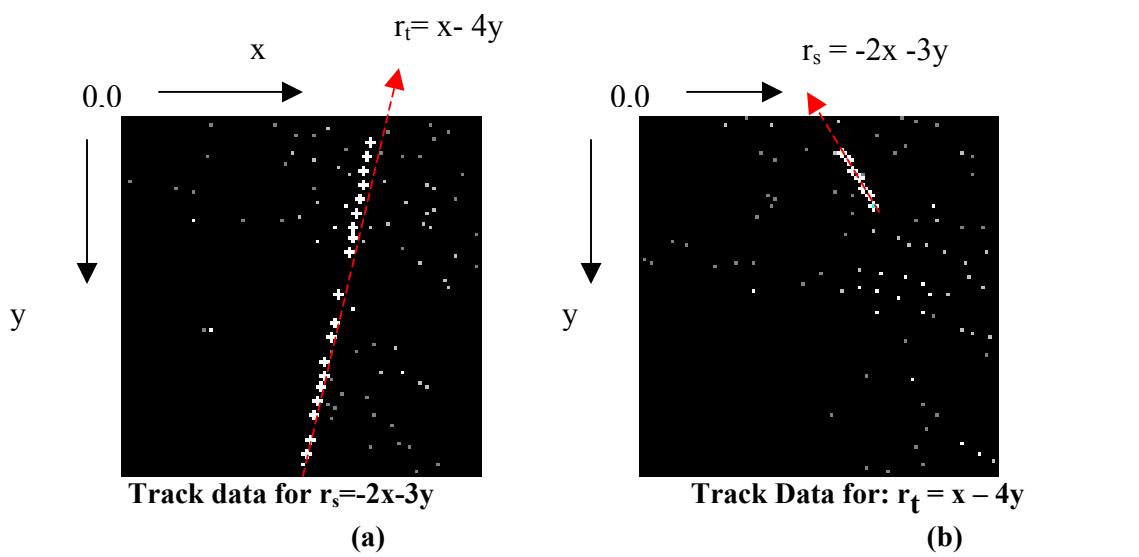


Figure 5: Track data for 25 sequential frames

The notion of detection being an event enables this approach to result in an assertion of target recognition bounded by the probabilities of type-I and type-II errors as follows. The data in figure 5 reveal that the filter detected an object on track r_t in $m=20$ of $N=25$ frames, and we employ the binomial probability law (1) to calculate the type-I and type-II errors, α and β respectively, for hypotheses of p_j on r_t .

$$P(x \geq m | p_j) = \sum_{i=m}^N \binom{N}{i} p_j^i q_j^{N-i} \quad (1)$$

$$\beta_j = 1 - P(x \geq m | p_j)$$

$$\alpha_j = P(x \geq m | p_j)$$

To find the maximum likelihood estimate of p_d , we plot the α and β errors as a function of $p_j = 0 - 1.0$ as shown in figure 6. From figure 6, we can see that a 99.6% confidence region of p_d is found by observing that the type-I error = 0.002 at $p=0.5$, and the type-II error = 0.002 at $p=0.945$, thus we have a probability of 0.996 that $0.5 < p_d < 0.945$. The maximum likelihood estimate of p_d is the value of p_j where the type-I error = the type-II error = 0.5, and that occurs for $p_j=0.78$.

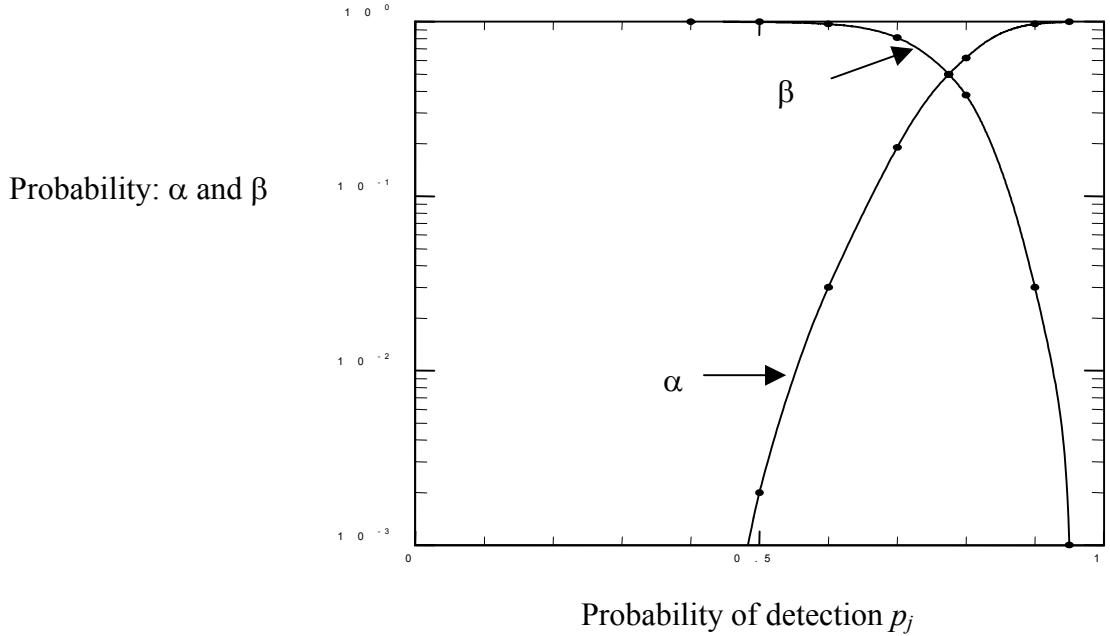


Figure 6: Type-I and Type-II probabilities of p_d

Estimating p_d is an intermediate step to asserting an object to be a target. We may combine the estimate of p_d with the notion that it is specific to a track and define target recognition as: $p_d > 0.5$ on a track. We may then accept or reject a null hypothesis that an object is *not* a target on that basis with the alternative hypothesis being that the object *is* a target; the decision determined by the respective probabilities of either hypothesis being true. The probability of the null hypothesis being true is given by the type-I error = 0.002 evaluated at $p_d = 0.5$, thus we may confidently reject the null hypothesis, knowing that there was only a 0.002 chance that it was true, i.e., a significance level of 0.002, and accept the alternative hypothesis that the object *is* a target on a track, or *it is recognized*, with a probability given by the type-II error = 0.998 of being true.

Tracks are found by hypotheses of a track on all pixels based on neighboring pixels within range of possible motion in subsequent frames. The hypotheses are confirmed in the manner shown here to be either satellite tracks and ignored, target tracks and recognized, or pixels that cannot be confidently associated with any track and ignored as false alarms. The development of this method is dependent on the fidelity of the simulation of the targets, and particularly dependent on the fidelity of the simulation of the imagery of a cloud cluttered earth produced by a satellite sensor.

Attribution:

This material is based upon work supported by, or in part, the US Army Research Laboratory and the US Army Research Office under contract/grant: G 44394-MA. Presented⁵: 8th Army Conference on Applied Statistics.

References

1. Wilburn, J. B., "Development of the local maximum variety of ranked-order filters", *Journal of the Optical Society of America A* (JOSA A), **19**, pp 1994-2004, 2002.
2. Wilburn, J. B., "Theory of ranked-order filters with applications to feature extraction and interpretive transforms", in *Advances in Imaging and Electron Physics*, ed. P. Hawkes, Harcourt-Brace Academic Press, **112**, pp233-332, 2000.
3. Wilburn, J. B., "Developments in generalized ranked-order filters", *Journal of the Optical Society of America A* (JOSA A), **15**, pp. 1084-1099, 1998.
4. Wilburn, J. B., "A possible worlds model of object recognition", *Synthese*, Kluwer Academic Publications, **116**(3), pp. 403-438, 1998
5. Wilburn, J. B., "Automatic Target Recognition is Satellite Imagery", Proc. 8th Army Conference on Applied Statistics, 31 October 2002.

Finding Clusters

Jon R. Kettenring, Telcordia Technologies

Abstract: Military intelligence and homeland defense problems often involve analyses of massive amounts of data. It is tempting to tackle such problems using methods of data mining, the most common of which is cluster analysis. Indeed, clustering methods appear on the surface to be just the right tools for breaking complex, difficult data down into manageable cohesive subsets. Yet the reality in practice is that these methods often fall short on performance. In this talk, I will offer a personal perspective on the difficulties of finding clusters and suggest opportunities where new research could help improve the situation.

Special Session 3

Assessing Uncertainty in Mesoscale Numerical Weather Prediction **Montserrat Fuentes, North Carolina State University; Adrian Raftery, University of Washington**

Abstract: Current methods of meteorological forecasting produce predictions with unknown levels of uncertainty, particularly in regions with few observational assets. Forecast errors and uncertainties also arise from shortcomings in model physics. With the ability to estimate the uncertainty in predictions, forecasters would have a powerful tool to make decisions and to judge the likelihood of mission success.

The goals of our work are to develop methods for evaluating the uncertainty of mesoscale meteorological model predictions, and to create methods for the integration and visualization of multisource information derived from model output, observations and expert knowledge. We do this by extending the recently developed Bayesian melding approach.

We also develop a new approach to assess the performance of mesoscale numerical models, and show how it can also be used to remove the bias in model output. We specify a simple model for both numerical model predictions and observations in terms of the unobserved ground truth, and estimate it in a Bayesian way.

Local Probability Propagation Algorithms for Approximate Inference in Graphical Models **Martin Wainwright, University of California-Berkeley, Tommi Jaakkola, Massachusetts Institute of Technology, Alan Willsky, Massachusetts Institute of Technology**

Abstract: Graphical models provide a flexible and powerful framework for modeling interactions among a collection of random variables. As such, they are studied and used for a variety of purposes, including statistical image processing, error-correcting coding, artificial intelligence, and communication theory. One important problem is to use a set of noisy observations to compute important statistical quantities, including posterior marginals, the maximum a posteriori (MAP) configuration, and the log likelihood of the data. Local probability algorithms (e.g., belief propagation; sum-product; max-product) have proven to be very useful for this purpose. If the graph is cycle-free (i.e., a tree), then the local message-passing updates are guaranteed to converge and compute the correct quantities. For a graph with cycles, on the other hand, convergence is no longer guaranteed, and the algorithms provide only approximations.

In this talk, we describe how these algorithms can be understood as seeking a particular “reparameterization” of the distribution on the graph with cycles. This perspective leads to an intuitive characterization of the possible fixed points, and also to understanding of the nature of the approximation error. We also briefly describe various extensions based on convex analysis, including: (a) a technique that provides efficiently computable

bounds on the log likelihood; (b) a method for computing provably exact MAP estimates for a subclass of problems, based on the idea of (hyper) tree agreement.

C4ISR and the Future Force

Monica Farah-Stapleton, Communications-Electronics Command

Abstract: The last decade has seen the U.S. Armed Forces engaged in an intensive effort to digitize the battlespace, by leveraging the Information Technology explosion of the Nineties. However, this approach did not address the warfighter's ability to be more mobile and responsive. The goal for the next decade is to exploit network centric technologies to support a lighter, more responsive, and more lethal force. Consistent with this trend, future tactical Army systems are expected to integrate Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) capabilities to an unprecedented degree. The concept of trading "information for armor" poses a significant challenge for the new generation of Command and Control (C2), Intelligence, Surveillance and Reconnaissance (ISR) (to include Target Acquisition) and Communications (Comm) systems. The US Army has performed a C4ISR systems engineering analysis focusing on the Dismounted Soldiers through the Unit of Employment (UoE) view. The intent of this study was to provide a process as well as a product that will facilitate discussion and collaboration among lead Government and Contractor organizations regarding the integration of previously disparate C2, ISR, and Comm domains. One of the by-products of a C4ISR systems engineering analysis is the source data employed in the methodology to represent C4ISR in Force on Force models. As the current focus is on Comm representation, the methodology is generating model propagation effects and model network effects. This talk will discuss the work being done by the Army on C4ISR systems of systems engineering analyses in support of future tactical Army systems, and the relationship of this work to C4ISR representation in Force on Force models.

Particle Filtering and Spatial Prediction in the Battlespace

Noel Cressie, Mark Irwin, and John Kornak, Ohio State University

Abstract: There are considerable difficulties in the integration, visualization, and overall management of battlespace information. One problem that we see as being very important is the combination of (typically digital) information from multiple sources in a dynamically evolving environment. In this paper, we present a spatial-temporal statistical approach to estimating the constantly changing battlefield, based on noisy data from multiple sources. The potential danger from an enemy's weapons is examined in the spatial domain and is extended to incorporate the temporal dimension. Statistical methods for estimating danger fields are discussed, and an application is given to a data set generated by a simple object-oriented combat-simulation program that we have developed. This research was carried out by a group of Ohio State University statisticians supported by ONR's Probability and Statistics Program.

General Session 4

A Microarray Lesson from Dear Old Dad (Design-Analyze-Display)

Russell Wolfinger, SAS Institute

Abstract: As our familiarity with microarray technologies matures, so should the sophistication of the ways we use them. Traditionally sound scientific practices, tempered with modern knowledge, remain the fundamental enablers of hypothesis-driven research. Adequate replication at population levels, balanced randomization of treatment assignments, and proper attention to sources of variation are all generic and time-tested methods that remain underutilized in many microarray studies. Good experimental design not only leads to the most efficient use of resources, but meshes naturally with well-chosen statistical models and inferences. Techniques such as the mixed-model analysis of variance have been employed successfully for decades in other disciplines such as agriculture, genetics, and clinical trials, and are now beginning to capture attention in the molecular biology community. A relatively new aspect is the need to visualize the high volumes of results from these analyses. We learn from DAD (Design-Analyze-Display) in the context of the peroxisome proliferator and phenobarbital data from the two papers of Hamadeh et al (2002, *Toxicological Sciences*, **67**, 219-240) and using software from the upcoming SAS Microarray Solution.

Contributed Session 9

Statistical Techniques for Breaking Steganography

R. Chandramouli, Stevens Institute of Technology

Abstract: Steganography is a relatively new area of information assurance research. It deals with hiding messages in plain looking signals (e.g., audio, image or video) such that their very presence is concealed. Recent press reports based on government intelligence sources suggest that steganography based covert communications were used in planning the 9/11 attacks. With the advances in Internet technologies steganography based security threats could become increasingly prevalent in the future. Therefore it is imperative to develop techniques to intercept and break malicious steganography. We believe that these techniques will help to develop early warning systems and trigger appropriate command and control for homeland security in which the U.S. Army plays a major role.

In this paper we discuss statistics based strategies to detect and estimate steganography based covert communications. First, breaking steganography (a.k.a. steganalysis) is formulated as a statistical decision fusion problem. This formulation is then used to detect if a signal under investigation contains any secret messages. Theoretical properties of this methodology are investigated along with practical applications. The second part of this paper considers estimating hidden messages based on a statistical blind separation technique. Within this theme, questions that are addressed include: what statistical properties must a steganography algorithm satisfy to escape detection and, steganography key estimation. A number of examples for image steganalysis applying our techniques will also be presented.

Classifier Optimization Via Graph Complexity Measures

J. L. Solka

D. A. Johannsen

Code B10

NSWCDD

Dahlgren, VA 22448

Code B10

NSWCDD

Dahlgren, VA 22448

Abstract

This paper examines the use of minimal spanning trees as an alternative measure of classifier performance. The ability of this measure to capture classifier complexity is studied through the use of a gene expression dataset. The effect of distance metric on classifier performance is also detailed within.

Keywords

minimal spanning trees, gene expression data, classifier complexity, distance metric

Introduction

Given a set of observations X along with a set of associated class labels C one is often interested in constructing a classifier which is essentially a mapping from X to C . The problem with constructing such a mapping comes when the inherent dimensionality of the observations is high. In fact many modern datasets may consist of less than one hundred observations in several thousand dimensions. Despite the requisite considerations for the curse of dimensionality the community that produced such data still wishes to be able to determine the classification mapping. In fact they are often interested in identifying a small, less than 100, subset of the collected data dimensions wherein the various classes are well separated. Given the fact that in the multiclass case the identified dimensions might be different for each of the particular classes or even different when we compare each class against every other class, one by one, the selection of these particularly fruitful dimensions might be particularly daunting.

Selection of the features is of course predicated on some sort of measure of their ability to contribute to the discriminant analysis of the particular class. One can formulate various figures of merit to use including cross-validated nearest neighbor classifier performance. This figure of merit while simple to implement is quite computationally intensive. Potential benefits may be obtained from the use of other figures of merit that might serve as an alternative to the cross-validated nearest neighbor performance, while still capturing the ability of the the cross-validated nearest neighbor performance to quantify the difficulty of the classification problem.

Another factor that can contribute to the classifier performance is the manner in which one measures the distance between observations. Our previous efforts (ref

CEP JSM 2000) indicated the benefit of varying the choice of Minkowski p parameter when one measures the distance between observations. In fact the creation of any sort of classification scheme is predicated on the ability to measure distances between observations. We are of course interested not only in how classifier performance might change as a function of the Minkowski p parameter but also how any sort of proposed complexity measure might change as one varies the p parameter.

Given these preliminary discussions the paper unfolds as follows. First we provide some background discussions as to our approach to the feature identification, classification complexity quantification, metric space adaptation methodology. Second we illustrate the application of the discussed approaches to an artificial nose and gene expression dataset. Finally we close our discussions with some indications of our future plans for continuing work on these issues.

Background

Given X a $n \times d$ matrix of observations and a set C of n associated class labels chosen from the set $1, 2, \dots, nc$ then one is interested in constructing a mapping $Y : X \rightarrow C$ that associates with each $x \in X$ a class label. The simplest way to measure the efficiency of said classifier is by the cross-validated single nearest neighbor error rate \hat{L} or alternatively by the cross-validated single nearest neighbor probability of correct classification $1 - \hat{L}$.

The performance of the classifier is in part determined by the way that we choose to measure distances between the observations. We choose to write a generalized distance function $d(x, y)$ between two observations in our original space as

$$d(x, y) = \left(\sum_{i=1}^d w_i |s(x_i) - s(y_i)|^p \right)^{\frac{1}{p}} \quad (1)$$

where w is a weighting vector that indicates how much each of the d dimensions contributes to the distance, s is a smoothing function, and p is a parameter that indicates the form of the Minkowski metric. We have previously presented results that examine the effects of varying these parameters on single nearest neighbor classifier performance as applied to an artificial nose dataset [Priebe et al., 2000].

The purpose of the current article is to continue these discussions and also to examine the use of a minimal spanning tree as a classifier complexity measure. Previously Friedman and Rafsky proposed a test based on the minimal spanning tree to determine whether two samples of observations were drawn from the same distribution [Friedman and Rafsky, 1979]. Recall that given a graph G a minimal spanning tree is a connected subgraph that has no cycles which covers each point in the graph and whose sum of edge weights is minimal [Gross and Yellen, 1999].

One can use this method to characterize classifier complexity as follows. First compute the minimal spanning tree based on the full set of observations. Second, count those edges in the minimal spanning tree that travel between disparate class types. The number of such edges is then used to characterize the complexity of the problem.

Figure 1 shows a representative spanning tree obtained based on a sample of 100 points from two bivariate normals, $\mu_1 = [1, 1], \mu_2 = [-1, -1]$, $\sigma_{xx}^1 = 1.5$, $\sigma_{yy}^1 = 1$, $\sigma_{xx}^2 = 1$, $\sigma_{yy}^2 = 1.5$. The red edges indicate edges between disparate classes.

At times in the results section we will be interested with which subset of d -dimensional features contributes in a positive manner to the discernibility of the two

classes. One classic way to characterize the manner in which a particular feature contributes is the scatter. There are numerous ways that this term has been defined in the literature but we will follow Duda, Hart, and Stork, 2000 [Duda et al., 2000]. We begin by considering two classes with \mathcal{C}_1 and \mathcal{C}_2 , with n_1 and n_2 members (respectively). The class means are given by

$$m_i = \frac{1}{n_i} \sum_{x \in \mathcal{C}_i} x.$$

The class scatter matrices are given by

$$S_i = \sum_{x \in \mathcal{C}_i} (x - m_i)(x - m_i)^t.$$

The within class scatter is defined by

$$S_W = S_1 + S_2.$$

The between class scatter is defined by

$$S_B = (m_1 - m_2)(m_1 - m_2)^t.$$

In the univariate case S_B and S_W are scalars and we choose those features with large values of S_B/S_W . In the multivariate case S_B and S_W are no longer scalars and we choose those features with large values of $\text{tr}(S_B)/\text{tr}(S_W)$.

Results

We have chosen to illustrate the performance of our algorithms using two datasets. The first dataset is an artificial nose dataset. The response of a nonlinear fiber optic artificial nose was measured when it was exposed to a target compound, trichloroethylene, in conjunction with various confusers, Coleman fuel, chloroform, and others, against just the confusers mixed with air. The dataset used for our analysis consisted of 80 observations with target compound and confuser along with 80 observations or just the confuser compound. Each observation consists of 19 fibers sampled at 2 wavelengths 60 times during the fiber exposure phase. Hence the response of the system consists of a point in a 2280 dimensional space.

The second dataset that we will be discussing is a gene expression dataset that was previously analyzed by Golub et al 1999, [Golub and Slonin, 1999]. This Afeymetix dataset consists of measurements of over 7000 genes on a group of 72 patients all of whom were currently ill with leukemia. The patient population was divided between those patients with the acute lymphoblastic variant (ALL), 47, and the acute myeloblastic variant (AML), 25. The discriminant analysis problem in this case was distinguishing between the ALL and AML varieties.

Figure 2 presents average cross-validated single nearest neighbor classifier performance, green curve, along with average cross-validated minimal spanning tree based complexity, blue curve, as a function of Minkowski p parameter obtained using the raw artificial nose data. We notice that the complexity curve is low when the performance curve is high, and that the complexity curve is high when the performance curve is low. We also note that the optimal performance is obtained at a Minkowski p parameter of 5. This type of deviation from the standard Euclidean parameter value of $p = 2$ is in keeping with our previous work on this dataset [Priebe et al., 2000].

Given an optimal p parameter value of 5 we would like to know how the performance varies as a function of the included scatter selected fibers. Figure 3 shows average cross-validated single nearest neighbor performance as a function of scatter selected fibers at $p = 5$ obtained using the raw artificial nose data. We notice that this plot indicates that one can improve upon the optimal performance of .75 obtained using all 38 fibers at a $p = 5$ by using the top 21 scatter selected fibers at a $p = 5$, measured performance level of .78125. We need to clarify that the ranking of the scatter selected fibers was obtained by computing the scatter value for each of the fibers individually and then ranking them based on the obtained scatter values. One again we point out the fact that the performance and complexity curves seem to be mirror images of one another. This is the type of behavior that we would expect if the minimal spanning tree complexity measure was truly capturing the difficulties associated with the classification problem.

The previous results detail the benefits of first choosing an optimal Minkowski p parameter and then choosing an optimal weighting or w parameter. Next we consider first choosing a smoothing function s , followed by a Minkowski p parameter, and finally an optimal w parameter. It is the hope that one can improve classifier performance by first choosing s , then p and finally w . First we smoothed the nose data using a spline-based smoother. Figure 4 presents average single nearest neighbor cross-validated performance and average complexity as a function of Minkowski p parameter for the spline smoothed nose data. The optimal p value of 29 with an associated performance of .85625 is in keeping with our previous studies of this data. The behavior of the complexity curve is as expected.

Proceeding as we did when we analyzed the raw nose data we next present a plot of performance as a function of scatter selected fibers at the associated optimal p parameter value. Figure 5 presents average single nearest neighbor cross-validated performance as a function of scatter selected fibers at $p = 29$ obtained using the spline smoothed nose data. In this case we were not able to improve upon the performance obtained using all 38 of the fibers

We next turn our attention to an analysis of the Golub gene expression dataset. Figure 6 presents average cross-validated single nearest neighbor performance, green curve, and average complexity, blue curve, as a function of Minkowski p parameter for the full Golub gene expression dataset. We note that the minimal spanning tree complexity measure seems still to perform quite well and that the optimal performance level of .8194 is obtained at a p parameter value of 4.

Figure 7 shows average cross-validated single nearest neighbor performance a function of scatter selected genes for the full Golub data at $p = 4$. We notice that we are able to improve upon the performance associated with using all 7000 of the genes by using the top 372 scatter selected genes, performance level = .84722.

The last analysis that we will present uses a reduced set of Golub genes obtained as follows. First we use only those genes that have an expression level of 20 or greater across all patients. Now consider a n_g genes by n_s patients data matrix. We first divide each column by its mean. Next we subject each row to a standard normalizing transformation. Our analysis then proceeds forward with this reduced set of 1753 genes. Figure 8 presents performance and complexity as a function of Minkowski p parameter for the reduced Golub gene expression dataset.

We notice that the optimal performance is obtained at a Minkowski p parameter value of 4. We also notice that once again the MST based complexity seems to capture the associated discriminant difficulty.

Finally we wish to examine the performance as a function of scatter selected genes for the reduced Golub dataset. Figure 9 shows performance as a function of

Table 1: Results Summary.

Data	Number of Features	p	$1 - \hat{L}$
Nose	38	2	.7
Nose	38	5	.75
Nose	21	5	.78125
Smoothed Nose	38	2	.70
Smoothed Nose	38	29	.85625
Smoothed Nose	13	29	.7825
Golub Gene	7129	2	.805
Golub Gene	7129	4	.8194
Golub Gene	372	4	.84722
Reduced Golub Gene	1853	2	.8
Reduced Golub Gene	1853	4	.84722
Reduced Golub Gene	1698	4	.8611

scatter selected genes for the reduced Golub data at $p = 4$.

In this particular case the full performance level is obtained when one uses 1698 of the original 1753 genes that appear in the reduced gene dataset. We do note that we can obtain a performance level of on the order of .82 utilizing fewer than 200 of the original reduced genes along with an associated Minkowski p parameter value of 4.

Conclusions

Our results are summarized in Table 1. We can see that one can make great improvements in classifier performance by merely adjusting the Minkowski p value. This is in keeping with our previous paper that just focused on the artificial nose data. We also note that in many cases one can improve performance not only by optimizing the p parameter value but by also following this optimization with a down select of the features utilizing a measure such as scatter for a figure of merit to use in our forward selection process. Finally we note that in the case of the artificial nose data, we demonstrated the advantage of using a smoother prior to the p value calculation.

We have not discussed how the magnitude/value of the Minkowski p parameter might be related in some sense to the structure of the data that we are dealing with. We have also not examined the simultaneous optimization of p parameter selection, smoother and feature down selection process. This research is relegated to future endeavors.

On the complexity front, we have shown that for the two applications at hand, the artificial nose data and the gene expression dataset that the MST based complexity measure does a good job of capturing the difficulty of a classification problem. Currently we merely provide this information as an observation. We have not to date examined the incorporation of the MST based complexity characterization into the general feature selection process. This too must be relegated to future work.

Acknowledgments

The authors would like to thank Wendy Martinez of the Office of Naval Research for funding this effort.

References

- [Duda et al., 2000] Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley-Interscience, ssecond edition.
- [Friedman and Rafsky, 1979] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717.
- [Golub and Slonin, 1999] Golub, T. R. and Slonin, D. K. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.
- [Gross and Yellen, 1999] Gross, J. and Yellen, J. (1999). *Graph Theory and Its Applications*. CRC.
- [Priebe et al., 2000] Priebe, C. E., J., M. D., and Solka, J. L. (2000). On the selection of distance for a high dimensional cla ssification problem. *ASA Proceedings of the Sections on Statistica and Statistical Graphics*, (58–63).

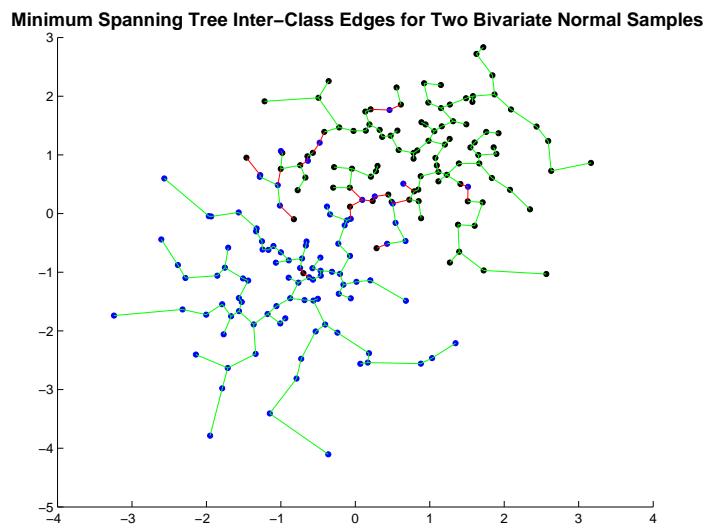


Figure 1: Example minimal spanning tree computed based on two samples from fairly well separated bivariate normals. The red edges are between disparate classes.

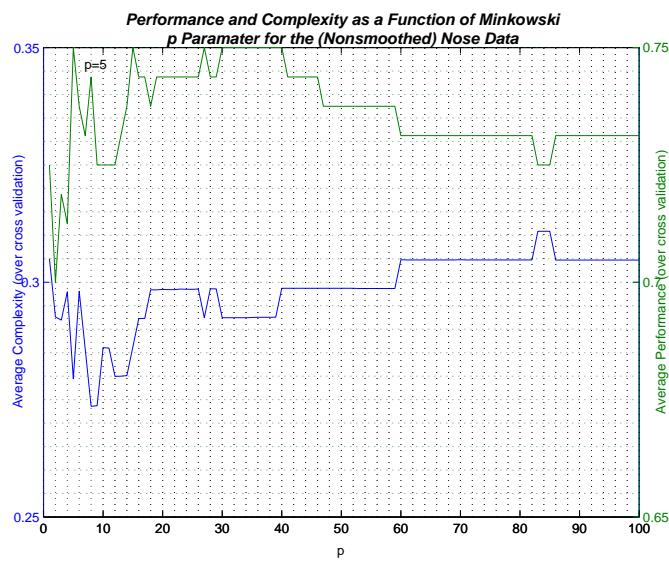


Figure 2: Average single nearest neighbor cross-validated performance, green curve, and average minimal spanning tree based complexity, blue curve, as a function of Minkowski p parameter based on the raw artificial nose data.

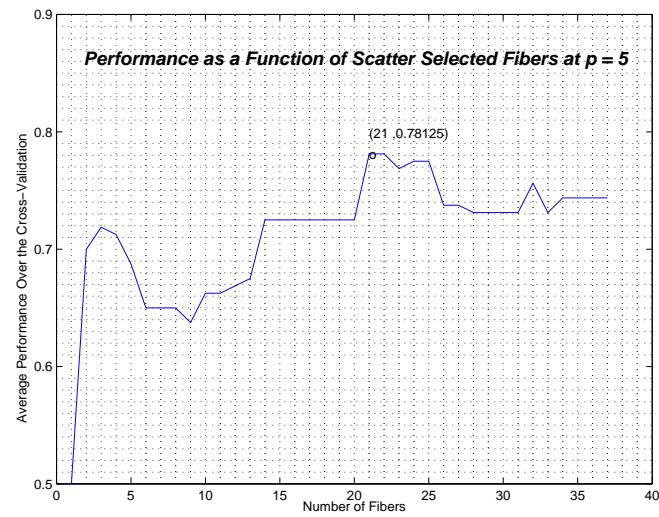


Figure 3: Average single nearest neighbor cross-validated performance as a function of number of scatter selected fibers at $p = 5$ for the raw artificial nose data.

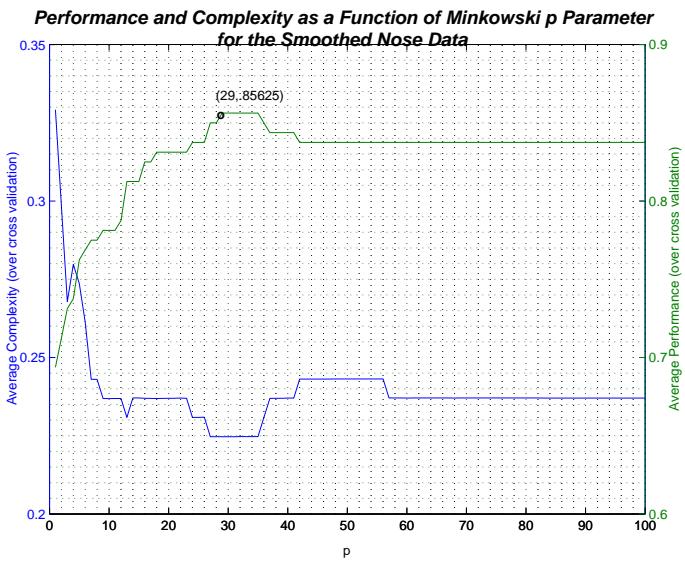


Figure 4: Average cross-validated single nearest neighbor performance, green curve, and average complexity, blue curve, as a function of Minkowski p parameter for the spline smoothed nose data.

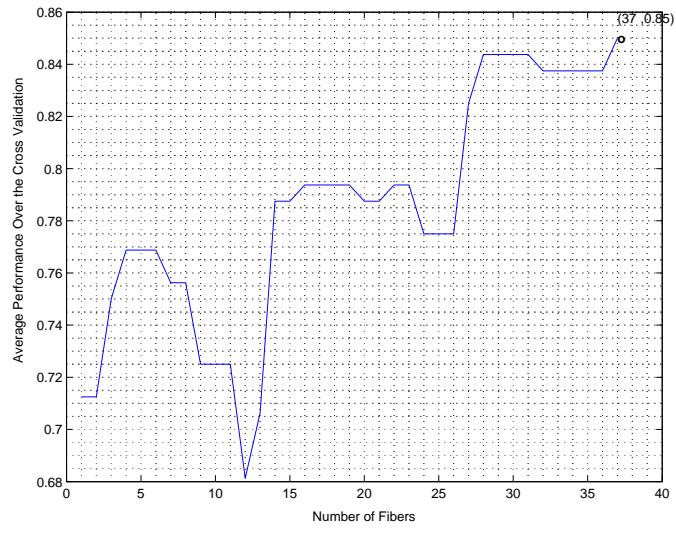


Figure 5: Average single nearest neighbor cross-validated performance as a function of number of scatter selected fibers at $p = 29$ obtained using the spline smoothed nose data.

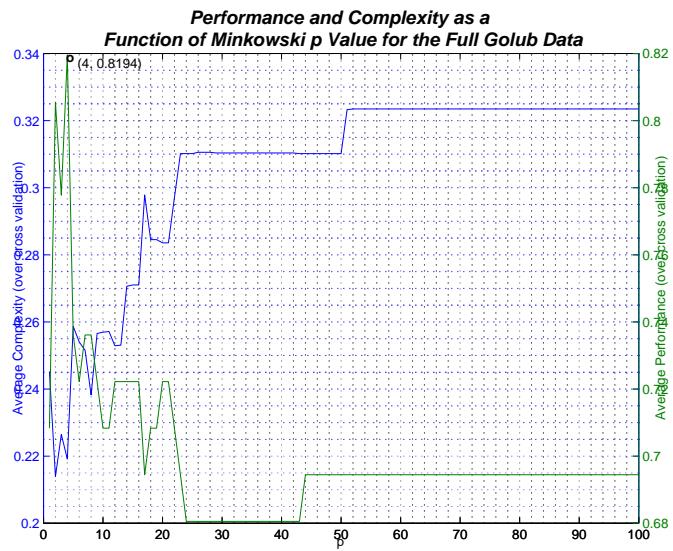


Figure 6: Average cross-validated single nearest neighbor performance, green curve, and average complexity, blue curve, as a function of the Minkowski p parameter for the full Golub gene expression dataset.

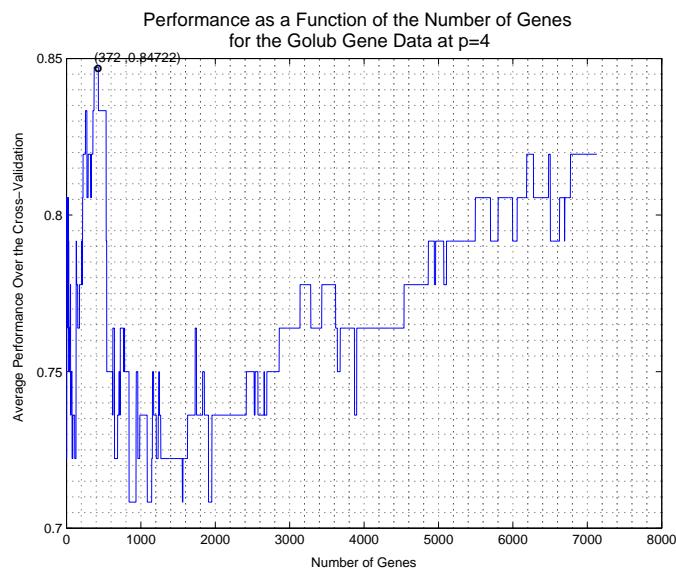


Figure 7: Average cross-validated single nearest neighbor performance as a function of scatter selected genes for the full Golub dataset at an optimal $p = 4$.

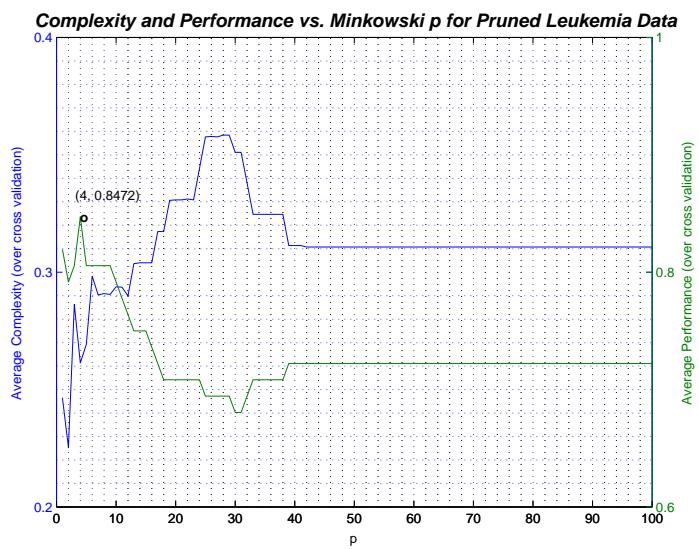


Figure 8: Performance and complexity as a function of the Minkowski p parameter for the reduced Golub gene expression dataset.

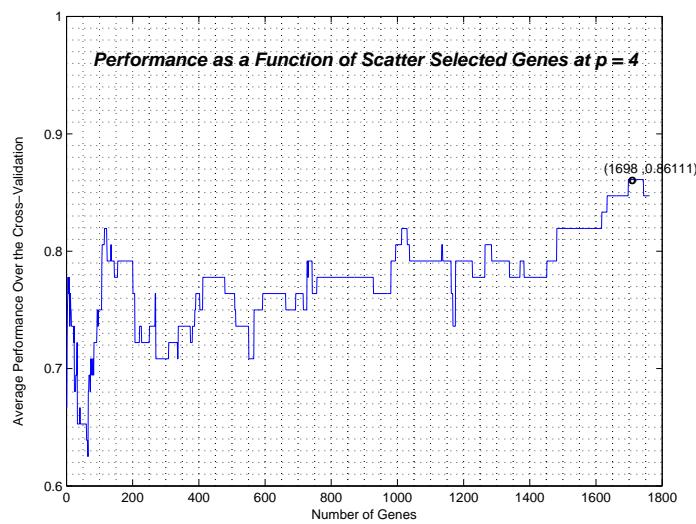


Figure 9: Performance as a function of scatter selected genes for the reduced Golub dataset at an optimal $p = 4$.

Statistical Classification Based on Contours

Mark Fitzgerald and Karen Kafadar, University of Colorado--Denver

Abstract: Paleontologists collect data on tracks and footprints of dinosaurs to gain information about their evolution, locomotion, and behavior. A particular goal is the classification of these footprints to achieve a taxonomy based on this physical evidence. Typically, an ``expert'' draws the outline of the footprint from a photograph which is then digitized and from which certain ``landmarks'' are identified (e.g., total length, total width, digit lengths and widths, angles between digits, heel-to-digit lengths, etc.), and a conventional classification algorithm is applied to these landmarks. Bookstein (1986) studied this approach in depth. Siegel (1982) developed a ``repeated median regression'' algorithm that was motivated by the comparison of skulls based on such identified landmarks. Kendall (1980) studied shape using an alternative approach based on geometric probability.

Most studies of shape start with the given data. However, these data generally arise from an ``expert'' who draws or outlines the footprint from a photograph. A less subjective approach would involve a contouring algorithm applied to the data from a digital photograph of the footprint. This approach raises two issues:

- (1) which contour? ``Experts'' can usually pick out the most likely contour representing the outline. Can the expert's intuition be modeled into an objectively-defined contour?
- (2) Comparison of contours. If a computer algorithm can identify a contour, or perhaps a set of plausible contours, can we model them in a way that allows us to compare the set of contours from one trackway to another? This would provide a taxonomy of dinosaurs based on footprints.

We describe some approaches to these two issues, along with examples from other fields where this problem of classification based on contours also arises.

Contributed Session 10

A Human Dimension Methodology for Assessing Future Combat Systems' C4ISR

Jock O. Grynovicki and Kragg P. Kysor, U.S. Army Research Laboratory

Abstract: As the U.S. Army invests in automation for battle command, it is important to determine how technology impacts (either positively or negatively) on operator, staff, and organizational performance. While many studies have taken a traditional task analysis approach to assessing technology's effects on battle command performance, this paper examines Future Combat Systems' (FCS) command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR) usability, functionality, and staff performance issues from a multiple-level systems analysis. Specifically, this paper identifies a number of tasks and behavioral characteristics that are associated with effective battle command performance at the operator, team, and organizational level. Each of these aspects, in turn, suggests a measurement taxonomy for assessing the enabling or degrading influence of digitization on battle command performance. Thus, a framework for assessing digital staff performance was developed that considers, hardware, software, battle command functions, soldier operator capabilities, as well as staff and leader dynamics. This Human Dimension of Battle Command initiative is intended to help the U.S. Army leadership assess the impact of FCS digitization on individual soldier, staff, and organizational performance. The lack of emphasis on the human component in the design and integration of automation can result in significant performance degradation, increased training requirements, and a lack of system acceptance by the soldier. This paper concludes with an example of a set of Likert-type scale metrics for use in assessing digitization.

Assessing and Removing Unexpected Collinearity in Designed Experiments

Trevor A. Craney, Pratt & Whitney

Abstract: When creating designed experiments, it is not always possible to run the experiment at the exact settings required to maintain orthogonal effects. However, this is not measurement error when precise measurements of the settings can be made once the experiment begins. A comparison is made for a 15-run Box-Behnken design using both the intended design settings and the actual design settings. Variance inflation factors are used to measure the induced collinearity in the effects. Two cutoff values are suggested for use to determine when an effect's variance inflation factor is too large to keep that effect in the model. This method is discussed relative to existing methods to offset collinearity in regression.

General Session 5

SiZer for Simple, Direct Inference in Exploratory Data Analysis
Steve Marron, University of North Carolina, Chapel Hill

Abstract: Smoothing methods for exploratory data analysis include histograms and scatterplot smoothers. These are powerful tools for finding structure in data, when used by knowledgeable practitioners. But traditional implementations can be dangerous when used by non-experts, since all too often one can interpret spurious sampling artifacts as important underlying structure. SiZer addresses the statistical inference problem of separating those features that are important and worth deeper investigation, from those that are mere noise artifacts. Results are presented via a novel visualization.

AUTHOR INDEX

- Agresti, Alan 2
Ath, Yontha 13
Bodt, Barry A. 11
Brodeen, Ann E. M. 11
Brucker, Duane E. 106
Bundy, Mark L. 10
Burkom, Howard S. 7
Chandramouli, R. 120
Cioppa, Thomas M. 5
Craney, Trevor A. 137
Cressie, Noel 118
Crosier, Ronald B. 76
Deason, Paul J. 106
Deininger, Paul 64
Dutoit, Gene 9
Elbert, Eugene 7
Evans, Diane 9
Farah-Stapleton, Monica 118
Fitzgerald, Mark 136
Frank, D. H. 5
Fuentes, Montserrat 117
Gentle, James 4
Genton, Marc G. 2
Ghosal, Subhashis 3
Ghosh, Sujit 2
Glen, Andrew 9
Graves, Todd L. 65
Grynovicki, Jock O. 137
Hamada, Michael 64
Harris, Bernard 10
Hengartner, Nicholas 64
Higdon, Dave 6
Irwin, Mark 118
Jaakkola, Tommi 117
Johannsen, D. A. 121
Jordan, Nikki 6
Kafadar, Karen 136
Kettenring, Jon R. 116
Kim, David B. 67
Kornak, John 118
Krabill, Robert 64
Kysor, Kragg P. 137
Leemis, Larry 9
Lettallier, Bruce 6
Li, Y. 8, 86, 87
Lucas, Thomas W. 5
Marchette, D. J. 15, 41
Marron, Steve 138
Martinez, Angel R. 27
McNulty, Mark 6
Nelson, Kevin 7
Niemuth, Nancy A. 87
Osborne, Jason A. 3
Powers, T. E. 8, 86, 87
Priebe, C. E. 15, 62
Raftery, Adrian 117
Reese, Shane 64
Rigsby, John 107
Rodriguez, Robert 4
Romano, Kevin P. 105
Samaniego, Francisco J. 62
Smith, Simon L. 4
Sobel, Milton 13
Solka, Jeffrey L. 1, 15, 107, 121
Sommerville, Douglas R. 76, 89
Thompson, James R. 61
Trofimovich, L. B. 8
Wainwright, Martin 117
Wasko, John A. 67
Webb, David W. 10
Wegman, Edward J. 1, 27
Wilburn, John Bart 108
Willsky, Alan 117
Wilson, Alyson 64
Wolfinger, Russell 119

Quality Assurance and OPSEC Review

This form is an approval record for ARL generated information to be presented or disseminated external to ARL. Note: Submit all manuscripts in electronic format or camera ready copy. See attached instructions. If more space is needed, use reverse of form (include block numbers).

A. General Information					
	1. Present Date	2. Unclassified Title			
	02/11/04	Proceedings of the Eighth Annual U.S. Army Conference on Applied Statistics			
3. Author(s) Barry A. Bodt Edward J. Wegman			4. Office Symbol(s) AMSRD-ARL-CI-CT George Mason University		5. Telephone Nr(s) (410) 278-6659 (703) 993-1691
6. Contractor generated <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes If yes, enter Contract No. and ARL COR			7. Type: <input checked="" type="checkbox"/> Report <input type="checkbox"/> Abstract <input type="checkbox"/> Publication <input type="checkbox"/> Presentation (<i>speech, briefing, video clip, poster, etc</i>) <input type="checkbox"/> Book <input type="checkbox"/> Book Chapter <input type="checkbox"/> Web		
8. Key Words Data Mining, Statistics, Experimental Design					
9. Distribution Statement (<i>required</i>) Is manuscript subject to export control? <input checked="" type="checkbox"/> No <input type="checkbox"/> Yes			Circle appropriate letter and number. (<i>see instructions for statement text</i>) <input checked="" type="checkbox"/> B <input type="checkbox"/> C <input type="checkbox"/> D <input type="checkbox"/> E <input type="checkbox"/> F <input checked="" type="checkbox"/> X 1 2 3 4 5 6 7 8 9 10 11		
B. Reports	11. Series TR	12. Type FINAL	13. No. of pages 139	14. Project No. P611102.H48	15. Period Covered 10/30-11/1/02
C. Publications					
17. Is MS an invited paper? <input type="checkbox"/> No <input type="checkbox"/> Yes			19. Material will be submitted for publication in		
18. Publication is a refereed journal? <input type="checkbox"/> No <input type="checkbox"/> Yes			Journal _____ Country _____		
D. Presentations		20. Conference Name/Location			21. Sponsor
22. Conference Date		23. Due Date	24. Conference is <input type="checkbox"/> Open to general public <input type="checkbox"/> Unclassified/controlled access <input type="checkbox"/> Classified		
25. For nonpublic meetings: Will foreign nationals be attending? <input type="checkbox"/> No <input type="checkbox"/> Yes (<i>If yes, list countries and identify International Agreement(s)</i>) <input type="checkbox"/> Don't know			26. Material will be <input type="checkbox"/> Oral presented only <input type="checkbox"/> Oral presented and published in <i>(If published, complete block 18 and 19, Section C.)</i>		
E. Authors Statement: 27. All authors have concurred in the technical content and the sequence of authors. All authors have made a substantial contribution to the manuscript and all authors who have made a substantial contribution are identified in Block 3.					
Barry A. Bodt <i>Barry A. Bodt</i> ARL Lead Author or COR			2/18/04 Date		
F. Approvals					
28. First line Supervisor of Senior ARL Author or COR <i>Patricia H. Jones</i> <i>2/18/04</i>		29. Reviewer(s) (<i>Technical/Editorial/NA</i>) <i>Barry A. Bodt (Tech)</i> <i>13 Feb 2004</i>			
Name _____ Date _____		Name(s) _____ Date _____			
30. Limited distribution information for release to foreign nationals		31. Classified Information Classified by _____ Declassified on _____			
Foreign Disclosure _____ Date _____		Command Security Manager _____ Date _____			

OPSEC REVIEW CHECKLIST

OPSEC POC: Complete and explain any positive responses in block 9.

Note: ARL must be the proponent of the proposed information for release.

1. Does this material contain Sensitive Information? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	8. Does this material contain:
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	a. Any contract proposals, bids, and/or proprietary information? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	b. Any information on inventions/patent application for which patent secrecy orders have been issued? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	c. Any weapon systems/component test results? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	d. Any ARL-originated studies or after action reports containing advice and recommendations? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	e. Weakness and/or vulnerability information? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	f. Any information on countermeasures? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	g. Any fielding/test schedule information? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	h. Any Force Protection, Homeland Defense (security) information? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	i. Information on subjects of potential controversy among military services or other federal agencies? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	j. Information on military applications in space, nuclear chemical or biological efforts: high energy laser information; particle beam technology; etc? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO
<input type="checkbox"/> YES <input checked="" type="checkbox"/> NO	k. Contain information with foreign policy or foreign relations implications? <input type="checkbox"/> YES <input checked="" type="checkbox"/> NO

OPSEC Approval Statement

I, the undersigned, am aware of the adversary's interest in DOD publications and in the subject matter of this material and that, to the best of my knowledge, the net benefit of this release out weights the potential damage to the essential security of all ARL, AMC, Army, or other DOD programs of which I am aware.

Fred S. Brundick

OPSEC Reviewer (Printed name/signature)

10 Feb 2004

Date

9. Space for explanations/continuations/OPSEC review comments

Final Release Clearances

32. Public/Limited release information

a. Material has been reviewed for OPSEC policy.

Douglas W. Hensley

b. The information contained in this material is / is not approved for public release/ has received appropriate tech/editorial review.

17 MAR 04

Date

KJ Division Chief

c. This information is accepted for public release.

Douglas W. Hensley
Public Affairs Office

10 March 04

Date

17 MAR 04

Date